University of Groningen

Segmentation in large-scale cellular electron microscopy with deep learning

Aswath, Anusha; Alsahaf, Ahmad; Giepmans, Ben; Azzopardi, George

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

[Link to publication in University of Groningen/UMCG research database](#)

Survey paper

Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey

Anusha Aswath [a,b,*], Ahmad Alsahaf [b], Ben N.G. Giepmans [b], George Azzopardi [a]

[a] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University Groningen, Groningen, The Netherlands*
[b] *Department of Biomedical Sciences of Cells and Systems, University Groningen, University Medical Center Groningen, Groningen, The Netherlands*

ARTICLE INFO

ABSTRACT

Electron microscopy (EM) enables high-resolution imaging of tissues and cells based on 2D and 3D imaging techniques. Due to the laborious and time-consuming nature of manual segmentation of large-scale EM datasets, automated segmentation approaches are crucial. This review focuses on the progress of deep learning-based segmentation techniques in large-scale cellular EM throughout the last six years, during which significant progress has been made in both semantic and instance segmentation. A detailed account is given for the key datasets that contributed to the proliferation of deep learning in 2D and 3D EM segmentation. The review covers supervised, unsupervised, and self-supervised learning methods and examines how these algorithms were adapted to the task of segmenting cellular and sub-cellular structures in EM images. The special challenges posed by such images, like heterogeneity and spatial complexity, and the network architectures that overcame some of them are described. Moreover, an overview of the evaluation measures used to benchmark EM datasets in various segmentation tasks is provided. Finally, an outlook of current trends and future prospects of EM segmentation is given, especially with large-scale models and unlabeled images to learn generic features across EM datasets.

1. Introduction

Electron microscopy (EM) is a widely used technique in life sciences for studying tissues, cells, subcellular components, and molecular complexes at the nanometer scale. EM captures snapshots of biological samples as either two-dimensional (2D) images or three-dimensional (3D) volumes to analyze the ultrastructure of various organelles and understand their complex spatial relationships. With advancements in EM technologies, various imaging methods in both 2D and 3D EM have emerged, Table 1. While 2D EM relies on biased regions of interest, automated pipelines for collecting, stitching, and publishing 2D EM images have been pioneered for both transmission EM (TEM) (Faas et al., 2012) and scanning TEM (STEM) (Sokol et al., 2015) for acquisition of areas up to 1 mm^2 at nanometer-range resolution. The large-scale images allow for open access worldwide data sharing, as evidenced by the nanotomy.org platform[1] hosting over 50 published studies and accessible EM data (Ravelli et al., 2013; de Boer et al., 2020; Dittmayer et al., 2021). This allows scientists to pan and zoom through different tissues or cellular structures in an unbiased manner, Fig. 1.

Advances in volume EM (vEM) or 3D EM have now enabled 3D analysis of ultra-structures in unprecedented detail (Peddie and Collinson

2014; Titze and Genoud, 2016; Peddie et al., 2022). 2D EM can be utilized to produce a sequence of slices, which can be stacked together to create a vEM dataset. Stable and automated imaging using 3D EM technologies has enabled the acquisition of massive volumes, leading to acquiring petabytes of data. For instance, the study by Zheng et al. (2018) imaged the complete brain volume of an adult fruit fly using serial section TEM (ssTEM), covering a volume of 1 mm^3 or 10,000 voxels, requiring 100 TB of storage. Large-scale 3D EM can also be imaged through cryo-electron tomography (cryo-ET), enabling the investigation of cellular architecture and macromolecular assemblies in their native environment.

With state-of-the-art EM technology, such as multibeam scanning EM (Eberle et al., 2015; Ren and Kruit, 2016; de Boer and Giepmans, 2021), up to hundred times faster acquisition and higher throughput allows for imaging of tissue-wide sections in the range of hours instead of days. Given the automated and faster image acquisition in EM a data avalanche (petabyte range per microscope/month) is becoming a reality. The manual segmentation and annotation of such large-scale EM datasets are prohibitively laborious. For instance, the manual annotation of a fraction (1 μm^3) of whole-cell volume annotation containing
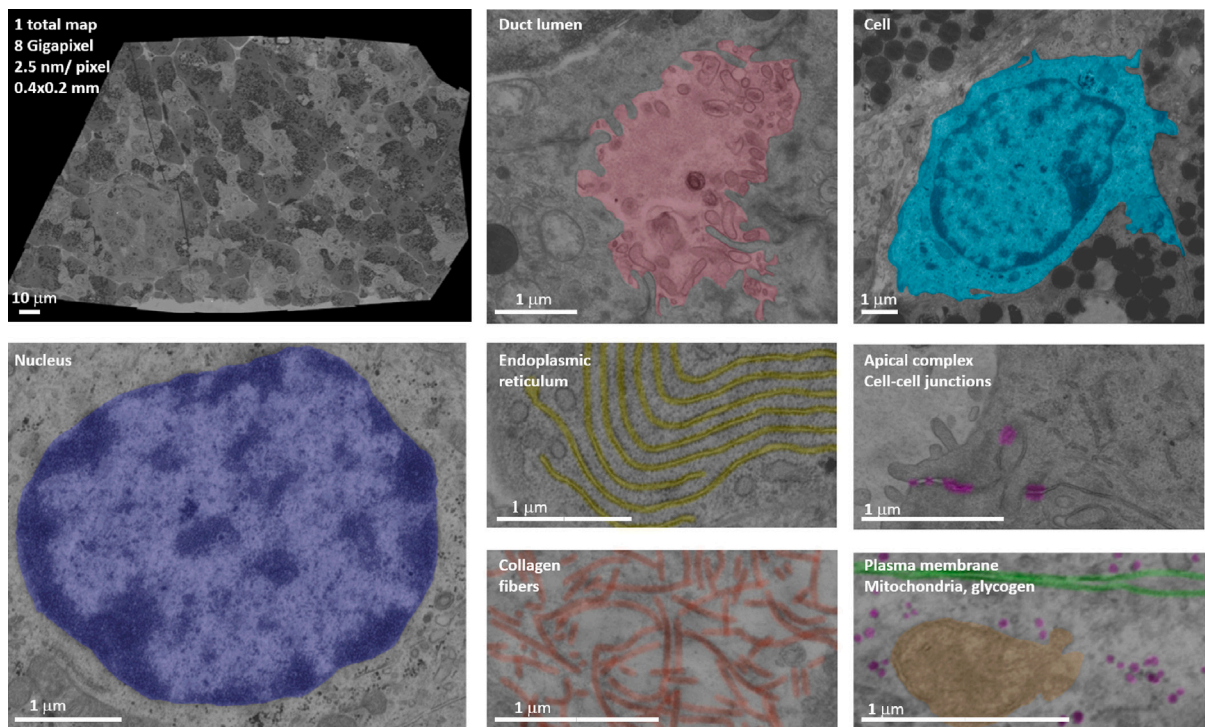

Fig. 1. Large-scale 2D EM of a section of human pancreas. Overview of a single dataset (top-left)[2] and snapshots from this total map at higher zoom showing several cellular, subcellular and macromolecular structures as indicated and annotated. Note the information density of these maps: millions of subcellular structures of a kind can be present per dataset (de Boer et al., 2020).

numerous instances of various organelles took two weeks, as demonstrated by Heinrich et al. (2021), and the whole-cell was estimated to take 60 person-years. This increase in the scale and acquisition speeds of data highlights the need for automated EM segmentation, for which both semantic and instance segmentation techniques are crucial, Fig. 2.

Semantic segmentation is a pixel-level image analysis task that involves partitioning an image into distinct and coherent regions, where each pixel is assigned a class label representing the semantic class it belongs to (e.g. nucleus, and mitochondrion). Instance segmentation assigns to each pixel of the semantic class label a unique instance identity for each structure. This is especially important for adjacent structures that need to be separated to analyze individual interactions.

Historically, conventional image analysis methods and shallow learning algorithms[3] were employed for segmenting EM images. These include statistical analysis of pixel neighborhoods (Kylberg et al., 2012), eigenvector analysis (Frangakis and Hegerl, 2002), watershed and hierarchical region merging (Liu et al., 2012, 2014), superpixel analysis and shape modeling (Karabağ et al., 2019), and random forest (Cao et al., 2019). However, in recent years, deep learning (DL) has emerged as the dominant approach in this field, mirroring the trends observed in segmentation techniques for light microscopy and other medical imaging modalities (Liu et al., 2021; Litjens et al., 2017).

Compared to traditional image analysis and machine learning methods that rely on handcrafted features, DL-based segmentation eliminates or significantly reduces the need for domain-specific knowledge to extract relevant features from the imaged sample (Liu et al., 2021). In particular, DL methods can capture complex and nonlinear relationships from raw data without significant preprocessing, handle diverse and large datasets, and provide robustness and scalability.

DL-based segmentation has gained popularity, leading to the development of plug-ins for commonly used biomedical image analysis tools like CellProfiler (Carpenter et al., 2006), ImageJ (Schindelin et al., 2012), Weka (Arganda-Carreras et al., 2017), and Ilastik (Berg et al., 2019), which were previously limited to traditional image processing or shallow learning.

We review the recent progress of automatic image segmentation in EM, with a focus on the last six years that marked significant progress in both DL-based semantic and instance segmentation, while also giving an overview of the main DL architectures that enabled this progress.

The manuscript is organized as follows: Section 2 describes the literature search strategy used for this review. Section 3 presents the key benchmark datasets, which have played a vital role in advancing segmentation methods. Section 4 lays the background about the main neural network architectures for 2D and 3D segmentation of EM datasets. Sections 5 and 6 review the papers that propose new methodologies for semantic and instance segmentation with different DL approaches. These are followed by Section 7, which describes the evaluation metrics used in the reviewed papers. Section 8 discusses the overall progress made so far along with presenting major limitations and an outlook for future research. Finally, we outline the conclusions in Section 9.

2. Strategy of literature search

Our survey strategy is motivated by the following questions:

- Which datasets are accessible for EM analysis, what are their challenges and what role do they play in DL research?
- How is EM image (semantic and instance) segmentation being addressed by fully/semi/un/self-supervised DL pipelines?
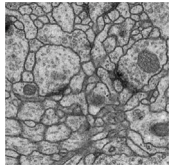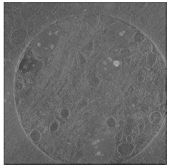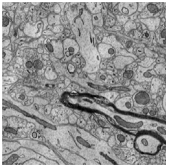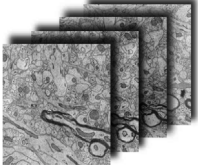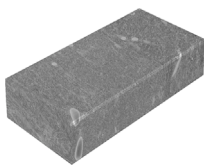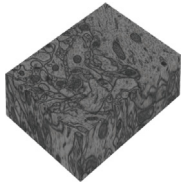
To answer these questions, the following search query was used in Pubmed, Web of Science, and Google Scholar on words in titles (TI) only, restricted to 2017–2022: TI = ((electron microscopy OR EM) AND (segmentation OR semantic OR instance OR supervised OR

---

[2] Full access to digital zoomable data at full resolution is via http://www.nanotomy.org/OA/nPOD/6153-2016-209/.

[3] Shallow learning in this context refers to machine learning with hand-crafted features as input.

Table 1

Main large-scale EM techniques. More information is given in the MyScope website[a] and the reviews by Peddie and Collinson (2014), Titze and Genoud (2016) and Kievits et al. (2022). The last row shows example 2D images and 3D stacks of such technologies except STEM, which is shown in Fig. 1.

2D EM	Data acquisition technique
Transmission Electron Microscopy (TEM)	A widefield electron beam illuminates an ultra-thin specimen and transmitted electrons are detected on the other side of the sample. The structure that is electron dense appears dark and others appear lighter depending on their (lack of) scattering.
Scanning Electron Microscopy (SEM)	The raster scanning beam interacts with the material and can result in backscattering or the formation of secondary electrons. Their intensity reveals sample information.
Scanning Transmission Electron Microscopy (STEM)	SEM on ultrathin sections and using a detector for the transmitted electrons.
3D EM	
Serial section TEM (ssTEM) or SEM (ssSEM)	A volume EM technique for examining 3D ultrastructure by scanning adjacent ultrathin (typical 60–80 nm) sections using TEM or SEM, respectively. Adjacent sections are obtained through serial sectioning. The sample is cut into ultrathin sections using an ultramicrotome and collected on grids (ssTEM) or tape (ssSEM) for imaging.
Serial Block-Face scanning EM (SBF-SEM)	The block face is scanned followed by removal of the top layer by a diamond knife (typically 20–60 nm) and the newly exposed block face is scanned. This can be repeated thousands of times.
Focused Ion Beam SEM (FIB-SEM)	Block face imaging as above, but sections are repeatedly removed by a focused ion beam that has higher precision than a knife (typically down to 4 nm), making it suitable for smaller samples.
Cryo-electron tomography (Cryo-ET)	It captures a series of 2D projection images of a flash-frozen specimen from different angles, and then uses computational reconstruction methods to generate a 3D model or tomogram.

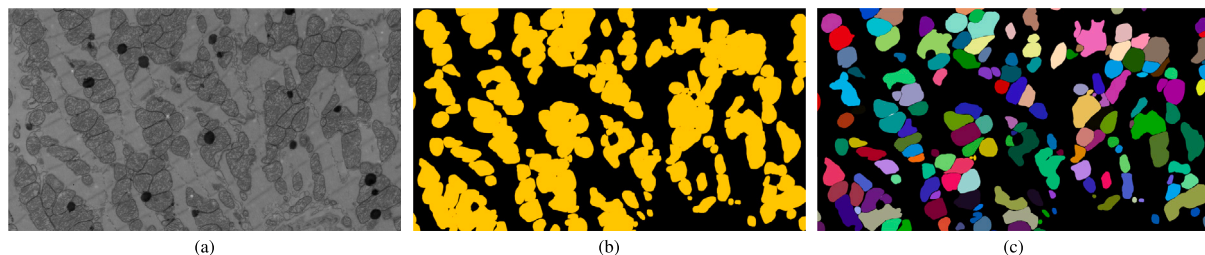| TEM 2D section	Cryo-ET 2D section	SEM 2D section	ssSEM volume - 2D sections	SBF-SEM volume	FIB-SEM volume
Ciresan et al. (2012)	Chen et al. (2017d)	Kasthuri et al. (2015)		Abdollahzadeh et al. (2021)	Lucchi et al. (2011)

[a] https://myscope.training/.

Fig. 2. Example of semantic and instance segmentation. (a) Original gray-scale image, and the corresponding (b) semantic and (c) instance segmentation maps of many apposed mitochondria. While semantic segmentation identifies all mitochondria as a single entity, instance segmentation accurately delineates and differentiates each instance even within the mitochondrion class.

unsupervised OR self-supervised OR semi-supervised)), and title or abstracts containing (deep learning, segmentation, electron microscopy) on Google Scholar. Results from the query that were outside the scope of this study, such as deep learning in material sciences and methods based on traditional image processing (pre-DL era), were excluded. The forward and backward snowballing technique was then used to compile the final list of 38 papers.

Fig. 3 summarizes this collection of 38 papers in terms of learning technique (fully supervised or not), segmentation type (semantic or instance), application (2D or 3D) and the underlying modeling backbone. Before reviewing these papers, we discuss the key EM datasets and describe the evolution of DL architectures, which are two crucial components that have been permitting the progress of EM segmentation analysis.

3. Collections of key EM datasets

Collections of labeled and unlabeled EM images have played a significant role in advancing DL research for EM segmentation, and some were associated with notable segmentation competitions and challenges. This section provides the details of all collections used by

the 38 papers in this survey. Table 2 reports the main properties of these datasets and below is an in-depth discussion of their characteristics and the challenges they address. The discussion is categorized according to the EM modality used to acquire the datasets.

3.1. Serial section TEM and SEM datasets

Serial-section transmission or scanning EM (ssTEM or ssSEM) is used for studying synaptic junctions and highly-resolved membranes in neural tissues. Advances in microscopy techniques in serial section EM have enabled the study of neurons with increased connectivity in complex mammalian tissues (such as mice and humans) and even whole brain tissues of smaller animal models, like the fruit fly and zebrafish. This imaging approach visualizes the generated volumes in a highly anisotropic manner, i.e. the x- and y-directions have a high resolution, however, the z-direction has a lower resolution, as it is reliant on serial cutting precision.

The Drosophila larvae dataset (#1)[4] of the ISBI 2012 challenge was the first notable EM dataset for automatic neuronal segmentation,

---

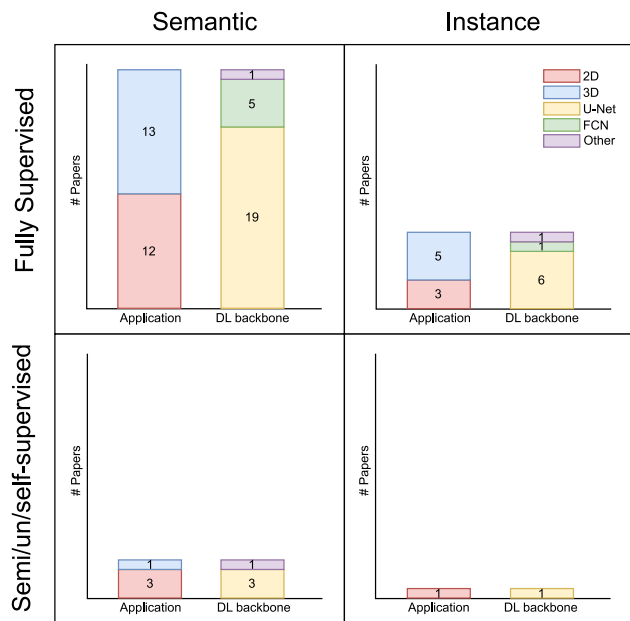[4] #n refers to the entry n in Table 2.

Fig. 3. Categorization of the 38 papers reviewed in this survey. The papers are first categorized on the learning paradigm (fully vs. semi/un/self-supervised) and on the segmentation type (semantic vs. instance). Each quadrant shows the distributions of applications (2D vs. 3D) and DL backbones (U-Net vs. FCN vs. Other) of the papers that use the corresponding learning and segmentation approaches. Note, U-Net is a specific type of fully convolutional network (FCN). The papers flagged as FCN use FCN architectures other than U-Net.

featuring two volumes with 30 sections each. The main challenge of that dataset is to develop algorithms that can accurately segment the neural structures present in the EM images. The success of deep neural networks as pixel classifiers in the ISBI 2012 challenge (Ciresan et al., 2012) paved the way for deep learning in serial section EM segmentation. Recently, a connectome of an entire brain of a Drosophila fruit fly has been published by Winding et al. (2023), and will serve as a new resource for various follow-up works.

The CREMI3D dataset (#2) consists of three large and diverse sub-volumes of neural tissue along with ground truth annotations for training and evaluation purposes, and was part of a competition at the MICCAI 2016 conference. The dataset comes from a full adult fly brain (FAFB) volume and contains 213 teravoxels. It was imaged at the synaptic resolution to understand the functioning of brain circuits (connectomics) and its goal was to segment neurons, synapses, and their pre-post synaptic partners. The CREMI3D dataset is part of the FlyEM project and since its inception, it has been used to evaluate various image analysis methods for neural circuit reconstruction, including DL approaches such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

The SNEMI3D dataset (#3) consists of a volume of 100 ssSEM images of the neural tissue from a mouse cortex. It is a subset of the largest mouse neocortex dataset imaged by Kasthuri et al. (2015) using an automated ssSEM technique and hence is also known as the Kasthuri dataset. The dataset was created as part of the ISBI 2013 challenge on segmentation of neural structures in EM images. The main challenge of this dataset is to develop algorithms that can accurately segment the neuronal membranes present in the EM images and reconstruct a 3D model of the tissue. This is a difficult task due to the large size of the dataset and the complexity of the neural structures, namely axons, dendrites, synapses, and glial cells.

The Kasthuri++ (#4) dataset, introduced by Casser et al. (2018), is an improved version of the original Kasthuri dataset for dense reconstructions of neuronal cells. It addresses the issue of inaccurate annotations related to the jaggedness between inter-slice components.

The Xiao (#5) dataset for mitochondria segmentation was collected from a rat brain by Xiao et al. (2018a) using advanced ssSEM technology. Automated cutting was used to produce 31 sections, each with an approximate thickness of 50 nm for segmenting mitochondria. The ground truth dataset was prepared through 2D manual annotation and image registration of serial-section images, which was made publicly available for accelerating neuroscience analysis.

Mito-EM (#6) (Wei et al., 2020) introduced the largest mammalian mitochondria dataset from humans (MitoEM-H) and adult rats (MitoEM-R). It is about 3600 times larger than the Lucchi dataset described below, which has become a standard dataset for mitochondria segmentation and contains mitochondria instances of at least 2000 voxels in size. Complex morphology such as mitochondria on a string (MOAS) connected by thin microtubules or instances entangled in 3D were captured using ssSEM. The MitoEM dataset was created to provide a comprehensive view of the ultrastructure of mitochondria and to facilitate a comparative study of mitochondrial morphology and function in rats and humans.

The NucMM dataset (#7) (Lin et al., 2021) contains two fully annotated volumes; one that contains almost a whole zebrafish brain with around 170,000 nuclei imaged using ssTEM, and another that contains part of a mouse visual cortex with about 7000 nuclei imaged using micro-CT. Micro-CT or micro-computed tomography uses X-rays to produce 3D images of objects at low resolution and hence is not a part of this review. The large-scale nuclei instance segmentation dataset from ssTEM covers $0.14\ mm^3$ of the entire volume of the zebrafish brain at $4 \times 4 \times 30$ nm/voxel. As most of the nuclei segmentation datasets are from light microscopy at the μm scale, the dataset was downsampled to $512 \times 512 \times 480$ nm/voxel.

3.2. FIB-SEM datasets

FIB-SEM offers high-resolution datasets with isotropic voxel size, providing equal resolution along the x, y, and z axes. This makes it an excellent tool for automated segmentation of neuronal cells and various sub-cellular structures, including mitochondria, vesicles, and Golgi apparatus, among others. FIB-SEM is used for examining tissues at resolutions lower than $10 \times 10 \times 10$ nm. The method can produce sections with a thickness of 4 nm, but the volumes are typically smaller in comparison to other techniques, due to their high z-resolutions.

The Lucchi dataset (#8) is an isotropic FIB-SEM volume imaged from the hippocampus of a mouse brain, and it has the same spatial resolution along all three axes. This dataset has now become the de facto standard for evaluating mitochondria segmentation performance. An enhanced version of this benchmark dataset, the Lucchi++ dataset (#9), was presented by Casser et al. (2018) with re-annotations that ensured consistent mitochondria boundaries and corrections of misclassifications.

Efforts to expand FIB-SEM to larger volumes were made by Takemura et al. (2015) who compiled the FIB-25 (#10) dataset by reconstructing the synaptic circuits of seven columns in the eye region of a Drosophila's brain. FIB-25 contains over 10,000 annotated neurons, including their synaptic connections, and is one of the most comprehensive EM datasets of the Drosophila brain to date. It was created to provide a detailed map of the neural circuits in the Drosophila brain and to facilitate the study of neural connectivity and information processing. The dataset is publicly available and can be accessed through the FlyEM project website. Enhanced FIB-SEM techniques have also enabled high-throughput and reliable long-term imaging for large-scale EM (10^3 to $3 \times 10^7\ \mu m^3$), such as the OpenOrganelle atlas (#11) of 3D whole cells and tissues of Xu et al. (2021). The datasets for the 3D reconstruction of cells were made open-source under the OpenOrganelle repository for exploring local cellular interactions and their intricate arrangements.

FIB-SEM datasets include the high-resolution analysis of organelles in critical tissues such as the heart muscle and urinary bladder. Cardiac mitochondria (#12) is a FIB-SEM dataset introduced to segment

Table 2

Key datasets from studies that perform high-resolution automated (volume) EM segmentation using deep learning. The abbreviations of the (sub) cellular structures are defined in the legend. Sub-cellular structures with instance labels are rendered in italics while the rest have semantic labels only.

#	Dataset	Acquisition	Model/Region	Voxel size (nm)	Volume size (voxels)	Labeled (sub) cellular structures	Public repository
1	ISBI 2012/ Drosophila VNC	ssTEM	Drosophila/Nervous cord	$4 \times 4 \times 50$	$512 \times 512 \times 30$	NM	https://imagej.net/events/isbi-2012-segmentation-challenge
2	MICCAI 2016/ CREMI3D	ssTEM	Drosophila/Adult fly brain	$4 \times 4 \times 40$	$1250 \times 1250 \times 125$	NM, S, SP	https://cremi.org
3	ISBI 2013/ SNEMI3D/Kasthuri	ssSEM	Mouse/Neocortex	$3 \times 3 \times 30$	$1024 \times 1024 \times 100$	NM	https://snemi3d.grand-challenge.org/
4	Kasthuri++	ssSEM	Mouse/Neocortex	$3 \times 3 \times 30$	$1643 \times 1613 \times 85$	M, NM	https://casser.io/connectomics
5	Xiao	ssSEM	Rat/Cortex	$2 \times 2 \times 50$	$8624 \times 8416 \times 20$	M	http://95.163.198.142/MiRA/mitochondria31/
6	MitoEM	ssSEM	Rat, Human/Cortex	$8 \times 8 \times 30$	$4096 \times 4096 \times 1000$	*M*	https://mitoem.grand-challenge.org/
7	NucMM	ssSEM	Zebrafish/Whole brain	$4 \times 4 \times 30$	$1450 \times 2000 \times 397$	*N*	https://nucmm.grand-challenge.org/
8	Lucchi/EPFL Hippocampus	FIB-SEM	Mouse/Hippocampus	$5 \times 5 \times 5$	$1024 \times 768 \times 165$	M	https://www.epfl.ch/labs/cvlab/data/data-em/
9	Lucchi++	FIB-SEM	Mouse/Hippocampus	$5 \times 5 \times 5$	$1024 \times 768 \times 165$	M	https://casser.io/connectomics
10	FIB-25	FIB-SEM	Drosophila/Optic lobe	$8 \times 8 \times 8$	$520 \times 520 \times 520$	N, S	http://research.janelia.org/FIB-25/FIB-25.tar.bz2
11	OpenOrganelle	FIB-SEM	Interphase HeLa, Macrophage, T-cells	$8 \times 8 \times 8$	Varying sizes	CN, CH, EN, ER, ERN, ERES, G, LP, L, MT, NE, NP, Nu, N, PM, R, V	https://openorganelle.janelia.org
12	Cardiac mitochondria	FIB-SEM	Mouse/Heart muscle	$15 \times 15 \times 15$	$1728 \times 2022 \times 100$	M	http://labalaban.nhlbi.nih.gov/files/SuppDataset.tif
13	UroCell	FIB-SEM	Mouse/Urothelial cells	$16 \times 16 \times 15$	5 subvolumes of $256 \times 256 \times 256$	G, L, *M*, *V*	https://github.com/MancaZerovnikMekuc/UroCell
14	Perez	SBF-SEM	Mouse/Brain	$7.8 \times 7.8 \times 30$	$16\,000 \times 12\,000 \times 1283$	L, M, Nu, N	https://www.sci.utah.edu/releases/chm_v2.1.367/
15	SegEM	SBF-SEM	Mouse/Cortex	$11 \times 11 \times 26$	279 volumes of $100 \times 100 \times 100$	NM	https://segem.rzg.mpg.de/webdav/SegEM_challenge/
16	CDeep3M-S	SBF-SEM	Mouse/Brain	$2.4 \times 2.4 \times 24$	$16\,000 \times 10\,000 \times 400$	M, NM, Nu, V	https://github.com/CRBS/cdeep3m
17	EMPIAR-10094	SBF-SEM	HeLa cells	$10 \times 10 \times 50$	$8192 \times 8192 \times 517$	Unlabeled	http://dx.doi.org/10.6019/EMPIAR-10094
18	Guay	SBF-SEM	Human/Platelets	$10 \times 10 \times 50$	$800 \times 800 \times 50$	Cell, CC, CP, GN, M	https://leapmanlab.github.io/dense-cell/
19	Axon	SBF-SEM	Mouse/White matter	$50 \times 50 \times 50$	$1000 \times 1000 \times 3250$	*A*, *M*, *My*, *N*	http://segem.brain.mpg.de/challenge/
20	CEM500K	All of the above	20 regions (10 organisms)	$2 \times 2 \times 2$ to $20 \times 20 \times 20$	$224 \times 224 \times 496\,544$	Unlabeled	https://www.ebi.ac.uk/empiar/EMPIAR-10592/
21	Cellular Cryo-ET	Cryo-ET	PC12 cells	$2.8 \times 2.8 \times 2.8$	$938 \times 938 \times 938$	M, MT, PM, R, V	https://www.ebi.ac.uk/emdb/EMD-8594
22	CDeep3M-C	Cryo-ET	Mouse/Brain	$1.6 \times 1.6 \times 1.6$	$938 \times 938 \times 938$	NM, V	https://github.com/CRBS/cdeep3m

A — Axons, CC — Canalicular channel, CH — Chromatin, CN — Centrosome, CP — Cytoplasm, D — Dendrites, EN — Endoplasmic Reticulum, ERES — Endoplasmic Reticulum Exit Site, G — Golgi, GC — Glial cells, GN — Granules, L — Lysosome, LP — Lipid Droplet, M — Mitochondria, MT — Microtubule, My — Myelin. N — Nucleus, NE — Nuclear Envelope, NM — Neuronal membrane, NP — Nuclear Pore, Nu — Nucleolus, PM — Plasma Membrane, R — Ribosome, S — Synapse, SP — Synaptic partners, V — Vesicle.

mitochondria in cardiomyocytes (Khadangi et al., 2021b). The FIB-SEM technique was needed to better characterize diffusion channels in mitochondria-rich muscle fibers. Isotropic voxels at 15 nm resolution were imaged according to the set of experiments performed by Glancy et al. (2015). The UroCell (#13) from FIB-SEM was imaged by Mekuč et al. (2022) to focus on mitochondria and endolysosomes and was further extended to Golgi apparatus and fusiform vesicles. The dataset is unique as it is publicly available for further analysis of the epithelium cells of the urinary bladder, where the organelles form an important

component in maintaining the barrier between the membrane of the bladder and the surrounding blood tissues.

3.3. SBF-SEM datasets

Connectomics research was also based on popular datasets imaged using SBF-SEM (Helmstaedter et al., 2013; Briggman et al., 2011). Imaging using SBF-SEM produces anisotropic sections but does not need

image registration and avoids missing sections in comparison to serial-sectioning TEM/SEM, as the technique images the sample intact on a block surface. Such a technique also enabled imaging large volumes for studying the organization of neural circuits and cells across hundreds of microns through millimeters of neurons in a z-stack.

The Perez dataset (#14) (Perez et al., 2014) involved the acquisition of 1283 serial images from the hypothalamus's suprachiasmatic nucleus (SCN), a small part of the mouse brain, to produce an image stack with tissue dimensions approximately measuring 450,000 μm^3. The large acquired volume was downsampled from 3.8 to 7.8 nm/pixel in the $x - y$ resolution to scale up the processing of these tetra-voxel-sized SBF-SEM images. It was introduced for the automatic segmentation of mitochondria, lysosomes, nuclei, and nucleoli in brain tissues.

SegEM (#15) introduced an EM dataset acquired using SBF-SEM from the mouse somatosensory cortex (Berning et al., 2015). The images in the SegEM dataset are provided with corresponding segmentation labels for dendrites, axons, and synapses. The labels were generated using a semi-automated approach which involved a combination of skeleton annotations and machine learning algorithms to trace long neurites accurately. Since then, SegEM has been used for benchmarking popular models like flood-filling networks that test the efficiency of algorithms on volume-spanning neurites.

CDeep3M proposed two new datasets from SBF-SEM and cryo-electron tomography (cryo-ET) for automatic segmentation. The first one, CDeep3M-S (#16), is a large SBF-SEM dataset for membrane, mitochondria, and synapse identification from the cerebellum and lateral habenula of mice. Imaged at 2.4 nm pixel size, a cloud implementation of the latest architecture for anisotropic datasets was used to segment structures such as the neuronal membrane, synaptic vesicles mitochondria, and nucleus in brain tissues. The second dataset, CDeep3M-C (#22), was from cryo-ET and is explained further in Section 3.4.

The EMPIAR-10094 dataset (#17) consists of EM images of cervical cancer "HeLa" cells imaged using SBF-SEM. The dataset is imaged at 8192×8192 pixels over a total of 518 slices, and consists of different HeLa cells distributed in the background of the embedding resin. The dataset has been made publicly available with no labels and has mostly been used for delineating structures such as plasma membranes and nuclear envelopes.

The Guay dataset (#18) is a fully annotated dataset of platelet cells from two human subjects and was designed for dense cellular segmentation (Guay et al., 2021). It has also been used for large-volume cell reconstruction along with mitochondria, nuclei, lysosomes, and various granules inside the cells.

The Axon dataset (#19) is a collection of SBF-SEM images of white matter tissue from rats, captured at a lower resolution of 50 nm/pixel (Abdollahzadeh et al., 2021). The low-resolution image stack of 130,000 μm^3 was enough to resolve structures like myelin, myelinated axons, mitochondria, and cell nuclei. A wide field of view employing low-resolution SBF-SEM stacks was considered important for quantifying metrics such as myelinated axon tortuosity, inter-mitochondrial distance, and cell density.

Unlabeled datasets, such as CEM500K, from various unrelated experiments and EM modalities for solving the segmentation of a particular structure seem promising. The CEM500K (#20) is an EM unlabeled dataset containing around 500,000 images from various unrelated experiments and different EM modalities for cellular EM. The images from different experiments were standardized to 2D images of size 512×512 pixels with pixel resolutions ranging from 2 nm in datasets from serial section EM and ~20 nm for SBF-SEM. The dataset was further filtered by removing duplicates and low-quality images in order to provide robustness to changes in image contrast and making it suitable for training modeling techniques.

3.4. Cryo-ET datasets

Electron tomography (ET) is used to obtain 3D structures of EM sections using the tilt-series acquisition technique. Cryo-ET does so at cryogenic temperatures to image vitrified biological samples. Attempts for segmentation on cryo-ET can be found by Moussavi et al. (2010) and in the review of Carvalho et al. (2018). The identification of macromolecular structures is beyond the scope of this review.

Cryo-ET presents challenges in visualizing and interpreting tomographic datasets due to two main factors. Firstly, sample thickness increases as the tilt angle increases, leading to an artifact known as the "missing wedge" and reduced resolution in the z-direction. Secondly, vitrified biological samples are sensitive to electron dose, resulting in a low signal-to-noise ratio and difficulties in distinguishing features of interest from background noise. As the resolution capacity of TEM decreases with the increase in sample thickness, focused ion beam (FIB) milling can be used to obtain a high-resolution tomogram. Cryo-FIB SEM is an evolving technology for cellular imaging that is rapidly being used in recent years. This is mainly attributable to its ability to image larger specimens that may be too thick for cryo-ET, such as whole cells or tissues.

The cellular cryo-ET dataset (#21) was acquired at low magnification for annotation and qualitative cellular analysis of organelles like mitochondria, vesicles, microtubules, and plasma membrane (Chen et al., 2017d). The PC12 cell line was reconstructed using 30 serial sections imaged at $850 \times 850 \times 81$ pixel size at 2.8 nm resolution. The tomograms of platelets and cyanobacteria utilized in that work are from previously published datasets (Wang et al., 2015; Dai et al., 2013). CDeep3M-C (#22) is a cryo-ET dataset for the segmentation of vesicles and membranes from the mouse brain Haberl et al. (2018). At a voxel size of 1.6 nm, it was used to digitally recreate a tiny section (approximately $1.5 \times 1.5 \times 1.5$ μm^3) of a high-pressure frozen tissue. The final volume was built from 7 sequential tomograms (serial sections), each created by tilting a sample every 0.5° in an electron beam from −60° to +60°.

4. Background of backbone deep learning networks for EM semantic and instance segmentation

The rapid progress of DL methods, in particular CNNs, has had a great impact on advancing segmentation of EM images, as well as other medical images of various modalities (Litjens et al., 2017; Shen et al., 2017), including light microscopy (Xing et al., 2017; Liu et al., 2021). Deep learning in EM analysis has also been addressed in the reviews by Treder et al. (2022) and Ede (2021). The former gives a broad overview of different EM applications in both physical and life sciences and the latter provides a practitioner's perspective focused on the hardware and software packages to perform DL-based EM analysis. In contrast, this review provides an in-depth view of fully/semi/self/un-supervised deep learning methods for the semantic and instance segmentation in (sub)cellular EM. This section covers the main milestones in the progression of network architectures and their key attributes, which are necessary to put in context the 38 papers that are reviewed in this work.

The rest of this section is structured as follows: Section 4.1 introduces the progress made in CNN architectures that have facilitated end-to-end learning for semantic segmentation of 2D EM images. Section 4.2 addresses the significant challenges in 3D EM analysis and categorizes the DL-based techniques into three main approaches. Lastly, Section 4.3 explores how advancements in segmentation networks have facilitated their application in instance segmentation.
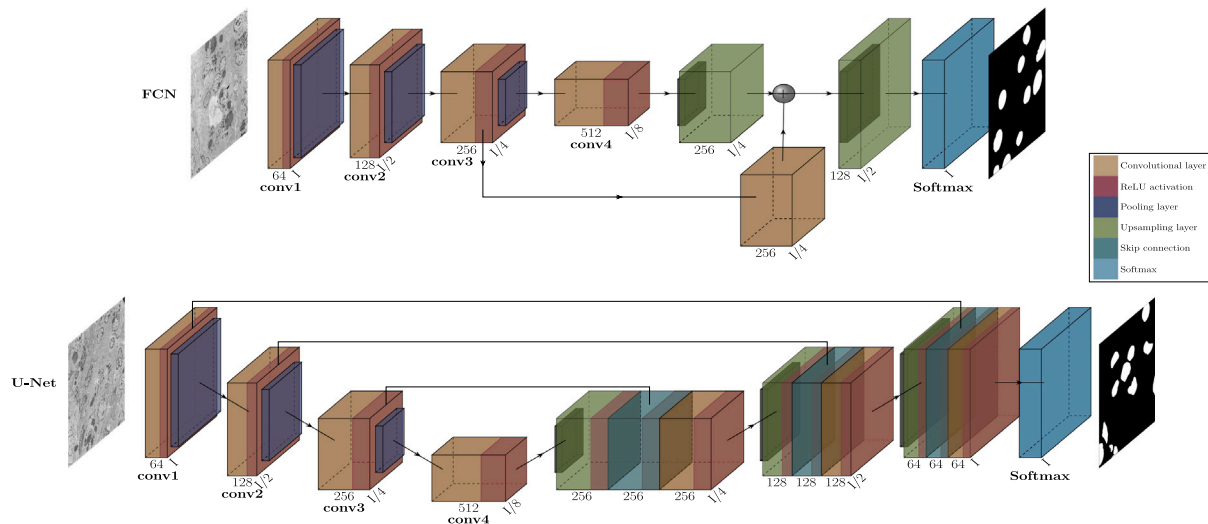
Fig. 4. Encoder–decoder architecture of FCNs (Long et al., 2015) and its symmetric version, popularly known as U-Net (Ronneberger et al., 2015). Each of the convolutional layers is followed by the nonlinear activation function ReLU and max pooling. The decoder upsamples features and combines them with the corresponding low-level features using skip connections. The last layer is a softmax function that assigns a probability class score to each pixel.

4.1. Semantic segmentation in 2D

The first notable application of CNNs for the segmentation of 2D EM was of neuronal membranes in serial sections (Ciresan et al., 2012). The images were segmented by predicting the label of each pixel centered on a local region or patch, covered by a convolutional filter in a sliding window approach. As indicated by Arganda-Carreras et al. (2015), it led to winning the ISBI 2012 neuronal segmentation challenge.[5] However, such a method suffered from two major limitations — firstly, the redundancy of processing large overlaps between adjacent patches, and secondly, the trade-off between the size of the patches (context) and localization accuracy. As the network's depth was an important factor for a larger receptive field (the size of the viewing field from which the network receives information), larger patches require deeper networks. However, downsampling caused by the several max-pooling layers resulted in a drop in localization ability with deeper networks, and the usage of smaller patches only enabled the network to observe a limited amount of context.

Improvements in the semantic segmentation of EM images started with the development of the Fully Convolutional Networks (FCN) (Long et al., 2015), among which is the popular U-Net architecture (Ronneberger et al., 2015), Fig. 4. FCNs are a family of network architectures that use fully convolutional layers instead of fully connected ones enabling the end-to-end training of models on a pixel-to-pixel basis for dense predictions. They enable the utilization of several CNN architectures, such as VGG16 or GoogLeNet, to generate coarse maps on input images of any size. These coarse maps approximate the locations of objects in the final convolutional layers and are subsequently upsampled to the input resolution using deconvolutions (or transposed convolutions). A skip architecture was introduced to make use of a feature spectrum by adding deep, coarse, semantic information with shallow, fine, appearance information before the upsampling process. The skip connections between the encoder–decoder layers bypass some of the neural network layers and as a result, an alternative and shorter path is provided for backpropagating the error of the loss function, which contributed to avoiding the vanishing gradient problem (Krizhevsky et al., 2012). Increased connectivity in the upsampling path within FCNs and the consideration of multi-level contexts were key to improving semantic segmentation (Badrinarayanan et al., 2017; Drozdzal et al., 2016). The U-Net architecture by Ronneberger et al.

(2015), extended an FCN network with a U-shaped topology to optimize the trade-off between localization and context. The contracting path (encoder) captures a larger context using the downsampled features and the expanding path (decoder) upsamples features to their original size with the same number of layers making it a symmetric or U-shaped network. The decoder network includes the concatenation of higher resolution feature maps from the encoder network followed by convolutions, to obtain more feature information during the upsampling process.

DeepLab is another family of semantic segmentation networks, which have the ability to achieve robustness for different of classes without increasing computational complexity (Chen et al., 2014, 2017b,c, 2018). DeepLab architectures are based on FCNs but extended with the use of dilated (or atrous) convolutions, which were originally proposed by Yu and Koltun (2016), and image-level features. The atrous dilations are used within Atrous Spatial Pyramid Modules (ASPP), which perform multi-scale feature extraction by using multiple atrous convolutions with different dilation rates. As a backbone network, the latest DeepLab architecture, namely DeepLab v3+, uses the Residual Neural Network (ResNet) to produce feature maps. The module performs parallel convolution on the feature map obtained from the ResNet backbone and outputs multiple feature maps, which are then concatenated and fed into the next layer. This allows the network to capture features of multiple scales, which is crucial for tasks like semantic segmentation. ResNet is notable for its ability to overcome the vanishing gradient problem and the degradation issue, simultaneously (He et al., 2016). This breakthrough was attributable to the introduction of residual connections, which allow the network to learn residual functions, or the difference between the desired output and the current output, rather than the full function. This helps the network to learn more effectively and avoid overfitting.

4.2. Semantic segmentation in 3D

Semantic segmentation in 3D involves the partitioning of a 3D volume into segments based on their semantic meaning, with each voxel receiving a corresponding label. Since many EM datasets consist of stacked 2D sections along the z-axis, performing volumetric segmentation becomes crucial to accurately capture the 3D structure and connectivity. The 3D segmentation of neurites in EM images was set as a challenge in ISBI 2013 for predicting the segmentation of voxels using 3D segmentation methods. The major challenges in analyzing volume EM datasets are spatial complexity, misalignments or missing

---

sections due to serial sectioning and volume anisotropy, which means different voxel resolutions in different directions. For serial-section EM, the anisotropic voxel resolution is due to the slicing of thicker sections which makes the z-axis resolution lower than the x–y plane.

The problem of 3D segmentation has been tackled through three types of approaches, each offering distinct solutions to obtain the 3D form. The first involves 2D segmentation of each image in the stack, followed by 3D reconstructions based on post-processing techniques, that may range from basic watershed to complex graph cuts algorithms. The second approach is based on 3D CNNs, which can learn representations of volumetric data that include 3D spatial context. One example of such 3D CNNs is the 3D U-Net by Çiçek et al. (2016), which was inspired by the original U-Net that uses local and larger contextual information. It was then extended into the V-Net model by Milletari et al. (2016) by adding residual stages. The HighRes3DNet is another 3D CNN based on the FCN architecture, with dilated and residual convolutions, and has been successful in obtaining accurate segmentations of neuronal mitochondria (Li et al., 2017). In terms of performance, both HighRes3DNet and V-Net have achieved state-of-the-art results on several medical image segmentation benchmarks. However, HighRes3DNet has been shown to have better performance on tasks involving high-resolution and multi-modal medical images, while V-Net has been shown to be more efficient in terms of computational resources and memory usage. A variant of the 3D network is the hybrid 2D-3D methodology as proposed by Lee et al. (2015) for the segmentation of anisotropic volumes. They utilize only 2D convolutions in the initial layers that downsample the input feature maps with high x – y resolution (independent of the z-axis) until they are roughly isotropic to be efficiently processed by 3D convolutions.

Graph analysis is the third approach for 3D segmentation. Graph-based methods typically involve partitioning a graph into regions or clusters based on properties such as color or intensity values, edge strength, or other image features such as shape. These methods often use graph theory algorithms, like graph cuts or minimum spanning trees, to identify regions that are distinct from one another. This may be coupled with structure-based analysis that uses certain geometrical properties to identify boundaries between objects. Global shape descriptors were used to learn the connectivity of 3D super voxels by Lucchi et al. (2013) for segmentation using graph-cuts, addressing issues with local statistics and distracting membranes. Turaga et al. (2010) suggested how CNNs can be used for directly predicting 3D graph affinities based on a structured loss function for neuronal boundary segmentation. The proposed loss function assigned scores to the edges between adjacent pixels based on their likelihood of belonging to same or different regions and also penalized their assignment for achieving incorrect predictions that violate the underlying structure of the image.

4.3. Instance segmentation

Instance segmentation involves classifying each pixel/voxel of a given image/volume to a particular class along with assigning a unique identity to pixels/voxels of individual objects. Instance segmentation using deep learning can be divided into proposal-based (top-down) and proposal-free (bottom-up) approaches. Proposal-based approaches such as RCNN, FastRCNN, and FasterRCNN are two-stage detection networks that use a deep neural network for feature extraction (encoder) and region proposals for the segmentation of objects of interest, followed by bounding box regression and classification to obtain instance segmentation (Liu et al., 2020b). Mask-RCNN (He et al., 2017) is a popular choice for generic object instance segmentation built upon FasterRCNN, which uses a branch of the network to predict a binary mask for each object instance. Top-down instance segmentation has also been accomplished using recurrent networks with attention mechanisms, either by extracting visual characteristics and producing instance labels one item at a time or by guiding the formation of bounding boxes followed by

a segmentation network (Ren and Zemel, 2017; Ghosh et al., 2019). The Flood Filling Network (FFN) uses this concept to obtain individual object masks directly from raw image pixels (Januszewski et al., 2018) and has also been used for EM segmentation as reviewed below.

The other approach is known as proposal-free, which aims to combine semantic and instance segmentation in a bottom-up approach. This was the strategy taken by Chen et al. (2017a) and Kirillov et al. (2017), where the prediction of contours/edges of objects along with semantic masks were incorporated into FCNs in a multi-task learning approach. Both contour/edge maps and semantic masks were then fused to obtain the instance segmentation maps. Other approaches use boundary-aware instance information (e.g. the distance between object boundaries or the amount of overlap between objects) to fuse edge features with intermediate layers of the network (Bai and Urtasun, 2017; Oda et al., 2018). Another bottom-up approach was proposed by De Brabandere et al. (2017), who introduced a discriminative loss function for learning clusters of pixel embeddings and demonstrated that it is superior to the cross-entropy and Dice loss function for instance segmentation. The effect of their discriminative loss function is that the feature embeddings of the pixels that belong to the same instance are mapped close to each other in the feature space. The discriminative loss function consists of three terms: a segmentation term, which penalizes incorrect class predictions; a boundary term, which penalizes incorrect boundary predictions; and a regularization term, which encourages smoothness in the predicted masks.

5. Fully supervised methods

Fully supervised methods use annotated images (training data) to learn computational models that can segment structures in unseen images from similar distributions (test data). The training set is used by the algorithm to determine the model's parameters in such a way as to maximize the model's generalization ability. Table 3 summarizes the 33 papers (of the 38) that have used supervised learning for the semantic and instance segmentation of (sub) cellular structures.

5.1. End-to-end learning — semantic segmentation

End-to-end learning is a machine learning approach where a single model learns to perform a task without relying on pre-defined intermediate steps or features. Instead, the model is trained to map the input data directly to the desired output, in a single end-to-end process. End-to-end learning has become increasingly popular in recent years due to advances in deep learning, which allow the creation of models with large numbers of layers that can learn complex representations of data. These models are trained using backpropagation, a method for updating the weights of the model based on the error generated by a given loss function between the predicted output and the true output, which allows the model to improve its performance during the learning process.

The 16 papers that fall within this category are focused on the semantic segmentation of two main cellular structures, namely NM — neuronal membranes (8 papers) and M — mitochondria (5 papers). Other structures include N — nuclei, NE — nuclear envelopes, and L — lysosome.

Neuronal membrane segmentation refers to the process of identifying and separating the neuronal membrane from other structures in an EM image. Segmenting neuronal membranes in EM volumes helps partition an image into distinct regions that represent different neuronal cells and processes. It is essential for studying the function of neurons along with their synaptic connections for understanding the different signaling pathways in the brain. Digital reconstruction or tracing of 3D neurons depends on the accuracy of neuronal membrane segmentation as discontinuities could lead to merge and split errors (see Section 7), which in turn affect the reconstruction.

Table 3

The list of 33 (out of 38) papers reviewed in this work that are based on fully supervised learning frameworks with 2D and 3D CNN architectures applied to both semantic and instance segmentation. The abbreviation Org. stands for the studied organelle/s. The Type (2D and/or 3D) column indicates the type of methods used and problems addressed. The studies that are marked as both 2D and 3D use a 2D backbone method coupled with some post-processing operations for 3D reconstruction. The other studies that are flagged as 2D or 3D only, use 2D or 3D only backbones to address 2D or 3D problems, respectively. The numbers in the Datasets column serve as correspondences to the identifiers in Table 2, and the definitions of the performance metrics are presented in Section 7.

Citation	Org.	Type		Datasets	Performance	Backbone[b]	Main methodological components
		2D	3D		metrics		
End-to-end learning — semantic segmentation							
Fakhry et al. (2017)	NM	✓	✓	1, 3	RE, WE, PE	2D FCN	Residual blocks, deconvolutions
Oztel et al. (2017)	M	✓	✓	1	Acc, P, R, F1, JI	2D FCN	Block processing, Z-filtering
Chen et al. (2017d)	MT, M, PM, V	✓		21	No evaluation	2D FCN	A CNN architecture with four layers
Xiao et al. (2018b)	NM	✓	✓	1	V^{Rand}, V^{Info}	2D FCN	Residual blocks, multi-level features
Casser et al. (2018)	M	✓		4, 9	Acc, P, R, JI	2D U-Net	Few parameters, light-weight model
Jiang et al. (2019)	N	✓		Private[a]	JI, Acc	2D FCN	Residual, atrous, multi-level fusion
Cao et al. (2020)	NM	✓		1	V^{Rand}	2D U-Net	Dense blocks, summation-skip
Quan et al. (2021)	NM	✓		1	V^{Rand}, V^{Info}	2D U-Net	Residual, summation-skip, multi-stage
Spiers et al. (2021)	NE	✓	✓	17	P, R, F1	2D U-Net	Tri-axis prediction
Cheng and Varshney (2017)	M		✓	8	P, R, JI	3D U-Net	Factorized convolutions
Lee et al. (2017)	NM		✓	3	RE	3D U-Net	3D graph affinity, hybrid 2D-3D, residual
Xiao et al. (2018a)	M		✓	5, 8	JI, DSC	3D U-Net	Hybrid 2D-3D, residual, auxiliary supervision
Funke et al. (2018)	NM		✓	2, 10, 15	V^{Info}, CREMI	3D U-Net	3D graph affinity prediction
Heinrich et al. (2018)	S		✓	2	CREMI	3D U-Net	Signed distance regression map, hybrid 2D-3D
Mekuč et al. (2020)	M, L		✓	13	TNR, R, DSC	3D FCN	HighRes3DZMNet, zero-mean, residual/atrous
Heinrich et al. (2021)	Many		✓	10	DSC	3D U-Net	Multi-class segmentation
Bailoni et al. (2022)	NM		✓	2	ARAND	3D U-Net	Signed 3D graph affinity prediction
End-to-end learning — instance segmentation							
Liu et al. (2020a)	M	✓		8	Acc, P, R, JI, DSC	Mask-RCNN	Recursive network, multiple bounding boxes
Yuan et al. (2021)	M	✓	✓	4, 8	JI, DSC, AJI, PQ	2D U-Net	Hierarchical view ensemble module, multi-task
Luo et al. (2021)	M	✓		4, 8	JI, DSC, AJI, PQ	2D U-Net	Residual blocks, two-stage, shape soft-labels
Wei et al. (2020)	M		✓	6, 8	JI, AP-75	3D U-Net	Mask, contour prediction, watershed
Abdollahzadeh et al. (2021)	A, N		✓	19	V^{Info}, ARAND	3D U-Net	Shape-based postprocessing
Lin et al. (2021)	N		✓	7	AP-50, AP-75, AP	3D U-Net	Hybrid 2D-3D module, residual blocks
Li et al. (2022)	M		✓	6	JI, DSC, AP	3D FCN	Hybrid 2D-3D module, multi-scale
Mekuč et al. (2022)	M		✓	13	TPR, TNR, JI, DSC	3D FCN	HighRes3DzNet, geodesic active contours
Ensemble learning — semantic segmentation							
Zeng et al. (2017)	NM		✓	3	RE	3D FCN	Hybrid 3D-2D, residual/inception/atrous
Haberl et al. (2018)	NM, M, N, V		✓	16, 22	A, P, R, F1	3D FCN	Hybrid 3D-2D, residual/inception/atrous
Guay et al. (2021)	C, M, GN		✓	18	Mean JI	3D U-Net	Hybrid 2D-3D, spatial pyramids
Khadangi et al. (2021b)	M	✓		12, 16	Acc, TPR, TNR, F1, JI, V^{Rand}, V^{Info}	2D U-Net	Ensemble of different networks
Transfer learning — semantic segmentation							
Dietlmeier et al. (2019)	M	✓		1	Acc, P, F1	VGG	Few shot, hypercolumn features, boosting
Bermúdez-Chacón et al. (2018)	M	✓		Private[b]	JI	2D U-Net	Deep domain adaptation, two-stream U-Net
Configurable networks — semantic segmentation							
Isensee et al. (2019)	S	✓	✓	2	CREMI	2D, 3D U-Net	nnU-Net, self-configuring method
Franco-Barranco et al. (2022)	M	✓	✓	4, 8	JI	2D, 3D U-Net	Stable networks, blended output, *z*-filtering

[a] Private indicates that the dataset used is not publicly available.

[b] The term U-Net is used to describe extended mechanisms that utilize U-shaped architectures, whereas other mechanisms are commonly referred to as FCN.

Similarly, mitochondria segmentation is the process of identifying and separating mitochondria, a type of organelle found in eukaryotic cells, from other structures in an EM image. Mitochondria segmentation is a challenging task due to the variability in their size, shape, and distribution within cells. Accurately segmenting mitochondria in 2D and 3D is important for studying the structure and function of these organelles, as well as investigating their role in various diseases.

Below we categorize the proposed approaches based on their underlying 2D or 3D CNN architectures.

5.1.1. Approaches based on 2D CNNs

Successes of DL networks for segmentation in EM were achieved using 2D architectures with deep contextual networks. These networks

are based on the FCN architecture, with many using its symmetric U-Net version. Their ability to capture larger receptive fields using deeper networks and integration of sufficient low-level information for pixel localization during the decoding process facilitates accurate prediction. Consequently, the need for a multi-step post-processing approach to attain precise 2D segmentation and subsequent 3D reconstruction based on the segmented regions is significantly reduced.

Furthermore the Residual Deconvolutional Networks (RDN) by Fakhry et al. (2017) extended the deconvolution network, by introducing residual connections between several stacks of convolutional and deconvolutional layers in the encoder and decoder respectively. The several unpooling and deconvolutional operations at the decoding stage of the deconvolution network are said to capture the shape information of multi-scale objects effectively (Noh et al., 2015). Additionally the authors employed summation-skip connections to fuse low-resolution feature maps with their corresponding resolution in the upsampled features, thereby achieving high-resolution pixel accuracy. The proposed method was evaluated on the ISBI 2012 and 2013 benchmark datasets and compared to several state-of-the-art segmentation methods. The results demonstrated that RDNs were superior in terms of segmentation accuracy and required a simple post-processing step such as watershed to segment/reconstruct neural circuits.

Oztel et al. (2017) introduced a highly effective method to reconstruct mitochondria from 2D segmentations. An FCN was used for delineating mitochondria from the background followed by median filtering along the *z* direction in the volume of images. Also known as *z*-filtering, this technique facilitated the removal of erroneous strokes and the recovery of regions of interest in cases where neighboring slices contained the missed component.

The deep contextual residual network (DCR) by Xiao et al. (2018b) is an extension of FCN with residual blocks and multi-scale feature fusion. They used the summation based skip connections which fuse high-level details from output of deconvolutions in the decoder and low-level information from ResNet encoder. The proposed post-processing method with a multi-cut approach and 3D contextual features proved important to reduce discontinuities (boundary splits or merges), which in turn helped to reduce false positives and false negatives in various 2D sections. DCR outperformed several state-of-the-art segmentation methods on the ISBI 2012 dataset.

Advanced networks for different tasks may be too computationally demanding to run on affordable hardware, leading users to modify macro-level design aspects. Examples of such modifications include downsampling input images and reducing network size or depth to ensure compatibility with computer hardware constraints. Casser et al. (2018) introduced a fast mitochondria segmentation method using a reduced number of layers and lightweight bilinear upsampling instead of transposed convolutions in the decoder of U-Net. Moreover, they introduced a novel data augmentation method that generates training samples on the fly by randomly applying spatial transformations to the original images, which leads to increased training efficiency and robustness to variations in image quality. Similar data augmentation operations also featured in some of the other reviewed studies, which employed cropping, flipping, rotations, scaling, resampling (Quan et al., 2016) and elastic deformations (Ronneberger et al., 2015). Casser et al. (2018) also incorporate a post-processing step based on *z*-filtering to reconstruct 3D mitochondria. The proposed approach was evaluated on several EM datasets and achieved state-of-the-art performance in terms of segmentation accuracy and speed.

A residual encoder module with ASPP for multi-scale contextual feature integration was investigated by Jiang et al. (2019). The decoder module included the fusion of previous low-level features and high-level features from the output of ASPP, followed by bi-linear upsampling to obtain the segmentation map. They achieved better performance compared to the baseline, U-Net, and Deeplabv3+ for the segmentation of cell bodies and cell nuclei.

The Dense U-Net model was proposed by Cao et al. (2020) as an extension of the popular U-Net architecture that incorporates densely connected blocks within the U-Net's skip connections. The densely connected blocks help to improve gradient flow and feature reuse, which leads to better feature representation and higher segmentation accuracy. Besides its outstanding results on the ISBI 2012 challenge, the model turned out to be highly robust to variations in noises and artifacts of neuronal membrane images, requiring no further post-processing.

FusionNet is a fully residual U-Net architecture that combines different levels of feature representations by fusing the output of multiple sub-networks with different receptive fields. It includes a residual learning framework along with deconvolutional layers to improve the training convergence and segmentation accuracy. The study by Quan et al. (2021) showed that an integrated multi-stage refinement process using four concatenated FusionNet units can effectively eliminate the requirement for any proofreading.[6]

A novel data augmentation strategy was also proposed by Spiers et al. (2021), which simulates realistic variations in the EM images to improve the robustness of their 2D CNN for the semantic segmentation of nuclear envelopes. The proposed approach based on 2D U-Net achieved high segmentation accuracy and can be used to extract meaningful biological information from the segmented nuclear envelope, such as the distribution of nuclear pores. Their model was run on each axis after transposing the stack, and the resulting three orthogonal predictions were merged to produce the ultimate segmentation.

Chen et al. (2017d) used a 2D CNN with only four layers for the segmentation of membranes, mitochondria, vesicles, and microtubules in cryo-ET. The architecture of the CNN layers was optimized to capture a large context by utilizing 15×15 pixel kernels in the first two layers. This design allowed for the use of a single max-pooling layer to downsample the output to half the input resolution, which aids in distinguishing intricate details of structures such as single (vesicle, microtubule) or double membrane (plasma membrane, mitochondria). A CNN for each of the four structures was trained with a few sections of the tomogram containing structures of interest. Subsequently, the obtained segmentation maps were employed for sub-tomogram classification and averaging, facilitating the determination of in-situ structures for the molecular components of interest.

5.1.2. Approaches based on 3D CNNs

Similar to 2D deep architectures, a 3D CNN consists of multiple layers of filters, including convolutional, pooling, and activation layers, to learn spatial features from the input data. The filters scan the input volume at different locations and scales to identify features that are relevant for segmentation. The key difference between 2D and 3D CNNs is the inclusion of an additional depth dimension in the input data. This allows the network to capture the spatial and depth relationships between adjacent slices in the volume. Due to the large amount of data and computational resources required for training 3D CNNs, such methods are typically used in high-end computing environments, such as specialized workstations or cloud computing platforms. Hybrid 2D-3D architectures have also been investigated that try to find the right trade-off between high computational demand and effectiveness.

In this review, there are three approaches that adopted complete 3D CNN architectures in a fully supervised way. The first is the work by Cheng and Varshney (2017) who proposed a 3D CNN for the segmentation of mitochondria in volumetric data. The authors also propose a novel data augmentation technique that uses stochastic sampling in the pooling layers to generate realistic variations in the feature space. In their thorough investigation, they conclude that the 3D CNNs outperform their 2D counterparts with a high statistical significance. The

---

[6] Proofreading refers to the manual validation of segmented (manual or automatic) image data.

improvement was mainly attributable to the introduced augmentations as well as to the factorized convolutions which not only permitted high efficiency, but also proved to be useful in FIB-SEM (isotropic) volumes.

Mekuč et al. (2020) also presented a 3D CNN-based method for the segmentation of mitochondria and endolysosomes in volumetric EM. The proposed method is based on the HighRes3DNet architecture, but it has the filters in the first layer constrained to having zero mean, and called it HighRes3DZMNet. The zero mean layer made the neural network robust to changes in the brightness of the volume inputs. The network is trained using the UroCell dataset for jointly segmenting mitochondria and endolysosomes due to similar morphologies of these biological structures. The method was also applied to segment mitochondria in the Lucchi++ dataset and achieved state-of-the-art segmentation results for FIB-SEM volumes.

Heinrich et al. (2021) also relied on a 3D CNN for the segmentation of 35 organelle classes in cells from FIB-SEM volumes. The multi-channel 3D U-Net was trained on 28 volumes from the open-source OpenOrganelle collection covering four different cell types. They investigated how one segmentation model that is trained with samples of all 35 organelles compares with more specific models that are trained with subsets of semantically-related organelle classes, such as the endoplasmic reticulum (ER) and its associated structures, namely ER exit sites, ER membrane, and ER lumen. It turned out, that the single model that is trained by all classes outperforms the more specific ones. This is attributable to the richer diversity in the training set which resulted in a model with better generalization abilities.

Hybrid 2D-3D approaches were adopted for the segmentation of volume datasets in order to reduce the computational cost of 3D convolutions in certain layers and achieve better convergence. Their main application lies in the ability to segment anisotropic volumes for efficiently processing their 3D context. For instance, both anisotropic and isotropic EM volumes could be processed using hybrid 2D-3D network architectures that include $3 \times 3 \times 1$ convolutions instead of $3 \times 3 \times 3$ to modify them to 2D ones. Xiao et al. (2018a) was the first to introduce a fully residual hybrid 2D-3D network with deep supervision to improve mitochondria segmentation. For reducing the number of parameters, 3D convolutions were used only in the first and last layers of a 3D U-Net. A deeply supervised strategy was proposed by injecting auxiliary branches into the initial layers of the decoder for avoiding the vanishing gradients problem. The fully residual architecture based on hybrid modules could efficiently handle anisotropic volume data in order to predict a correctly segmented output. As a result, a simple connected component analysis method was effective for 3D reconstruction on both isotropic and anisotropic EM datasets.

Lee et al. (2017) adapted the hybrid 2D-3D model of Turaga et al. (2010) to predict 3D affinity maps for the segmentation of neuronal membranes in 3D volumes. The proposed CNN model incorporated multi-slice inputs along with long-range affinity-based auxiliary supervision along the x, y, and z directions. The process of long-range affinity monitoring involves utilizing a larger affinity neighborhood as an auxiliary objective to improve the accuracy of the main task, the nearest-neighbor affinity prediction. Long-range affinities were assigned to membrane voxels and voxels further apart by connecting them with extended edges. They utilized a hybrid 2D-3D U-Net for segmenting anisotropic volumes and post-processing with a simple mean-affinity agglomeration strategy for segmenting neuronal regions. The proposed affinity supervision simulates the use of boundary maps with different thicknesses in the DeepEM3D (Section 5.3), outperforming it in the SNEMI3D competition.

A structured loss that favors high affinities between 3D voxels was used to obtain topologically correct segmentations by Funke et al. (2018). The affinity predictions were accurate enough to be used with a simple agglomeration to efficiently segment both isotropic and anisotropic (CREMI, FIB, and SegEM) data, outperforming methods with more elaborate post-processing pipelines. Bailoni et al. (2022) used signed graphs to anticipate both attractive and repulsive forces

among 3D voxels, enabling graph prediction through a 3D U-Net, in a manner similar to the method proposed by Funke et al. (2018).

Building on the concept of long-range affinities for boundary detection, Heinrich et al. (2018) used neighboring context to predict a signed distance transform of the binary synapse labels. They assigned positive distances to the pixels in the synapse region and negative distances to the exterior pixels, relative to the boundary of the binary mask. The proposed approach gathered information from a broader context by transforming the voxel-wise classification into a voxel-wise regression problem. The distance predictions, when thresholded, generated precise binary segmentations for synapses. Such distance prediction maps with simple thresholding allowed scaling the prediction at high-throughput speeds (3 megavoxels per second) for a full adult fly brain volume of 50 teravoxels in size.

5.2. End-to-end learning — instance segmentation

End-to-end learning approaches are also the most popular ones for instance segmentation, which require the delineation of each instance within the same class of structures. This is particularly important for classes of structures that tend to be apposed with each other, such as mitochondria.

CNN-based methods for instance segmentation were grouped into two categories by Wei et al. (2020): top-down and bottom-up. Top-down methods typically utilize region proposal networks followed by precise delineation in each region. Conversely, bottom-up approaches aim to predict a binary segmentation mask, an affinity map, or a binary mask with instance boundary followed by several post-processing steps to distinguish instances. Due to the undefined scale of bounding boxes in EM images, bottom-up approaches have been the preferred methodology for 2D and 3D instance segmentation.

For neuronal region segmentation, instance segmentation is essentially transformed into an image partitioning task where every pixel in the image is assigned to a specific instance, thereby forming partitions. Each partition represents an instance of a neuronal region. This partitioning approach enables the reconstruction of individual neuronal structures through post-processing techniques. Fig. 5 shows examples of semantic and instance segmentation of mitochondria along with an illustration of neuronal 3D reconstruction after image partitioning.

5.2.1. Approaches based on 2D CNNs

The only top-down approach from the reviewed works in this paper is the one proposed by Liu et al. (2020a). They introduced a pipeline that complements Mask-RCNN. In particular, they proposed a mechanism that refines undersegmented mitochondria in the output of Mask-RCNN, by iteratively enhancing the field of view that preserves the previous segmentation states. They systematically demonstrated that their approach outperformed competing methods that rely on U-Net, FFN, and Mask-RCNN in instance segmentation of mitochondria.

Shape prior turned out to be important for some techniques to improve the quality of instance segmentation. Shape prior refers to the incorporation of prior knowledge about the expected shape or structure of an object of interest into segmentation algorithms. For example, Yuan et al. (2021) proposed the Hive-Net CNN, which was designed to overcome the challenges posed by the high variability in mitochondria shapes and sizes, as well as the presence of other cellular structures in the images. The network consists of multiple view-specific sub-networks that process different views of the image, and a centerline-aware hierarchical ensemble module that combines the outputs of the sub-networks to generate the final segmentation result. The centerline-aware module uses a new type of loss function that encourages the network to learn the topology of mitochondria and to segment them along their centerlines. The proposed network was evaluated on two publicly available datasets, and an ablation
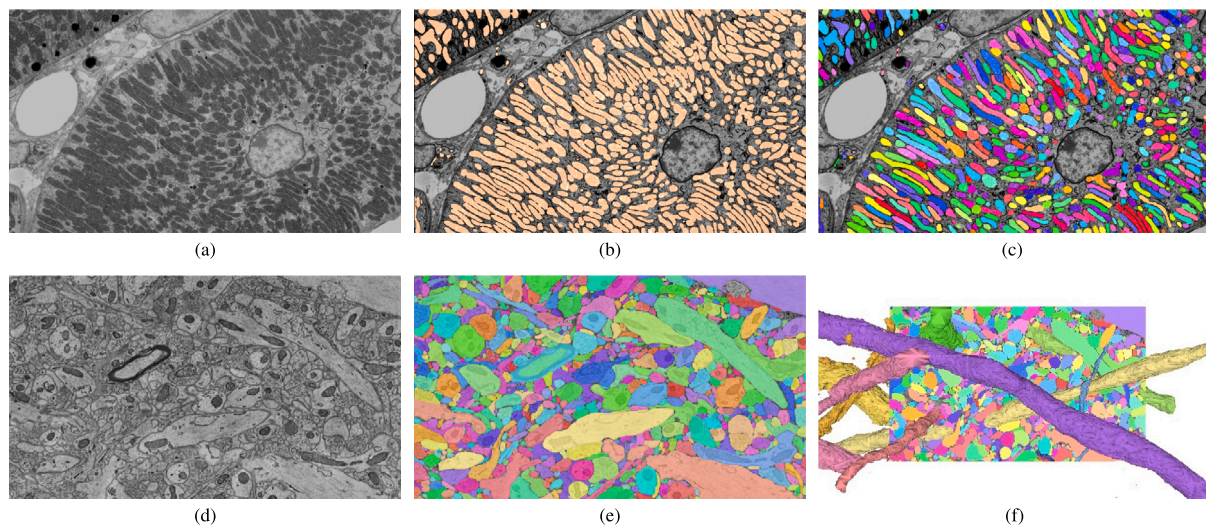
Fig. 5. Example of (top row) semantic and instance segmentation of mitochondria and (bottom row) neuronal membrane segmentation followed by 3D reconstruction of neuronal objects from a volumetric EM image. (a) Raw EM 2D section extracted from a FIB-SEM volume of a mouse kidney from the OpenOrganelle jrc_mus-kidney dataset.[7] (b, c) Ground truth labels for semantic and instance segmentation. The instance segmentation map identifies each individual mitochondria with a unique color. (d) Raw EM 2D section extracted from the SNEMI3D (#3) dataset for the task of neuronal membrane segmentation and reconstruction. (e) The ground truth map of the neuronal membrane segmentation, which is used to partition the image completely. (f) 3D reconstruction of selected neuronal structures that pass through the given 2D section from adjacent sections of the EM volume. The information from multiple images is used to create a 3D reconstruction through various post-processing methods, such as clustering, watershed, or graph-based methods.

study concluded that the centerline-aware module and the view-specific sub-networks were critical for achieving high segmentation accuracy.

Shape information has also been exploited by the hierarchical encoder–decoder network (HED-Net) for the instance segmentation of mitochondria (Luo et al., 2021). That strategy used the shape information available in the manual labels to train the model more effectively. Instead of relying solely on the ground truth label maps for model training, an additional subcategory-aware supervision was introduced. That was achieved by decomposing each manual label map into two complementary label maps based on the ovality of the mitochondria. The resulting three-label maps were used to supervise the training of the HED-Net. The original label map was used to guide the network to segment all mitochondria of varying shapes, while the auxiliary label maps guided the network to segment subcategories of mitochondria with circular and elliptic shapes, respectively. The experiments conducted on two publicly available benchmarks show that the proposed HED-Net outperforms state-of-the-art methods.

The inclusion of apriori knowledge about shape in segmentation algorithms contributes to increased specificity as they become more selective in delineating the structures of interest and keep false positives to a minimum. They can also improve generalization ability especially when the training data is limited. Methods that use shape priors, however, are more structure-specific and, therefore, different methods may need to be designed for the segmentation of distinct organelles.

5.2.2. Approaches based on 3D CNNs

The largest instance segmentation dataset for mitochondria (MitoEM) proposed by Wei et al. (2020) benchmarks the dataset by proposing a 3D U-Net. It is trained with binary masks and contours using two separate decoders, followed by a marker-controlled watershed to obtain instance segmentations, and is called U3D-BC +MW for short. Wei et al. (2020) introduced two networks, MitoEM-R and MitoEM-H, citing variations in sizes, shapes, and noise content for serial sections from rat and human samples. The MitoEM-R network can generalize on the human dataset as the rat samples have complex mitochondrial

morphologies. The simpler U3D-BC +MW method was shown to be more effective than FFNs, as they were not able to capture the fine geometry of mitochondria with complex shapes or in close contact to each other.

The DeepACSON approach by Abdollahzadeh et al. (2021), which was proposed for the instance segmentation of axons and nuclei in 3D volumes, is supported by a postprocessing method that relies on shape features. To correct for topological errors of axons, a cylindrical shape decomposition (CSD) algorithm is used as a postprocessing step to identify any erroneously detected axons and to correct under-segmented ones at their cross-overs. The CSD is a shape-analysis algorithm that decomposes an object into its semantic components based on the object's skeleton curve and cross-sectional analysis. The CSD slices objects at their cross-overs based on geometrical changes in cross-sectional shapes and then reconstructs semantic objects from the cut sections using generalized cylinders. A generalized cylinder is a solid object created by sweeping a 2D contour along a curve in space, allowing for varying cross-sections along its length. The circularity of the cell nucleus is corrected using the level-set-based geometric deformable model, which approximates the initial shape of the object with a curve. This is then adjusted to minimize an energy function associated with the curve when it fits perfectly to the object's boundaries. Energy functions enable the inclusion of shape information, whether it is a vague concept like smoothness constraints or a precise idea like shape constraints (strict adherence to a particular shape).

Nuclei instance segmentation on a large-scale EM dataset was proposed by Lin et al. (2021). Their network, U3D-BCD, was inspired by the U3D-BC above but involved the additional learning of a signed Euclidean distance transform map along with foreground masks and instance contours to capture the structure of the background for segmentation. The Euclidean distance transform calculates the distance of each pixel in a binary image to the nearest boundary pixel. If a pixel is part of a foreground object then it has a positive Euclidean distance, otherwise negative. To locate the seeds for object centers, their method starts by thresholding the predictions to identify markers with high foreground probability and distance value, but low contour probability. Next, the marker-controlled watershed transform algorithm is applied with the predicted distance map and seeds to generate masks. This approach has two advantages over the U3D-BC model (Wei et al., 2020), which also utilizes marker-controlled watershed transform for

---

[7] https://open.quiltdata.com/b/janelia-cosem-datasets/tree/jrc_mus-kidney/.

decoding. Firstly, the consistency among the three representations is used to locate the seeds, which makes it more robust than the U3D-BC method that relies only on two predictions. Secondly, it uses the smooth signed distance map in the watershed decoding process, which is more effective in capturing instance structure than the foreground probability map used in U3D-BC.

Li et al. (2022) addressed 3D mitochondria instance segmentation with two supervised deep neural networks, namely ResUNet-H and ResU-Net-R, for the rat and human samples on the MitoEM dataset, respectively. Both networks produce outputs in the form of a semantic and instance boundary masks. Due to the increased difficulty of the human sample, Res-UNet-H has an additional decoder path to separately predict the semantic mask and instance boundary, while Res-UNet-R has only one path. Once the semantic mask and instance boundary are obtained, a seed map is synthesized, and the mitochondria instances are obtained using connected component labeling. To enhance the networks' segmentation performance, a simple but effective anisotropic convolution block is designed, and a multi-scale training strategy is deployed. The MitoEM dataset has sparsely distributed imaging noise, with the human sample having a stronger subjective noise level than the rat sample. To reduce the influence of noise on segmentation, an interpolation network was utilized to restore the regions with noise, which were coarsely marked by humans. Besides mitochondria instance segmentation, the proposed method was demonstrated to have superior performance for mitochondria semantic segmentation.

Mekuč et al. (2022) extended their previous approach based on the HighRes3DZMNet with post-processing steps based on active contours, to separate apposing mitochondria and thus achieve instance segmentation. By means of experiments on the extended UroCell dataset, they demonstrated that this new approach is more effective than the U3D-BC +MW method.

5.3. Ensemble learning

Ensemble learning methods combine outputs of multiple algorithms or models to obtain better predictive performance in terms of accuracy and generalization. Pixel- or voxel-wise averaging and the majority or median voting are among the main aggregation methods.

An ensemble technique was in fact investigated by Zeng et al. (2017) for the segmentation of neuronal membranes in the brain volumes. They trained several variations of their DeepEM3D network, which could process different numbers of input slices and inputs with varying thicknesses of object boundaries. The DeepEM3D network extended the FCN architecture by introducing a hybrid network with 3D convolutions in the first two layers to enable integrating anisotropic information in the early stages, and 2D layers afterwards. DeepEM3D employed inception and residual modules, multiple dilated convolutions, and combined the result of three models that integrated one, three, and five consecutive serial sections. Employing an ensemble strategy for enhancing boundaries (by maximum superposition) within the probability maps generated by these models proved essential for performing with near-human accuracy in the SNEMI3D challenge.

CDeep3M is a cloud implementation of DeepEM3D to segment various anisotropic SBF-SEM and cryo-ET datasets (Haberl et al., 2018). Trained by a few sub-volumes of the cryo-ET tomogram, the resulting network was able to segment vesicles and membranes with high accuracy in other tomograms. The network implementation proved efficient for segmenting large-volume EM datasets such as SBF-SEM making it easier to analyze enormous amounts of imaging data.

The strengths of the ensemble paradigm was also confirmed by Guay et al. (2021) for the segmentation of cytoplasm, mitochondria, and four types of granules in platelet cells. They demonstrated that the best segmentation performance (in terms of intersection over union) was achieved by combining the output of the top k performing weak classifiers, with each such classifier learned by a small portion of the training data. Similar to above, each model was a hybrid 2D-3D

network used to segment anisotropic SBF-SEM volumes. They also highlighted that besides its effectiveness, their ensemble paradigm ensured better reproducibility of the results in comparison to individual models that were sensitive to initialization.

Multiple network outputs were also combined with a workflow for binary EM segmentation provided by the EM-stellar platform (Khadangi et al., 2021b). Unlike the above two approaches, Khadangi et al. (2021b) used the ensemble paradigm to aggregate the output of different types of networks, namely CDeep3EM (Haberl et al., 2018), EM-Net (Khadangi et al., 2021a), PReLU-Net (He et al., 2015), ResNet, SegNet, U-Net, and VGG-16. A cross-evaluation using a heatmap of different evaluation metrics revealed that no single deep architecture performs consistently well across all segmentation metrics. This is why ensemble approaches have an edge over individual methods as they use the strengths of each underlying model as was demonstrated in the evaluation of two different datasets for mitochondria segmentation in cardiac and brain tissue.

5.4. Transfer learning

Transfer learning is a framework that adapts the knowledge acquired from one dataset to another, and is generally used when an application has an insufficient amount of training samples. A pre-trained model is fine-tuned, usually in the final layers, with the training samples of a new dataset. This technique was used by Mekuč et al. (2020) for the segmentation of mitochondria and endolysosomes from the background in EM images. Since mitochondria and endolysosomes share similar texture and mitochondria are more in abundance a binary segmentation model was first learned to segment mitochondria from the background. Subsequently, transfer learning was used to adapt the learned model for the segmentation of endolysosomes too. This was achieved by freezing all layers of the network except for the last one, which was fine-tuned by a smaller training set that included examples of endolysosomes. This approach is a demonstration how transfer learning can be used when the availability of a certain structure is limited.

Fine-tuning a pre-trained network comes with the risk of overfitting to the few labeled training examples of the new dataset or application. This challenge has opened up new research avenues, namely few-shot learning and domain adaptation. The former can be a meta-learning approach that "learns to learn" from a given pre-trained model when conditioned on a few training examples (referred to as the support set) to perform well on new queries passed through a fixed feature extractor (Shaban et al., 2017).

Few-shot learning was the focus of the work by Dietlmeier et al. (2019), who proposed a few-shot hypercolumn-based approach for mitochondria segmentation in cardiac and outer hair cells. The idea behind hypercolumn feature extraction was to extract features from different levels of a pre-trained CNN and combine them to form a single, high-dimensional feature representation for each pixel. The VGG-16 model pre-trained on the ImageNet dataset was used to extract hypercolumns, which were then passed through a linear regressor for actively selecting features. Only 20 labeled patches (2%–98% train-test split) were used from a FIB-SEM stack for training a gradient-based boosting classifier (XGBoost). They showed how high segmentation accuracy on the Drosophila VNC dataset could be achieved by actively selecting features and learning using far less training data and even by using a single training sample (single-shot).

Domain adaptation is another form of transfer learning, where the source and target datasets share the same labels (classes) but have a different data distribution. Changes in data distribution can be due to slightly different experimental parameters during EM imaging or due to the imaging of different tissue types or body locations. Bermúdez-Chacón et al. (2018) proposed the two-stream U-Net architecture, where the weights are related, yet different for each of the two domains, for supervised training on a few target labels. Only 10% of labeled target data was required for domain adaptation to achieve state-of-the-art performance when compared to a U-Net trained on a fully annotated dataset.

Table 4

The list of 5 (out of 38) papers reviewed in this work and that are based on semi-, un- and self-supervised learning frameworks. The abbreviation Org. stands for the studied organelle/s. The Type (2D and/or 3D) column indicates the type of methods used and problems addressed. The studies that are marked as both 2D and 3D use a 2D backbone method coupled with some post-processing operations for 3D reconstruction. The other studies that are flagged as 2D or 3D only, use 2D or 3D only backbones to address 2D or 3D problems, respectively. The numbers in the Datasets column serve as correspondences to the identifiers in Table 2, and the definitions of the performance metrics are presented in Section 7.

Citation	Org.	Type		Datasets	Performance	Backbone	Main methodological components
		2D	3D		metrics		
Semi-supervised learning — The superscripts S and I indicate semantic and instance segmentation							
Takaya et al. (2021)^S	NM	✓		1	V^{Rand}, V^{Info}	2D FCN	Sequential semi-supervised learning
Wolny et al. (2022)^I	M	✓		1, 6	AP-50, AP	2D U-Net	Positive unlabeled, momentum encoder
Unsupervised learning — Semantic segmentation							
Bermúdez-Chacón et al. (2019)	M, S	✓		1, 3, 8	JI	2D U-Net	Two stream U-Net, domain adaptation
Peng et al. (2020)	M	✓		3, 8	JI, DSC	2D U-Net	Domain discriminators for adversarial loss
Self-supervised learning — Semantic segmentation							
Conrad and Narayan (2021)	M	✓	✓	2, 4, 8, 10, 13, 16	JI	3D U-Net	Self-supervised learning, fine-tuning

5.5. Configurability and reproducibility

A key challenge in designing CNNs is the determination of the right architecture for the problem at hand. This has motivated research effort in what are known self-configurable networks that can automatically determine certain design choices. A self-configurable network is thus a type of artificial neural network that is capable of dynamically adapting its structure and parameters based on the input data and task concerned. This concept was used by Isensee et al. (2019), who proposed the no-new-UNet (nnU-Net) framework that consists of a 2D U-Net, 3D U-Net and a cascade of two 3D U-Nets. Self-configuration based on cross-validation was used to automatically determine some hyperparameters, such as the patch size, batch size and number of pooling operations. While it was shown to be very effective in various semantic segmentation problems in medical image benchmark datasets, its generalization ability in EM datasets has yet to be evaluated thoroughly.

An experimental study by Franco-Barranco et al. (2022) uncovered substantial reproducibility issues of different networks proposed for mitochondria segmentation in EM data. Additionally, it distinguished the impact of innovative architectures from that of training choices (such as pre-processing, data augmentation, output reconstruction, and post-processing strategies) by conducting multiple executions of the same configurations. Their systematic analysis enabled the identification of stable and lightweight models that consistently deliver state-of-the-art performance on publicly available datasets.

## 6. Semi-, un- and self-supervised methods

Semi-supervised and unsupervised learning are two types of machine learning methods, whose main difference is in the amount of labeled data used to train the model.

Unsupervised learning is a type of machine learning that deals with finding patterns and relationships in unlabeled training data. In this case, the algorithm learns to identify patterns and relationships in the data by clustering or grouping similar data points together. Semi-supervised learning, on the other hand, is a combination of supervised and unsupervised learning. It uses both labeled and unlabeled data to train the model. The labeled data is used to train the model on specific tasks, while the unlabeled data is used to help the algorithm learn patterns and relationships in the data (Zhu and Goldberg, 2009). In self-supervised learning, a model is trained on a dataset with labels that are automatically generated from the data itself. The goal is to learn useful representations of the data that can be used for downstream tasks, such as segmentation.

A common strategy for semi-supervised learning is to use label propagation through self-training. The process begins by training a classifier on labeled samples and then classifying the unlabeled samples. A selection of these samples based on an active selection strategy or learned classifier is then added to the training set and the process is repeated multiple times (Cheplygina et al., 2019). This can be performed either inductively or transductively. The former refers to training a model on unseen targets to add new information to the previously trained model so that it can generalize on new unseen data, and the latter to training a model based on a selected subset of labeled and unlabeled data to be able to predict correctly on a limited set of seen targets. Table 4 summarizes the 5 papers (of the 38) that have employed semi, un- and self-supervised approaches for the semantic and instance segmentation of (sub) cellular structures.

A semi-supervised approach was proposed by Takaya et al. (2021) for the segmentation of neuronal membranes. They called their approach 4S that stands for sequential semi-supervised segmentation. It was based on the fact that adjacent images in a volume are strongly correlated. The goal of their method is to have a model that can only generalize to the next few slices instead of to the whole volume. This was achieved by starting with a few labeled slices that are used to train the first model. Then, in an iterative approach the model was used to infer the segmentation maps of a small set of subsequent images and the resulting segmentation maps were used as pseudolabels to retrain the model. Label propagation from labeled to the available unlabeled data was performed by predicting pseudo labels on the subsequent sections which represent the same targets and whose predictions could be included in the next round of model training as ground truth labels. It allowed the training to weigh the most recent inputs heavily unlike transfer learning where the goal is to generalize well on all use cases of the unlabeled dataset.

Another semi-supervised method was introduced by Wolny et al. (2022) for the segmentation of mitochondria. In contrast to the above, their goal was to train a model with a few manually annotated structures in some images, which can generalize for the whole dataset. In particular, they employed a training dataset comprising labeled (i.e. masks) samples of only a limited number of mitochondria. All unlabeled mitochondria and other unlabeled structures were treated as background. As there is no direct supervision on the unlabeled part of the image, an embedding consistency term was introduced by training two networks on different data-augmented versions of each pixel. This was coupled with a push-pull loss function that they proposed to enforce constraints between different instances. It was realized by using anchor projections in the embedding space of a point in each instance to derive a soft label based on the set of surrounding pixels in the projected space. The instance segmentation was then achieved by
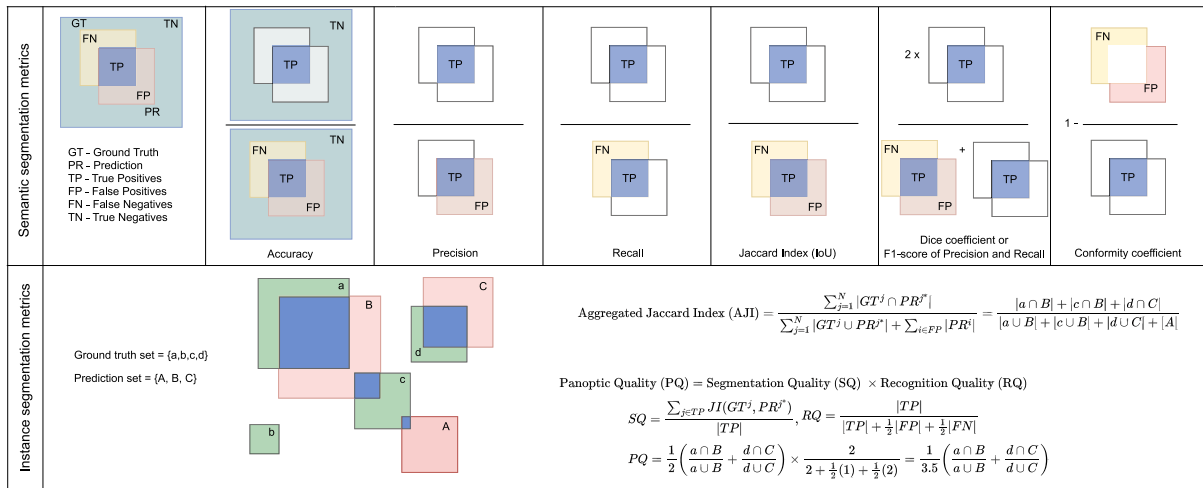
Fig. 6. Common performance metrics for segmentation methods. For semantic segmentation, the overall overlap of the ground truth (GT) mask with the prediction (PR) is compared without differentiating between objects of the foreground class. As to instance segmentation, each GT component is matched with only one PR component, the one with which it has the largest intersection. In the above example, the GT component 'c' overlaps with two PR components, 'A' and 'B', but is matched only with 'B' due to a larger overlap. The Aggregated Jaccard Index (AJI) is the ratio of the sum of all intersections of the matched pairs of GT and PR components to the sum of the unions of such pairs plus the sum of all pixels in the unmatched PR components. The Panoptic Quality (PQ) captures both semantic and instance segmentation performance. The former is the sum of all IoUs between the matched GT and PR components divided by the number of matched components (TPs), and the latter is the number of TPs divided by the number of TPs plus half of the FPs and FNs together. The symbol $|.|$ indicates the cardinality of the set concerned.

grouping the pixel embeddings. This semi-supervised method is notable for a good tradeoff between segmentation performance and effort in manual annotation.

Unsupervised learning was explored by Bermúdez-Chacón et al. (2019), who investigated the unsupervised domain adaptation strategy for mitochondria segmentation to demonstrate how a model trained on one brain structure (source: mouse striatum) could be adapted to another brain structure (target: mouse hippocampus). Labeled data was only available to train the model on the source dataset (striatum). Visual correspondences were then used to determine pivot locations in the target dataset to characterize regions of mitochondria or synapses. These locations were then aggregated through a voting scheme to construct a consensus heatmap, which guided their model adaptation in two ways: (a) optimizing model parameters to ensure agreement between predictions and their sets of correspondences, or (b) incorporating high-scoring regions of the heatmap as soft labels in other domain adaptation pipelines. These unsupervised techniques yielded high-quality segmentations on unannotated volumes for mitochondria and synapses, consistent with results obtained under full supervision, without the need for new annotation effort.

In the case of severe domain shifts such as from a FIB-SEM to an ssSEM dataset as investigated by Peng et al. (2020), adversarial learning may be used for domain adaptation in different tissues of various species. Adversarial learning is a machine learning paradigm that trains a model with an adversarial loss function that encourages the model to learn domain-invariant features. Peng et al. (2020) combined the geometrical cues from annotated labels with visual cues latent in images of both the source and target domains using adversarial domain adaptive multi-task learning. Instead of manually-defined shape priors, they learned geometrical cues from the source domain through adversarial learning, while jointly learning domain-invariant and discriminative features. By doing so, the model learned features that were useful for both source and target domains, and could perform well on the target domain despite having only labeled data in the source domain. The method was evaluated extensively on three benchmarks under various settings through ablations, parameter analysis, and comparisons, demonstrating its superior performance in segmentation accuracy and visual quality compared to state-of-the-art methods.

Contrastive learning is a self-supervised paradigm where a model is trained to learn useful representations of input data by contrasting similar and dissimilar samples. The basic idea is to take a set of positive pairs (e.g., two different augmentations of the same image) and a set of negative pairs (e.g., two images containing different types of objects), and train the model to assign higher similarity scores to positive pairs and lower similarity scores to negative pairs. This results in a model that captures the underlying structure of the data and can be used for downstream tasks like classification, object detection, and semantic segmentation. Conrad and Narayan (2021) used contrastive learning, specifically moment contrast, He et al. (2020), to learn useful feature representation from the unlabeled CEM500K dataset followed by transfer learning on given datasets. The heterogeneity of CEM500k coupled with the unsupervised initialization of a segmentation model contributed to achieving overall state-of-the-art results on six benchmark datasets that concern different types of organelles.

7. Segmentation evaluation metrics

Segmentation methods are evaluated by measuring the extent of overlap between the ground truth (GT) and prediction (PR) segmentation maps.

For semantic segmentation, all GT connected components are considered as one object, and similarly all PR connected components are treated as one object. This reduces the problem to binary classification. Typical performance measures include Accuracy, Precision and Recall and their harmonic mean, also called F-score (or $F1$ when Precision and Recall are given the same weight) or Dice similarity coefficient (DSC), the Pixel Error (PE), Jaccard Index (JI), also known as the Intersection over Union (IoU), and the Conformity coefficient (Chang et al., 2009), Fig. 6. They are defined as:

$$Accuracy\ (Acc) = (TP + TN)/(TP + FP + FN + TN)$$
$$Precision\ (P) = TP/(TP + FP)$$
$$Recall\ (R) = TP/(TP + FN)$$
$$F1\ (or\ DSC) = 2PR/(P + R) \tag{1}$$
$$Pixel\ Error\ (PE) = 1 - maximal\ F1$$
$$JI\ (or\ IoU) = TP/(TP + FP + FN)$$
$$Conformity\ (CF) = 1 - (FN + FP)/TP$$

where TP, FP, FN, and TN are the number of true positives, false positives, false negatives, and true negatives at pixel level. The Accuracy

measure is a ratio of all correctly classified pixels to all pixels. It is a simple and a good global measure but it is only suitable when the class distribution is balanced (TPs and TNs are balanced). Precision is the ratio of all TPs to the number of positive predictions made by the algorithm, and Recall (Sensitivity or True Positive Rate) is the ratio of all TPs to the number of all positive pixels in GT. The PE measure, which was used in the ISBI 2012 challenge, is the error version of the maximal $F1$-score. The maximal $F1$-score is determined by iteratively computing the P and R from the binarized output of a segmentation algorithm with different thresholds, and finally finding the combination that yields the highest $F1$-score. The JI (or IoU) and DSC measure the similarity between the predicted class labels and the true class labels, while the Conformity coefficient measures the ratio of the number of misclassified pixels to the number of true positive pixels subtracted from 1. A negative Conformity value indicates that the number of misclassified pixels is higher than the true positive ones, and vice-versa. In case there are multiple classes, as in the work by Guay et al. (2021), the mean JI is computed by first determining the JI for each class and subsequently combining all JIs by their mean.

Segmentation of neuronal regions through image partitioning involves identifying regions following membrane delineation. In such tasks, where the ground truth labels of the partitions are unavailable, measures that rely on counting the number of matches between the PR and GT maps, such as accuracy, are not suitable. Instead, the Rand Index (RI) is a more appropriate measure as it is invariant to the permutation of regions. Originally proposed in statistics for measuring the similarity between two data clusterings, the RI is defined as:

$$RI = \frac{2(a+b)}{n(n-1)} \qquad (2)$$

where a represents the number of pairs of pixels that are assigned to the same PR partition and to the same GT partition, b represents the number of pairs of pixels that are assigned to separate partitions in both PR and GT, and n represents the total number of pixels in the given image. The error version of RI referred to as the Rand Error (RE) is computed as $1 - RI$, which quantifies the degree of disagreement between the PR and GT partitions. The RI primarily assesses the overall similarity between two sets of clustered data. In image segmentation terms it considers both foreground and background partitions. The foreground-restricted RI, as used in the ISBI 2012, is a constrained version of the RI, which measures the agreements of the foreground segments only.

Another metric that was part of the ISBI 2012 challenge is the warping error (WE). WE is a measure of topological disagreements between PR and GT, which evaluates the number of topologically-relevant boundary labeling errors, including geometric labeling errors if a geometric mask is used. It provides an upper bound on the number of errors that would cause topological changes if the values of pixels in one segmentation were flipped to match the other segmentation. While small shifts affect the RI mildly, they have no affect at all on the WE. On the other hand, the PE metric solely focuses on whether a specific pixel is correctly classified as a boundary pixel, disregarding the overall impact of that prediction on the resulting topology. For instance, manipulating the boundary between two neurons through expansion, shrinkage, or translation would not lead to splits or mergers, but it would result in a significant PE. Additionally, even a minor gap of a single pixel in the boundary between two neurons would cause a merge error, yet it might only contribute a very small fraction of the total PE compared to the entire image.

While the foreground-restricted RI focuses on evaluating the agreement between the foreground regions in PR and GT, it does not explicitly consider split and merge errors. Split errors occur when a single GT region is segmented into multiple smaller regions, and merge errors happen when multiple GT regions are merged into a single segmented region. Although the foreground-restricted RI does not directly address split and merge errors, it can indirectly capture

some aspects of these errors if they affect the agreement between the foreground segmentations. For example, if one segmentation algorithm consistently splits objects while another merges them, there will be a disagreement in the foreground regions, which can be reflected in a lower RI score. To explicitly evaluate split and merge errors, there are alternative metrics that are more suitable, such as the V^{Rand}, V^{Info}, and Variation of Information (VOI) (Arganda-Carreras et al., 2015).

The V^{Rand} measure involves quantifying the split and merge scores and combining them by weighted harmonic mean. Before defining the split score let us denote by p_{ij} the probability that a randomly selected pixel belongs simultaneously to segment i in PR and segment j in GT. This joint probability distribution adheres to the normalization condition $\sum_{ij} p_{ij} = 1$. Moreover, the probability that a random pixel belongs to segment i in PR is denoted by s_i and the probability that a random pixel is assigned segment j in GT is denoted by t_j. When two pixels are randomly chosen, the likelihood of them belonging to the same segments in both PR and GT is determined by $\sum_{ij} p_{ij}^2$. This value is expected to increase with increasing similarity between PR and GT. The Rand split and merge scores of the V^{Rand} measure are then defined as:

$$V_{split}^{Rand} = \frac{\sum_{ij} p_{ij}^2}{\sum_k t_k^2}, \quad V_{merge}^{Rand} = \frac{\sum_{ij} p_{ij}^2}{\sum_k s_k^2} \qquad (3)$$

where V_{split}^{Rand} is the probability that two randomly selected pixels are assigned to the same cluster in PR, given that they belong to the same cluster in GT, and V_{merge}^{Rand} is the inverted conditional probability; it is the probability that two randomly selected pixels belong to the same segment in GT, given that they are assigned the same cluster in PR.

The V_α^{Rand} score is then the weighted harmonic mean of the split and merge errors:

$$V_\alpha^{Rand} = \frac{\sum_{ij} p_{ij}^2}{\alpha \sum_k s_k^2 + (1-\alpha) \sum_k t_k^2} \qquad (4)$$

The Rand F-score is defined as $\alpha = 0.5$, yielding equal importance to the split and merge errors. This metric was used in the SNEMI 3D challenge to compute the adapted Rand error (ARAND), defined as ARAND = 1 − RandF−score.

On the other hand, the V^{Info} uses information theory to compute the split and merge scores:

$$V_{split}^{Info} = \frac{I(PR;GT)}{H(PR)}, \quad V_{merge}^{Info} = \frac{I(PR;GT)}{H(GT)} \qquad (5)$$

where $I(PR;GT) = \sum_{ij} p_{ij} \log p_{ij} - \sum_i s_i \log s_i - \sum_i t_i \log t_i$ is the mutual information between the PR and GT maps, and $H(PR)$ and $H(GT)$ are the PR and GT entropy values, respectively. Finally, the V^{Info} measure is the weighted harmonic mean of the information theoretic split and merge scores:

$$V_\alpha^{Info} = \frac{I(PR;GT)}{(1-\alpha)H(PR) + \alpha H(GT)} \qquad (6)$$

Similar to the Rand F-score, the information theoretic F-score is defined as $\alpha = 0.5$, where the split and merge scores are given the same weighting, and it is closely related to the VOI. The VOI quantifies the distance between PR and GT segmentation maps by measuring the amount of information that is lost or gained when one segmentation is transformed into the other (Arbelaez et al., 2010). The VOI between the GT and PR components is the sum of two conditional entropies: the first one, $H(GT \mid PR)$, represents the degree of under-segmentation by measuring the information loss, while the second one, $H(PR \mid GT)$, measures the degree of over-segmentation by quantifying the information gained when transitioning from GT to PR. These measures are referred to as the VOI split or merge error, respectively. The VOI and ARAND were also combined to form the CREMI score in neuron segmentation by first taking the sum of the VOI split and VOI merge and then combining the result with ARAND using geometric mean.

For the evaluation of object detection, where different connected components are treated as different objects, the above measures are

also applicable. The main difference is the way a true positive is considered. In object detection a PR region is considered a TP if it overlaps with more than a given threshold (e.g. 50%) a GT component in terms of IoU, otherwise, it is an FP. The unmatched GT components are then considered FNs. A popular metric in object detection is average precision (AP), which is essentially the area under the precision–recall curve that is determined by systematically changing the detection threshold. The default AP measure uses a 50% IoU overlap threshold, but other variations of the AP can be used depending on how strict the evaluation must be. The term AP-α denotes the average precision at a given IoU threshold α. The higher the α the stricter the evaluation is. In problems with more than two classes, the mean AP (mAP) can be used to aggregate all the APs of all the classes involved by taking their average.

Instance segmentation requires more detailed measures to quantify the segmentation mask accuracy along with the detection performance. Metrics such as the aggregated Jaccard index (AJI) and the Panoptic Quality (PQ), which were originally proposed by Kumar et al. (2017) and Kirillov et al. (2019), respectively, have also been used in EM (Luo et al., 2021; Yuan et al., 2021) to evaluate instance segmentation algorithms more comprehensively. See Fig. 6 for an example. These two metrics are defined as:

$$
\begin{aligned}
AJI &= \frac{\sum_{j=1}^{N} |GT^j \cap PR^{j^*}|}{\sum_{j=1}^{N} |GT^j \cup PR^{j^*}| + \sum_{i \in FP} |PR^i|} \\
PQ &= \frac{\sum_{j \in TP} JI(GT^j, PR^{j^*})}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}
\end{aligned}
\tag{7}
$$

where N is the number of GT regions, and j^* is the index of the connected region in PR that is matched with the largest overlapping region (in terms of JI) with ground truth segment GT^j; FP is the set of false positive segments in PR without the corresponding ground truth regions in GT, FN is the set of false negative segments in GT that have been left unmatched with any regions in PR and TP is the set of all matched regions in GT and PR with at least 50% overlap in JI. The symbol $|.|$ indicates the cardinality of a given set. A GT component can only be used once to match with a PR component. In case there are multiple PR components overlapping the same GT component, the GT component will only be matched with the PR component having the largest IoU. The AJI is an object-level performance metric which measures the ability of a segmentation algorithm to accurately identify and delineate individual objects within an image. It takes into account both the segmentation quality and the accuracy of object identification. For problems where many GT regions are apposing or in very close proximity with each other (e.g. mitochondria in 2D EM), there is a high risk that one PR region overlaps multiple GT regions. Such cases are overpenalised by the AJI measure. Overpenalization is prevented to happen with PQ because the matching of PR and GT regions are only valid when they overlap with more than 50% in JI.

Table 5 presents the achieved results from the reviewed papers based on the above measures, showcasing the state-of-the-art performance.

8. Discussion and open challenges

8.1. Overview

CNNs have emerged as the preferred option for automated feature extraction and segmentation in EM data, with notable backbone networks based on FCNs, including the popular U-Net architecture. These networks use deeper architectures along with incorporating various levels of image context to generate effective 2D predictions, which can subsequently be simply integrated for 3D reconstruction without the need for explicit post-processing procedures. Additionally, researchers have explored the benefits of using dilated or atrous convolutions,

particularly in the initial layers, to expand the receptive field of convolutions. Recent works have also employed spatial pyramid pooling to capture multi-scale contextual information, enabling the acquisition of global information at higher levels. Notably, architectures like DeepEM3D have achieved high accuracy in anisotropic EM datasets by employing 3D operations exclusively in the initial layers and predicting pixel probabilities on the central slice, while utilizing multiple input sections for prediction.

8.2. 2D and 3D segmentation

3D CNNs have become a prominent component in EM segmentation workflows for volume EM datasets, offering overall improved precision compared to 2D CNNs by using voxel representation. The effectiveness of 3D CNNs is particularly notable for isotropic voxels, enabling precise segmentation of diverse organelles. The hybrid 2D-3D network is now the de facto approach for addressing anisotropy in serial-section EM. The nnU-Net, a self-configuring method for EM segmentation in both 2D and 3D, provides a good starting point for the configuration of network depth or hyperparameter tuning based on the dataset characteristics and a set of empirical experiments. However, certain limitations persist due to the dataset characteristics and the unique challenges with various organelles.

The primary constraints with 3D networks are the increased computational complexity, as the number of operations grows cubically with input size, and the increased demand for high computational memory. To address these concerns, smaller block sizes have been investigated as inputs during the training phase, resulting in a diminished context that causes imprecise predictions. Moreover, the large number of parameters in 3D networks requires larger datasets for effective model training. The training and inference on 2D data require fewer computational resources compared to 3D, which is advantageous when high-performance computing resources are limited. Another challenge for 3D models is dealing with organelles that lose their shape continuity through slices, as demonstrated by Franco-Barranco et al. (2022). This motivates the investigation of shape priors in 3D models, similar to what has been done in 2D models, where shape priors were used for regularization.

Both 2D and 3D approaches share a common challenge with respect to the instance segmentation of apposing organelles, especially when the organelles contain structures that are similar to their membranes. The cristae within cardiac mitochondria are one such example. One direction that may be taken in the future is the investigation of how surround suppression filtering can be effectively embedded in CNNs to suppress responses to cristae, thus allowing better delineation of the membranes. The potential of this approach is showcased in the recent study conducted by Aswath et al. (2023), where they introduce surround suppression filtering as a post-processing technique for mitochondria instance segmentation. Surround suppression has been demonstrated to be very effective in low-level image processing for the suppression of spurious strokes in contour detection tasks (Melotti et al., 2020).

Overall, future research should aim to overcome the current limitations of 3D CNNs, explore innovative techniques for incorporating more context, and advance the understanding and application of shape priors. These advancements will contribute to more robust and accurate EM segmentation, unlocking further insights and discoveries in the field of EM.

8.3. Data annotations and learning methods

Most advances in EM segmentation have been achieved with fully supervised approaches, which strongly rely on the availability and quality of finely annotated data. However, fine annotations are rarely available for modern large-scale EM datasets in biology. Instead, most

Table 5

State-of-the-art results. Abbreviations: D, un, CR, V_r, and V_i represent Dataset, unspecified, CREMI, V^{Rand}, and V^{Info}, respectively. Underlined results are from non-supervised methods. Structure(s) of interest are superscripted in column one, except for the Mitochondria section.

D	Evaluation metric					Citation
Neuronal membrane (NM), Synaptic clefts (SC)						
	RE	CR[a]	V^R	V^I	Others	
1^{NM}	.030				$.002^\vartheta$, $.094^\iota$	Fakhry et al. (2017)
1^{NM}			.983	.990		Xiao et al. (2018b)
1^{NM}			.983			Cao et al. (2019)
1^{NM}			.978	.990		Quan et al. (2021)
2^{NM}	.025					Lee et al. (2017)
2^{NM}	.060					Zeng et al. (2017)
3^{NM}	.091					Fakhry et al. (2017)
3^{NM}		.221			$.030^\kappa$, $.454^\lambda$	Bailoni et al. (2022)
3^{SC}		50				Heinrich et al. (2018)
3^{SC}		74.9				Isensee et al. (2019)
Mitochondria						
	JI	DSC	P	R	Others	
3	<u>.628</u>	<u>.770</u>				Peng et al. (2020)
4	.846		.932	.902	$.995^\alpha$	Casser et al. (2018)
4	.928	.962			$.915^\beta$, $.866^\gamma$	Yuan et al. (2021)
4	.926	.961			$.916^\beta$, $.866^\gamma$	Luo et al. (2021)
4	<u>.915</u>					Conrad and Narayan (2021)
4	.937					Franco-Barranco et al. (2022)
5	.918	.957	.911	.934	$.910^\delta$	Xiao et al. (2018a)
6					$.605^\epsilon_{rat}$, $.521^\epsilon_{human}$	Wei et al. (2020)
6					$.917^\epsilon_{rat}$, $.828^\epsilon_{human}$	Li et al. (2021)
6					$.560^\epsilon$, $.429^\zeta$	Wolny et al. (2022)
8	.906		.956	.946	$.996^\alpha$	Oztel et al. (2017)
8	.900	.947	.882	.938	$.887^\delta$	Xiao et al. (2018a)
8	.994		.950	.933		Cheng and Varshney (2017)
8	.890		.946	.937	$.994^\alpha$	Casser et al. (2018)
8	<u>.600</u>	<u>.747</u>				Peng et al. (2020)
8	.887				$.812^\epsilon$	Wei et al. (2020)
8	.864					Liu et al. (2020a)
8	.901	.948			$.890^\beta$, $.839^\gamma$	Yuan et al. (2021)
8	.899	.947			$.897^\beta$, $.850^\gamma$	Luo et al. (2021)
8	.895	.945				Li et al. (2021)
8	.893					Franco-Barranco et al. (2022)
9	.900		.974	.922	0.993^α	Casser et al. (2018)
9	<u>.895</u>					Conrad and Narayan (2021)
9	.926					Franco-Barranco et al. (2022)
12			.816	un	$.967^\alpha$	Dietlmeier et al. (2019)
12			.982	.985	$.984^\alpha$	Khadangi et al. (2021b)
13		.942		.921	$.999^\eta$	Mekuč et al. (2020)
13			.877	.887	$.820^\gamma$	Mekuč et al. (2022)
14	<u>.884</u>					Conrad and Narayan (2021)
16			.892	.992	$.984^\alpha$	Khadangi et al. (2021b)
20	.770					Conrad and Narayan (2021)
Nucleus (N), Nuclear envelope (NE), Lysosomes (L), Axons (A), Vesicles (V), Mitochondria and Lysosomes as one class (ML), Five structures (C)[b]						
7^N					$.978^\epsilon$, $.809^\epsilon$, $.894^\zeta$	Lin et al. (2021)
13^L		.822		.852	$.999^\eta$	Mekuč et al. (2020)
13^{ML}		.882				Mekuč et al. (2020)
13^{ML}	<u>.729</u>					Conrad and Narayan (2021)
14^N	<u>.988</u>					Conrad and Narayan (2021)
16^V			.979	.977		Haberl et al. (2018)
17^{NE}			.792	.628		Spiers et al. (2021)
18^C	.935				$.446^\theta$	Guay et al. (2021)
18^C					<u>$.429^\theta$</u>	Conrad and Narayan (2021)
19^A			.965	.877		Abdollahzadeh et al. (2021)

Symbols: α = Acc, β = AJI, γ = PQ, δ = CF, ϵ = AP50, ϵ = AP75, ζ = mAP, η = TNR, θ = mean JI, ϑ = WE, ι = PE, κ = ARAND, λ = VOI
[a] The CREMI scores for synaptic cleft and neuronal membranes are as per https://cremi.org/leaderboard/.
[b] The five organelle classes are mitochondrion, canalicular channel, alpha granule, dense granule, and dense granule core.

datasets are released with rough masks that are semi-automatically generated by pre-trained networks and proofreading. Such large-scale EM datasets need to be explored using weakly-, semi-, and self-supervised techniques in an end-to-end manner for improving EM segmentation (Papandreou et al., 2015; Kirillov et al., 2023).

In addition to manual annotation, EM images can be labeled using specialized imaging modalities that target specific structures in the sample. For instance, CLEM (Correlative light electron microscopy) is used to label structures targeted with fluorescent probes at (sub)cellular scales (de Boer et al., 2015; Heinrich et al., 2021). Other EM modalities include analytical imaging at the nanoscale-range to provide element-guided identification of various organelles (Pirozzi et al., 2018). These methods can reduce the bias in human annotation but may require longer sample preparation and acquisition times, specialized equipment, and additional post-processing to produce segmentations.

Due to the lack of labels, developing new training procedures for each imaging experiment is impractical. When confronted with unseen

samples, the performance of supervised methods is often negatively impacted and requires expensive redesign and retraining efforts. Various approaches have been taken to adapt models with few labeled samples and improve the generalization capabilities of CNNs. Both transfer learning and self-supervised techniques have been used for segmenting unseen EM datasets. These approaches enable the segmentation of EM images with minimal or no annotations. While overall fully supervised methods are the most effective approaches, preliminary results with few-shot and domain adaptation strategies look promising and have shown comparable performance in the segmentation of certain organelles.

The rise of self-supervised learning, which is attributable to its ability to learn generic representations from unlabeled data, also holds great promise in the field of EM segmentation. In fact, the availability of CEM500K, a comprehensive repository of unlabeled data encompassing organelles from diverse cell types, tissues, and preparation methods, played a significant role in achieving generic representations through self-supervised learning. By fine-tuning networks pre-trained on CEM500K, the resulting models demonstrated an ability to handle a broad spectrum of organelles, with comparable performance to specialized supervised networks.

An example of the potential of self-supervised learning is the recent breakthrough of the Segment Anything Model (SAM) (Kirillov et al., 2023), which has good performance in various segmentation applications across different imaging domains. The model has a transformer-based architecture and zero-shot transfer across new image distributions and tasks. This method has the potential to allow for the investigation of many more diverse datasets and for the incorporation of supplementary inputs which can be used as a prompt to highlight regions of interest. Looking ahead, the potential of using transformers in self-segmentation is particularly noteworthy, especially for managing large-scale EM data.

8.4. Experimental designs and performance evaluation

Multi-class segmentation has been investigated by only one of the reviewed papers (Heinrich et al., 2018), while the remaining methods focus on binary segmentation of a specific organelle. With this limitation, the potential of CNNs is not fully exploited. The utilization of multi-class segmentation leverages global and local features to improve the overall accuracy. The global features help in learning contextual information by capturing inter-class differences and the spatial context in which these structures occur. The training of models in a diverse range of classes leads to improved generalization ability, making them more adaptable and transferable to new datasets.

As to performance measures, there is a lack of standardized evaluation protocols in quantifying the effectiveness of segmentation in EM. While various performance measures have been proposed, the methods' performance is not consistently compared with the same measures. This is especially the case in mitochondria (Table 5). This makes it difficult to draw strong conclusions about existing investigations. A standard evaluation protocol is thus crucial for enabling fair and meaningful comparisons between different algorithms.

Furthermore, it is important to consider the specific needs and requirements of biologists when evaluating segmentation methods. While computer scientists may prioritize precise contour delineation, biologists may place greater emphasis on measuring morphological properties such as diameter, area, or minor–major axis ratio. Consequently, future datasets could potentially include not only segmentation masks but also ground truth for the morphological analysis performed by biologists after segmentation. In addition, biologists are interested in both common and rare ultra-structural alterations in tissues. The goal of segmentation from a biological perspective is to identify not only all organelles but also certain outliers or unmodeled phenomena such as fission events, cell destruction, or disease progression. The current metrics for segmentation are inadequate in capturing specific types of errors within segmentation results or the spatial distribution of these errors. The spatial distribution of errors provides insights into the localization and extent of inaccuracies or discrepancies in the segmented regions which can help characterize tissues. The involvement of biologists in the evaluation process to report specific errors or anomalies that deviate from the normal appearance of structures is required for assessing and quantifying such events.

9. Conclusion

In this survey, we describe the role of CNNs in large-scale cellular EM segmentation. End-to-end learning based on advanced CNN architectures using labeled data has achieved human-level accuracy in semantic segmentation tasks whereas the problem of instance segmentation still requires efforts, especially in the case of highly crowded structures. Despite this notable progress, certain challenges in 3D segmentation still remain, mainly due to the increased computational complexity and misclassification of voxels that hinder reconstruction.

As we move towards larger EM datasets, the process of obtaining consistent annotations becomes more challenging. Given that the lack of fully annotated data in EM will persist, the use of semi- and self-supervised learning will become more common. Previously, the emphasis was primarily on designing network architectures that address specific tasks related to individual structures, like synapses in neuronal regions or the instance segmentation of mitochondria. However, in the coming years, we expect a shift towards more general-purpose segmentation models, using the large-scale networks and learning methods discussed in our review for extracting generic features in a task-independent manner. These features could allow the unsupervised discovery of new structures and regions of interest or could be adapted to specific supervised segmentation tasks.

Data availability

No data was used for the research described in the article.

References

Abdollahzadeh, A., Belevich, I., Jokitalo, E., Sierra, A., Tohka, J., 2021. DeepACSON automated segmentation of white matter in 3D electron microscopy. Commun. Biol. 4 (1), 1–14.

Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., 2010. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 898–916.

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K.W., Schindelin, J., Cardona, A., Sebastian Seung, H., 2017. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. Bioinformatics 33 (15), 2424–2426.

Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Cireşan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J.M., et al., 2015. Crowdsourcing the creation of image segmentation algorithms for connectomics. Front. Neuroanat. 142.

Aswath, A., Alsahaf, A., Westenbrink, B.D., Giepmans, B.N., Azzopardi, G., 2023. COFI - Coarse-semantic to fine-instance unsupervised mitochondria segmentation in EM. In: Proceedings of the 20th International Conference on Computer Analysis of Images and Patterns. CAIP, Springer.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bai, M., Urtasun, R., 2017. Deep watershed transform for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5221–5229.

Bailoni, A., Pape, C., Hütsch, N., Wolf, S., Beier, T., Kreshuk, A., Hamprecht, F.A., 2022. GASP, a generalized framework for agglomerative clustering of signed graphs and its application to Instance Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11645–11655.

Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., et al., 2019. Ilastik: interactive machine learning for (bio) image analysis. Nature Methods 16 (12), 1226–1232.

Bermúdez-Chacón, R., Altingövde, O., Becker, C., Salzmann, M., Fua, P., 2019. Visual correspondences for unsupervised domain adaptation on electron microscopy images. IEEE Trans. Med. Imaging 39 (4), 1256–1267.

Bermúdez-Chacón, R., Márquez-Neila, P., Salzmann, M., Fua, P., 2018. A domain-adaptive two-stream U-Net for electron microscopy image segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 400–404.

Berning, M., Boergens, K.M., Helmstaedter, M., 2015. SegEM: efficient image analysis for high-resolution connectomics. Neuron 87 (6), 1193–1206.

de Boer, P., Giepmans, B.N., 2021. State-of-the-art microscopy to understand Islets of Langerhans: what to expect next? Immunol. Cell Biol. 99 (5), 509–520.

de Boer, P., Hoogenboom, J., Giepmans, B., 2015. Correlated light and electron microscopy: ultrastructure lights up!. Nature Methods 12 (6), 503–513.

de Boer, P., Pirozzi, N.M., Wolters, A.H., Kuipers, J., Kusmartseva, I., Atkinson, M.A., Campbell-Thompson, M., Giepmans, B.N., 2020. Large-scale electron microscopy database for human type 1 diabetes. Nature Commun. 11 (1), 1–9.

Briggman, K.L., Helmstaedter, M., Denk, W., 2011. Wiring specificity in the direction-selectivity circuit of the retina. Nature 471 (7337), 183–188.

Cao, Y., Liu, S., Peng, Y., Li, J., 2020. DenseUNet: densely connected UNet for electron microscopy image segmentation. IET Image Process. 14 (12), 2682–2689.

Cao, L., Lu, Y., Li, C., Yang, W., 2019. Automatic segmentation of pathological glomerular basement membrane in transmission electron microscopy images with random forest stacks. Comput. Math. Methods Med. 2019.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al., 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 7 (10), 1–11.

Carvalho, L., Sobieranski, A.C., von Wangenheim, A., 2018. 3D segmentation algorithms for computerized tomographic imaging: a systematic literature review. J. Digit. Imaging 31, 799–850.

Casser, V., Kang, K., Pfister, H., Haehn, D., 2018. Fast mitochondria segmentation for connectomics.

Chang, H.-H., Zhuang, A.H., Valentino, D.J., Chu, W.-C., 2009. Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage 47 (1), 122–135.

Chen, M., Dai, W., Sun, S.Y., Jonasch, D., He, C.Y., Schmid, M.F., Chiu, W., Ludtke, S.J., 2017d. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nature Methods 14 (10), 983–985.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017b. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017c. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P.-A., 2017a. DCAN: Deep contour-aware networks for object instance segmentation from histology images. Med. Image Anal. 36, 135–146.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Cheng, H.-C., Varshney, A., 2017. Volume segmentation using convolutional neural networks with limited training data. In: 2017 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 590–594.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Ciresan, D., Giusti, A., Gambardella, L., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. Adv. Neural Inf. Process. Syst. 25, 2843–2851.

Conrad, R., Narayan, K., 2021. CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. Elife 10, e65894.

Dai, W., Fu, C., Raytcheva, D., Flanagan, J., Khant, H.A., Liu, X., Rochat, R.H., Haase-Pettingell, C., Piret, J., Ludtke, S.J., et al., 2013. Visualizing virus assembly intermediates inside marine cyanobacteria. Nature 502 (7473), 707–710.

De Brabandere, B., Neven, D., Gool, L.V., 2017. Semantic instance segmentation with a discriminative loss function. arXiv:1708.02551.

Dietlmeier, J., McGuinness, K., Rugonyi, S., Wilson, T., Nuttall, A., O'Connor, N.E., 2019. Few-shot hypercolumn-based mitochondria segmentation in cardiac and outer hair cells in focused ion beam-scanning electron microscopy FIB-SEM data. Pattern Recognit. Lett. 128, 521–528.

Dittmayer, C., Goebel, H.-H., Heppner, F.L., Stenzel, W., Bachmann, S., 2021. Preparation of samples for large-scale automated electron microscopy of tissue and cell ultrastructure. Microsc. Microanal. 27 (4), 815–827.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 179–187.

Eberle, A., Mikula, S., Schalek, R., Lichtman, J., Tate, M.K., Zeidler, D., 2015. High-resolution, high-throughput imaging with a multibeam scanning electron microscope. J. Microsc. 259 (2), 114–120.

Ede, J.M., 2021. Deep learning in electron microscopy. Mach. Learn.: Sci. Technol. 2 (1), 011004.

Faas, F.G., Avramut, M.C., M. van den Berg, B., Mommaas, A.M., Koster, A.J., Ravelli, R.B., 2012. Virtual nanoscopy: generation of ultra-large high resolution electron microscopy maps. J. Cell Biol. 198 (3), 457–469.

Fakhry, A., Zeng, T., Ji, S., 2017. Residual deconvolutional networks for brain electron microscopy image segmentation. IEEE Trans. Med. Imaging 36 (2), 447–456.

Franco-Barranco, D., Muñoz-Barrutia, A., Arganda-Carreras, I., 2022. Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes. Neuroinformatics 20 (2), 437–450.

Frangakis, A.S., Hegerl, R., 2002. Segmentation of two-and three-dimensional data from electron microscopy using eigenvector analysis. J. Struct. Biol. 138 (1–2), 105–113.

Funke, J., Tschopp, F., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., Turaga, S.C., 2018. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. 41 (7), 1669–1680.

Ghosh, S., Das, N., Das, I., Maulik, U., 2019. Understanding deep learning techniques for image segmentation. ACM Comput. Surv. 52 (4), 1–35.

Glancy, B., Hartnell, L.M., Malide, D., Yu, Z.-X., Combs, C.A., Connelly, P.S., Subramaniam, S., Balaban, R.S., 2015. Mitochondrial reticulum for cellular energy distribution in muscle. Nature 523 (7562), 617–620.

Guay, M.D., Emam, Z.A., Anderson, A.B., Aronova, M.A., Pokrovskaya, I.D., Storrie, B., Leapman, R.D., 2021. Dense cellular segmentation for EM using 2D–3D neural network ensembles. Sci. Rep. 11 (1), 1–11.

Haberl, M.G., Churas, C., Tindall, L., Boassa, D., Phan, S., Bushong, E.A., Madany, M., Akay, R., Deerinck, T.J., Peltier, S.T., et al., 2018. CDeep3M—Plug-and-Play cloud-based deep learning for image segmentation. Nature Methods 15 (9), 677–680.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heinrich, L., Bennett, D., Ackerman, D., Park, W., Bogovic, J., Eckstein, N., Petruncio, A., Clements, J., Pang, S., Xu, C.S., et al., 2021. Whole-cell organelle segmentation in volume electron microscopy. Nature 1–6.

Heinrich, L., Funke, J., Pape, C., Nunez-Iglesias, J., Saalfeld, S., 2018. Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete drosophila brain. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 317–325.

Helmstaedter, M., Briggman, K.L., Turaga, S.C., Jain, V., Seung, H.S., Denk, W., 2013. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500 (7461), 168–174.

Isensee, F., Petersen, J., Kohl, S.A., Jäger, P.F., Maier-Hein, K.H., 2019. nnu-net: Breaking the spell on successful medical image segmentation. p. 2, arXiv preprint arXiv:1904.08128, 1.

Januszewski, M., Kornfeld, J., Li, P.H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., Jain, V., 2018. High-precision automated reconstruction of neurons with flood-filling networks. Nature Methods 15 (8), 605–610.

Jiang, Y., Xiao, C., Li, L., Chen, X., Shen, L., Han, H., 2019. An effective encoder-decoder network for neural cell bodies and cell nucleus segmentation of EM images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 6302–6305.

Karabağ, C., Jones, M.L., Peddie, C.J., Weston, A.E., Collinson, L.M., Reyes-Aldasoro, C.C., 2019. Segmentation and modelling of the nuclear envelope of hela cells imaged with serial block face scanning electron microscopy. J. Imaging 5 (9), 75.

Kasthuri, N., Hayworth, K.J., Berger, D.R., Schalek, R.L., Conchello, J.A., Knowles-Barley, S., Lee, D., Vázquez-Reina, A., Kaynig, V., Jones, T.R., et al., 2015. Saturated reconstruction of a volume of neocortex. Cell 162 (3), 648–661.

Khadangi, A., Boudier, T., Rajagopal, V., 2021a. EM-net: Deep learning for electron microscopy image segmentation. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 31–38.

Khadangi, A., Boudier, T., Rajagopal, V., 2021b. EM-stellar: benchmarking deep learning for electron microscopy image segmentation. Bioinformatics 37 (1), 97–106.

Kievits, A.J., Lane, R., Carroll, E.C., Hoogenboom, J.P., 2022. How innovations in methodology offer new prospects for volume electron microscopy. J. Microsc. 287 (3), 114–137.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P., 2019. Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C., 2017. Instancecut: from edges to instances with multicut. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5008–5017.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.

Kylberg, G., Uppström, M., Hedlund, K.-O., Borgefors, G., Sintorn, I.-M., 2012. Segmentation of virus particle candidates in transmission electron microscopy images. J. Microsc. 245 (2), 140–147.

Lee, K., Zlateski, A., Ashwin, V., Seung, H.S., 2015. Recursive training of 2D-3D convolutional networks for neuronal boundary prediction. Adv. Neural Inf. Process. Syst. 28.

Lee, K., Zung, J., Li, P., Jain, V., Seung, H.S., 2017. Superhuman accuracy on the SNEMI3D connectomics challenge. arXiv:1706.00120.

Li, M., Chen, C., Liu, X., Huang, W., Zhang, Y., Xiong, Z., 2022. Advanced deep networks for 3d mitochondria instance segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Li, Z., Chen, X., Zhao, J., Xiong, Z., 2021. Contrastive learning for mitochondria segmentation. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, IEEE, pp. 3496–3500.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 348–360.

Lin, Z., Wei, D., Petkova, M.D., Wu, Y., Ahmed, Z., Zou, S., Wendt, N., Boulanger-Weill, J., Wang, X., Dhanyasi, N., et al., 2021. NucMM dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 164–174.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, Z., Jin, L., Chen, J., Fang, Q., Ablameyko, S., Yin, Z., Xu, Y., 2021. A survey on applications of deep learning in microscopy image analysis. Comput. Biol. Med. 134, 104523.

Liu, T., Jones, C., Seyedhosseini, M., Tasdizen, T., 2014. A modular hierarchical approach to 3D electron microscopy image segmentation. J. Neurosci. Methods 226, 88–102.

Liu, T., Jurrus, E., Seyedhosseini, M., Ellisman, M., Tasdizen, T., 2012. Watershed merge tree classification for electron microscopy image segmentation. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, pp. 133–137.

Liu, J., Li, L., Yang, Y., Hong, B., Chen, X., Xie, Q., Han, H., 2020a. Automatic reconstruction of mitochondria and endoplasmic reticulum in electron microscopy volumes by deep learning. Front. Neurosci. 14, 599.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020b. Deep learning for generic object detection: A survey. Int. J. Comput. Vis. 128, 261–318.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lucchi, A., Li, Y., Fua, P., 2013. Learning for structured prediction using approximate subgradient descent with working sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1987–1994.

Lucchi, A., Smith, K., Achanta, R., Knott, G., Fua, P., 2011. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. IEEE Trans. Med. Imaging 31 (2), 474–486.

Luo, Z., Wang, Y., Liu, S., Peng, J., 2021. Hierarchical encoder-decoder with soft label-decomposition for mitochondria segmentation in EM images. Front. Neurosci. 15.

Mekuč, M.Ž., Bohak, C., Boneš, E., Hudoklin, S., Marolt, M., et al., 2022. Automatic segmentation and reconstruction of intracellular compartments in volumetric electron microscopy data. Comput. Methods Programs Biomed. 223, 106959.

Mekuč, M.Ž., Bohak, C., Hudoklin, S., Kim, B.H., Kim, M.Y., Marolt, M., et al., 2020. Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data. Comput. Biol. Med. 119, 103693.

Melotti, D., Heimbach, K., Rodríguez-Sánchez, A., Strisciuglio, N., Azzopardi, G., 2020. A robust contour detection operator with combined push-pull inhibition and surround suppression. Inform. Sci. 524, 229–240. http://dx.doi.org/10.1016/j.ins.2020.03.026, URL: https://www.sciencedirect.com/science/article/pii/S0020025520302073.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Moussavi, F., Heitz, G., Amat, F., Comolli, L.R., Koller, D., Horowitz, M., 2010. 3D segmentation of cell boundaries from whole cell cryogenic electron tomography volumes. J. Struct. Biol. 170 (1), 134–145.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1520–1528.

Oda, H., Roth, H.R., Chiba, K., Sokolić, J., Kitasaka, T., Oda, M., Hinoki, A., Uchida, H., Schnabel, J.A., Mori, K., 2018. BESNet: boundary-enhanced segmentation of cells in histopathological images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer, pp. 228–236.

Oztel, I., Yolcu, G., Ersoy, I., White, T., Bunyak, F., 2017. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 1195–1200.

Papandreou, G., Chen, L.-C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1742–1750.

Peddie, C.J., Collinson, L.M., 2014. Exploring the third dimension: volume electron microscopy comes of age. Micron 61, 9–19.

Peddie, C.J., Genoud, C., Kreshuk, A., Meechan, K., Micheva, K.D., Narayan, K., Pape, C., Parton, R.G., Schieber, N.L., Schwab, Y., et al., 2022. Volume electron microscopy. Nat. Rev. Methods Primers 2 (1), 1–23.

Peng, J., Yi, J., Yuan, Z., 2020. Unsupervised mitochondria segmentation in em images via domain adaptive multi-task learning. IEEE J. Sel. Top. Sign. Proces. 14 (6), 1199–1209.

Perez, A.J., Seyedhosseini, M., Deerinck, T.J., Bushong, E.A., Panda, S., Tasdizen, T., Ellisman, M.H., 2014. A workflow for the automatic segmentation of organelles in electron microscopy image stacks. Front. Neuroanat. 8, 126.

Pirozzi, N.M., Hoogenboom, J.P., Giepmans, B.N., 2018. ColorEM: analytical electron microscopy for element-guided identification and imaging of the building blocks of life. Histochem. Cell Biol. 150 (5), 509–520.

Quan, T.M., Hildebrand, D.G., Jeong, W.-K., 2016. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics.

Quan, T.M., Hildebrand, D.G.C., Jeong, W.-K., 2021. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. Front. Comput. Sci. 34.

Ravelli, R.B., Kalicharan, R.D., Avramut, M.C., Sjollema, K.A., Pronk, J.W., Dijk, F., Koster, A.J., Visser, J.T., Faas, F.G., Giepmans, B.N., 2013. Destruction of tissue, cells and organelles in type 1 diabetic rats presented at macromolecular resolution. Sci. Rep. 3 (1), 1–6.

Ren, Y., Kruit, P., 2016. Transmission electron imaging in the Delft multibeam scanning electron microscope 1. J. Vac. Sci. Technol. B 34 (6), 06KF02.

Ren, M., Zemel, R.S., 2017. End-to-end instance segmentation with recurrent attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6656–6664.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al., 2012. Fiji: an open-source platform for biological-image analysis. Nature Methods 9 (7), 676–682.

Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B., 2017. One-shot learning for semantic segmentation. arXiv:1709.03410.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Sokol, E., Kramer, D., Diercks, G.F., Kuipers, J., Jonkman, M.F., Pas, H.H., Giepmans, B.N., 2015. Large-scale electron microscopy maps of patient skin and mucosa provide insight into pathogenesis of blistering diseases. J. Invest. Dermatol. 135 (7), 1763–1770.

Spiers, H., Songhurst, H., Nightingale, L., de Folter, J., Community, Z.V., Hutchings, R., Peddie, C.J., Weston, A., Strange, A., Hindmarsh, S., et al., 2021. Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations. Traffic.

Takaya, E., Takeichi, Y., Ozaki, M., Kurihara, S., 2021. Sequential semi-supervised segmentation for serial electron microscopy image with small number of labels. J. Neurosci. Methods 351, 109066.

Takemura, S.-y., Xu, C.S., Lu, Z., Rivlin, P.K., Parag, T., Olbris, D.J., Plaza, S., Zhao, T., Katz, W.T., Umayam, L., et al., 2015. Synaptic circuits and their variations within different columns in the visual system of Drosophila. Proc. Natl. Acad. Sci. 112 (44), 13711–13716.

Titze, B., Genoud, C., 2016. Volume scanning electron microscopy for imaging biological ultrastructure. Biol. Cell 108 (11), 307–323.

Treder, K.P., Huang, C., Kim, J.S., Kirkland, A.I., 2022. Applications of deep learning in electron microscopy. Microscopy 71 (Supplement_1), i100–i115.

Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S., 2010. Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Comput. 22 (2), 511–538.

Wang, R., Stone, R.L., Kaelber, J.T., Rochat, R.H., Nick, A.M., Vijayan, K.V., Afshar-Kharghan, V., Schmid, M.F., Dong, J.-F., Sood, A.K., et al., 2015. Electron cryotomography reveals ultrastructure alterations in platelets from patients with ovarian cancer. Proc. Natl. Acad. Sci. 112 (46), 14266–14271.

Wei, D., Lin, Z., Barranco, D., Wendt, N., Liu, X., Yin, W., Huang, X., Gupta, A., Jang, W., Wang, X., Arganda-Carreras, I., Lichtman, J., Pfister, H., 2020. MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images. In: International Conference on Medical Image Computing and Computer Assisted Intervention.

Winding, M., Pedigo, B.D., Barnes, C.L., Patsolic, H.G., Park, Y., Kazimiers, T., Fushiki, A., Andrade, I.V., Khandelwal, A., Valdes-Aleman, J., Li, F., Randel, N., Barsotti, E., Correia, A., Fetter, R.D., Hartenstein, V., Priebe, C.E., Vogelstein, J.T., Cardona, A., Zlatic, M., 2023. The connectome of an insect brain. Science 379 (6636), eadd9330.

Wolny, A., Yu, Q., Pape, C., Kreshuk, A., 2022. Sparse object-level supervision for instance segmentation with pixel embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4402–4411.

Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., Han, H., 2018a. Automatic mitochondria segmentation for EM data using a 3D supervised convolutional network. Front. Neuroanat. 12, 92.

Xiao, C., Liu, J., Chen, X., Han, H., Shu, C., Xie, Q., 2018b. Deep contextual residual network for electron microscopy image segmentation in connectomics. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 378–381.

Xing, F., Xie, Y., Su, H., Liu, F., Yang, L., 2017. Deep learning in microscopy image analysis: A survey. IEEE Trans. Neural Netw. Learn. Syst. 29 (10), 4550–4568.

Xu, C.S., Pang, S., Shtengel, G., Müller, A., Ritter, A.T., Hoffman, H.K., Takemura, S.-y., Lu, Z., Pasolli, H.A., Iyer, N., et al., 2021. An open-access volume electron microscopy atlas of whole cells and tissues. Nature 599 (7883), 147–151.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: Bengio, Y., LeCun, Y. (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. URL: http://arxiv.org/abs/1511.07122.

Yuan, Z., Ma, X., Yi, J., Luo, Z., Peng, J., 2021. HIVE-Net: Centerline-aware hierarchical view-ensemble convolutional network for mitochondria segmentation in EM images. Comput. Methods Programs Biomed. 200, 105925.

Zeng, T., Wu, B., Ji, S., 2017. DeepEM3D: approaching human-level performance on 3D anisotropic EM image segmentation. Bioinformatics 33 (16), 2555–2562.

Zheng, Z., Lauritzen, J.S., Perlman, E., Robinson, C.G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C.B., Sharifi, N., et al., 2018. A complete electron microscopy volume of the brain of adult Drosophila melanogaster. Cell 174 (3), 730–743.

Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. 3 (1), 1–130.