FISBe: A real-world benchmark dataset for instance segmentation of long-range thin filamentous structures

Lisa Mais[1,2,4,*,✉], **Peter Hirsch**[1,2,*], **Claire Managan**[3], **Ramya Kandarpa**[3], **Josef Lorenz Rumberger**[1,2], **Annika Reinke**[2,5], **Lena Maier-Hein**[2,5], **Gudrun Ihrke**[3], **Dagmar Kainmueller**[1,2,4,✉]

[1] Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC) [2] Helmholtz Imaging [3] HHMI Janelia Research Campus [4] University of Potsdam [5] German Cancer Research Center (DKFZ)

✉ {firstname.lastname}@mdc-berlin.de * shared first authors

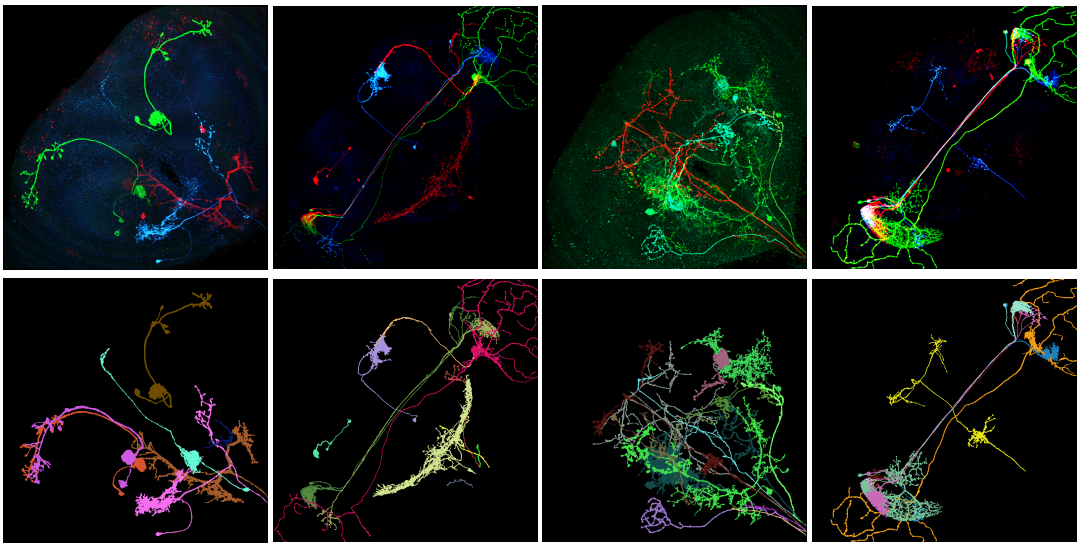

Figure 1. We release the FlyLight Instance Segmentation Benchmark (FISBe) dataset, a 3d multi-color light microscopy dataset of neuronal morphologies in the brain of the fruit fly Drosophila melanogaster, together with high-quality pixel-wise instance segmentations of individual neurons. To the best of our knowledge, FISBe constitutes the first publicly available real-world benchmark dataset for instance segmentation of wide-spanning, thin filamentous and tightly interweaving objects. Top row: Exemplary FISBe images (3d) visualized in 2d via maximum intensity projection (MIP). Bottom row: Corresponding 2d projections of ground truth instance segmentation masks (3d).

Abstract

Instance segmentation of neurons in volumetric light microscopy images of nervous systems enables groundbreaking research in neuroscience by facilitating joint functional and morphological analyses of neural circuits at cellular resolution. Yet said multi-neuron light microscopy data exhibits extremely challenging properties for the task of instance segmentation: Individual neurons have long-ranging, thin filamentous and widely branching morphologies, multiple neurons are tightly inter-weaved, and partial volume effects, uneven illumination and noise inherent to light microscopy severely impede local disentangling as well as long-range tracing of individual neurons. These properties reflect a current key challenge in machine learning research, namely to effectively capture long-range dependencies in the data. While respective methodological research is buzzing, to date methods are typically benchmarked on synthetic datasets. To address this gap, we release the FlyLight Instance Segmentation Benchmark (FISBe) dataset, the first publicly available multi-neuron light microscopy dataset with pixel-wise annotations. In addition, we define a set of instance segmentation metrics for benchmarking that we designed to be meaningful with regard to downstream analyses. Lastly, we provide three baselines to kick off a competition that we envision to both advance the field of machine learning regarding methodology for capturing long-range data dependencies, and facilitate scientific discovery in basic neuroscience. Project page: https://kainmueller-lab.github.io/fisbe

1. Introduction

Most existing instance segmentation benchmarks in computer vision are collections of natural images [9, 35, 59]. These are often suitably addressed with proposal-based methods like Mask R-CNN [22, 24], as the assumption that shapes of objects are well approximated by bounding boxes mostly holds. However, this key assumption is violated in a range of highly relevant application domains, including neuroscience [18, 57], the domain the FISBe dataset we contribute in this work stems from. Here, objects can span large parts of an image and have complex (e.g., tree-like) and intertwined shapes. Consequently, multiple instances may have very similar, very large bounding boxes.

Benchmarks from the neuroscientific domain, namely on neuron instance segmentation in electron microscopy (EM) data [1], have greatly facilitated the development of instance segmentation methodology that is applicable in the face of complex, wide-ranging object shapes. Beyond sophisticated object shapes, said benchmarks also call for methodology that applies to very large images, way beyond what current GPU memory can hold [18, 55]. On these benchmarks, proposal-free methods based on CNN backbones [7, 10, 19, 31, 32, 58] constitute the current state of the art, and mostly also lend themselves to arbitrarily large images thanks to tile-and-stitch inference.

However, an object category that generally renders image segmentation very challenging is not represented in the above-mentioned benchmarks, namely objects that exhibit very thin (down to single-pixel width), filamentous structures. Benchmarks have been established for *semantic* segmentation of very thin filamentous structures in a range of real-world applications, including neuron segmentation in light microscopy (LM) data [3, 40, 45], blood vessel segmentation in various medical imaging modalities [37, 50], and road extraction from satellite imagery [2, 14]. However, respective *instance* segmentation benchmarks are currently lacking despite the high relevance of the task, e.g., in basic neuroscience [42]. The closest related publicly available dataset that exhibits thin and complex object shapes is [57]. Yet it lacks tightly inter-weaved objects by design, and furthermore does not come with pixel-wise ground truth segmentations nor recommended metrics for benchmarking.

Consequently, there is at present a lack of methodology applicable for instance segmentation of wide-ranging thin filamentous intertwined shapes: Only very few deep-learning approaches are potentially suitable, among which Flood Filling Networks [26] and PatchPerPix [39]. Most proposal-free instance segmentation methods do not appear suitable: Three-label models [4] degenerate in the face of very thin instances because their interior equals their boundary; models predicting pixel affinities [16, 54] become inappropriate if they rarely encounter foreground in their fixed pixel neighborhoods (as compared to the dense-foreground EM data they were designed for); metric learning models [7, 10] lack the capacity to capture long-range connectivity beyond their receptive fields. Likewise, methods proposed specifically for cell instance segmentation do not appear suitable: Cellpose [51] assumes locally (i.e., within receptive field) visible cues towards some semantically meaningful center point which does not hold true for our dataset; Stardist [56] employs star-convex polygons/polyhedrons as proposals, which do not provide viable approximations of neurons. As for non-learnt, application-specific methods for neuron separation, some approaches rely on user-defined [47] or pre-detected anchor points, in particular on cell body detection [34, 46, 60]. This renders these methods not directly applicable to our data, where cell bodies may lie outside of the imaged volume (namely in the ventral nerve cord). Other non-learnt application-specific approaches are based on color clustering [13, 52], which is technically applicable, yet the underlying assumption that each neuron has a unique color is often violated on our data.

A promising recent alternative are query-based methods [6, 8, 29], which operate without explicit prior assumptions on object sizes or shapes. However, e.g., SAM [29] is not directly applicable as it has not yet been extended to full 3d and it is unclear if and how tile-and-stitch prediction, as would be necessary given the size of individual FISBe images, could be achieved in a seamless manner. We deem respective potential extensions of SAM a very interesting research topic for which, albeit out of scope, FISBe can serve as benchmark. Further recent trends towards explicit modelling of long-range data dependencies appear promising [23, 30, 44], yet so far these models have only been benchmarked on synthetic data [28, 36], sequence data [53], and image classification [11], and thus, their potential for improving instance segmentation of long, complex, intertwined objects in real-world tasks has not been assessed.

Our work addresses the gap that, to date, solely synthetic data is available to facilitate methods development towards capturing long-range data dependencies. To this end, we herewith release the FISBe dataset, a 3d multicolor light microscopy dataset of wide-ranging and tightly interweaving neuronal morphologies in the brain of the fruit fly Drosophila melanogaster, together with high quality expert instance segmentations of individual neurons. The dataset comprises 101 large, expert-labeled 3d images, of which 30 are completely- and 71 partly labeled, with a total of $\sim$600 pixel-wise neuron instance masks. Exemplary images and instance masks are shown in Fig. 1. The novelty of our data entails a gap in evaluation metrics: Metrics commonly employed for benchmarking instance segmentation methodology do not appropriately account for the long, very thin and filamentous object shapes; e.g., mean average precision (mAP) with pixel-level IoU for localization is not appropriate for thin structures [38, 48]. Thus standard metrics may

not provide meaningful quantification of segmentation performance. To this end, we identify a set of informative evaluation metrics, and contribute a novel aggregate score that we recommend for method benchmarking. Given our metrics, we evaluate three baseline methods, namely the two learnt methods that are, to our knowledge, technically able to handle the intricacies of our data to date [26, 39], as well as one non-learnt application-specific method that is technically applicable [13]. In summary, we contribute:

- The FISBe dataset, to our knowledge the first public benchmark dataset for instance segmentation of real-world, wide-ranging, thin filamentous and tightly interweaving objects.
- A set of metrics and a novel ranking score for respective meaningful method benchmarking.
- An evaluation of three baseline methods in terms of the above metrics and score.

Concerning the size of our dataset, on the one hand, latest 2d natural image datasets are orders of magnitude bigger than ours [29] and thus pave the way for particularly data-hungry methods development. Such size is, however, far beyond reach for 3d data, let alone for data from the life sciences where expert knowledge is required for annotation, and acquiring pixel-wise ground truth for image data alike ours has been deemed difficult or infeasible in related work [13, 17]. On the other hand, numerous benchmark datasets similarly sized as FISBe have proven to greatly boost methods development in the machine learning community, and have likewise boosted respective application-specific scientific discovery [18, 27, 55]. We thus foresee our work to be of impact both in advancing the field of machine learning regarding methodology for capturing long-range data dependencies, and in streamlining cell-level analyses of brain function towards advances in basic neuroscience. We release our data through zenodo (https://zenodo.org/doi/10.5281/zenodo.10875063) and our project page https://kainmueller-lab.github.io/fisbe.

2. Dataset

The FISBe dataset consists of 101 3d multicolor LM images of the central nervous system of the fruit fly Drosophila melanogaster. The images originate from a large pre-existing resource of LM acquisitions [42]. Similar data has already contributed to breakthrough neuroscientific findings, e.g., towards a mechanistic understanding of memory formation and -retrieval in Drosophila on a cellular level [12]. Our work aims at facilitating such findings at scale. For a more elaborate introduction to the biological background of our data, we refer the interested reader to Suppl. Sec. A.4. In the following, we describe the imaging resource FISBe stems from in Sec. 2.1, the selection and annotation process for our dataset in Sec. 2.2, and recommended data splits and evaluation for benchmarking in Sec. 2.3.

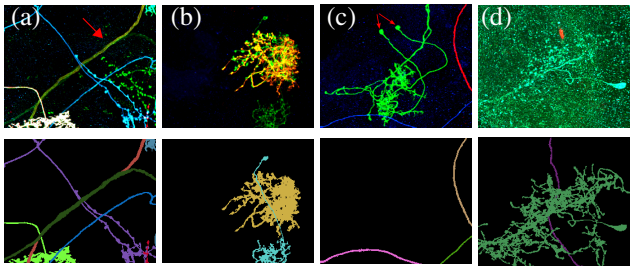

Figure 2. Exemplary challenging cases for disentangling neurons in FISBe images (top row), and respective expert annotations (bottom row). (a) Long overlap of two neurons running in parallel, (b) two almost completely overlapping neurons in different color (only one could successfully be annotated), (c) two inter-weaved neurons of same color that could not be separated (clearly identified by two somata), and (d) dim neuron in noisy background.

2.1. Image Data Acquisition and Characteristics

Our dataset originates from the FlyLight project [42], where confocal microscopy images of the nervous systems of ∼74,000 flies were acquired with a technique called MultiColor FlpOut (MCFO) [43]. This image collection was previously released[1]. We selected images from the "40x Gen1" subset where images have an isotropic resolution of $0.44\mu m$, an average size of $\sim 400\times 700\times 700$ pixels, and three color channels. Fig. 1 (top row) shows exemplary MCFO images. Note that visualizations of image data shown in this paper are maximum intensity projections along the z axis, if not noted otherwise. For exemplary orthographic views see Suppl. Fig. 9. For more information on the dataset see our datasheet [20] in Suppl. Sec. A.1.

MCFO images capture the very thin, tree-like morphology of individual neurons as well as the intertwining of multiple neurons. The number of neurons expressed in individual images varies from extremely dense (>50) to very sparse (1-2). The FlyLight project has sorted images into five categories according to expression density, where 18% of images express up to 10 neurons (cat. 1 and 2) and 55% express around 20 neurons (cat. 3, cf. Suppl. Fig. 1 in [42]).

MCFO imagery has sparse, unbalanced foreground signal, low signal-to-noise ratio, and exhibits artifacts like broken structures and intensity shifts. Intensity varies strongly, per image, per neuron and within neurons. Thus neurons may appear in very different quality, ranging continuously from clearly visible neurons to very dim neurons that are partly indistinguishable from noise. The MCFO technique causes individual neurons to exhibit random colors, though color diversity per image is often not sufficient to allow for distinguishing all neurons by color. Multiple neurons in very close proximity may appear as overlapping due to partial volume effects[2]. Neurons in MCFO images are par-

[1] Download (CC BY 4.0 license): https://gen1mcfo.janelia.org

[2] Multiple instances occupy the same 3d pixel (distinct from occlusion)

ticularly hard to distinguish if neurons of same or similar color form dense clusters or overlap (see Fig. 2).

2.2. Image Selection and Labeling Process

Labeling 3d data is generally cumbersome as objects often need to be viewed with different angles, scales and color settings. As for the FlyLight data, limited resolution inherent in confocal microscopy makes isolating individual similarly colored, close-by neurons particularly difficult. Furthermore, in some cases poor signal-to-noise ratio makes it difficult to identify the complete wide-spanning extent of neurons. In both cases, expert anatomical knowledge of fruit fly neurons is often crucial for successful annotation.

To form FISBe, expert annotators chose 101 samples from the FlyLight data for which they determined by eye that manual annotation is feasible. Compared to the full FlyLight resource, this introduces a bias in our dataset towards sparser expression densities: Of our 101 images, one image is of density cat. 1, 72 are of cat. 2, and 28 of cat. 3. Two expert annotators manually segmented and proof-read each other to label as many neurons as possible in these 101 images using the interactive rendering tool VVD Viewer[49]. Annotators were able to segment a total of 590 neurons. Labeling a single neuron took 30-60 min on average, yet for a difficult neuron it could take up to 4 hours. Not every neuron in every sample could be annotated successfully, thus yielding completely- as well as partly labeled images. A third annotator performed a final visual inspection of all labeled neurons and revised the categorization into completely- and partly labeled images.

Our completely labeled dataset comprises 30 MCFO images and a total of 139 labeled neurons; see Fig. 1 (1st and 2nd example) and Suppl. Fig. 10. Our partly labeled dataset comprises 71 MCFO images and a total of 451 labeled neurons; see Fig. 1 (3rd and 4th example) and Suppl. Fig. 12. These images exhibit unlabeled neuronal morphologies because expert annotators were either unable to disentangle multiple neurons of the same color in a dense cluster, or unable to annotate very dim neurons that are partly indistinguishable from noisy background. Note, 61 images contain labeled neurons that overlap due to partial volume effects.

Complementing our new annotated data, the large trove of previously released non-annotated images in the FlyLight resource may serve for self-supervised pre-training.

2.3. Benchmarking Setup

We split the completely labeled data into train, validation and test sets with 18, 5 and 7 samples respectively as defined in Suppl. Table 5. We split the partly labeled data into train, validation and test sets with 43, 12 and 16 samples respectively as defined in Suppl. Table 6. We recommend evaluation on the combined data (i.e., the union of completely and partly labeled data) as the main benchmarking scenario. To assess training stability, we recommend to report summary statistics over three training runs in each evaluated scenario.

3. Evaluation Metrics

Our dataset constitutes the first benchmark dataset for *instance* segmentation of thin filamentous structures. Consequently, we need to assess which evaluation metrics are suitable for benchmarking. The main requirements for a suitable metric are: (r1) To account for thin filamentous structures, (r2) to be able to handle overlapping instances (both in ground truth and predicted instances), and (r3) to be meaningful with respect to downstream tasks.

Some existing benchmarks for *semantic* segmentation of filamentous structures have employed topology-based metrics, which assess the similarity between graph representations of ground truth- and predicted objects [2, 21]. We deem these not suitable for our data, as obtaining topologically correct (tree) graph representations of neurons is infeasible due to the limited resolution of light microscopy.

Instead, we follow the Metrics Reloaded [38] recommendation for thin filamentous instance segmentation and apply an instance-level F1 score as one of our main evaluation metrics. Moreover, we propose to complement the F1 score with a custom metric that we design towards satisfying r3, namely a centerline recall with one-to-many matching, which we refer to as *average ground truth coverage*. We combine both metrics to derive an aggregate benchmark score. Finally, we define a set of easily interpretable error measures that may provide additional insight to methods developers and practitioners. E.g., we extend existing work [5, 41] by defining false split (FS) and false merge (FM) error counts for overlapping instances. We define our selected metrics in Sec. 3.1, ensuring that all apply not only to completely- but also to partly labeled data, and discuss properties and suitability in Sec. 3.2. Suppl. Table 2 summarizes all metrics with their localization and matching.

3.1. Metrics Definitions

Instance segmentation can be phrased as a pixel labeling problem, where pixels with same label form instances. Note that in FISBe, one pixel can be assigned multiple labels due to overlapping instances. Segmentation quality is generally assessed via evaluation metrics that capture how well predicted instances overlay with given ground truth (gt) instances. We denote a set of gt instances $G = \{g_k\}_{k \in L_G}$ and a set of predicted instances $P = \{p_l\}_{l \in L_P}$ where L_G and L_P represent the sets of labels identifying gt and predicted instances, and $|G|$ and $|P|$ denote the total number of instances of the respective set. With subscript i the set is limited to a single image, e.g., G_i, otherwise it refers to the set over all images I. Gt- and prediction sets exclude background (bg) as instance label if not stated otherwise.

Average F1 score $avF1$**.** Following [38], a metric consists of three steps: localization, matching and computation. The localization step employs some function to compute how well each pair of predicted and gt instances are co-localized. The matching step selects subset of these pairs, resulting in a match of predicted to gt instances. The computation step computes the value of the metric based on the quality of the previously computed subset of matched instances.

It has been shown that pixel-level IoU or F1 are not suitable for thin structures as small variations on boundaries can have a large effect [15, 38]. Thus we employ clDice [48], a variation of the Dice score that operates on object centerlines, for the localization step. Following [48], we use medial surface axial thinning algorithm [33] to skeletonize volumetric instance masks and denote it with function $skeletonize(\cdot)$. Given ground truth and prediction we compute clDice as follows:

$$\text{clPrecision}(p,g) = \frac{|\text{skeletonize}(p) \cap g|}{|\text{skeletonize}(p)|} \quad \forall p \in P_i, \forall g \in G_i \cup \{bg\} \tag{1}$$

$$\text{clRecall}(g,p) = \frac{|\text{skeletonize}(g) \cap p|}{|\text{skeletonize}(g)|} \quad \forall g \in G_i, \forall p \in P_i \cup \{bg\} \tag{2}$$

$$\text{clDice}(g,p) = 2 * \frac{\text{clPrecision}(p,g) * \text{clRecall}(g,p)}{\text{clPrecision}(p,g) + \text{clRecall}(g,p)} \quad \forall g \in G_i, \forall p \in P_i \tag{3}$$

clDice is only computed for foreground pairs as we do not skeletonize the background, but we, e.g., include it for clPrecision to detect predictions mainly located in the gt background.

For the matching step we follow the greedy strategy recommended in [38]. To this end we compute clDice for all pairs of predicted and gt instances $\{\text{clDice}(p,g) \mid \forall i \in I : \forall p \in P_i, \forall g \in G_i\}$. We iterate through all scores in descending order and match the corresponding (p,g)-pair if neither has been assigned before. Similarly to [41], we denote $p \Subset g$, if the two instances have been matched. In the computation step, we derive true positives (TP), false positives (FP) and false negatives (FN) for all clDice thresholds th in the range [0.1,0.9] with step size 0.1:

- TP: all predicted instances that are assigned to a gt instance with clDice $> th$:

$$\text{TP} = |\{p \in P \mid \exists g : p \Subset g \wedge \text{clDice}(p,g) > \text{th}\}|$$

- FP: all unassigned predicted instances $\text{FP} = |P| - \text{TP}$
- FN: all unassigned gt instances $\text{FN} = |G| - \text{TP}$

Based on these values we compute the F1 score $F1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$ for each threshold. Note that TP, FP, FN are thus computed across all images. The final *avF1* score is the average of all F1 scores.

Average ground truth coverage C**.** We compute clPrecision scores for all pairs of predicted and gt instances as localization criterion. We match each prediction to the gt instance with the highest clPrecision score (one-to-many matching). Then we average clRecall for all gt instances and the union of their matched predictions (to avoid double-counting of pixels with overlapping predictions):

$$C = 1/|G| \sum_{g \in G} \text{clRecall}(g, \bigcup_{\forall p \in P : p \Subset g} p) \tag{4}$$

Aggregate benchmark score S**.** We propose to combine the average F1 score $avF1$ and the average ground truth coverage C to form a primary benchmark ranking score S. We average both measures to obtain the final ranking score: $S = 0.5avF1 + 0.5C$. Note that we do not multiply them, as a linear increase in segmentation quality should lead to linear increase in the score function [15].

False splits FS **and false merges** FM**.** False split errors occur if one gt instance is covered by multiple predicted instances. False merge errors occur if one predicted instance covers more than one gt instance. We propose to use a greedy many-to-many matching algorithm that naturally handles overlapping instances and based on which we can compute FM and FS directly in a unified way. For the matching, we iteratively assign predicted and gt instances with the highest clRecall value while keeping track of already matched pixels (see Algorithm 1). Remaining clRecall values are constantly updated to only include *free* pixels, which are available for further matching. By doing so, we avoid that predicted instances in overlapping gt regions are assigned multiple times; or that mostly overlapping predicted instances are assigned to the same gt instance (see Suppl. Fig. 5). Note that we monitor centerline pixels for gt instances and complete pixelwise masks for predicted instances due to the definition of clRecall.

We then count for each gt instance the additional number of assigned predicted instances apart from one correctly matched instance to compute false splits (with th $= 0.05$)

$$\text{FS} = \sum_{g \in G} \max((\sum_{p \in P : p \Subset g} 1) - 1, 0), \tag{5}$$

and analogously for false merges (with th $= 0.1$)

$$\text{FM} = \sum_{p \in P} \max((\sum_{g \in G : g \Subset p} 1) - 1, 0). \tag{6}$$

True positives clDice. We report average centerline Dice for uniquely matched instances $\text{clDice}_{\text{TP}}$ to provide a measure of how well true positives are reconstructed. We re-use the matching computed for $avF1$, employ a threshold of 0.5 for the definition of TP and define

$$\text{clDice}_{\text{TP}} = \frac{1}{|\text{TP}_{0.5}|} \sum_{(p,g) : p \in \text{TP}_{0.5} \wedge p \Subset g} \text{clDice}(p,g). \tag{7}$$

Algorithm 1: Greedy Many-to-many Matching

input : G: set of gt instances g_k,
P: set of predicted instances p_l,
th: clRecall threshold
output: M: set of matched (g_k, p_l)-instances

initialize $M \leftarrow \emptyset$
initialize $G_{\text{free}} \leftarrow \{\text{skeletonize}(g_k) \mid \forall g_k \in G\}$
initialize $P_{\text{free}} \leftarrow \text{copy}(P)$
clR $\leftarrow$ sort($\{$clRecall(skeletonize$(g_k), p_l) \mid \forall g_k \in G, \forall p_l \in P\}, \downarrow$)
while top(clR) $>$ th **do**
  $g_k, p_l \leftarrow$ pop(clR)
  $M \mathrel{+}= \{(g_k, p_l)\}$
  update $g_{\text{free}_k} \leftarrow g_{\text{free}_k} \setminus p_l$
  update $p_{\text{free}_l} \leftarrow p_{\text{free}_l} \setminus g_k$
  forall $(g_m, p_n) \in$ clR **do**
    if $g_m = g_k$ **then**
      update clR$(g_m, p_n) = \frac{|g_{\text{free}_m} \cap p_n|}{|\text{skeletonize}(g_m)|}$
    if $p_n = p_l$ **then**
      update clR$(g_m, p_n) = \frac{|\text{skeletonize}(g_m) \cap p_{\text{free}_n}|}{|\text{skeletonize}(g_m)|}$
  sort(clR, $\downarrow$)

Evaluating partly labeled samples. In partly labeled samples only a subset of neurons is annotated. For unlabeled pixels we do not know if there is background or other neuronal structures and for labeled neurons if they partly overlap with a non-annotated one. This has no influence on average ground truth coverage as well as false split and false merge counts, although the reported measures only reflect parts of the whole volume. However, for the F1 score the false positive count cannot be computed. Therefore, we only count predicted instances that are not one-to-one matched based on clDice, but that primarily lie within a gt instance and not the background:

$$\text{FP}_{\text{partly}} = |\{p \in P \mid \nexists g : p \Subset g \wedge \arg\max_{g \in G \cup \{\text{bg}\}} \text{clPrecision}(p, g) \neq \text{bg}\}| \quad (8)$$

We use this error count as an approximation of false positives and adapt the formula to $F1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}_{\text{partly}}+\text{FN}}$. All other calculation steps remain unchanged. Note that thus the F1 scores of completely and partly labeled image sets cannot be compared directly. To evaluate the full dataset, we average the results for the completely and partly labeled set (for normalized measures, counting measures such as FS, FM and TP are summed up).

Evaluating challenging cases. Main challenges are dim and overlapping neurons (cf. Fig. 2). To evaluate how well such subsets of neurons are segmented, we report gt coverage C and the relative number of TPs (tp $= \frac{\text{TP}_{S,0.5}}{|G_S|}$, with greedy one-to-one matching and clDice > 0.5) for the respective subsets (G_{dim}/G_{ovlp}). Dim neuron instances of validation and test sets are listed within the dataset.

3.2. Discussion

We satisfy requirement (r1) by using clPrecision, clRecall and clDice in all of our measures. They are defined in such a way that they handle overlaps in both predicted and gt instances, thus satisfying requirement (r2).

The $avF1$ score considers all error types equally. However, FP and FS errors are more likely to occur in MCFO segmentations due to low signal-to-noise ratio and broken structures. They are often induced by only a limited number of incorrect pixels, whereas FN errors are limited by the number of neurons and require that all or large parts of neurons are missing. This can lead to disproportionately low scores that do not reflect how we would visually rate segmentation quality (see Fig. 3 (a)+(b)).

To mitigate this, we complement the $avF1$ score by the average gt coverage C, resulting in an improved balancing of the different types of errors. C provides an intuitive way for measuring how comprehensively gt instances are segmented by the model. C strongly penalizes FN and FM errors, an important property for downstream tasks. Consider the edge case of a perfect foreground segmentation. If FM are not penalized, assigning the same label to each connected component would lead to a perfect score, despite merges occurring at every overlap and point of contact. Thus, as desired, if a model achieves to split a previously occurring FM correctly, this will lead to a large improvement as an additional gt instance will get matched (see Fig. 3 (c)).

However, C does not incorporate FP and FS errors. Consider the same edge case as before, but now we assign a separate instance label to each foreground pixel. This, as a consequence of the one-to-many matching still leads to a perfect score for C. But the $avF1$ score will be exceedingly low. Similarly if the gt instances are indeed segmented perfectly, but the outside noise is segmented as well (see Fig 3 (d)). Note that this highlights why C should not be used as a standalone metric. FP and FS errors are accounted for by the $avF1$ score, and C penalizes FN and FM errors. Thus C and $avF1$ nicely complement each other. For additional edge cases and the corresponding quantitative evaluation please see Suppl. Fig. 6.

Concerning (r3) there are two highly relevant downstream tasks we took into account: 1. Morphological analysis of neurons, and 2. the task of searching for a given neuronal morphology within all MCFO images. For the first task neurons of interest need to be reconstructed in their entirety. In this regard, FN and only partly annotated neurons are critical, as their manual curation is very time-consuming. FP, FM and FS are less relevant as, up

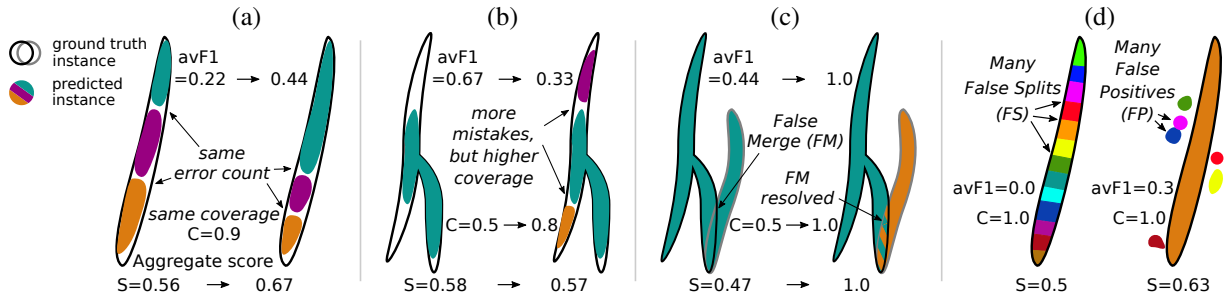

Figure 3. Visualization of segmentation examples to assess suitable evaluation metrics: (a) Depending on the split position, avF1 can vary significantly at identical gt coverage and error count. (b) Using the avF1 score alone would favor lower coverage over more false split errors. This might be disadvantageous in downstream analysis tasks. In (c) resolving the merge leads to a large improvement in the overall score. In (d) both cases achieve a perfect score wrt. the gt coverage C. By penalizing FP and FS errors in the F1 score the limitations of these predictions are reflected in the overall score. For more edge cases and the full quantitative numbers please see Suppl. Fig. 6.

to a point, they can relatively easily be corrected with a few clicks. This importance of completeness is strongly reflected in C. For the second task, the above considerations still hold, except that FP are more problematic. This is accounted for by the F1 score, which, as the number of FP can often be much higher than the number of TP and FN (see Table 1), heavily penalizes such cases.

Overall, our benchmark score S is in accordance with all our requirements. However, as a single benchmark score most likely cannot do justice to every possible use case, we report additional metrics that can be resorted to for alternative downstream tasks as well as for considerations on method improvements. The additional metrics all satisfy r1 and r2. Finally, our proposed many-to-many matching algorithm is very generic, despite adhering to our quite specific requirements. There is no special treatment of overlapping regions which strongly facilitates the matching. The algorithm can be applied to other object shapes by replacing clRecall with other overlap-based metrics like IoU or IoR.

4. Baselines

To showcase the FISBe dataset together with our selection of metrics, we provide evaluation results for three baseline methods, namely PatchPerPix [39], Flood Filling Network (FFN) [25, 26] and a non-learnt application-specific color clustering from Duan et al. [13, 52]. For information on all models, including training and validation details, see Suppl. Sec. A.3. Quantitative results for the full dataset are shown in Table 1, for completely and partly sets in Suppl. Table 3 and 4. Qualitative results are shown in Fig. 4, Suppl. Fig. 7 and 8. The results show that all three baseline methods yield large fragments for clearly visible and easily separable neurons. There are, however, many segmentation errors as reflected by the low avF1 scores (maximum value of $0.34_{\pm 0.01}$). PatchPerPix achieves best results for all metrics, except for false merges. Inspecting PatchPerPix results visually shows that many touching neurons are falsely merged even with different color. FFNs and Duan et al. perform similar, although Duan et al. has the highest number of false splits and lowest number of false merges. Thus it separates best touching neurons.

As the training of PatchPerPix is not directly applicable to the partly labeled data we only report results for models trained on the completely labeled data. FFNs, on the other hand, operate in a one-versus-all fashion and can thus by design train on partly labeled data without any modifications. Training FFN on the full dataset shows increases in all metrics, especially on the test set. PatchPerPix intrinsically handles overlapping instances but it can only bridge small overlaps up to the used patch size. FFN does not support overlaps out of the box but could, with some modifications for efficient inference, be extended to this end. None of the learnt methods models long-range data dependencies. In summary, all three baseline methods do yield some true positive neuron reconstructions, but extensive further method development is necessary to be able to achieve high quality instance segmentation on this dataset.

5. Conclusion

With this work we release the FISBe dataset, which is, to the best of our knowledge, the first real-world benchmark dataset for instance segmentation of wide-ranging thin filamentous intertwined objects. In addition to the data we contribute a set of metrics for meaningful method benchmarking and three baselines. A limitation of FISBe is its bias towards sparser samples from the FlyLight MCFO image resource it stems from. This entails that in general, benchmarking on FISBe does not serve to gauge method performance on denser samples. A possible avenue to mitigate this issue is to define proxy evaluation metrics on denser samples with the help of domain-specific downstream tasks for which higher-level annotations exist (see Sec. 3.2 for examples of such tasks). Main limitations of our baseline methods are that they handle no or only small overlaps and are computationally very demanding. In future work, we are excited to see recently proposed methods for

Table 1. Quantitative results on the full FISBe validation and test sets (i.e., completely and partly labeled data combined; for results on the respective subsets see Suppl. Table 3). We compare PatchPerPix (ppp, [39]), Flood Filling Networks (FFN, [26]) trained on the completely labeled and the full training dataset (+partly) and Duan et al.'s color clustering [13]. We report mean and standard deviation (±) over three independent runs (except for Duan et al.'s as it is non-learnt). For all scores except FS and FM higher values are better.

Split	Method	S	$avF1$	C	$clDice_{TP}$	FS	FM	C_{dim}	C_{ovlp}	tp	tp_{dim}	tp_{ovlp}
Val	ppp	$0.38_{\pm 0.02}$	$0.41_{\pm 0.02}$	$0.35_{\pm 0.01}$	$0.75_{\pm 0.02}$	$6.0_{\pm 0.8}$	$24_{\pm 1.6}$	$0.12_{\pm 0.01}$	$0.38_{\pm 0.04}$	$0.46_{\pm 0.01}$	$0.16_{\pm 0.04}$	$0.39_{\pm 0.03}$
	FFN	$0.25_{\pm 0.01}$	$0.27_{\pm 0.01}$	$0.23_{\pm 0.01}$	$0.79_{\pm 0.01}$	$7.0_{\pm 2.9}$	$12_{\pm 2.0}$	$0.03_{\pm 0.01}$	$0.30_{\pm 0.01}$	$0.32_{\pm 0.01}$	$0.04_{\pm 0.01}$	$0.37_{\pm 0.02}$
	FFN+partly	$0.27_{\pm 0.01}$	$0.29_{\pm 0.01}$	$0.24_{\pm 0.01}$	$0.79_{\pm 0.01}$	$7.7_{\pm 2.6}$	$14_{\pm 0.8}$	$0.02_{\pm 0.01}$	$0.33_{\pm 0.02}$	$0.34_{\pm 0.03}$	$0.03_{\pm 0.00}$	$0.38_{\pm 0.04}$
	Duan et al.	0.24	0.26	0.22	0.70	14	13	0.02	0.28	0.37	0.03	0.42
Test	ppp	$0.35_{\pm 0.00}$	$0.34_{\pm 0.01}$	$0.35_{\pm 0.01}$	$0.80_{\pm 0.00}$	$19_{\pm 2.9}$	$52_{\pm 3.4}$	$0.16_{\pm 0.03}$	$0.27_{\pm 0.04}$	$0.36_{\pm 0.01}$	$0.19_{\pm 0.04}$	$0.19_{\pm 0.03}$
	FFN	$0.25_{\pm 0.03}$	$0.22_{\pm 0.04}$	$0.29_{\pm 0.02}$	$0.80_{\pm 0.01}$	$17_{\pm 1.7}$	$39_{\pm 5.3}$	$0.03_{\pm 0.01}$	$0.26_{\pm 0.03}$	$0.32_{\pm 0.03}$	$0.00_{\pm 0.00}$	$0.24_{\pm 0.05}$
	FFN+partly	$0.27_{\pm 0.01}$	$0.24_{\pm 0.02}$	$0.31_{\pm 0.00}$	$0.80_{\pm 0.01}$	$18_{\pm 3.7}$	$36_{\pm 3.6}$	$0.04_{\pm 0.01}$	$0.28_{\pm 0.01}$	$0.36_{\pm 0.01}$	$0.03_{\pm 0.00}$	$0.28_{\pm 0.01}$
	Duan et al.	0.30	0.27	0.33	0.77	45	29	0.03	0.36	0.37	0.03	0.34

Test	$F1_{0.1}$	$F1_{0.2}$	$F1_{0.3}$	$F1_{0.4}$	$F1_{0.5}$	$F1_{0.6}$	$F1_{0.7}$	$F1_{0.8}$	$F1_{0.9}$
ppp	$0.50_{\pm 0.01}$	$0.48_{\pm 0.01}$	$0.44_{\pm 0.01}$	$0.41_{\pm 0.02}$	$0.35_{\pm 0.02}$	$0.29_{\pm 0.02}$	$0.26_{\pm 0.01}$	$0.19_{\pm 0.02}$	$0.12_{\pm 0.01}$
FFN	$0.34_{\pm 0.05}$	$0.31_{\pm 0.04}$	$0.28_{\pm 0.04}$	$0.25_{\pm 0.05}$	$0.22_{\pm 0.04}$	$0.20_{\pm 0.04}$	$0.17_{\pm 0.03}$	$0.12_{\pm 0.01}$	$0.07_{\pm 0.01}$
FFN+partly	$0.36_{\pm 0.02}$	$0.32_{\pm 0.02}$	$0.30_{\pm 0.02}$	$0.27_{\pm 0.03}$	$0.25_{\pm 0.03}$	$0.21_{\pm 0.03}$	$0.18_{\pm 0.02}$	$0.15_{\pm 0.02}$	$0.09_{\pm 0.01}$
Duan et al.	0.43	0.38	0.35	0.33	0.31	0.29	0.20	0.12	0.06

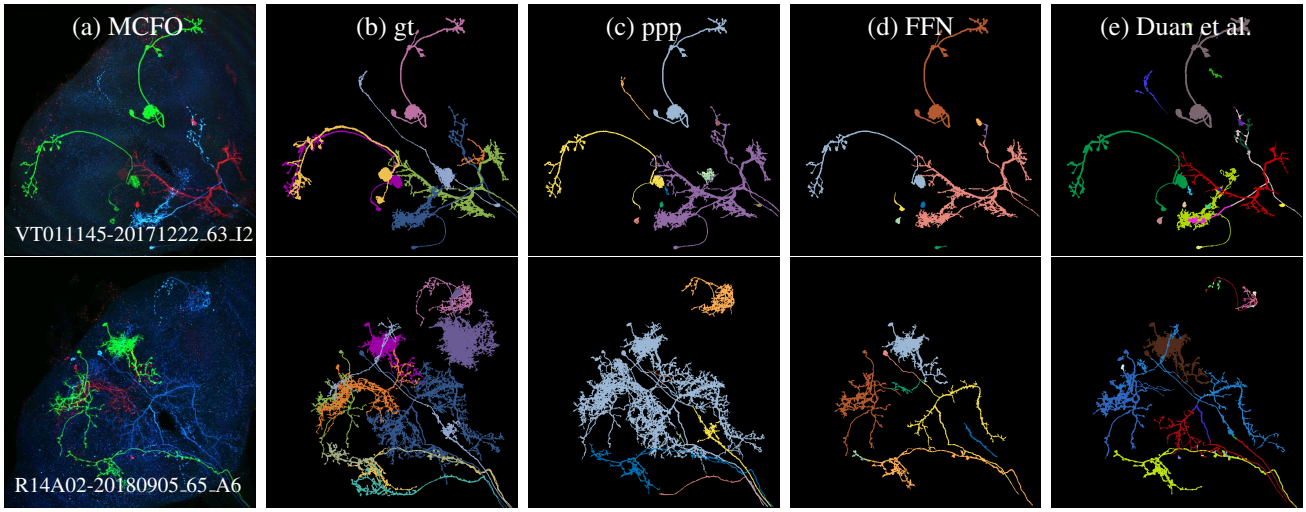

Figure 4. Qualitative results for our three baseline methods: PatchPerPix (ppp), Flood Filling Networks (FFN) and Duan et al.'s color clustering. In the top row all three methods yield few correctly segmented neurons (the two green neurons), but ppp and FFN merge the blue and the red one, and Duan et al.'s splits the blue neuron while nicely segmenting the red one. In the bottom row ppp merges many neurons of different color; FFN segments three neurons, but has low coverage; and Duan et al.'s also merges different colored neurons.

long-range data dependencies, such as structured state space models [23, 44] and continuous CNNs [30], applied to this real-world dataset. In addition, the huge set of already available non-annotated images should lend itself perfectly for self-supervised pretraining. We believe that the new challenging FISBe dataset is a great resource to the computer vision community as it might reveal blind spots of current methods. Thus, we hope that it will lead to new methods development for capturing long range data dependencies, while at the same time advancing cell-level analyses in basic neuroscience.

Acknowledgements. We thank Aljoscha Nern for providing unpublished MCFO images as well as Geoffrey W. Meissner and the entire FlyLight Project Team for valuable discussions. P.H., L.M. and D.K. were supported by the HHMI Janelia Visiting Scientist Program.

References

[1] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015. 2

[2] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4720–4728, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 2, 4

[3] Kerry M Brown, Germán Barrionuevo, Alison J Canty, Vincenzo De Paola, Judith A Hirsch, Gregory S X E Jefferis, Ju Lu, Marjolein Snippe, Izumi Sugihara, and Giorgio A Ascoli. The DIADEM data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions. *Neuroinformatics*, 9(2-3):143–157, 2011. 2

[4] Juan C. Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W. Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, and Anne E. Carpenter. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A*, 95(9):952–965, 2019. 2

[5] Juan C. Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W. Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, and Anne E. Carpenter. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A*, 95(9):952–965, 2019. 4

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 2

[7] Long Chen, Martin Strauch, and Dorit Merhof. Instance segmentation of biomedical images with an object-aware embedding learned with local constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–459. Springer, 2019. 2

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[10] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[12] Michael-John Dolan, Ghislain Belliart-Guérin, Alexander Shakeel Bates, Shahar Frechter, Aurélie Lampin-Saint-Amaux, Yoshinori Aso, Ruairí JV Roberts, Philipp Schlegel, Allan Wong, Adnan Hammad, et al. Communication from learned to innate olfactory processing centers is required for memory retrieval in drosophila. *Neuron*, 100(3):651–668, 2018. 3

[13] Bin Duan, Logan A Walker, Douglas H Roossien, Fred Y Shen, Dawen Cai, and Yan Yan. Unsupervised neural tracing in densely labeled multispectral brainbow images. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1122–1126, 2021. 2, 3, 7, 8

[14] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series. *CoRR*, abs/1807.01232, 2018. 2

[15] Adrien Foucart, Olivier Debeir, and Christine Decaestecker. Panoptic quality should be avoided as a metric for assessing cell nuclei segmentation and classification in digital pathology. *Scientific Reports*, 13(1):8614, 2023. 5

[16] C. Fowlkes, D. Martin, and J. Malik. Learning affinity functions for image segmentation: combining patch-based and gradient-based approaches. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–54, 2003. 2

[17] Drew Friedmann, Albert Pun, Eliza L Adams, Jan H Lui, Justus M Kebschull, Sophie M Grutzner, Caitlin Castagnola, Marc Tessier-Lavigne, and Liqun Luo. Mapping mesoscale axonal projections in the mouse brain using a 3d convolutional network. *Proceedings of the National Academy of Sciences*, 117(20):11068–11075, 2020. 3

[18] J. Funke, E. Perlman, S. Turaga, D. Bock, and S. Saalfeld. Cremi challenge, 2016. 2, 3

[19] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP: 1–1, 2018. 2

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. 3

[21] Todd A Gillette, Kerry M Brown, and Giorgio A Ascoli. The DIADEM metric: comparing multiple reconstructions of the same neuron. *Neuroinformatics*, 9(2-3):233–245, 2011. 4

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, Washington, DC, USA, 2014. IEEE Computer Society. 2

[23] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *CoRR*, abs/2111.00396, 2021. 2, 8

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017. cite arxiv:1703.06870Comment: open source; appendix on more results. 2

[25] Michal Januszewski, Jeremy Maitin-Shepard, Peter Li, Jörgen Kornfeld, Winfried Denk, and Viren Jain. Flood-filling networks. *CoRR*, abs/1611.00421, 2016. 7

[26] Michał Januszewski, Jörgen Kornfeld, Peter H. Li, Art Pope, Tim Blakely, Larry Lindsey, Jeremy Maitin-Shepard, Mike Tyka, Winfried Denk, and Viren Jain. High-precision automated reconstruction of neurons with flood-filling networks. *Nature Methods*, 15(8):605–610, 2018. 2, 3, 7, 8

[27] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732, 2022. 3

[28] Junkyung Kim, Drew Linsley, Kalpit Thakkar, and Thomas Serre. Disentangling neural mechanisms for perceptual grouping. *CoRR*, abs/1906.01558, 2019. 2

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3

[30] David M Knigge, David W Romero, Albert Gu, Efstratios Gavves, Erik J Bekkers, Jakub M Tomczak, Mark Hoogendoorn, and Jan-Jakob Sonke. Modelling long range dependencies in nd: From task-specific to a general purpose cnn. *arXiv preprint arXiv:2301.10540*, 2023. 2, 8

[31] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[32] Kisuk Lee, Ran Lu, Kyle Luther, and H Sebastian Seung. Learning and segmenting dense voxel embeddings for 3d neuron reconstruction. *IEEE Transactions on Medical Imaging*, 40(12):3801–3811, 2021. 2

[33] T.C. Lee, R.L. Kashyap, and C.N. Chu. Building skeleton models via 3-d medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6): 462–478, 1994. 5

[34] Rui Li, Muye Zhu, Junning Li, Michael S Bienkowski, Nicholas N Foster, Hanpeng Xu, Tyler Ard, Ian Bowman, Changle Zhou, Matthew B Veldman, et al. Precise segmentation of densely interweaving neuron clusters using g-cut. *Nature communications*, 10(1):1549, 2019. 2

[35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[36] Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 2

[37] Xingzheng Lyu, Li Cheng, and Sanyuan Zhang. The reta benchmark for retinal vascular tree analysis. *Scientific Data*, 9(1):397, 2022. 2

[38] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Büttner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Pitfalls and recommendations for image analysis validation, 2022. 2, 4, 5

[39] Lisa Mais, Peter Hirsch, and Dagmar Kainmueller. Patchperpix for instance segmentation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*, page 288–304, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 3, 7, 8

[40] Linus Manubens-Gil, Zhi Zhou, Hanbo Chen, Arvind Ramanathan, Xiaoxiao Liu, Yufeng Liu, Alessandro Bria, Todd Gillette, Zongcai Ruan, Jian Yang, et al. Bigneuron: a resource to benchmark and predict performance of algorithms for automated tracing of neurons in light microscopy datasets. *Nature Methods*, pages 1–12, 2023. 2

[41] Pavel Matula, Martin Maška, Dmitry V. Sorokin, Petr Matula, Carlos Ortiz-de Solórzano, and Michal Kozubek. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLOS ONE*, 10(12):1–19, 2015. 4, 5

[42] Geoffrey W Meissner, Aljoscha Nern, Zachary Dorman, Gina M DePasquale, Kaitlyn Forster, Theresa Gibney, Joanna H Hausenfluck, Yisheng He, Nirmala A Iyer, Jennifer Jeter, Lauren Johnson, Rebecca M Johnston, Kelley Lee, Brian Melton, Brianna Yarbrough, Christopher T Zugates, Jody Clements, Cristian Goina, Hideo Otsuna, Konrad Rokicki, Robert R Svirskas, Yoshinori Aso, Gwyneth M Card, Barry J Dickson, Erica Ehrhardt, Jens Goldammer, Masayoshi Ito, Dagmar Kainmueller, Wyatt Korff, Lisa Mais, Ryo Minegishi, Shigehiro Namiki, Gerald M Rubin, Gabriella R Sterne, Tanya Wolff, Oz Malkesman, and FlyLight Project Team. A searchable image resource of *Drosophila* gal4 driver expression patterns with single neuron resolution. *eLife*, 12:e80660, 2023. 2, 3

[43] Aljoscha Nern, Barret D. Pfeiffer, and Gerald M. Rubin. Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proceedings of the National Academy of Sciences*, 112(22): E2967–E2976, 2015. 3

[44] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. In *Advances in Neural Information Processing Systems*, 2022. 2, 8

[45] Hanchuan Peng, Michael Hawrylycz, Jane Roskams, Sean Hill, Nelson Spruston, Erik Meijering, and Giorgio A Ascoli. BigNeuron: Large-Scale 3D neuron reconstruction from optical microscopy images. *Neuron*, 87(2):252–256, 2015. 2

[46] Tingwei Quan, Hang Zhou, Jing Li, Sw Li, Anan Li, Yuxin Li, Xiaohua Lv, Qingming Luo, Hui Gong, and Shaoqun Zeng. Neurogps-tree: Automatic reconstruction of large-scale neuronal populations with dense neurites. *Nature Methods*, 13, 2015. 2

[47] Douglas H Roossien, Benjamin V Sadis, Yan Yan, John M Webb, Lia Y Min, Aslan S Dizaji, Luke J Bogart, Cristina Mazuski, Robert S Huth, Johanna S Stecher, Sriakhila Akula, Fred Shen, Ye Li, Tingxin Xiao, Madeleine Vandenbrink, Jeff W Lichtman, Takao K Hensch, Erik D Herzog, and Dawen Cai. Multispectral tracing in densely labeled mouse brain with nTracer. *Bioinformatics*, 35(18):3544–3546, 2019. 2

[48] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. cldice - a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, 2021. 2, 5

[49] Janelia Scientific Computing Software. VVDViewer, 2019. open-source software funded by NIH grant R01-GM098151-01NIH grant R01-GM098151-01, https://github.com/JaneliaSciComp/VVDViewer. 4

[50] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. 2

[51] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021. 2

[52] Uygar Sümbül, Douglas Roossien, Dawen Cai, Fei Chen, Nicholas Barry, John P Cunningham, Edward Boyden, and Liam Paninski. Automated scalable segmentation of neurons from multispectral images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 2, 7

[53] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *CoRR*, abs/2011.04006, 2020. 2

[54] Srinivas C. Turaga, Kevin L. Briggman, Moritz Helmstaedter, Winfried Denk, and H. Sebastian Seung. Maximin affinity learning of image segmentation. *CoRR*, 2009. 2

[55] Donglai Wei, Kisuk Lee, Hanyu Li, Ran Lu, J. Alexander Bae, Zequan Liu, Lifu Zhang, Márcia dos Santos, Zudi Lin, Thomas Uram, Xueying Wang, Ignacio Arganda-Carreras, Brian Matejek, Narayanan Kasthuri, Jeff Lichtman, and Hanspeter Pfister. Axonem dataset: 3d axon instance segmentation of brain cortical regions. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 175–185, Cham, 2021. Springer International Publishing. 2, 3

[56] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[57] Johan Winnubst, Erhan Bas, Tiago A. Ferreira, Zhuhao Wu, Michael N. Economo, Patrick Edson, Ben J. Arthur, Christopher Bruns, Konrad Rokicki, David Schauder, Donald J. Olbris, Sean D. Murphy, David G. Ackerman, Cameron Arshadi, Perry Baldwin, Regina Blake, Ahmad Elsayed, Mashtura Hasan, Daniel Ramirez, Bruno Dos Santos, Monet Weldon, Amina Zafar, Joshua T. Dudman, Charles R. Gerfen, Adam W. Hantman, Wyatt Korff, Scott M. Sternson, Nelson Spruston, Karel Svoboda, and Jayaram Chandrashekar. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell*, 179(1):268–281.e13, 2019. 2

[58] Steffen Wolf, Constantin Pape, Alberto Bailoni, Nasim Rahaman, Anna Kreshuk, Ullrich Köthe, and Fred A. Hamprecht. The mutex watershed: Efficient, parameter-free image partitioning. In *ECCV (4)*, pages 571–587. Springer, 2018. 2

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 2

[60] Hang Zhou, Shiwei Li, Anan Li, Qing Huang, Feng Xiong, Ning Li, Jiacheng Han, Hongtao Kang, Yijun Chen, Yun Li, et al. Gtree: an open-source tool for dense reconstruction of brain-wide neuronal population. *Neuroinformatics*, 19:305–317, 2021. 2