Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quality Aware Candidate Selection for Silver Truth in Cell Segmentation

David Číž

david.ciz@vsb.cz

VSB - Technical University of Ostrava

João Manuel Da Silva Carias

VSB - Technical University of Ostrava

Monika Weissová

VSB - Technical University of Ostrava

Vladimír Ulman

VSB - Technical University of Ostrava

Research Article

Keywords: cell segmentation, silver truth, label fusion, quality assessment, instance segmentation, Cell Tracking Challenge

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11515932/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quality Aware Candidate Selection for Silver Truth in Cell Segmentation

David Číž[1*], João Manuel Da Silva Carias[1], Monika Weissová[1] and Vladimír Ulman[1]

[1*]IT4Innovations National Supercomputing Center, VSB – Technical University of Ostrava, 17. listopadu 2172/15, 708 00, Ostrava-Poruba, Czech Republic.

*Corresponding author(s). E-mail(s): david.ciz@vsb.cz;
Contributing authors: joao.carias@vsb.cz; monika.weissova@vsb.cz; vladimir.ulman@vsb.cz;

Abstract

Background. Quantitative microscopy requires accurate cell masks, but cell segmentation accuracy varies across individual cells and across segmentation models. Combining candidate masks available for a cell can yield more accurate segmentations, especially when poor-quality candidates are identified and eliminated. We test whether instance-level quality ranking and selection of candidate masks improves label fusion accuracy. **Methods.** We evaluate segmentation on five Cell Tracking Challenge datasets. Using reference tracking markers to collect candidate masks from competitor submissions for each cell, we extract aligned crops of the candidate masks and the underlying microscopy image. A deep quality estimator predicts the intersection over union (IoU) for each candidate, and a pairwise ranker prunes degraded masks to retain the most promising candidates. Our ENSEMBLE pipeline combines the retained candidates using learned Deep Sets weighting and a boundary-aware vote. For a controlled comparison, we also apply the established SIMPLE fusion method to the same selected candidates as well as unfiltered pools. Using a two-fold cross-validation scheme across paired video sequences, we evaluate IoU on aligned cell crops and on fully reconstructed segmentation images. **Results.** The Deep Sets ENSEMBLE achieves 0.8676 mean IoU. SIMPLE with the same quality-based candidate selection achieves 0.8654, compared

with 0.8431 for the descriptive best fusion baseline without candidate selection. Reconstructing Deep Sets predictions as full images yields 0.8664 cell-weighted mean IoU with zero cell loss. The small difference between ENSEMBLE and SIMPLE indicates that instance-level candidate selection produces most of the improvement. The descriptive reference formed by selecting the best individual competitor in each fold averages 0.8712. **Conclusions.** Ranking and pruning candidate masks for each individual cell substantially improves label fusion when candidate quality varies. ENSEMBLE achieves the leading fusion result, while the comparable performance of Selected SIMPLE demonstrates that the primary practical gain stems from quality-aware candidate selection prior to fusion.

Keywords: cell segmentation, silver truth, label fusion, quality assessment, instance segmentation, Cell Tracking Challenge

1 Introduction

Quantitative cell biology depends on measurements such as cell size, shape, proliferation rate, or migration speed. Extracting these measurements from microscopy images requires cell segmentation, in which each cell is assigned a mask that separates it from the background and neighboring cells. Creating these masks manually is typically tedious, thus slow, and requires an expert to annotate every cell. Segmentation models, on the other hand, make analysis at a larger scale possible, but cannot guarantee high segmentation accuracy for every single cell. Typically, a model shows different accuracy with changes of local image contrast, cell shape or cell density.

The average performance of a model would improve, if the less accurate masks could be refined, or eliminated. For a new dataset, majority of models may be improved by retraining (or fine-tuning) them. But this requires to have annotated training data prepared and that is often deemed impractical as expert annotation capacity is usually limited. Eliminating low quality masks from a segmentation result is theoretically more feasible approach when an expert is available to swiftly review and discard the wrong masks, or when

a computational mask quality classifier is available. However, cell biology quantification cannot be carried out on a cell whose masks have been removed.

If multiple segmentation results—acting as candidate masks—from different models are available for a dataset, it becomes possible to retain a complete, dense segmentation of the dataset even when poor individual segmentation masks are eliminated. Whenever one model produces an inaccurate mask or misses a cell entirely, alternative candidate masks from other models remain available to preserve coverage.

Combining multiple segmentation outputs is a well-established paradigm in bioimage analysis, represented by classic label fusion algorithms such as STAPLE [1] and SIMPLE [2]. A prominent recent example of this paradigm is the creation of the silver truth reference annotations in the Cell Tracking Challenge (CTC) [3], where combining multiple algorithm submissions—which were not necessarily complete on their own—achieved impressive spatial coverage and annotation quality.

Crucially, however, naive fusion of all available predictions is vulnerable to majority bias, which occurs when several models make correlated errors on a challenging cell and outvote a minority of accurate predictions. Indeed, the CTC silver truth annotations were generated by fusing suitably selected subsets of available segmentations rather than blindly merging all submissions [3]. The authors reported that their silver truth annotations needed to be obtained using a different subset per video sequence in order to avoid the majority bias. Furthermore, their subsets could have been decided only on the few video frames for which the precise ground truth annotations were available, keeping the subset fixed for the rest of the frames. In the absence of ground truth, selecting candidates per each cell is a true challenge. Therefore, to apply this principle automatically at the level of individual cells, an independent quality

control mechanism is needed to estimate the quality of individual candidate masks and discard poor predictions before fusion takes place.

In this paper, we address these challenges by focusing on two critical steps: instance-level mask quality assessment (QA) and candidate mask fusion. First, we use an independent quality estimator to score and rank individual candidate masks in the context of the underlying microscopy image, eliminating low-quality candidates on a per-cell basis. Second, we fuse the surviving top-ranked candidates into a single consensus mask. To disentangle how much performance gain originates from quality-aware candidate selection versus the fusion algorithm itself, we systematically compare the effect of our QA filtering on both a classical fusion baseline (SIMPLE [2]) and our novel learned, permutation-invariant Deep Sets fusion method.

We evaluate the pipeline on five Cell Tracking Challenge datasets [3]. For each dataset, one annotated sequence supplies training and validation data while the other is reserved for testing. The sequences are then switched to form the second fold. This design tests transfer between videos with different cell shapes, densities, image quality, and patterns of failure among candidate sources. To evaluate the first video of a dataset, the quality estimator and ranker were trained on the second video of the same dataset. Deep Sets pools training rows across all five datasets, excluding the evaluated video (Section 4.2).

Our contributions are:

1. **Instance-level quality-aware candidate pruning:** We introduce an image-conditioned quality assessment and ranking stage that evaluates individual candidate masks and filters out degraded predictions prior to fusion, preventing a “bad majority” of erroneous masks from corrupting the consensus while preserving dense cell coverage.

2. **Learned permutation-invariant mask fusion:** We propose a neural fusion architecture based on Deep Sets [4] that aggregates an arbitrary subset of top-ranked candidate masks alongside local image-support and contextual features into a refined instance mask.
3. **Controlled ablation of selection versus fusion:** Through experiments on Cell Tracking Challenge datasets, we decouple the impact of candidate pruning from the choice of the fusion backend. We demonstrate that quality-aware candidate selection accounts for the majority of the improvement (+0.0223 intersection over union (IoU) over the descriptive best unfiltered fusion baseline of 0.8431), raising classical SIMPLE fusion to 0.8654 IoU (Selected SIMPLE), while our learned Deep Sets ENSEMBLE further advances instance-level performance to 0.8676 IoU (+0.0021 over Selected SIMPLE) and achieves 0.8664 cell-weighted mean IoU in full-image reconstruction with zero cell loss.

2 Related Work

2.1 Candidate Generation for Cell Segmentation

Deep learning is now widely used for image segmentation [5]. Segment Anything (SAM) introduced prompt based segmentation for natural images [6], and MedSAM adapted this approach to medical images [7]. More recent systems focus on microscopy. The μSAM model supports several forms of microscopy data [8], while CellSAM is designed specifically for cell segmentation [9]. A recent benchmark of foundation models for cell instance segmentation shows that candidate generators continue to develop rapidly [10]. These systems can provide useful candidate masks, but they do not decide how several masks for the same cell should be selected and combined.

2.2 Fusion of Multiple Segmentations

Methods for combining several segmentations have a long history in medical image analysis. Simultaneous Truth and Performance Level Estimation (STAPLE) estimates both a hidden reference segmentation and the reliability of each source [1]. The Selective and Iterative Method for Performance Level Estimation (SIMPLE) repeatedly removes unreliable sources before estimating a consensus [2].

The Cell Tracking Challenge (CTC) creates silver truth by collecting cell masks from strong submitted methods and combining them with the help of reference tracking markers [3]. This process provides dense annotations, and the tracking markers make it possible to align masks from different sources with individual cells. We use this alignment to study candidate selection and fusion as separate stages.

Learned fusion methods can make selection more local. In multi atlas brain segmentation, VoteNet predicts which registered atlas labels are trustworthy at each location before combining the selected labels by voting [11]. DeepFuse instead learns to correct and combine full image masks from several sources. It was evaluated on CTC data and improved on STAPLE, SIMPLE, and voting methods [12]. DeepFuse is the closest prior work aimed directly at creating biomedical silver truth, but its main operation is full image refinement and fusion rather than selecting candidate masks separately for each cell.

2.3 Quality Assessment and Selection

Quality assessment estimates the accuracy of a segmentation when ground truth is unavailable at inference. Reverse Classification Accuracy estimates quality by training a reverse classifier from the proposed segmentation and testing it on reference images [13]. Other methods learn to predict an overlap

score directly from the image, the proposed mask, or uncertainty information [14–16]. Mask Scoring R-CNN brought this idea to instance segmentation by predicting the intersection over union of each mask and using the estimate to rank detections [17].

Predicted quality can also choose among complete segmentation results. Hann et al. select the best result from a group of cardiac segmentation networks for each case [18]. EvanySeg predicts Dice scores and uses them to select among outputs from SAM based models [19]. Recent work has also explored quality assessment that transfers across organs and imaging datasets [20].

These studies show that a quality score can support ranking, weighting, or filtering, but the three uses place different demands on the score. Ranking only requires the order of candidates to be useful. A fixed filter also requires scores to have the same meaning across datasets. Our method therefore uses relative ranking within each candidate cluster before learned weighting.

Candidates aligned to one cell form a set whose order has no meaning. Deep Sets provides a general architecture for learning from such inputs by processing each member and then pooling information across the complete set [4]. This makes it suitable for assigning candidate weights using both the properties of each mask and the context supplied by competing masks.

2.4 Research Gap

Prior work has selected reliable atlas labels at each image location, selected complete segmentations for individual cases, and learned to refine several full image masks. Less is known about selecting masks separately for each cell when the selection method must transfer between microscopy videos. We test this setting across five CTC datasets and learn candidate weights with a set based model. ENSEMBLE is our primary pipeline, while applying SIMPLE to

the same selected candidates shows how much of the improvement comes from selection rather than the final combination method.

3 Methods

3.1 Framework Overview

ENSEMBLE combines candidate masks separately for each cell. The input consists of segmentation masks from several candidate sources and reference tracking markers that identify the cells. The pipeline has five stages. First, masks from different sources are aligned with the tracking markers to form one candidate cluster for each cell. Second, every candidate is described using predicted quality, source reliability, agreement, image, shape, and temporal evidence. Third, a pairwise ranker orders the candidates and retains a small pool that contains no more than one mask from each source. Fourth, a Deep Sets model assigns a weight to every retained mask using information from the whole candidate cluster. Finally, a weighted vote produces the cell mask. The cell masks can be scored as aligned crops or returned to their original image coordinates to form a complete label map. Figure 1 summarizes these stages and the controlled comparison with SIMPLE.

3.2 Candidate Alignment

For every reference tracking marker, we examine the segmentation supplied by each candidate source. The source label with the largest marker overlap is assigned to that cell when at least 50% of the marker pixels coincide with the label. The matched label is converted into a binary candidate mask and cropped around the cell. Candidates assigned to the same marker form a candidate cluster. This gives every cluster a consistent cell identity while preserving the different boundaries proposed by the candidate sources.

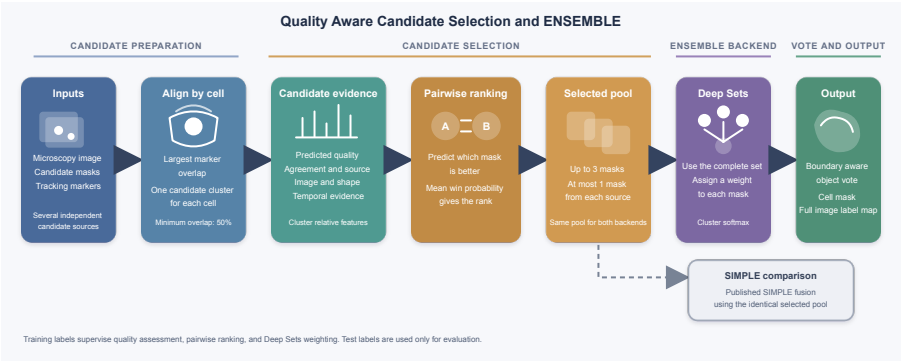


Figure 1 The ENSEMBLE pipeline. Candidate masks are aligned with reference tracking markers, described by quality and context evidence, ranked, and reduced to a pool containing different sources. Deep Sets assigns candidate weights before a boundary aware object vote. SIMPLE receives the identical selected pool as a controlled comparison.

The alignment step also records the candidate source, crop coordinates, cell center, frame number, agreement count, and the location of the original image. These values support ranking and allow every predicted cell mask to be placed back into its source image.

3.3 Candidate Evidence and Quality Assessment

Quality assessment (QA) estimates candidate accuracy without using its test label. Each candidate is described by the following evidence:

- **predicted quality:** the estimated IoU of the candidate mask;
- **agreement:** the number and rank of sources that propose a compatible cell near the same location;
- **source reliability:** the mean IoU of the candidate source on training cells;
- **shape and crop geometry:** mask area, extent, perimeter, connected components, centroid offset, border contact, and visible crop fraction;
- **image support:** intensity contrast inside the mask and along its boundary, measured in the crop and in original image coordinates;
- **temporal evidence:** distance and support from candidates in adjacent frames when those candidates are available.

Features whose scale depends on the image or cell type are also expressed as ranks or differences relative to the other candidates in the same cluster. For example, mask area is recorded both directly and as its log ratio to the median area in the cluster.

The quality estimator is a ResNet18 regression model [21]. Its two input channels contain the raw microscopy crop and the binary candidate mask. The first convolution accepts two channels, and the final sigmoid output estimates IoU on the interval $[0, 1]$. The target is the IoU between the candidate and ground truth cell masks, and the network is trained with mean squared error using Adam and weight decay 10^{-4}. Ground truth is used to form this target only for training and validation cells.

3.4 Pairwise Ranking and Candidate Selection

Candidate ranking is local to each cell. During training, two candidates from the same cluster form a pair when their true IoU values differ by at least 0.03. A logistic regression model receives the difference between their feature vectors and predicts which candidate has the higher IoU. At inference, every candidate is compared with every other candidate in its cluster. Its pairwise score is the mean predicted probability that it wins these comparisons.

Candidates are ordered by the pairwise score. The selection rule retains at most three candidates per cluster and at most one candidate from each source. This fixed rule preserves useful disagreement between sources while preventing one source from occupying several positions in the retained pool. If fewer than three sources provide matched masks, all available distinct sources are retained.

3.5 Deep Sets Candidate Weighting

The primary ENSEMBLE pipeline uses Deep Sets [4] to assign weights to the retained candidates. Let x_i denote the feature vector for candidate i in

a cluster C. A shared network ϕ maps every candidate to a 64 dimensional representation. The mean representation of the cluster provides context for the complete candidate pool:

$$h_i = \phi(x_i), \qquad \bar{h}_C = \frac{1}{|C|} \sum_{j \in C} h_j. \tag{1}$$

The scoring network ρ receives the candidate representation and the cluster mean, then produces a scalar score:

$$z_i = \rho\big([h_i, \bar{h}_C]\big)\,. \tag{2}$$

Both networks use fully connected layers with ReLU activations. The shared network has two 64 unit layers. The scoring network has one 64 unit layer and a sigmoid output. Numeric features are standardized using the training data; dataset and candidate source are one-hot encoded and concatenated with the numeric features.

The training target gives the best candidate in each cluster a value of one and decreases smoothly for candidates with lower IoU, using temperature parameter $\tau = 0.08$ to sharpen target weights toward the best candidate:

$$y_i = \exp\left(\frac{\mathrm{IoU}_i - \max_{j \in C} \mathrm{IoU}_j}{0.08}\right). \tag{3}$$

The model minimizes mean squared error between z_i and y_i. It is trained for 80 epochs with AdamW, a learning rate of 10^{-3}, weight decay 10^{-4}, batches of 32 candidate clusters, and random seed 13. For each target fold, retained candidates are pooled across all five datasets. Every row from the target dataset–video pair is excluded before fitting, and only rows marked train supply labels. Section 4.2 reports the resulting fitting pools.

The scores are converted to positive weights by a softmax within each cluster, using temperature $T = 1$:

$$w_i = \frac{\exp(z_i/T)}{\sum_{j \in C} \exp(z_j/T)}. \tag{4}$$

Because the cluster mean is unchanged by candidate order, the resulting weights are permutation invariant.

3.6 Boundary Aware Object Vote

For candidate mask M_i and pixel p, the vote uses the learned weight w_i. On the one pixel inner boundary of the mask, the weight is multiplied by the candidate's local image support s_i. This support is the mean rank of boundary contrast and normalized intensity contrast measured locally and in original image coordinates. It is limited to the interval $[0.75, 1]$ for the vote. The pixel vote is

$$a_i(p) = \begin{cases} s_i, & p \in \partial M_i, \\ 1, & p \notin \partial M_i. \end{cases} \tag{5}$$

$$V(p) = \sum_{i \in C} w_i a_i(p) M_i(p). \tag{6}$$

The predicted cell mask contains pixels for which $V(p) \geq 0.48$. SIMPLE is evaluated on the same retained candidates, but uses its published iterative source selection and fusion procedure rather than the ENSEMBLE weights [2]. This comparison holds the candidate pool constant and changes only the final combination method.

3.7 Cell Crops and Full Image Reconstruction

The controlled primary evaluation compares methods on aligned cell crops. For full image reconstruction, every predicted crop is returned to its recorded coordinates and assigned its tracking marker label. When two cells claim the

same pixel, the label with the larger continuous vote $V(p)$ is retained. The reconstructed label map is then compared with the full ground truth image. This additional evaluation measures interactions between cells that are absent when each crop is scored independently.

4 Experimental Setup

4.1 Datasets

We evaluate the pipeline on five datasets from the Cell Tracking Challenge [3]: **BF-C2DL-HSC** with 64 px crops, **BF-C2DL-MuSC** with 256 px crops, **DIC-C2DH-HeLa** with 256 px crops, **Fluo-C2DL-MSC** with 1024 px crops, and **Fluo-N2DH-GOWT1** with 160 px crops. The fixed crop sizes match the cell scale and surrounding image context in each dataset. They are part of preprocessing and were not selected from test results. The datasets include bright field, differential interference contrast, and fluorescence microscopy, with both crowded, low contrast cells and settings where individual candidate sources are already strong.

We use a leave one sequence out evaluation. For each dataset, sequence 01 is used for training and validation while sequence 02 is tested, and then the roles are reversed. The sequence used for model development has frame-level train and validation assignments. This produces two test-video folds for each dataset. The quality estimator and ranker use the paired video; Deep Sets uses pooled training rows with the target video excluded, as detailed in Section 4.2.

Within a fold, IoU is averaged over all test cells. The score for a dataset is the mean of its two fold scores, giving both sequences equal weight. The overall score is the mean of the five dataset scores, giving each dataset equal weight regardless of its number of cells.

4.2 Data Used by Each Component

The sequence split has different roles at different stages. The quality estimator and pairwise ranker use the dataset-specific fold, whereas the Deep Sets model uses a pooled training collection. Table 1 states these roles explicitly.

Table 1 Data use during fitting, selection, and evaluation

Component	Fitting data	Selection or evaluation
Quality estimator	Candidate crops marked train in the other video of the same dataset; raw image and candidate mask are inputs, SEG mask IoU is the target.	Validation crops select the checkpoint. Test crops are scored without their SEG masks as inputs.
Source reliability and pairwise ranker	Source IoU summaries and candidate pairs from train rows of the dataset-specific fold.	The fitted ranker scores validation and test candidates. The retained pool contains at most three candidates from distinct sources.
Deep Sets	Train rows pooled from all ten retained candidate tables; every row from the target dataset–video pair is excluded before fitting. Other datasets contribute train rows from both video directions.	The fitted model assigns weights to the target fold's retained candidates. Validation summaries select the final configuration.
SIMPLE and object vote	No additional fitting to SEG labels in the test video; both receive the same retained candidate pool.	Test SEG masks are used to calculate the reported cell crop and reconstructed image IoU.

The recovered retained tables contain 4282 train rows before target exclusion. The Deep Sets fitting pool therefore varies from 3499 to 4048 candidate rows across folds (Table 2). This is pooled training with a held out target video, rather than training only on its paired video. Ground truth segmentation (SEG) supplies quality targets and evaluation masks; tracking markers (TRA) supply cell identities for alignment.

Additional file 1 records the exact annotated frame assignments recovered from these tables. Train and validation annotations are assigned at frame level;

the assignments need not form a single chronological cut. The original unpruned crop inventories are not reconstructed from the retained tables, so the counts below do not assert complete quality-estimator or ranker sample counts.

Table 2 Cell counts and pooled Deep Sets fitting rows by test video

Dataset	Test video	Train cells	Validation cells	Test cells	Deep Sets fitting rows
BF-C2DL-HSC	02	204	51	309	3499
BF-C2DL-HSC	01	261	48	255	3670
BF-C2DL-MuSC	02	174	43	298	3568
BF-C2DL-MuSC	01	238	60	217	3762
DIC-C2DH-HeLa	02	78	20	97	4048
DIC-C2DH-HeLa	01	78	19	98	4048
Fluo-C2DL-MSC	02	90	22	116	4003
Fluo-C2DL-MSC	01	93	23	112	4012
Fluo-N2DH-GOWT1	02	115	29	121	3991
Fluo-N2DH-GOWT1	01	97	24	144	3937

Cell counts describe annotated cells represented in the retained candidate tables, not the number of candidate crops used to train the quality estimator or ranker. Train and validation cells belong to the other video of the same dataset. Deep Sets fitting rows are retained candidates pooled across the five datasets after excluding the test video; several rows can describe one cell. The ten test folds contain 1767 cells in total. Additional file 1 lists the observed annotated frame assignments.

4.3 Candidate Sources

For each dataset, the candidate masks come from a fixed collection of available Cell Tracking Challenge competitor submissions, present consistently across both sequence 01 and sequence 02 of that dataset. The available sources differ among datasets, but each collection includes several independently developed segmentation methods. This diversity provides alternative cell boundaries for the ranking and fusion stages. Reference tracking markers align these alternatives to the same cells, allowing the experiment to focus on candidate selection and boundary fusion.

Table 3 describes the range of source quality in the evaluation data. The best value is descriptive: it identifies the strongest individual competitor separately in each fold after evaluation and is not a source selection rule for use on new data.

Table 3 Individual competitor quality by dataset

Dataset	Best	Median	Worst
BF-C2DL-HSC	0.8739	0.8156	0.6537
BF-C2DL-MuSC	0.8007	0.7125	0.5889
DIC-C2DH-HeLa	0.9608	0.8757	0.6138
Fluo-C2DL-MSC	0.7747	0.6637	0.5137
Fluo-N2DH-GOWT1	0.9460	0.8670	0.7312
Dataset mean	0.8712	0.7869	0.6202

Values are mean IoU across the two test sequence folds. The best column selects the strongest competitor separately in each fold before averaging.

The spread between the columns shows that candidate source quality varies widely. The descriptive best competitor reference in Table 4 uses the same values as the best column above. ENSEMBLE uses the variation among sources for each cell by ranking the candidate masks before fusion.

4.4 Comparison Hierarchy

We organize the comparisons by the question that each one answers:

1. **Descriptive best fusion baseline:** The strongest baseline based only on fusion is selected separately in each fold after evaluation. Across the ten folds, unpruned SIMPLE supplies nine of the ten fold values, while flat voting (BICv2) supplies the BF-C2DL-MuSC fold 1 value. It provides strong context but is not a method selection rule for new data.
2. **Descriptive best competitor reference:** The strongest individual competitor for each fold shows the quality of the best source in the evaluated collection. Because it is identified from evaluation results, it is context rather than a selection strategy.
3. **Quality based selection with SIMPLE:** Applying SIMPLE to the candidates retained by our ranking stage isolates the contribution of candidate selection. It uses exactly the same selected pool as the primary method.
4. **Deep Sets ENSEMBLE:** Our primary method applies learned candidate weights and the boundary aware object vote to that selected pool. Its

comparison with the preceding method isolates the contribution of the ENSEMBLE backend.

5. **CTC silver truth reference:** The challenge reference uses tracking markers, a selected group of strong methods, and thresholds tuned for each video against ground truth [3]. It is reported as a strong reference produced with additional challenge information, not as a directly comparable method for use on unseen data.

4.5 Evaluation Metric

We report the main comparisons using intersection over union (IoU), a standard overlap metric for cell segmentation [3]. Higher values are better. For a predicted cell mask $\hat{S}$ and ground truth cell mask S^*, IoU is

$$\mathrm{IoU}(\hat{S}, S^*) = \frac{|\hat{S} \cap S^*|}{|\hat{S} \cup S^*|} \tag{7}$$

IoU penalizes both missing and excess mask area. The controlled primary evaluation computes it on the aligned crop for each cell. The full image reconstruction check reports a cell weighted mean after all predicted cells have been returned to the complete label map and overlap conflicts have been resolved. F1 is also computed as a diagnostic but is not reported in the main comparison.

4.6 Implementation Details

The quality estimator and Deep Sets model are implemented in PyTorch. The evaluation scripts reconstruct the selected candidate masks, compute cell scores for each fold, and aggregate the fold and dataset means as described above.

Quality estimator. The modified ResNet18 is trained for at most 50 epochs with Adam, a learning rate of 10^{-4}, weight decay of 10^{-4}, and a batch

size of 16. Training stops when validation performance has not improved for 10 epochs. Augmentation uses random horizontal and vertical flips and random rotations in 90 degree steps.

Ranking and weighting. The pairwise ranker is fitted on labeled training cells from the sequence available for model development. The primary Deep Sets model uses pooled train rows from all five datasets, with every row from the target dataset–video pair excluded before fitting. The other datasets contribute training rows from both fold directions. Its training settings, weight conversion, and vote parameters are given in Section 3.5. The reported configuration is selected from validation summaries using mean validation IoU minus twice its standard deviation across folds, which favors both accuracy and stability. The fixed pipeline constants—including the ranking pair margin (0.03), candidate pool cap (3), target temperature sharpening ($\tau = 0.08$), boundary support floor (0.75), and object vote threshold (0.48)—were established via this validation stability sweep on the development sequences, without reference to test evaluation.

4.7 Use of generative artificial intelligence tools

OpenAI Codex was used during manuscript preparation to assist with literature searches, language editing, consistency checks, and code based verification of reported values. The authors reviewed the resulting material and take full responsibility for the accuracy, originality, and conclusions of the work.

5 Results

5.1 Main Five Dataset Comparison

Table 4 reports the controlled cell crop comparison. Deep Sets ENSEMBLE achieves the highest mean IoU among the tested fusion pipelines, reaching

0.8676. SIMPLE applied to the same selected candidates reaches 0.8654, while the descriptive best fusion baseline reaches 0.8431.

Table 4 Mean cell crop IoU by dataset

Dataset	Fusion	Best source	Selected SIMPLE	ENSEMBLE	CTC reference
BF-C2DL-HSC	0.8726	0.8739	0.8712	0.8718	0.8753
BF-C2DL-MuSC	0.7307	0.8007	0.8066	0.8091	0.8103
DIC-C2DH-HeLa	0.9307	0.9608	0.9525	0.9531	0.9647
Fluo-C2DL-MSC	0.7463	0.7747	0.7495	0.7560	0.8045
Fluo-N2DH-GOWT1	0.9350	0.9460	0.9474	0.9477	0.9546
Dataset mean	0.8431	0.8712	0.8654	**0.8676**	0.8819

Each dataset value is the mean of its two test sequence folds, and the final row gives equal weight to all five datasets. Fusion is the descriptive best fusion baseline, selected separately in each fold. Selected SIMPLE and Deep Sets ENSEMBLE use the same candidate pool. The best source is also selected separately in each fold. The CTC silver truth column is a descriptive reference.

Deep Sets ENSEMBLE is above selected SIMPLE on all five datasets. The gains range from 0.0003 IoU on GOWT1 to 0.0065 on MSC. This consistent numerical lead supports ENSEMBLE as the primary method, although the mean difference of 0.0021 is modest. A bootstrap over folds, stratified by dataset, gives a 95% confidence interval of [0.8552, 0.8799] for the ENSEMBLE mean, reflecting cross-fold variation. For the paired difference from selected SIMPLE, a paired bootstrap across the ten folds yields a 95% confidence interval of [+0.0006, +0.0043] (two-sided Wilcoxon signed-rank test $W = 1.0, p = 0.0039$, with Deep Sets higher in 9 of 10 folds and on all 5 datasets), indicating that the modest lead over selected SIMPLE is statistically consistent across evaluation folds.

The descriptive best competitor reference averages 0.8712. It is assembled by identifying the strongest individual competitor in each evaluated fold, so it provides context rather than a rule that could select a source for unseen data. ENSEMBLE exceeds this reference on MuSC and GOWT1 and is within 0.0021 on HSC. The CTC silver truth reference remains highest at 0.8819 and uses the additional challenge information described in Section 4.4.

5.2 Contribution of Selection and Fusion

Table 5 Contribution of candidate selection and the ENSEMBLE backend

Pipeline	Mean IoU	Change
Descriptive best fusion baseline	0.8431	reference
Selected candidates with SIMPLE	0.8654	+0.0223
Deep Sets ENSEMBLE	**0.8676**	+0.0021

Changes are absolute IoU differences from the preceding row, computed from unrounded values. The descriptive best fusion baseline draws nine fold values from unpruned SIMPLE and one from flat voting (BICv2).

Applying SIMPLE after quality based candidate selection improves the mean by 0.0223 IoU over the fusion baseline. Replacing SIMPLE with the Deep Sets weights and boundary aware vote adds 0.0021, for a total improvement of 0.0245 over the fusion baseline. Candidate selection therefore accounts for most of the observed improvement, while the ENSEMBLE backend provides a smaller gain and the leading final result.

To further separate the roles of candidate selection and feature representations: selecting merely the single top-ranked candidate per cell without fusion yields 0.8582 mean IoU (+0.0151 over unpruned fusion), showing that ranking alone captures most of the available signal. Furthermore, an ablation that removes explicit QA feature columns (predicted IoU and image ranks) from the final Deep Sets weighting stage achieves 0.8669 mean IoU (a difference of only −0.0007 from the full model). This indicates that while upstream pairwise candidate ranking benefits from QA predictions, the final Deep Sets combination relies predominantly on candidate-set context, source diversity, and geometric consensus.

5.3 Full Image Reconstruction

Returning the cell predictions to the full images changes the dataset mean from 0.8676 to 0.8664. All test cells are present in the reconstructed label maps.

Table 6 Cell crop and full image IoU for Deep Sets ENSEMBLE

Dataset	Cell crops	Full image	Difference
BF-C2DL-HSC	0.8718	0.8724	+0.0006
BF-C2DL-MuSC	0.8091	0.8061	−0.0031
DIC-C2DH-HeLa	0.9531	0.9533	+0.0002
Fluo-C2DL-MSC	0.7560	0.7526	−0.0034
Fluo-N2DH-GOWT1	0.9477	0.9477	0.0000
Dataset mean	0.8676	0.8664	−0.0011

Full image scores are cell weighted.

The difference of 0.0011 shows that resolving overlap between cells has only a small effect on the overall result. The largest changes occur on MuSC and MSC, where reconstruction reduces IoU by 0.0031 and 0.0034, respectively.

5.4 An Illustrative Cell Example

Figure 2 illustrates candidate selection for one held out BF-C2DL-MuSC cell (video 02, frame t0624, label 3; fold 1). Of four available candidates, the pairwise ranker retains CALT-US, MU-Lux-CZ, and KIT-Sch-GE and rejects KTH-SE (3). The rejected candidate has the lowest predicted IoU (0.6334), lowest ranking score (0.0180), and lowest measured IoU (0.5447). The retained candidates have predicted IoUs of 0.7658–0.8183 and measured IoUs of 0.7252–0.8539.

SIMPLE applied to all four candidates reaches IoU 0.7342, whereas SIMPLE applied to the three retained candidates and the final ENSEMBLE both reach 0.8316 with identical output masks. This example illustrates the contribution of upstream candidate selection; Deep Sets provides no additional improvement for this particular cell. The best individual candidate reaches IoU 0.8539, so the improvement here is relative to unselected fusion. The cell was chosen retrospectively to illustrate the mechanism and does not estimate typical performance. Measured IoU uses the SEG reference only for evaluation.

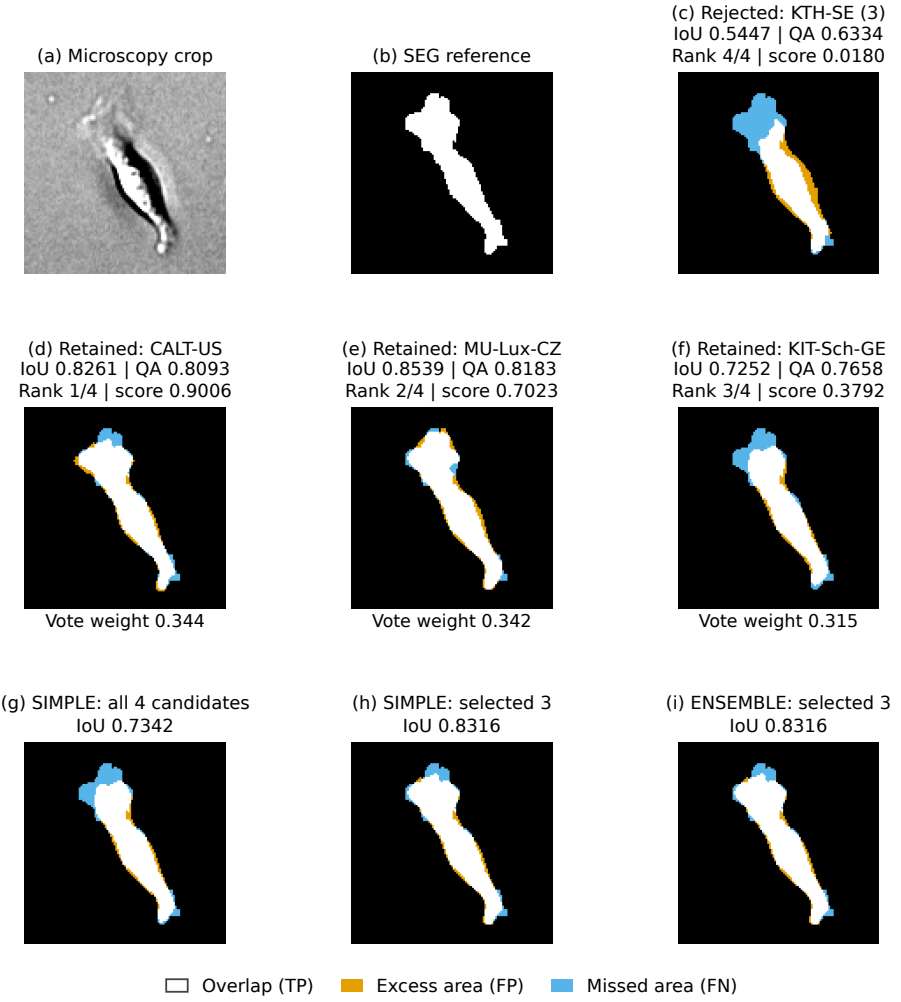


Figure 2 Candidate selection for one held out MuSC cell. (a) Microscopy crop, (b) SEG reference, and (c) rejected candidate. (d–f) The three retained candidates, with measured IoU, QA predicted IoU, pairwise ranking score and position within the four-candidate pool, and normalized final vote weight. (g) SIMPLE applied to all four candidates, (h) SIMPLE applied to the selected pool, and (i) ENSEMBLE applied to the same selected pool. White indicates foreground overlap with the reference, orange excess predicted area, and blue missed reference area. All panels are zoomed to the same display window; IoU is evaluated over the full native 256×256 crop. The SEG reference and measured IoUs are for evaluation and are not inputs to candidate ranking or Deep Sets.

5.5 Differences Among Datasets

MuSC shows the largest benefit from the complete pipeline. ENSEMBLE improves over the fusion baseline by 0.0784 IoU and exceeds the descriptive

best competitor by 0.0084. On GOWT1, it improves over the fusion baseline by 0.0127 and exceeds the descriptive best competitor by 0.0017. HeLa also gains 0.0224 over the fusion baseline, although its strongest individual competitor remains 0.0077 higher.

HSC is already a strong and closely grouped setting. ENSEMBLE is 0.0008 below the fusion baseline and 0.0021 below the descriptive best competitor, while remaining 0.0006 above SIMPLE on the selected candidates. MSC has the largest gain from the ENSEMBLE backend, 0.0065 above selected SIMPLE, but remains 0.0187 below the descriptive best competitor. This makes MSC the clearest opportunity for further work on candidate generation, selection, and boundary refinement. The present comparison does not isolate which of these stages is responsible for the remaining gap.

6 Discussion

6.1 Interpretation of the Main Result

The main result separates the effect of candidate selection from the effect of the final fusion method. Applying SIMPLE to the selected candidates improves mean IoU by 0.0223 over the descriptive best fusion baseline. Deep Sets ENSEMBLE then adds 0.0021 and produces the leading fusion result. Candidate selection is therefore the main source of improvement, but the ENSEMBLE backend contributes a small and consistent advantage: it is above selected SIMPLE on all five datasets.

This distinction matters for the design of silver truth pipelines. Combining all available masks gives weak sources the opportunity to influence every cell. Ranking candidates separately for each cell instead allows the pipeline to use a source where it is strong without assuming that it is equally reliable throughout

a video. The comparison with selected SIMPLE shows that a strong candidate pool benefits more than one fusion method.

The small difference between the two selected pool methods should not be read as a large separation between Deep Sets and SIMPLE. It shows that the learned weights improve the result after selection, while the larger practical gain comes from deciding which candidate masks should enter fusion.

6.2 Variation Among Datasets

The benefit of candidate selection depends on the dataset. MuSC shows the largest improvement over the fusion baseline, and ENSEMBLE also exceeds the descriptive best competitor reference. GOWT1 shows the same ordering with a smaller gain. These results demonstrate the value of choosing candidates for each cell when no single source is uniformly best.

HSC is a closely grouped case in which the fusion baseline and strongest individual source are already competitive. HeLa benefits from the complete pipeline but retains a gap to its strongest individual source. On MSC, the ENSEMBLE backend provides its largest gain over selected SIMPLE, yet the complete pipeline remains below the descriptive best competitor reference. The current experiment does not determine whether this gap comes from missing candidate masks, ranking errors, or boundary fusion. Separating these error types is an important next analysis.

6.3 Scope of the Comparisons

The CTC silver truth reference reaches 0.8819, but it is constructed with information beyond that used by the evaluated fusion pipelines. It uses reference tracking markers, a chosen group of strong methods, and thresholds tuned for each video against ground truth [3]. ENSEMBLE also uses the tracking markers to align cells, but its weights and vote parameters are learned or

selected without labels from the test sequence. The CTC silver truth result is therefore useful as a strong reference, while selected SIMPLE provides the direct controlled comparison for the ENSEMBLE backend.

The same approach can extend beyond the CTC setting. For example, several nucleus segmentation models can provide alternative masks in digital pathology. Overlap based matching can align these masks, after which image, shape, agreement, and source evidence can support the same ranking and fusion stages. In such settings, automatic matching would take the role of the tracking markers used in this study.

6.4 Limitations

The evaluation contains five datasets with two annotated sequences each. Giving every dataset equal weight supports a broad comparison, but ten test folds provide limited evidence about variation across new laboratories, microscopes, and cell types. Evaluation on additional datasets is needed to measure that transfer more precisely.

The candidate sources are existing CTC competitor outputs, and their number and quality differ among datasets. The results therefore describe selection and fusion for this candidate collection rather than every possible collection of segmentation models. Reference tracking markers also provide cell identity before candidate alignment. This controlled design isolates boundary selection and fusion, but it does not measure errors from automatic cell detection or matching.

The Deep Sets configuration is chosen with a validation stability rule, but the family of configurations was developed within the same study. Testing the fixed configuration on an independent benchmark would provide stronger confirmation. Finally, the 0.0021 gain over selected SIMPLE is consistent across

the five datasets but small. The present analysis supports a numerical lead, not a claim of a large difference between the two fusion methods.

6.5 Future Directions

Several directions emerge naturally from this work.

Choose whether to select or fuse. Some cells have one clearly superior candidate, while others benefit from combining several compatible masks. A future model could choose between retaining one candidate, fusing a group, or marking an uncertain cell for review. This would avoid applying the same action to every candidate cluster.

Refine boundaries from the image. Deep Sets currently assigns one weight to each complete mask. A later stage could use the raw image crop and the weighted mask as inputs, then correct local boundary errors directly. The weighting and refinement stages should remain separate so that their contributions can be measured.

Use information across frames. Cell position, area, and boundary shape usually change gradually. Evidence from the preceding and following frames could help distinguish stable candidates from masks that produce unlikely changes, merges, or splits. This information is present in the image sequences but is only partly used by the current pipeline.

Train for complete images and broader candidate pools. The present training target measures one cell crop at a time. Adding a validation or training objective for reconstructed images could penalize conflicts between neighboring cells directly. Further experiments should also include more datasets and candidate sources with genuinely different behavior, especially methods designed for low contrast or touching cells. MSC is the clearest dataset on which to study these additions.

7 Conclusions

This study examined whether selecting candidate masks separately for each cell can improve silver truth produced from several segmentation sources. Across five Cell Tracking Challenge datasets, Deep Sets ENSEMBLE reaches 0.8676 mean cell crop IoU, compared with 0.8654 for SIMPLE using the same selected candidates and 0.8431 for the descriptive best fusion baseline. ENSEMBLE is above selected SIMPLE on all five datasets. Full image reconstruction gives 0.8664, confirming that overlap between reconstructed cells has little effect on the overall result.

The results show that candidate selection is the main source of improvement. The Deep Sets weights and boundary aware vote add a smaller gain of 0.0021 and produce the leading fusion result. The central lesson is therefore clear: a silver truth pipeline benefits most from choosing a small, diverse group of candidate masks for each cell before fusion. This provides a practical basis for future work with automatic candidate matching, information from adjacent frames, and direct boundary refinement.

8 Declarations

8.1 Ethics approval and consent to participate

Not applicable. This study used publicly available benchmark microscopy data and did not involve human participants, identifiable human data, animals, or newly collected biological material.

8.2 Consent for publication

Not applicable.

8.3 Availability of data and materials

The publicly released Cell Tracking Challenge microscopy datasets analysed in this study are available from the challenge dataset repository [3] at https://celltrackingchallenge.net/datasets/. The competitor outputs were provided by the Cell Tracking Challenge team for this study and include third party competition submissions. They are not publicly available from the authors and will not be redistributed without permission. Access requests should be directed to the Cell Tracking Challenge team and remain subject to its approval and any applicable permissions from the submitting competitors. The aggregated results needed to interpret the findings are included in this article.

The source code is available in the silver-truth GitHub repository [22] at https://github.com/David-Ciz/silver-truth. The ten trained Deep Sets weighting models used in this study are included in that repository and supplied as Additional file 2.

8.4 Competing interests

The authors declare that they have no competing interests.

8.5 Funding

This work was funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project “InnovAIte Slovakia, Illuminating Pathways for AI-Driven Breakthroughs” No. 09I02-03-V01-00029. The funding body had no role in the design of the study; the collection, analysis, or interpretation of the data; or the writing of the manuscript.

8.6 Authors' contributions

VU proposed the original research idea, provided access to the study data, contributed scientific advice, and reviewed the manuscript. DC developed the methods and software used in the reported study, performed the experiments and analyses, prepared the visualizations, and wrote the manuscript. JMC conducted earlier exploratory ensemble experiments. MW contributed to earlier exploratory quality-assessment work. The latter exploratory approaches were subsequently substantially revised or superseded in the reported study. All authors read and approved the final manuscript.

8.7 Acknowledgements

The authors thank the Cell Tracking Challenge team for providing controlled access to the competitor submissions analysed in this study.

Additional files

Additional file 1. Annotated frame assignments and retained candidate counts (CSV). Dataset, fold, video sequence, split, frame identifier, represented cell count, and retained candidate-mask count recovered from the retained tables.

Additional file 2. Trained Deep Sets weighting models (ZIP). Ten fold-specific models, including fitted preprocessing and training configurations, with a README and reference fold scores. This file contains no competitor masks or original microscopy data.

References

[1] Warfield SK, Zou KH, Wells WM. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image

Segmentation. IEEE Transactions on Medical Imaging. 2004;23(7):903–921. https://doi.org/10.1109/TMI.2004.828354.

[2] Langerak TR, van der Heide UA, Kotte ANTJ, Viergever MA, van Vulpen M, Pluim JPW. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). IEEE Transactions on Medical Imaging. 2010;29(12):2000–2008. https://doi.org/10.1109/TMI.2010.2057442.

[3] Maska M, Ulman V, Delgado-Rodriguez P, Gomez-de Mariscal E, Necasova T, Guerrero Pena FA, et al. The Cell Tracking Challenge: 10 Years of Objective Benchmarking. Nature Methods. 2023;20:1010–1020. https://doi.org/10.1038/s41592-023-01879-y.

[4] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R, Smola AJ. Deep Sets. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 3391–3401. Available from: https://papers.nips.cc/paper/6931-deep-sets.

[5] Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;44(7):3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968.

[6] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 4015–4026.

[7] Ma J, He Y, Li F, Han L, You C, Wang B. Segment Anything in Medical Images. Nature Communications. 2024;15:654. https://doi.org/10.1038/

s41467-024-44824-z.

[8] Archit A, Freckmann L, Nair S, Khalid N, Hilt P, Rajashekar V, et al. Segment Anything for Microscopy. Nature Methods. 2025;22:579–591. https://doi.org/10.1038/s41592-024-02580-4.

[9] Marks M, Israel U, Dilip R, Li Q, Yu C, Laubscher E, et al. CellSAM: A Foundation Model for Cell Segmentation. Nature Methods. 2025;22:2585–2593. https://doi.org/10.1038/s41592-025-02879-w.

[10] Archit A, Pape C. Revisiting Foundation Models for Cell Instance Segmentation. In: Proceedings of the 9th International Conference on Medical Imaging with Deep Learning. vol. 315 of Proceedings of Machine Learning Research. PMLR; 2026. p. 3464–3495. Available from: https://proceedings.mlr.press/v315/archit26a.html.

[11] Ding Z, Han X, Niethammer M. VoteNet: A Deep Learning Label Fusion Method for Multi-Atlas Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. vol. 11766 of Lecture Notes in Computer Science. Springer; 2019. p. 202–210.

[12] Akbas CE, Ulman V, Maska M, Kozubek M. DeepFuse: A Multi-Rater Fusion and Refinement Network for Computing Silver-Standard Annotations. Computers in Biology and Medicine. 2025;192:110186. https://doi.org/10.1016/j.compbiomed.2025.110186.

[13] Valindria VV, Lavdas I, Bai W, Kamnitsas K, Aboagye EO, Rockall AG, et al. Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. IEEE Transactions on Medical Imaging. 2017;36(8):1597–1606. https://doi.org/10.1109/TMI.2017.2665165.

[14] DeVries T, Taylor GW. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. arXiv preprint. 2018;arXiv:1807.00502.

[15] Robinson R, Oktay O, Bai W, Valindria VV, Sanghvi MM, Aung N, et al. Real-Time Prediction of Segmentation Quality. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. vol. 11073 of Lecture Notes in Computer Science. Springer; 2018. p. 578–585.

[16] Arbelle A, Elul E, Riklin Raviv T. QANet: Quality Assurance Network for Image Segmentation. arXiv preprint. 2019;arXiv:1904.08503.

[17] Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask Scoring R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 6409–6418.

[18] Hann E, Popescu IA, Zhang Q, Gonzales RA, Barutcu A, Neubauer S, et al. Deep Neural Network Ensemble for On-the-Fly Quality Control-Driven Segmentation of Cardiac MRI T1 Mapping. Medical Image Analysis. 2021;71:102029. https://doi.org/10.1016/j.media.2021.102029.

[19] Senbi A, Huang T, Lyu F, Li Q, Tao Y, Shao W, et al. Coherence-Based Segmentation Quality Evaluator Trained on a Large Collection of Annotated Medical Images. In: Pattern Recognition and Computer Vision. vol. 16284 of Lecture Notes in Computer Science. Springer; 2026. p. 238–253.

[20] Marciano V, Chaptoukaev H, Fernandez V, Cardoso MJ, Antonelli M, Ourselin S, et al. Diffusion-Based Quality Control of Medical Image Segmentations across Organs. IEEE Transactions on Medical Imaging. 2026;https://doi.org/10.1109/TMI.2026.3714697.

[21] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770–778.

[22] Číž D.: Silver-truth: Code and Trained Deep Sets Models for Cell Mask Selection and Fusion. Accessed 26 October 2026. Available from: https://github.com/David-Ciz/silver-truth.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1description.txt
- Additionalfile1annotatedframesplits.csv
- Additionalfile2DeepSetsmodels.zip
- Additionalfile2description.txt