Unsupervised Source-Free Ranking of Biomedical Segmentation Models Under Distribution Shift

Joshua Talks[1], Kevin Marchesini[2], Luca Lumetti[2]
Federico Bolelli[2], and Anna Kreshuk[1]

[1] EMBL Heidelberg, Meyerhofstrasse 1, Germany
{name.surname}@embl.de
[2] University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it

Abstract. Model reuse offers a solution to the challenges of segmentation in biomedical imaging, where high data annotation costs remain a major bottleneck for deep learning. However, although many pre-trained models are released through challenges, model zoos, and repositories, selecting the most suitable model for a new dataset remains difficult due to the lack of reliable model ranking methods. We introduce the first black-box-compatible framework for unsupervised and source-free ranking of semantic and instance segmentation models based on the consistency of predictions under perturbations. While ranking methods have been studied for classification and a few segmentation-related approaches exist, most target-related tasks such as transferability estimation or model validation and typically rely on labelled data, feature-space access, or specific training assumptions. In contrast, our method directly addresses the repository setting and applies to both semantic and instance segmentation, for zero-shot reuse or after unsupervised domain adaptation. We evaluate the approach across a wide range of biomedical segmentation tasks in both 2D and 3D imaging, showing that our estimated rankings strongly correlate with true target-domain model performance rankings. https://github.com/kreshuklab/model_ranking.

Keywords: Model Ranking · Distribution Shift · Medical Segmentation

1 Introduction

Segmentation is a ubiquitous problem in biomedical images, vital for the interpretation of biological and anatomical structures. While modern deep learning has significantly improved segmentation accuracy, practical use is limited by the labour-intensive annotation of training data. Model transfer [59,94], with or without fine-tuning, offers a potential solution by reusing pre-trained "source" networks for new "target" datasets of interest. In biomedical imaging, where comprehensive public training datasets are rare, and new imaging modalities continue to emerge, practitioners often train domain-specific models [37]. Community initiatives such as the BioImage Model Zoo (BMZ) [63], public challenges

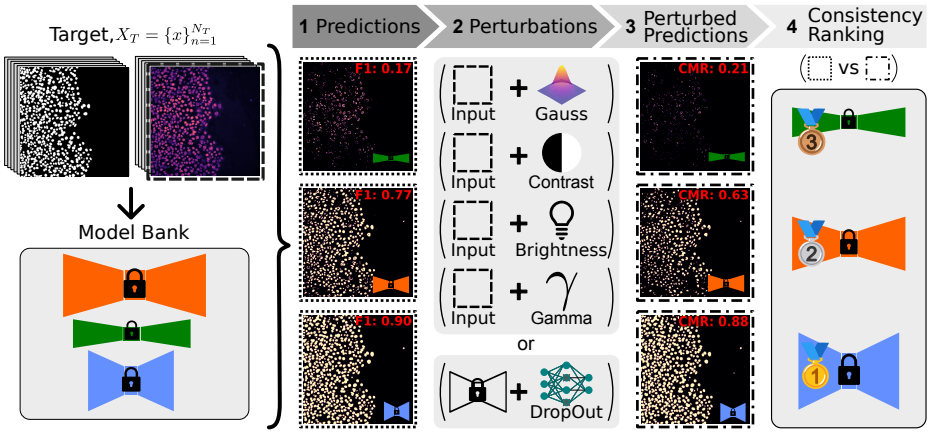


Fig. 1: Unsupervised consistency-based model ranking (CMR) for a set of models via pairwise consistency between perturbed and unperturbed predictions.

such as ToothFairy [12], and emerging large generalist models [3, 65, 80, 93] have created a growing ecosystem of reusable models. This raises the practical challenge of model selection when multiple candidate models—trained on the same task but with varying settings, architectures, or source data—are available.

Transferability estimation [1] aims to address this task, often in a source-free, model-agnostic setting appropriate for repository model ranking. Most prior work addresses the task of image classification, where multiple supervised methods have been introduced [8, 23, 36, 47, 61, 71, 91, 100] to rank post-fine-tuning performance. It has also been suggested that transferability ranking for semantic segmentation can be achieved by simple extension of classification approaches [1]. In our experiments, this claim does not hold for biomedical imaging data, where classification-derived methods fail to reliably correlate with model target data performance. Furthermore, while semantic segmentation can be expressed as pixel classification, instance segmentation cannot. Instance segmentation outputs a permutation invariant set of labels that arise from model-dependent post-processing (e.g., connected components analysis), so the labels have no stable, fixed-dimensional correspondence to instance representations in feature space. The problem of instance segmentation transferability ranking remains largely unaddressed for both natural and biomedical images.

While supervised transferability estimation can save fine-tuning costs, annotation, especially in 3D, can take weeks of expert time without guaranteeing that a representative set has been labelled. Unsupervised ranking, therefore, has significant practical advantages. Also, recent advances in Unsupervised Domain Adaptation (UDA) enable potential improvements to model target data performance without labels, but still require unsupervised ranking of the adapted models. First such methods have been introduced [58, 75, 98], again largely for classification. However, in our experiments, the state-of-the-art (SOTA) for UDA validation [98] fails to reliably rank semantic segmentation models and, like other classification-derived approaches, cannot be applied to instance segmentation.

While model repositories exist, HuggingFace alone hosts over 2.7 million models, practical ranking methods remain scarce because real-world deployment requires them to be unsupervised, source-free, and model-agnostic—constraints that few existing approaches satisfy. To address this gap, we therefore consider the problem of selecting the best segmentation model from a repository for a new dataset without access to target labels.

Contribution. In summary, our contribution is fourfold:

(i) We introduce the first unsupervised, source-free method for ranking both semantic and instance segmentation models based on prediction consistency under feature- or input-space perturbations (Fig. 1);
(ii) The method is training- and architecture-agnostic and operates only on model predictions. It can be applied to black-box models, making it compatible with heterogeneous model zoos even if using inference APIs or containerised deployments;
(iii) We evaluate on realistic biomedical transfer scenarios spanning nuclei, cell, and mitochondria segmentation in light and electron microscopy. Our method successfully ranks models ranging from domain-specific architectures to large generalist segmentation models (e.g., μSAM [3] and CellposeSAM [65]), both for direct application and after unsupervised domain adaptation;
(iv) We further demonstrate the practical utility of our approach by ranking ToothFairy challenge [13] submissions, a benchmark for multi-class 3D CBCT semantic segmentation. Without access to test dataset labels, our method successfully reproduces the challenge's average model DICE ranking.

2 Related Work

In the following, we review supervised transferability metrics, UDA validation, performance estimation, generalisation, and uncertainty estimation while highlighting the distinct challenges of unsupervised model ranking. These fields address distinct yet related problems. While they share many methodological approaches, they use different terminology and rarely cross-cite.

Supervised Transferability Metrics. Transferability metrics (TM) aim to rank a set of models by their post-finetuned target performance. Existing work follows two main categories: *(i)* label-comparison-based methods, such as LEEP [61], and *(ii)* feature-embedding-based methods, which analyse the target data embedding space with guidance from target labels ($\mathcal{N}$LEEP [47], LogME [100], H-score [8], Regularised H-score [36], GBC [66], NCTI [91]). They all require target domain labels and were initially designed for classification, except CC-FV [99], which focuses on semantic segmentation. Our task and method are unsupervised, but fundamentally aim to answer a very similar question of selecting the most suitable source models for a given domain-shifted target dataset [1]. Hence, many TM methods satisfy our requirements of being source-free and model agnostic and can serve as strong baselines for ranking 'zero-shot transferability'. *Our key difference:* label-free, no fine-tuning assumption.

UDA Validation. Many popular UDA methods rely on self or adversarial training [15,44] and are notoriously unstable [89,92]. Hence, UDA validation addresses a problem close to ours: how to select the best adapted model without target domain labels. The SOTA in this field is the Transfer Score (TS) [98], developed for classification, but also tested on segmentation. It measures the post-UDA model classifier bias in the target domain along with the feature space transferability and discriminability. We use TS as a further baseline in our experiments. *Our key difference:* ranking across heterogeneous models, not only UDA setting.

Performance Estimation. Another related field is unlabelled Performance Estimation (PE), also known as Quality Control (QC), where the aim is to directly estimate performance metrics. However, in PE, the unit of analysis is typically a single model, and methods aim to assess the quality of the model's predictions in deployment, e.g., under distribution drift. In contrast, our method estimates the relative success of a candidate set of models applied to specific new domain-shifted target data. Thus we do not require the score to equate to a direct performance metric for each model, as we aim to recover a relative ranking across the model set. PE methods need to rely on assumptions about the nature of the domain shift to uniquely identify the target conditional [28]. At the same time, PE methods are not required to be target label-free [85], source-free [11, 28] or model-training agnostic [34, 42, 49, 56, 72, 74], making most of them inapplicable for model selection in the unsupervised model repository setting. Using a recent benchmark [101], we identify a few exceptions that satisfy our requirements, namely Nuclear Norm [21] and Dispersion [97]. The first seeks to quantify the confidence and dispersity of model predictions via the nuclear norm, while the second, similarly to [66, 99] seeks to use inter-class feature dispersion as a surrogate for model performance. We implement both as baselines. Ensemble-based PE methods [6, 78] incur a large computational overhead, but can in principle also be applied to ranking, with [78] specifically proposed for segmentation. However, [78] reduces the evaluation to object centroids, making the method insensitive to object shape, especially for irregularly shaped objects. Missing pixelwise object evaluation, the method is closer to object detection than instance segmentation ranking, but still provides a useful baseline as the only other approach to address instance segmentation. *Our key difference:* repository ranking vs single-model QC, source-free, model-training agnostic.

Generalisation and Uncertainty Estimation. Generalisability is another term used to describe the ability of a model to perform across new datasets. While often used for In-Domain (ID) data, Out-Of-Distribution (OOD) generalisability also exists. Typically, it is considered as an intrinsic model property that can be improved, e.g., by learning more robust and broadly applicable features. Our aim is not to improve target performance, yet we can exploit similar approaches to those measuring generalisability, but in a specific paired model-dataset setting. Ranking differs from generalisability as it explicitly measures the compatibility between a model and a target dataset either directly or post-UDA and is not a property of the model alone. Consistency-based approaches [2,20,77] provide potentially suitable unsupervised methods and have been explored for

classification. Early works [2, 77] suggest that prediction invariance under Test-Time-Augmentation (TTA) can be linked to model generalisability, but do not explore feature perturbations. Furthermore, they only focus on the ID setting, relying on carefully weighted penalty scores [2], or directly using ground-truth [77] to evaluate prediction invariance. Effective Invariance (EI) [20] addresses OOD generalisability, measuring both the consistency and confidence of probabilistic model outputs. However, EI cannot be extended to instance segmentation.

Consistency to perturbations has also been investigated in the Uncertainty Estimation (UE) field, where rich theory ties the variance of predictions to an approximation of Bayesian Neural Networks (BNNs) [27, 45, 53, 57, 87, 88]. Calibrating model confidence to correlate with true performance is a key challenge and often requires access to source/target labelled validation sets [45, 53] or restriction of the training procedure [27]. The UE field itself is not concerned with assessing model applicability to target data, but rather aims to improve model interpretability via the quantification of uncertainty. However, UE has been widely used in many PE, QC, or failure detection methods [22, 34, 39, 42, 49, 56, 74, 102] that explicitly conflate estimating performance and uncertainty, thus suffering the same limitations. These approaches are typically developed to assist medical experts in segmentation and provide per-image evaluation of a single model's predictions. While important, such methods are not suitable for model selection in model zoos, adding unnecessary, hard to satisfy uncertainty calibration requirements (i.e., restricting the model training [22, 34, 42, 56, 74]). *Our key difference:* no uncertainty calibration, semantic and instance segmentation, model-agnostic.

Following source-free, model-agnostic, unsupervised constraints of a model repository setting, we use the building blocks from the above fields to create a ranking-specific approach, stripping away complexities associated with confidence calibration or feature space reasoning, and explicitly do not use perturbations for UE, as described in Sec. 3. Our method, for the first time, can be applied in a unified framework for both semantic and instance segmentation ranking.

3 Methods

We rank model performance via the consistency of model outputs under perturbation. Focusing on output analysis enables a fully unsupervised and source-free ranking. Since the output spaces of candidate models are fixed and directly comparable—unlike their feature spaces, which may differ in composition and dimensionality—our approach provides a uniform evaluation basis. Moreover, our method, for the first time, directly assesses pixel-wise instance segmentation performance across arbitrary instantiation methods. SEG [78] has previously attempted to address instance segmentation, but reduces the problem to object detection based on centroids. This constrains SEG to circular objects and disregards pixel-wise information. We motivate our consistency-based approach as follows: a source model learns decision boundaries to partition data by class. When applied to a distribution-shifted target domain, these boundaries may fail to separate classes effectively or may intersect embedding clusters. To assess model-

dataset compatibility, we introduce small perturbations—either to the model input or its features—and observe predicted class flips as pixel embeddings cross decision boundaries. Measuring such pixel class flips provides a model-agnostic estimate of compatibility, free from feature space-based reasoning used in prior methods (e.g., feature-space compactness or separability [21, 47, 66, 91, 97–99]). This is advantageous for semantic/instance segmentation, where common losses (e.g., Dice and cross-entropy) do not explicitly enforce feature compactness.

Prediction consistency can be viewed as a proxy for the margin of a model's decision boundaries on target data. Models that transfer well are expected to exhibit larger margins, leading to better-separated classes and less perturbation sensitive predictions. This intuition aligns with prior discussions on manifold smoothness and margin distribution [38,60]. Without access to labels we rely on perturbation-induced class changes to indicate separability. From this perspective, input-image perturbations (common in generalization) and feature-space perturbations (popular in uncertainty estimation) are conceptually equivalent.

Consistency perturbations need not represent realistic image changes between source and target distributions; e.g., feature-space perturbations do not necessarily have realistic image interpretations. Their purpose is to perturb pixel embeddings, testing the suitability of a model's decision boundary for target data by measuring the number of output class label flips. We explicitly do not use perturbations to estimate uncertainty, and do not use MC-DropOut [27], which is used in many PE methods for calibrated confidence scores and must be applied during both training and inference. Instead, we employ Test-Time-Dropout (TTD) [45, 83] which can be applied to arbitrary layers at inference time only. Thus remaining model agnostic and setting no model training requirements.

3.1 Problem Definition

Considering a pre-trained model $M_S : X_S \rightarrow Y_S$, trained on a source domain $\mathcal{D}_S = \{(x_n, y_n)\}_{n=1}^{N_S}$ we apply M_S to an unlabelled target dataset X_T, sampled from a shifted target domain $\mathcal{D}_T = \{x_n\}_{n=1}^{N_T}$. Inference on X_T can be performed directly or after UDA, which further trains M_S on a subset of $\mathcal{D}_T$. We enforce that the source and target task are the same (e.g., nuclei segmentation). Given a set of candidate source models $\mathcal{M} = \{M_m\}_{m=1}^{N_M}$, our goal is to rank the models on X_T so that the ranking correlates with true performance $\mathcal{P}_{M_m(X_T)}$ (e.g., F1 score) ranking, which is not available without labels. We introduce an unsupervised Consistency-based Model Ranking method $\mathcal{R}_{M_m(X_T)}$ (CMR) that aims to preserve the ranking $\mathcal{R}_{M_i(X_T)} > \mathcal{R}_{M_j(X_T)}$ if and only if $\mathcal{P}_{M_i(X_T)} > \mathcal{P}_{M_j(X_T)}$.

3.2 Our Consistency Model Ranking (CMR)

Considering a target dataset $X_T = \{x_n\}_{n=1}^{N_T}$, we rank the transfer performance of a set of models $\mathcal{M} = \{M_m\}_{m=1}^{N_M}$ on X_T by measuring pixelwise prediction consistency between perturbed and unperturbed predictions (Fig. 1). Firstly, we propose to extend the Effective Invariance (EI) [20], originally formulated for

classification tasks, to semantic segmentation. CMR-EI measures the per-class invariance of a single segmentation prediction $\hat{y}_n$ to test-time perturbations,

$$\text{CMR-EI}_n^{(c)} = \frac{1}{|\mathcal{I}_n^{(c)}|} \sum_{i \in \mathcal{I}_n^{(c)}} \sqrt{\hat{p}_{n,i}^{(c)} \cdot \tilde{p}_{n,i}^{(c)}} \cdot \mathbb{1}[\tilde{y}_{n,i}^{(c)} = \hat{y}_{n,i}^{(c)}], \tag{1}$$

where $\hat{p}_{n,i}^{(c)}$ and $\hat{y}_{n,i}^{(c)}$ represent the pixelwise softmax and thresholded output of a model for class c, with $\tilde{p}_{n,i}^{(c)}$, $\tilde{y}_{n,i}^{(c)}$ the perturbed equivalents. To counteract high-class imbalance in biomedical data and allow for multiclass problems, we calculate a foreground restricted per-class consistency score, limiting iteration to the pixel set $\mathcal{I}_n^{(c)} = \{i : \tilde{y}_{n,i}^{(c)} = 1 \text{ or } \hat{y}_{n,i}^{(c)} = 1\}$, thus preventing background pixels from dominating. CMR-EI is a "soft" metric as it utilises both output consistency and confidence, introducing a reliance on confidence calibration which is not guaranteed in transfer [64]. We therefore also propose a "hard" consistency measure based on the thresholded per-class binary model output. Using the Normalised Hamming Distance (NHD), we count the number of foreground restricted pixelwise label changes between the perturbed and unperturbed predictions,

$$\text{CMR-NHD}_n^{(c)} = 1 - \frac{1}{|\mathcal{I}_n^{(c)}|} \sum_{i \in \mathcal{I}_n^{(c)}} \mathbb{1}[\tilde{y}_{n,i}^{(c)} \neq \hat{y}_{n,i}^{(c)}]. \tag{2}$$

This is equivalent to the IoU, but considering pixel flips adds intuition. The consistency score for a single prediction $\hat{y}_n$ is the average of the per-class scores.

For instance segmentation, direct individual pixel comparison is impossible as instance labels are arbitrary. Instead, we measure the consistency between $\tilde{y}_n$ and $\hat{y}_n$ via the agreement of paired pixels in each prediction, as measured by the Rand Index [70,84]. Intuitively, if a pair of pixels is assigned to a single instance in $\tilde{y}_n$, then to be consistent, this should also be true in $\hat{y}_n$. To account for class imbalance, we use the foreground-restricted Adapted Rand Score (ARS) [4],

$$\text{CMR-ARS}_n = \frac{\sum_{k,\ell} w_{k\ell}^2}{\alpha \sum_k \tilde{s}_k^2 + (1-\alpha) \sum_\ell \hat{s}_\ell^2}, \tag{3}$$

where $w_{k\ell}$ is the proportion of pixels that belong to instance k in $\tilde{y}_n$ and instance ℓ in $\hat{y}_n$, restricted to the pairwise union of foreground pixels in $\tilde{y}_n$ and $\hat{y}_n$. $\tilde{s}_k = \sum_\ell w_{k\ell}$ represents the proportion of pixels with label k in $\tilde{y}_n$ and $\hat{s}_\ell = \sum_k w_{k\ell}$ is similarly defined for $\hat{y}_n$. Thus, $\sum_{k,\ell} w_{k\ell}^2$ is the proportion of pixel pairs belonging to a single instance in both $\hat{y}_n$ and $\tilde{y}_n$. From the perspective of either prediction, Eq. (3) can then be considered as the weighted harmonic mean between two terms:

$$\text{ARS}_{\text{split}} = \frac{\sum_{k,\ell} w_{k\ell}^2}{\sum_\ell \hat{s}_\ell}, \qquad \text{ARS}_{\text{merge}} = \frac{\sum_{k,\ell} w_{k\ell}^2}{\sum_k \tilde{s}_k}. \tag{4}$$

$\text{ARS}_{\text{split}}$ penalises split disagreements and is the proportion of pixels pairs in one instance in $\tilde{y}_n$ given they are in one instance in $\hat{y}_n$. $\text{ARS}_{\text{merge}}$ penalises merge disagreements and is the proportion of pixel pairs in one instance in $\hat{y}_n$ given they

are in one instance in $\tilde{y}_n$. The weighting $\alpha = 0.5$ is used by default. Following convention, every background pixel is treated as an instance containing a single pixel. The consistency score for a single transfer $M_m(X_T)$ is then calculated as:

$$\mathrm{CMR}_{M_m(X_T)} = \underset{n=1,\dots,N_T}{\mathrm{median}} \left(\frac{1}{N_{\mathrm{pert}}} \sum_{j=1}^{N_{\mathrm{pert}}} \mathrm{consis}\Big(\hat{y}_{n,m}, \tilde{y}_{n,m}^{(j)}\Big) \right), \tag{5}$$

where $\mathrm{consis}\Big(\hat{y}_{n,m}, \tilde{y}_{n,m}^{(j)}\Big)$ is calculated using one of the above metrics (CMR-EI, CMR-NHD, CMR-ARS), and we average over N_{pert} perturbations. The final ranking within $\mathcal{M}$ is the descending order of $\mathrm{CMR}_{M_m(X_T)}$.

3.3 Model Perturbation

Perturbations can be applied to inputs via TTA or directly to feature space, provided model access. We aim to assess model decision boundaries via target prediction consistency. This differs from evaluating the model's learned invariance to perturbations within the source domain (e.g., due to training augmentations). To examine this, we rank models trained both with and without augmentation, showing perturbations still expose informative variations on target data. Perturbations must remain tolerable [57]: if too strong, even good models become inconsistent, while if overly weak all models stay stable. Empirically, we find a broad range of perturbations enable effective ranking via CMR (*Supp. 11 and 12*). Furthermore, all CMR scores are computed from perturbations sampled over a range, rather than a single fixed value. Thus, rankings are inherently based on a distribution of perturbations rather than a specific tuned parameter.

Input Space Perturbations. We explored several popular image transformations: additive Gaussian noise, gamma correction, and changes in brightness and contrast. In our experiments, we found similar performance across all transformations. Thus, for brevity, in the main text we mostly only report Gaussian noise, with strength controlled by σ. See *Supp. 11.2 and 12.2* and Tab. 2 for additional TTA results and *Supp. 9* for perturbation definitions.

Feature Perturbation. Perturbations can be applied directly to intermediate feature layers, while keeping the network weights frozen. Motivated by findings in semi-supervised learning [62] and uncertainty estimation [57], we apply Test-Time Dropout [45, 83] (TTD). Specifically, we apply TTD *(i)* across all layers of the network (Tabs. 1 and 3 mitochondria experiments), *(ii)* only at the bottleneck, where information is the most redundant [33] (Tabs. 1 and 4 nuclei and cell experiments) or *(iii)* at the bottleneck and skip connection layers (Tab. 2).

4 Experimental Setting

4.1 Datasets

We evaluate our approach on a range of public segmentation datasets, covering six distinct tasks and many modalities. For each model ranking run, we fix one target dataset and evaluate all task-specific models (see *Supp. 7* for detail).

Electron Microscopy (EM), Mitochondria. We use four semantic segmentation datasets with neural tissues from different species and EM modalities: [25, 54, 68]. We also reformulated these datasets as a per-patch binary classification task, where a positive label indicates a mitochondrion is present. This auxiliary task allows us to disentangle biomedical domain effects [17] from the intrinsic suitability of classification-based approaches for semantic segmentation.
Light Microscopy (LM), Nuclei. We use nine nuclei datasets for instance and semantic segmentation, taken from a variety of species and LM modalities: [5, 10, 14, 16, 18, 19, 31, 40, 43, 52, 82, 86, 103].
Light Microscopy, Cells. We consider four instance cell segmentation datasets, taken from fruit fly, *Arabidopsis thaliana* and human samples: [26, 67, 95, 96].
Cone-Beam Computed Tomography (CBCT), ToothFairy2. We use the ToothFairy2 datasets taken from the MICCAI2024 challenge [12,13], comprising 3D CBCT volumes of the human jaw with 42 labelled classes. Model ranking is assessed on a held-out test set acquired at another institution.

4.2 Models and Correlation Evaluation Metrics

Our diverse set of models include self-trained and pre-trained candidates. For classification: ResNet18/50 [32], DenseNet121 [35], MobileNetV2/V3 [69,76]. For binary semantic and instance segmentation: 2D U-Net [73], Residual U-Net [46], UNETR [30], μSAM [3], Cellpose-SAM [65], and BMZ [63] model, id: 5092850 'powerful-chipmunk' [67]. For multi-class semantic segmentation, we select top-models from the ToothFairy2 challenge [90] (Isensee *et al.* and Wang *et al.*) and further add architectures covering CNNs (2D nnU-Net, residual-encoder 3D nnU-Net [37]), Transformers (SwinUNETR [29], TotalSegmentator [93]) and State Space Models (UMamba [55], VMamba [51]). See *Supp. 8* for detailed per-experiment model sets.

For all ranking methods $\mathcal{R}_{M_m(X_T)}$, we calculate the correlation between the metric ranking and the performance score $\mathcal{P}_{M_m(X_T)}$ ranking. As standard [1], we evaluate the correlation scores using three complementary measures: Pearson correlation coefficient (P_r) [9], which captures linear relationships; Spearman's rank correlation coefficient (S_ρ) [79], which measures monotonicity; and Kendall tau (K_τ) [41], which assesses ordinal association and is more sensitive to local pairwise rank disagreements. All three metrics range from $[-1, 1]$ with $K_\tau = P_r = S_\rho = 1$ indicating perfect positive correlation. While a perfect $\mathcal{R}_{M_m(X_T)}$ should be monotonically related to $\mathcal{P}_{M_m(X_T)}$, linear correlation is not required.

5 Results

5.1 Semantic Segmentation

Previous work [1, 66, 99] suggested that supervised transferability metrics developed for classification can be extended to semantic segmentation by sampling a class-balanced set of pixels and treating them as a classification dataset. However, for our biomedical datasets this approach fails: existing transfer metrics,

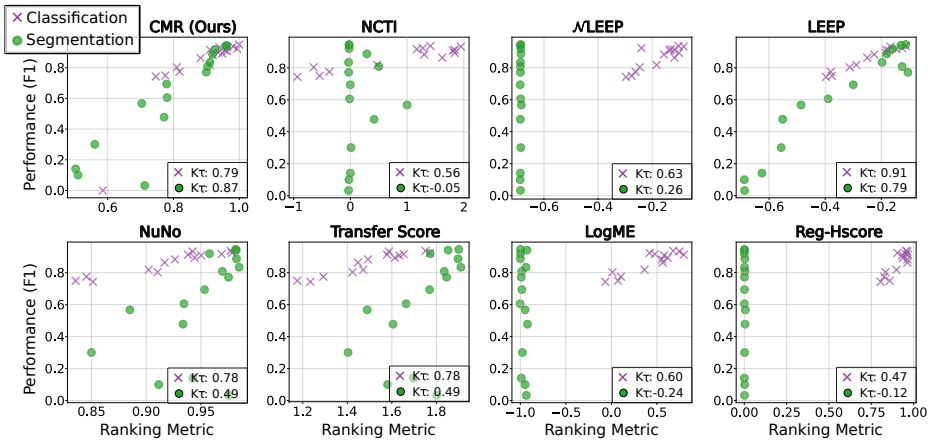


Fig. 2: Classification & semantic segmentation ranking (EPFL [54]). Point per model.

even with supervision, do not correlate with either direct transfer performance (Tab. 1b) or post-fine-tuning performance (*Supp. 11.6*). To determine whether this limitation arises from the biomedical domain [17] or from the shift to segmentation, we additionally evaluated the metrics on mitochondria classification. Fig. 2 shows model performance (F1) against each metric for a single target dataset (EPFL [54]) for both classification and semantic segmentation. While classification ranking performance is high (violet crosses), the metrics fail for segmentation (green circles), where only CMR and supervised LEEP remain reliable.

Tab. 1 reports the mean correlation across two sets of 4 target datasets, firstly for mitochondria segmentation in EM [25,54,68] and secondly, for nuclei segmentation in LM [5, 14, 16, 52] (see *Supp. 11* for per-dataset results). We evaluate CMR using both "hard" (CMR-NHD) and "soft" (CMR-EI) consistency metrics under input (Gauss) and feature (DropOut) perturbations. In all cases, CMR shows strong correlation with the F1-score based model ranking, outperforming the other unsupervised baselines-Transfer Score (TS), NuNo, and Dispersion. Most supervised transferability metrics also fail despite access to target labels, with LEEP as the only consistent exception. Most baseline methods operate in feature space and rely on assumptions about feature space geometry, such as inter-class separability or intra-class compactness [21, 66, 91, 97, 98]. The contrast between LEEP and $\mathcal{N}$LEEP, which adds feature-space reasoning, illustrates the difficulty of extending such assumptions from classification to segmentation. In contrast, CMR operates directly in the output space, comparing predictions rather than labels and thus enabling fully unsupervised ranking. Among the evaluated approaches, CMR provides the most reliable method for ranking models in practical repository settings.

Table 1: Semantic-segmentation. Mean correlation to F1 ranking over 4 target datasets each. Mito [25, 54, 68] ($|\mathcal{M}|$=15) and Nuclei [5, 14, 16, 52]($|\mathcal{M}|$=7). † notes our metrics.

Metric		Mito			Nuclei		
		Kτ	Sρ	Pr	Kτ	ρ	Pr
CMR-EI † (Gauss)	μ	**0.77**	0.85	0.83	0.62	0.77	**0.98**
	σ	**±0.1**	±0.1	±0.1	±0.1	±0.1	**±0.0**
CMR-NHD † (Gauss)	μ	0.71	0.83	0.79	0.69	0.82	0.97
	σ	±0.1	±0.1	±0.1	±0.1	±0.1	±0.0
CMR-EI † (DropOut)	μ	0.73	**0.86**	**0.86**	**0.74**	**0.85**	0.90
	σ	±0.1	**±0.1**	**±0.1**	**±0.2**	**±0.1**	±0.1
CMR-NHD † (DropOut)	μ	0.69	0.85	0.84	0.71	0.84	0.62
	σ	±0.1	±0.1	±0.1	±0.2	±0.1	±0.4
TS	μ	0.25	0.30	0.23	0.02	0.10	-0.09
	σ	±0.2	±0.3	±0.3	±0.3	±0.4	±0.4
NuNo	μ	0.17	0.20	0.13	0.09	0.17	0.08
	σ	±0.2	±0.3	±0.3	±0.5	±0.5	±0.5
Dispersion	μ	-0.03	-0.07	-0.17	-0.18	-0.29	-0.19
	σ	±0.0	±0.1	±0.1	±0.3	±0.4	±0.5

(a) Unsupervised

Metric		Mito			Nuclei		
		Kτ	Sρ	Pr	Kτ	Sρ	Pr
CCFV	μ	-0.11	-0.16	-0.18	-0.03	-0.05	-0.16
	σ	±0.1	±0.1	±0.1	±0.3	±0.4	±0.5
NLEEP	μ	0.41	0.49	0.39	0.11	0.12	0.32
	σ	±0.4	±0.5	±0.3	±0.1	±0.1	±0.2
LEEP	μ	0.89	0.96	0.97	0.67	0.77	0.94
	σ	±0.1	±0.0	±0.0	±0.4	±0.3	±0.0
GBC	μ	0.39	0.52	0.47	-0.13	-0.16	-0.08
	σ	±0.1	±0.1	±0.2	±0.3	±0.4	±0.4
LogME	μ	0.08	0.10	0.06	-0.20	-0.28	-0.49
	σ	±0.2	±0.3	±0.3	±0.3	±0.4	±0.4
NCTI	μ	0.21	0.28	0.37	0.29	0.25	0.30
	σ	±0.2	±0.2	±0.2	±0.5	±0.6	±0.3
RegHscore	μ	0.24	0.34	0.37	0.35	0.36	0.34
	σ	±0.2	±0.4	±0.3	±0.4	±0.4	±0.3

(b) Supervised

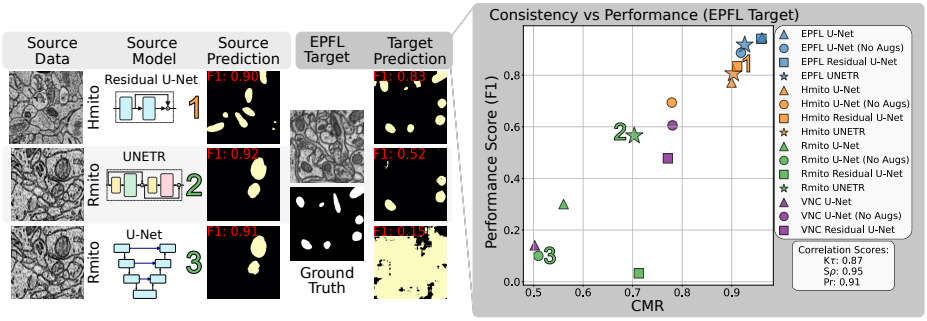


Fig. 3: Correlation between F1 & CMR for Semantic segmentation (EPFL [54] target).

Tab. 1 shows all CMR variants exhibit strong linear and monotonic correlations with performance across all target datasets. Comparable results between input- and feature-space perturbations confirm both approaches provide effective ranking. For feature-space perturbations, we examined where in the network they should be applied. Prior work [33, 57] suggests intermediate layers, where representations contain the most redundant information. To test generality, we applied TTD either to all layers (Mito) or only to the bottleneck (Nuclei). As shown in Tab. 1 and *Supp. 11*, both strategies yield similarly strong correlations, but with different effective p_d ranges.

Our model sets include networks trained both with and without data augmentations; the consistently high CMR correlations indicate that ranking is unaffected by source-domain augmentation strategies. "Hard" (CMR-NHD) and "soft" (CMR-EI) metrics perform equivalently well for semantic segmentation likely because the large number of pixel-wise predictions per image provides a strong signal for the ranking score. UE methods [57] typically require multiple

Table 2: Correlation on multi-class semantic segmentation of the ToothFairy2 dataset $|\mathcal{M}| = 8$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Class		CMR-EI (Gauss)	CMR-NHD (Gauss)	CMR-EI (DropOut)	CMR-NHD (DropOut)	CMR-EI (Gamma)	CMR-NHD (Gamma)	TS	NuNo	Dispersion
IACs	$K\tau$	0.90 (**)	0.81 (**)	0.87 (*)	0.87 (*)	**1.00** (**)	0.90 (*)	0.21 (0.2)	0.14 (0.4)	0.07 (0.4)
	$S\rho$	0.96 (**)	0.93 (*)	0.94 (**)	0.94 (**)	**1.00** (**)	0.96 (**)	0.22 (0.1)	0.14 (0.2)	0.17 (0.3)
	Pr	**0.99** (**)	0.98 (**)	0.83 (*)	0.80 (0.1)	**0.99** (**)	0.98 (**)	0.45 (0.1)	0.01 (0.4)	0.08 (0.3)
Teeth	$K\tau$	0.81 (**)	**1.00** (**)	0.67 (0.1)	0.75 (0.1)	0.71 (*)	0.90 (**)	0.36 (0.1)	0.07 (0.3)	0.43 (*)
	$S\rho$	0.89 (*)	**1.00** (**)	0.77 (0.1)	0.77 (0.1)	0.86 (*)	0.96 (**)	0.55 (0.1)	0.21 (0.2)	0.60 (*)
	Pr	0.95 (**)	**0.98** (**)	0.85 (**)	0.94 (**)	0.89 (**)	0.95 (**)	0.55 (*)	-0.01 (0.4)	0.49 (0.1)
Mand.	$K\tau$	0.62 (*)	0.52 (0.2)	**0.87** (*)	**0.87** (**)	0.43 (0.3)	0.81 (**)	0.09 (0.3)	0.07 (0.3)	0.18 (0.2)
	$S\rho$	0.82 (*)	0.68 (0.1)	0.87 (*)	**0.94** (*)	0.54 (0.2)	0.89 (**)	0.17 (0.2)	0.12 (0.3)	0.25 (0.2)
	Pr	0.73 (0.1)	0.65 (0.1)	0.89 (**)	**0.96** (**)	0.74 (0.1)	0.80 (*)	0.02 (0.3)	0.23 (0.2)	0.38 (0.1)
Sinus.	$K\tau$	0.43 (0.3)	0.81 (*)	0.87 (*)	0.83 (*)	0.43 (0.2)	**0.90** (**)	0.57 (*)	-0.07 (0.3)	0.34 (0.1)
	$S\rho$	0.50 (0.3)	0.89 (*)	0.84 (*)	0.94 (*)	0.57 (0.2)	**0.96** (**)	0.64 (*)	-0.05 (0.4)	0.35 (0.2)
	Pr	0.38 (0.4)	0.73 (0.1)	0.88 (**)	0.90 (**)	0.38 (0.4)	**0.98** (**)	0.53 (0.1)	-0.01 (0.4)	0.54 (*)
Overall	$K\tau$	0.71 (*)	0.90 (**)	0.93 (*)	0.73 (0.1)	0.81 (**)	**1.00** (**)	0.33 (0.1)	0.25 (0.1)	0.28 (0.1)
	$S\rho$	0.86 (*)	0.96 (**)	0.95 (*)	0.83 (0.1)	0.89 (*)	**1.00** (**)	0.49 (0.2)	0.28 (0.2)	0.50 (0.1)
	Pr	0.93 (**)	**0.98** (**)	0.93 (**)	0.93 (**)	0.90 (**)	0.96 (**)	0.55 (0.2)	0.26 (0.2)	0.44 (0.1)

inference runs (≈ 10). In contrast, we find that measuring consistency with a single perturbation per test image is sufficient for reliable model ranking, requiring only two inference passes per image. For both input and feature-space perturbations, a broad range of perturbation strengths produces stable rankings (see *Supp. 11*); therefore all main tables report results for a single perturbation setting.

Fig. 3 illustrates CMR performance on the EPFL [54] mitochondria segmentation dataset across a diverse set of architectures (U-Net, Residual U-Net, UNETR) trained on multiple source datasets with and without input augmentations. Although models perform similarly on their respective source datasets, transfer performance varies substantially, and simple source-target similarity does not reliably predict ranking. For example, UNETR (2) and U-Net (3), trained on the same Rmito dataset, show markedly different target performance. CMR captures both architectural properties and source-domain effects through the learned weights, directly assessing model suitability for the target dataset. While CMR generally correlates strongly with transfer performance, pathological cases can cause mis-rankings. These occur primarily among low-performing transfers where models exhibit "mode collapse" (predicting all foreground or background), producing artificially high consistency. Such cases could be mitigated by combining consistency with cross-model prediction diversity to distinguish stable models from collapsed predictions.

Tab. 2 shows highly multi-class 3D experiments on a public challenge dataset ToothFairy2 [12,13]. Here, both our CMR-EI and CMR-NHD metrics remain reliable, although CMR-NHD is slightly better, likely due to poor model calibration on the challenge test set [64]. Correlation scores are computed after aggregating semantically similar classes, i.e., averaging left/right Inferior Alveolar Canals into IACs, the 32 teeth into Teeth, and left/right Maxillary Sinuses into Sinus. The Overall score shows the mean across all 42 classes in the dataset. Tab. 2 also reports three competitor metrics adapted to the 3D multi-class setting; since these metrics are inherently multi-class, we report the correlation between the over-

Table 3: Post-UDA correlation to semantic F1-score $|\mathcal{M}| = 12$. † notes our metric. Best per group (Mean Teacher/AdaBN) and dataset in bold. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

	Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
Mean Teacher	CMR-EI † *Gauss*	$K\tau$	**0.78**	(**)	**0.75**	(**)	**0.64**	(*)	**0.21**	(0.4)
		$S\rho$	**0.91**	(**)	**0.83**	(**)	**0.81**	(**)	**0.24**	(0.4)
		Pr	**0.93**	(**)	**0.76**	(**)	**0.70**	(*)	**0.28**	(0.4)
	TS	$K\tau$	0.45	(0.1)	0.42	(0.1)	0.38	(0.1)	0.00	(1.0)
		$S\rho$	0.57	(0.1)	0.57	(0.1)	0.50	(0.1)	0.05	(0.9)
		Pr	0.44	(0.2)	0.56	(0.1)	**0.70**	(*)	0.03	(0.9)
AdaBN	CMR-EI † *Gauss*	$K\tau$	**0.70**	(**)	**0.73**	(**)	**0.55**	(*)	**0.39**	(0.2)
		$S\rho$	**0.86**	(**)	**0.88**	(**)	**0.73**	(**)	**0.47**	(0.2)
		Pr	0.76	(**)	**0.91**	(**)	0.87	(**)	**0.71**	(*)
	TS	$K\tau$	0.50	(0.1)	0.50	(0.1)	0.39	(0.2)	0.17	(0.6)
		$S\rho$	0.67	(0.1)	0.73	(*)	0.65	(0.1)	0.27	(0.5)
		Pr	**0.90**	(**)	0.81	(**)	**0.89**	(**)	0.34	(0.4)

Table 4: Mean correlation to Instance mAP over target datasets. † notes ours. $|\mathcal{M}_{Cells}| = 8$ to [26, 95, 96]), $|\mathcal{M}_{Nuclei}| = 5$ to [5,16,43,52]).

Metric		Cells		Nuclei	
		μ	σ	μ	σ
CMR-ARS † (Gauss)	$K\tau$	0.69	±0.15	0.72	±0.25
	$S\rho$	0.83	±0.09	0.83	±0.17
	Pr	0.90	±0.04	0.79	±0.21
CMR-ARS † (DropOut)	$K\tau$	**0.69**	**±0.09**	**0.80**	**±0.28**
	$S\rho$	**0.83**	**±0.06**	**0.85**	**±0.24**
	Pr	**0.91**	**±0.04**	**0.84**	**±0.11**
SEG	$K\tau$	0.12	±0.49	0.48	±0.42
	$S\rho$	0.16	±0.68	0.53	±0.49
	Pr	0.31	±1.02	0.13	±0.72

all metric ranking and the corresponding per-group model performance. Once again, both input and feature space perturbations employed by our method yield the strongest and most coherent rank correlations. Splitting the class-grouped results reveals our metrics are effective and stable across a morphologically heterogeneous set of classes. Our results largely reproduce the challenge rankings without access to target labels, suggesting a practical approach for self-assessing challenge submissions when the test dataset is available. Notably, our metrics are substantially more memory efficient than competitors: also, the CMR-EI variant only requires the maximum probability of each voxel, i.e., predicted class, whereas TS, NuNo, and Dispersion require storing all softmax probabilities, which is prohibitive in 3D (e.g., more than 200GB for ToothFairy2 dataset).

UDA Validation. UDA may be applied to improve model performance, but still requires ranking the adapted models. Tab. 3 shows ranking correlation scores across four mitochondria datasets and two UDA methods: Mean Teacher (MT) [81] self-supervised training and adapted batch normalisation (AdaBN) [48]. We observe strong correlation between CMR and post-UDA F1-score on EPFL, Hmito and Rmito for both MT and AdaBN, outperforming the previous SOTA for UDA validation, Transfer Score (TS) [98]. VNC exhibits lower correlation scores for both CMR and TS metrics, particularly under the MT approach. For CMR, correlation scores are reduced by a few pathological models, all with EPFL source, that consistently segment non-mitochondrial structures in the target data, yielding stable but incorrect predictions and, consequently, low task performance. This 'class confusion' likely arises from specific distributional differences between EPFL and VNC and from the inherent visual similarity of distinct structures in microscopy images inducing class confusion during transfer (see *Supp. 11.5* for more detail).

Consistency regularisation in MT training reinforces this class confusion between visually similar, but distinct structures present in EPFL and VNC, violating the assumption that source and target models address the same task. Hence, models can be ranked well under direct transfer but exhibit outlier be-

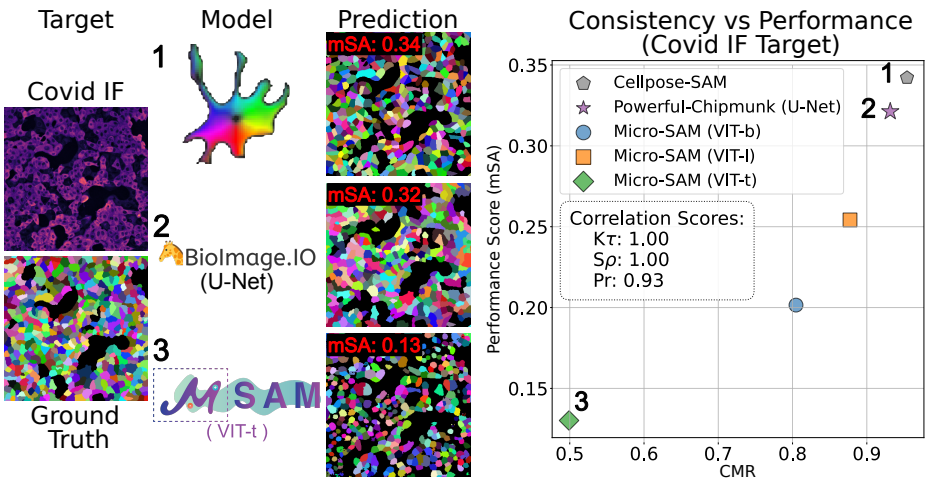


Fig. 4: Instance segmentation: Correlation to mSA on Covid-IF [67] target.

haviour after MT self-training. This is further highlighted by the AdaBN UDA approach, which lacks a consistency regularisation component and thus is less affected, resulting in higher correlation scores. Post-UDA ranking is inherently more challenging, as target data becomes part of the model's adapted 'source' distribution, potentially reducing the impact of perturbations used in CMR estimation, yet our approach maintains good correlation.

5.2 Instance Segmentation

Instance segmentation model ranking remains largely unexplored. Unlike semantic segmentation, instance outputs consist of unordered sets of masks with no fixed correspondence to feature representations, making previously compared transferability metrics inapplicable. Tab. 4 reports correlation with mean Average Precision (mAP@[0.5:0.95]) [50] for our CMR-ARS method and the ensemble-based SEG method across two tasks and seven target datasets (see *Supp. 8* for model details). SEG requires two hyperparameters: the agreement ratio a_r defining ensemble consensus for pseudo ground truth and the centroid dilation radius r. Following [78] we set $a_r = 0.75$ and estimate r as half the mean instance size from a small random sample, reflecting a realistic unsupervised setting. However, SEG proves highly sensitive to these parameters, as reflected by the large variance (σ) values (see *Supp. 12.3*). In contrast, CMR-ARS shows consistently strong correlations with performance under both input and feature-space perturbations, outperforming SEG across all targets. A key advantage of CMR-ARS is that it evaluates instance segmentations directly at the pixel level using the Rand score. In contrast, SEG reduces each instance to a circularly dilated centroid and measures object detections only, which is insensitive to pixel-level quality differences and problematic for elongated or irregular biomedical objects.

Since CMR relies only on output consistency, it can compare models with arbitrary instantiation strategies. Fig. 4 illustrates a realistic ranking scenario on the Covid-IF cell segmentation dataset using several widely used models: μSAM [3] with 3 transformer backbones (ViT-T, ViT-B, ViT-L), Cellpose-SAM [65], and the specialist U-Net model "Powerful-Chipmunk" [67] from the BioImage Model Zoo [63]. Despite substantial differences in architecture and instance generation strategies, CMR accurately reproduces the true mean Segmentation Accuracy (mSA) [24] ranking across this diverse model set (Fig. 4), demonstrating effective unsupervised model selection in a realistic repository setting.

6 Limitations and Conclusion

This work introduces a Consistency-based Model Ranking (CMR) method for semantic and instance segmentation that is source-free, unsupervised, and model-agnostic. Our rankings strongly correlate with true target-domain model performance across diverse segmentation tasks and model types, both under direct application and after unsupervised domain adaptation, while requiring only two inference passes on unlabelled target data. The resulting combination of accuracy, efficiency, and conceptual simplicity makes CMR well suited for practical model selection in repository settings with heterogeneous pre-trained models.

Like other transferability metrics, CMR assumes alignment between source and target tasks, which may be violated for highly generalist models without explicit task specification (e.g., μSAM and Cellpose-SAM). A more detailed discussion on cases of 'task misalignment' can be found in *Supp. 12.5*. In addition, the current formulation aligns most closely with pixel-wise evaluation metrics such as F1 or mAP; boundary-sensitive metrics could be incorporated through pixel-importance weighting within the consistency calculation. Importantly, we never observed low consistency, but high-performance models, supporting our method's validity. More broadly, we believe that unsupervised ranking methods such as CMR can enable scalable model reuse and benchmarking in emerging model repositories, supporting wider adoption of segmentation models in domains where labelled data remains scarce.

Acknowledgements

The research was carried out as part of the AI4Life consortium. AI4Life receives funding from the European Commission through the Horizon Europe program (AI4LIFE project, grant agreement 101057970-AI4LIFE). Computational resources were provided by the High-Performance Computing (HPC) cluster at EMBL. The authors acknowledge the support of these resources, including technical assistance and computational infrastructure, which were essential for this work. Special thanks also to Fynn Beuttenmueller for many constructive conversations.

References

1. Agostinelli, A., P'andy, M., Uijlings, J., Mensink, T., Ferrari, V.: How stable are Transferability Metrics evaluations? European Conference on Computer Vision (2022). https://doi.org/10.48550/arxiv.2204.01403
2. Aithal K, S., Kashyap, D., Subramanyam, N.: Robustness to augmentations as a generalization metric (2021). https://doi.org/10.48550/arXiv.2101.06459
3. Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Teuber, C., Spitzner, M., Tapia Contreras, C., Buckley, G., von Haaren, S., Gupta, S., Grade, M., Wirth, M., Schneider, G., Dengel, A., Ahmed, S., Pape, C.: Segment Anything for Microscopy. Nature Methods **22**(3), 579–591 (2025). https://doi.org/10.1038/s41592-024-02580-4
4. Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Cireşan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J.M., Liu, T., Seyedhosseini, M., Tasdizen, T., Kamentsky, L., Burget, R., Uher, V., Tan, X., Sun, C., Pham, T.D., Bas, E., Uzunbas, M.G., Cardona, A., Schindelin, J., Seung, H.S.: Crowdsourcing the creation of image segmentation algorithms for connectomics. Frontiers in Neuroanatomy **9**, 142 (2015). https://doi.org/10.3389/fnana.2015.00142
5. Arvidsson, M., Rashed, S.K., Aits, S.: An annotated high-content fluorescence microscopy dataset with Hoechst 33342-stained nuclei and manually labelled outlines. Data in Brief **46**, 108769 (2023). https://doi.org/10.1016/j.dib.2022.108769
6. Baek, C., Jiang, Y., Raghunathan, A., Kolter, Z.: Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In: Advances in Neural Information Processing Systems. vol. 35 (2022). https://doi.org/10.48550/arXiv.2206.13089
7. Bailoni, A., Pape, C., Hütsch, N., Wolf, S., Beier, T., Kreshuk, A., Hamprecht, F.A.: GASP, a generalized framework for agglomerative clustering of signed graphs and its application to instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11645–11655 (2022). https://doi.org/10.48550/arXiv.1906.11713
8. Bao, Y., Li, Y., Huang, S.L., Zhang, L., Zheng, L., Zamir, A., Guibas, L.: An information-theoretic approach to transferability in task transfer learning. In: IEEE International Conference on Image Processing. pp. 2309–2313 (2019). https://doi.org/10.1109/ICIP.2019.8803726
9. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing, pp. 1–4. Springer (2009). https://doi.org/10.1007/978-3-642-00296-0_5
10. Bhatia, H.S., Brunner, A.D., Öztürk, F., Kapoor, S., Rong, Z., Mai, H., Thielert, M., Ali, M., Al-Maskari, R., Paetzold, J.C., Kofler, F., Todorov, M.I., Molbay, M., Kolabas, Z.I., Negwer, M., Hoeher, L., Steinke, H., Dima, A., Gupta, B., Kaltenecker, D., Caliskan, Ö.S., Brandt, D., Krahmer, N., Müller, S., Lichtenthaler, S.F., Hellal, F., Bechmann, I., Menze, B., Theis, F., Mann, M., Ertürk, A.: Spatial proteomics in three-dimensional intact specimens. Cell **185**(26), 5040–5058.e19 (2022). https://doi.org/10.1016/j.cell.2022.11.021
11. Białek, J., Kivimäki, J., Kuberski, W., Perrakis, N.: Estimating Model Performance Under Covariate Shift Without Labels (2025). https://doi.org/10.48550/arXiv.2401.08348

12. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., R. Rokuss, M., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles Ballester, Vicent amd Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging (2024). https://doi.org/10.1109/TMI.2024.3523096
13. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volumes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
14. Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., Rohban, M., Singh, S., Carpenter, A.E.: Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature Methods **16**(12), 1247–1253 (2019). https://doi.org/10.1038/s41592-019-0612-7
15. Cao, H., Xu, Y., Yang, J., Yin, P., Yuan, S., Xie, L.: Multi-modal continual test-time adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023). https://doi.org/10.48550/arXiv.2303.10457
16. von Chamier, L., Laine, R.F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., Lerche, M., Hernández-Pérez, S., Mattila, P.K., Karinou, E., Holden, S., Solak, A.C., Krull, A., Buchholz, T.O., Jones, M.L., Royer, L.A., Leterrier, C., Shechtman, Y., Jug, F., Heilemann, M., Jacquemet, G., Henriques, R.: Democratising deep learning for microscopy with ZeroCostDL4Mic. Nature Communications **12**(1), 2276 (2021). https://doi.org/10.1038/s41467-021-22518-0
17. Chaves, L., Bissoto, A., Valle, E., Avila, S.: The performance of transferability metrics does not translate to medical tasks. In: Domain Adaptation and Representation Transfer – DART 2023. pp. 105–114. Springer (2023). https://doi.org/10.1007/978-3-031-45857-6_11
18. Chen, Y., Al-Maskari, R., Horvath, I., Ali, M., Hoher, L., Yang, K., Lin, Z., Zhai, Z., Shen, M., Xun, D., Wang, Y., Xu, T., Goubran, M., Wu, Y., Mori, K., Paetzold, J.C., Erturk, A.: SELMA3D challenge: Self-supervised learning for 3D light-sheet microscopy image segmentation (2025). https://doi.org/10.48550/arXiv.2501.03880
19. De, T., Urbanski, A., Thangamani, S., Wyrzykowska, M., Yakimovich, A.: HeLa-CytoNuc: fluorescence microscopy dataset with segmentation masks for cell nuclei and cytoplasm (2024). https://doi.org/10.14278/rodare.3001
20. Deng, W., Gould, S., Zheng, L.: On the strong correlation between model invariance and generalization. In: Advances in Neural Information Processing Systems. vol. 35, pp. 28052–28067 (2022)
21. Deng, W., Suh, Y., Gould, S., Zheng, L.: Confidence and Dispersity Speak: Characterising Prediction Matrix for Unsupervised Accuracy Estimation (2023). https://doi.org/10.48550/arXiv.2302.01094
22. DeVries, T., Taylor, G.W.: Leveraging Uncertainty Estimates for Predicting Segmentation Quality (2018). https://doi.org/10.48550/arXiv.1807.00502
23. Ding, Y., Jiang, B., Yu, A., Zheng, A., Liang, J.: Which Model to Transfer? A Survey on Transferability Estimation (2024). https://doi.org/10.48550/arXiv.2402.15231

24. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
25. Franco-Barranco, D., Lin, Z., Jang, W.D., Wang, X., Shen, Q., Yin, W., Fan, Y., Li, M., Chen, C., Xiong, Z., Xin, R., Liu, H., Chen, H., Li, Z., Zhao, J., Chen, X., Pape, C., Conrad, R., Nightingale, L., de Folter, J., Jones, M.L., Liu, Y., Ziaei, D., Huschauer, S., Arganda-Carreras, I., Pfister, H., Wei, D.: Current Progress and Challenges in Large-Scale 3D Mitochondria Instance Segmentation. IEEE Transactions on Medical Imaging **42**(12), 3956–3971 (2023). https://doi.org/10.1109/TMI.2023.3320497
26. Funke, J., Mais, L., Champion, A., Dye, N., Kainmueller, D.: A Benchmark for Epithelial Cell Tracking. In: Computer Vision – ECCV 2018 Workshops, pp. 437–445. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-11024-6_33
27. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proceedings of The 33rd International Conference on Machine Learning. pp. 1050–1059 (2016). https://doi.org/10.48550/arXiv.1506.02142
28. Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging Unlabeled Data to Predict Out-of-Distribution Performance (2022). https://doi.org/10.48550/arXiv.2201.04234
29. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: International MICCAI Brainlesion Workshop. pp. 272–284 (2021). https://doi.org/10.1007/978-3-031-08999-2_22
30. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1748–1758 (2022). https://doi.org/10.1109/WACV51458.2022.00181
31. Hawkins, R., Balaghi, N., Rothenberg, K.E., Ly, M., Fernandez-Gonzalez, R.: ReSCU-Nets: Recurrent u-nets for segmentation of three-dimensional microscopy data. Journal of Cell Biology **224**(11), e202506102 (2025). https://doi.org/10.1083/jcb.202506102
32. 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. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
33. He, T., Fan, Y., Qian, Y., Tan, T., Yu, K.: Reshaping deep neural network for fast decoding by node-pruning. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 245–249 (2014). https://doi.org/10.1109/ICASSP.2014.6853595
34. Hoebel, K., Andrearczyk, V., Beers, A.L., Patel, J.B., Chang, K., Depeursinge, A., Mueller, H., Kalpathy-Cramer, J.: An exploration of uncertainty information for segmentation quality assessment. In: Medical Imaging 2020: Image Processing. p. 55 (2020). https://doi.org/10.1117/12.2548722
35. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017). https://doi.org/10.48550/arXiv.1608.06993

36. Ibrahim, S., Ponomareva, N., Mazumder, R.: Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In: Machine Learning and Knowledge Discovery in Databases – ECML PKDD 2022, vol. 13713, pp. 693–709. Springer (2023). https://doi.org/10.1007/978-3-031-26387-3_42
37. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
38. Jiang, Y., Krishnan, D., Mobahi, H., Bengio, S.: Predicting the generalization gap in deep networks with margin distributions. In: International Conference on Learning Representations (2019). https://doi.org/10.48550/arXiv.1810.00113
39. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Frontiers in Neuroscience **14**, 282 (2020). https://doi.org/10.3389/fnins.2020.00282
40. Kaltenecker, D., Al-Maskari, R., Negwer, M., Hoeher, L., Kofler, F., Zhao, S., Todorov, M., Rong, Z., Paetzold, J.C., Wiestler, B., Piraud, M., Rueckert, D., Geppert, J., Morigny, P., Rohm, M., Menze, B.H., Herzig, S., Berriel Diaz, M., Ertürk, A.: Virtual reality-empowered deep-learning analysis of brain cells. Nature Methods **21**(7), 1306–1315 (2024). https://doi.org/10.1038/s41592-024-02245-2
41. Kendall, M.G.: Rank correlation methods. In: 4th edition (1970) (1972)
42. Kivimäki, J., Białek, J., Nurminen, J.K., Kuberski, W.: Confidence-based Estimators for Predictive Performance in Model Monitoring. Journal of Artificial Intelligence Research **82**, 209–240 (2025). https://doi.org/10.1613/jair.1.16709
43. Kromp, F., Bozsaky, E., Rifatbegovic, F., Fischer, L., Ambros, M., Berneder, M., Weiss, T., Lazic, D., Dörr, W., Hanbury, A., Beiske, K., Ambros, P.F., Ambros, I.M., Taschner-Mandl, S.: An annotated fluorescence image dataset for training nuclear segmentation methods. Scientific Data **7**(1), 262 (2020). https://doi.org/10.1038/s41597-020-00608-w
44. Kumar, A., Ma, T., Liang, P.: Understanding Self-Training for Gradual Domain Adaptation. In: Proceedings of the 37th International Conference on Machine Learning. pp. 5468–5479 (2020). https://doi.org/10.48550/arXiv.2002.11361
45. Ledda, E., Fumera, G., Roli, F.: Dropout injection at test time for post hoc uncertainty quantification in neural networks. Information Sciences **645**, 119356 (2023). https://doi.org/10.1016/j.ins.2023.119356
46. Lee, K., Zung, J., Li, P., Jain, V., Seung, H.S.: Superhuman Accuracy on the SNEMI3D Connectomics Challenge (2017). https://doi.org/10.48550/arXiv.1706.00120
47. Li, Y., Jia, X., Sang, R., Zhu, Y., Green, B., Wang, L., Gong, B.: Ranking neural checkpoints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2663–2673 (2021). https://doi.org/10.1109/CVPR46437.2021.00269
48. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive Batch Normalization for practical domain adaptation. Pattern Recognition (2018). https://doi.org/10.1016/j.patcog.2018.03.005
49. Lin, Q., Chen, X., Chen, C., Garibaldi, J.M.: A Novel Quality Control Algorithm for Medical Image Segmentation Based on Fuzzy Uncertainty. IEEE Transactions on Fuzzy Systems **31**(8), 2532–2544 (2023). https://doi.org/10.1109/TFUZZ.2022.3228332

50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
51. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual State Space Model. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024). https://doi.org/10.48550/arXiv.2401.10166
52. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature Methods **9**(7), 637–637 (2012). https://doi.org/10.1038/nmeth.2083
53. Loquercio, A., Segù, M., Scaramuzza, D.: A General Framework for Uncertainty Estimation in Deep Learning. IEEE Robotics and Automation Letters **5**(2), 3153–3160 (2020). https://doi.org/10.1109/LRA.2020.2974682
54. Lucchi, A., Li, Y., Fua, P.: Learning for Structured Prediction Using Approximate Subgradient Descent with Working Sets. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1987–1994 (2013). https://doi.org/10.1109/CVPR.2013.259
55. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv (2024). https://doi.org/10.48550/arXiv.2401.04722
56. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020). https://doi.org/10.1109/TMI.2020.3006437
57. Mi, L., Wang, H., Tian, Y., He, H., Shavit, N.N.: Training-free uncertainty estimation for dense regression: Sensitivity as a surrogate. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10042–10050 (2022). https://doi.org/10.1609/aaai.v36i9.21243
58. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: International Conference on Learning Representations (2018). https://doi.org/10.48550/arXiv.1711.10288
59. Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S.M., Sieniek, M., Winkens, J., Liu, Y., Bui, P., Prabhakara, S., Telang, U., Karthikesalingam, A., Houlsby, N., Natarajan, V.: Supervised Transfer Learning at Scale for Medical Imaging. arXiv.org (2021). https://doi.org/10.48550/arXiv.2101.05913
60. Ng, N., Hulkund, N., Cho, K., Ghassemi, M.: Predicting Out-of-Domain Generalization with Neighborhood Invariance (2023). https://doi.org/10.48550/arXiv.2207.02093
61. Nguyen, C.V., Hassner, T., Seeger, M., Archambeau, C.: LEEP: A new measure to evaluate transferability of learned representations. In: Proceedings of the International Conference on Machine Learning (2020). https://doi.org/10.48550/arXiv.2002.12462
62. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684 (2020). https://doi.org/10.48550/arXiv.2003.09005
63. Ouyang, W., Beuttenmueller, F., Gómez-de Mariscal, E., Pape, C., Burke, T., Garcia-López-de Haro, C., Russell, C., Moya-Sans, L., de-la Torre-Gutiérrez,

C., Schmidt, D., Kutra, D., Novikov, M., Weigert, M., Schmidt, U., Bankhead, P., Jacquemet, G., Sage, D., Henriques, R., Muñoz-Barrutia, A., Lundberg, E., Jug, F., Kreshuk, A.: BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis. bioRxiv (2022). https://doi.org/10.1101/2022.06.07.495102

64. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems (2019). https://doi.org/10.48550/arXiv.1906.02530
65. Pachitariu, M., Rariden, M., Stringer, C.: Cellpose-SAM: superhuman generalization for cellular segmentation (2025). https://doi.org/10.1101/2025.04.28.651001
66. Pándy, M., Agostinelli, A., Uijlings, J., Ferrari, V., Mensink, T.: Transferability estimation using bhattacharyya class separability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022). https://doi.org/10.1109/CVPR52688.2022.00896
67. Pape, C., Remme, R., Wolny, A., Olberg, S., Wolf, S., Cerrone, L., Cortese, M., Klaus, S., Lucic, B., Ullrich, S., Anders-Össwein, M., Wolf, S., Cerikan, B., Neufeldt, C.J., Ganter, M., Schnitzler, P., Merle, U., Lusic, M., Boulant, S., Stanifer, M., Bartenschlager, R., Hamprecht, F.A., Kreshuk, A., Tischer, C., Kräusslich, H.G., Müller, B., Laketa, V.: Microscopy-based assay for semi-quantitative detection of SARS-CoV-2 specific antibodies in human sera: A semi-quantitative, high throughput, microscopy-based assay expands existing approaches to measure SARS-CoV-2 specific antibody levels in human sera. BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology **43**(3), e2000257 (2021). https://doi.org/10.1002/bies.202000257
68. Phelps, J.S., Colburn Hildebrand, D.G., Graham, B.J., Kuan, A.T., Thomas, L.A., Nguyen, T.M., Buhmann, J., Azevedo, A.W., Sustar, A., Agrawal, S., Liu, M., Shanny, B.L., Funke, J., Tuthill, J.C., Allen Lee, W.C.: Reconstruction of motor control circuits in adult Drosophila using automated transmission electron microscopy. Cell **184**(3), 759–774.e18 (2021). https://doi.org/10.1016/j.cell.2020.12.013
69. Qian, S., Ning, C., Hu, Y.: Mobilenetv3 for image classification. In: 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). pp. 490–497 (2021). https://doi.org/10.1109/ICBAIE52039.2021.9389905
70. Rand, W.M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association **66**(336), 846–850 (1971). https://doi.org/10.1080/01621459.1971.10482356
71. Renggli, C., Pinto, A.S., Rimanic, L., Puigcerver, J., Riquelme, C., Zhang, C., Lucic, M.: Which model to transfer? finding the needle in the growing haystack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9205–9214 (2022). https://doi.org/10.1109/CVPR52688.2022.00899
72. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Kainz, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Rueckert, D., Glocker, B.: Real-Time Prediction of Segmentation Quality: 21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018. Medical Image Computing and Computer Assisted Intervention

– MICCAI 2018 - 21st International Conference, 2018, Proceedings pp. 578–585 (2018). https://doi.org/10.1007/978-3-030-00937-3_66
73. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
74. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage **195**, 11–22 (2019). https://doi.org/10.1016/j.neuroimage.2019.03.042
75. Saito, K., Kim, D., Teterwak, P., Sclaroff, S., Darrell, T., Saenko, K.: Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9184–9193 (2021). https://doi.org/10.1109/ICCV48922.2021.00905
76. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474
77. Schiff, Y., Quanz, B., Das, P., Chen, P.Y.: Predicting deep neural network generalization with perturbation response curves. In: Advances in Neural Information Processing Systems. vol. 34, pp. 21176–21188 (2021). https://doi.org/10.48550/arXiv.2106.04765
78. Sims, Z., Strgar, L., Thirumalaisamy, D., Heussner, R., Thibault, G., Chang, Y.H.: SEG: Segmentation Evaluation in absence of Ground truth labels. bioRxiv p. 2023.02.23.529809 (2023). https://doi.org/10.1101/2023.02.23.529809
79. Spearman, C.: The Proof and Measurement of Association between Two Things. Am. J. Psychol. **15** (1904). https://doi.org/10.2307/1412159
80. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: A generalist algorithm for cellular segmentation. Nature Methods **18**(1), 100–106 (2021). https://doi.org/10.1038/s41592-020-01018-x
81. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30 (2017). https://doi.org/10.48550/arXiv.1703.01780
82. Todorov, M.I., Paetzold, J.C., Schoppe, O., Tetteh, G., Shit, S., Efremov, V., Todorov-Völgyi, K., Düring, M., Dichgans, M., Piraud, M., Menze, B., Ertürk, A.: Machine learning analysis of whole mouse brain vasculature. Nature methods **17**(4), 442–449 (2020). https://doi.org/10.1038/s41592-020-0792-1
83. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656 (2015). https://doi.org/10.1109/CVPR.2015.7298664
84. Unnikrishnan, R., Hebert, M.: Measures of Similarity. In: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1. pp. 394–394 (2005). https://doi.org/10.1109/ACVMOT.2005.71
85. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. IEEE Transactions on Medical Imaging **36**(8), 1597–1606 (2017). https://doi.org/10.1109/TMI.2017.2665165

86. Vijayan, A., Mody, T.A., Yu, Q., Wolny, A., Cerrone, L., Strauss, S., Tsiantis, M., Smith, R.S., Hamprecht, F.A., Kreshuk, A., Schneitz, K.: A deep learning-based toolkit for 3D nuclei segmentation and quantitative analysis in cellular and tissue context. Development **151**(14), dev202800 (2024). https://doi.org/10.1242/dev.202800
87. Wang, G., Li, W., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing (2019). https://doi.org/10.1016/j.neucom.2019.01.103
88. Wang, H., Ji, Q.: Epistemic uncertainty quantification for pretrained neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11052–11061 (2024). https://doi.org/10.48550/arXiv.2404.10124
89. Wang, M., Deng, W.: Deep Visual Domain Adaptation: A Survey. Neurocomputing **312**, 135–153 (2018). https://doi.org/10.1016/j.neucom.2018.05.083
90. Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F.: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data: MICCAI 2024 Challenges: ToothFairy 2024, 3DTeethLand 2024, and STS 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings. Springer Nature (2025). https://doi.org/10.1007/978-3-031-88977-6
91. Wang, Z., Luo, Y., Zheng, L., Huang, Z., Baktashmotlagh, M.: How Far Pre-trained Models Are from Neural Collapse on the Target Dataset Informs their Transferability. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5549–55558 (2023). https://doi.org/10.1109/iccv51070.2023.00511
92. Wang, Z., Dai, Z., Póczos, B., Carbonell, J.: Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11293–11302 (2019). https://doi.org/10.48550/arXiv.1811.09751
93. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
94. Weiss, K.R., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data (2016). https://doi.org/10.1186/s40537-016-0043-6
95. Willis, L., Refahi, Y., Wightman, R., Landrein, B., Teles, J., Huang, K.C., Meyerowitz, E.M., Jönsson, H.: Cell size and growth regulation in the Arabidopsis thaliana apical stem cell niche. Proceedings of the National Academy of Sciences of the United States of America **113**(51), E8238–E8246 (2016). https://doi.org/10.1073/pnas.1616768113
96. Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A.V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S.S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G.W., Lohmann, J.U., Tsiantis, M., Hamprecht, F.A., Schneitz, K., Maizel, A., Kreshuk, A.: Accurate and versatile 3D segmentation of plant tissues at cellular resolution. eLife **9**, e57613 (2020). https://doi.org/10.7554/eLife.57613
97. Xie, R., Wei, H., Feng, L., Cao, Y., An, B.: On the importance of feature separability in predicting out-of-distribution error. In: Advances in Neural Information Processing Systems. vol. 36, pp. 27783–27800 (2023). https://doi.org/10.48550/arXiv.2303.15488

98. Yang, J., Qian, H., Xu, Y., Wang, K., Xie, L.: Can we evaluate domain adaptation models without target-domain labels? In: International Conference on Learning Representations (2024). https://doi.org/10.48550/arXiv.2305.18712
99. Yang, Y., Wei, M., He, J., Yang, J., Ye, J., Gu, Y.: Pick the best pre-trained model: Towards transferability estimation for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 674–684. Springer (2023). https://doi.org/10.1007/978-3-031-43907-0_64
100. You, K., Liu, Y., Long, M., Wang, J.: LogME: Practical Assessment of Pre-trained Models for Transfer Learning. International Conference on Machine Learning (2021). https://doi.org/10.48550/arXiv.2102.11005
101. Yu, H., Li, K., Li, D., He, Y., Zhang, X., Cui, P.: ODP-Bench: Benchmarking out-of-distribution performance prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025). https://doi.org/10.48550/arXiv.2510.27263
102. Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P.F., Maier-Hein, K.: Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. Medical Image Analysis **101**, 103392 (2025). https://doi.org/10.1016/j.media.2024.103392
103. Zhao, S., Todorov, M.I., Cai, R., AI-Maskari, R., Steinke, H., Kemter, E., Mai, H., Rong, Z., Warmer, M., Stanic, K., Schoppe, O., Paetzold, J.C., Gesierich, B., Wong, M.N., Huber, T.B., Duering, M., Bruns, O.T., Menze, B., Lipfert, J., Puelles, V.G., Wolf, E., Bechmann, I., Ertürk, A.: Cellular and Molecular Probing of Intact Human Organs. Cell **180**(4), 796–812.e19 (2020). https://doi.org/10.1016/j.cell.2020.01.030

Unsupervised Source-Free Ranking of Biomedical Segmentation Models Under Distribution Shift

Supplementary Material

7 Datasets

The general experimental setup for source model ranking under transfer to a target dataset is as follows, a set of 'source' models $\mathcal{M} = \{M_i\}_{i=1}^{N_m}$ are trained for the same task (i.e nuclei semantic segmentation). The models can have a range of architectures and source datasets. A single dataset is then selected as 'target' and all of the models from $\mathcal{M}$ are applied to the test set of the target data. We then rank the models $M_{i=1...N_m}$ and compare this ranking to the true transfer performance ranking, obtained using ground-truth. In this section we further detail the employed datasets, and in Sec. 8 the tested models.

We experimented applying our consistency based model ranking (CMR) method across a wide range of datasets and modalities from the microscopy and biomedical domains. We focused on ranking the target performance of semantic and instance segmentation models across several distinct tasks. The tasks covered arguably the most common problems in microscopy image analysis; nuclei, cell and mitochondria segmentation. We additionally investigated a popular biomedical segmentation challenge (i.e., ToothFairy2) for human jaw segmentation. Our wide range of experiments across such a diverse set of datasets and tasks highlights the general applicability of our approach for unsupervised ranking of segmentation models.

7.1 Electron Microscopy (EM), Mitochondria

Semantic Segmentation. We use four datasets with semantic segmentation ground-truth, all are neural tissues taken from different species and EM modalities:

- EPFL (rat) [54]
  - The dataset represents a section taken from the CA1 hippocampus region of the brain. The dataset has the rescaled dimensions $330 \times 480 \times 640$.
  - The volume was split with the first 165 z-slices used for training, the next 40 z-slices used for validation and the remaining 125 z-slices used for testing.
- Hmito, Rmito [25]
  - The dataset comprises two 3D EM image stacks taken from the brain tissue of a human (Hmito) and a rat (Rmito). Each volume has pixel dimensions $500 \times 4096 \times 4096$.
  - The volumes were split with the first 300 z-slices used for training, the next 50 z-slices for validation and the remaining 150 z-slices for testing.
- VNC (fruit fly) [68]

- Transmission Electron Microscopy of ventral nerve cord (VNC) of an adult female *Drosophila melanogaster*. Once voxels were rescaled the dataset had pixel dimensions $20 \times 589 \times 589$.
- The VNC dataset is by far the smallest dataset, therefore, in order to maximise available patches for testing we did not perform a typical test train split. For VNC source models we used the first 18 z-slices for training and the remaining 2 z-slices for validation. We did not transfer VNC source trained models to a VNC test set, removing this transfer from our set of evaluated transfers. For models trained on alternative sources and transferred to VNC we thus tested on the entire 20 z-slices of VNC.

All datasets were resized to match the MitoEM voxel dimensions. For all datasets we split the volumes into 2D 256×256 image patches for training and evaluation.

Classification. In addition to direct use for semantic segmentation, we reformulated these datasets as a per-patch binary classification task, labelling a patch as positive if mitochondria are present. The volumes were split into 80×80 2D patches. To remove positive images with very little mitochondria present, we additionally filtered images based on proportion of mitochondria present in each ground-truth patch. For each dataset we used the corresponding foreground ratios; EPFL = 0.1 foreground, Hmito = 0.1 foreground, Rmito = 0.1 foreground, VNC = 0.05 foreground.

This auxiliary classification task allows us to disentangle the effects of domain shift from the intrinsic suitability of classification-based ranking methods for semantic segmentation.

7.2 Light Microscopy (LM), Nuclei

Semantic Segmentation. We trained nuclei segmentation models on the following fluorescent datasets;

- BBBC039 [52]
  - A high-throughput chemical screen on U2OS cells.
  - The dataset comprises of a set of 2D images. We used 100 images for training, 50 for validation and 48 for testing. (See github repo for exact image id split).
  - From each image 2D 256×256 patches were extracted for training and images were cropped to a standard 512×512 for evaluation.
- Go-Nuclear [86]
  - Fluorescent dataset of *Arabidopsis thaliana*.
  - The dataset comprises of 5 separate data volumes. Three volumes ('id: 1135' ($263 \times 1024 \times 1024$), 'id: 1136' ($268 \times 1120 \times 1120$) and 'id: 1137' ($262 \times 1080 \times 1080$) were used as training data. For each volume the 40-210 z-slices were taken to avoid imaging artefacts at the beginning and end of the volumes. One volume 'id: 1139' ($280 \times 1120 \times 1120$) was used for validation and the last volume 'id: 1160' ($241 \times 753 \times 1672$) was used for testing.

  - From each volume 2D 256×256 patches were extracted for training and evaluation.
- HeLaCytoNuc [19]
  - A dataset of HeLa cell nuclei.
  - The dataset consists of 1873 training images, 535 validation images and 267 test images. Each image has pixel dimensions 520×696.
  - From each image overlapping 2D 256×256 patches were extracted for training and whole images were taken for evaluation.
- Hoechst [5]
  - A modified U2OS osteosarcoma cell line.
  - The dataset consists of 30 training images and 10 test images. Each image has pixel dimensions 1104×1104
  - From each image overlapping 2D 256×256 patches were extracted for training and whole images were taken for evaluation.
- S_BIAD895/SB-895 [16]
  - The dataset consists of 47 images with pixel dimensions 1024×1024. We did not split the dataset into test and train splits as we did not apply models trained on S_BIAD895 to S_BIAD895 data, removing this transfer from our set of evaluated transfers. We only considered S_BIAD895 models under transfer to other target datasets.
  - Thus for models transferred to S_BIAD895, the full image set could be used for testing.
  - From each image 2D 256×256 patches were extracted for training and whole images were taken for evaluation.
- The SELMA3D 2024 challenge [10, 18, 40, 82, 103].
  - The dataset consists of eight $200 \times 200 \times 200$ training volumes, one $200 \times 200 \times 200$ validation volume and three $200 \times 200 \times 200$ test volumes.
  - We took 2D 200×200 patches from each z-slice for training and evaluation.
- S_BIAD1410 [31]
  - A dataset of fruit fly embryonic development.
  - The dataset comprises of ten training volumes with the following dimensions ($(401 \times 512 \times 512)$, $(240 \times 512 \times 512)$, $(400 \times 512 \times 512)$, $(380 \times 512 \times 512)$ and six times $(121 \times 512 \times 512)$).
  - From each volume 2D 256×256 patches were extracted for training and evaluation.

The set of source models trained on the seven available datasets were then transferred to four target datasets; BBBC039 [52], Hoechst [5], S_BIAD895 [16] and DSB2018 [14] (a fluorescent only subset of the data science bowl 2018 challenge). The target datasets were chosen based on two criteria, firstly, for the quality of their ground-truth annotation, to allow for realistic performance evaluation scores. Secondly, the performance of models transferred to the target dataset should have some differentiation. Model differentiation allows for proper investigation of ranking metrics. If all the models transferred to a dataset perform equally then ranking becomes obsolete and hard to evaluate.

Instance Segmentation. We trained nuclei instance segmentation models on five datasets; BBBC039 [52], HeLaCytoNuc [19], Hoechst [5], S_BIAD895/SB-895 [16] and SBIAD1410 [31]. The source models were then transferred to the test sets of four target datasets BBBC039 [52], Hoechst [5], S_BIAD895/SB-895 [16] and one additional dataset S_BIAD634/SB-634 [43] was introduced. S_BIAD634/SB-634 is a dataset of nuclei images of normal and cancer cells from different tissue origins and sample preparation types. The dataset comprises of 42 1024×1280 training images and 37 1024×1280 test images. Once again target datasets were chosen with the same two criteria for instance ground-truth quality and differentiable model performance under transfer. For details on training of instance segmentation models see Sec. 8.

7.3 Light Microscopy (LM), Cells

Instance Segmentation. We consider four dense instance segmentation problems for cell datasets:

- FlyWing [26]
  - Fluorescent volume of a developing fruit fly wing.
  - The dataset comprises of four training volumes with the following pixel dimensions ($(160 \times 773 \times 881)$, $(160 \times 773 \times 881)$, $(148 \times 765 \times 877)$ and $(180 \times 776 \times 893)$), two validation volumes both with dimensions $(160 \times 590 \times 773)$ and two test volumes both with dimensions $(160 \times 639 \times 765)$.
  - From each volume 2D 256×256 patches were extracted for training and full z-slices taken for evaluation.
- Ovules [96]
  - Fluorescent volume of a *Arabidopsis thaliana*
  - We used 20 volumes (see data for dimensions) for training, a further single volume $(397 \times 880 \times 1332)$ for validation and a single volume $(320 \times 960 \times 1000)$ for testing.
  - From each volume 2D 256×256 patches were extracted for training and full z-slices taken for evaluation.
- PNAS [95]
  - Fluorescent volume of the *Arabidopsis thaliana* apical stem cell niche.
  - For training we used 103 volumes collected at 4 hour time points between 0-84hrs from four plants (see data for dimensions). Volumes from two separate plants were used for validation and testing.
  - From each volume 2D 256×256 images were extracted for training and full z-slices taken for evaluation.
- Covid-IF [67]
  - Immunofluorescence imaging of human serum.
  - The dataset comprises of 48 1024×1024 images, which are all used for testing since we didn't train any models on this dataset and only use it to evaluate publicly available pre-trained models.

7.4 Cone-Beam Computed Tomography (CBCT), ToothFairy2

Multiclass Semantic Segmentation. For multiclass semantic segmentation experiments, we use ToothFairy2 from a MICCAI2024 challenge [12,13]. It comprises 530 volumetric scans with voxel-wise annotations of maxillofacial anatomy, of which 480 are publicly available for training, while 50, acquired by a different institution, are held out for evaluation via the grand-challenge platform. The dataset provides annotations for 42 classes, including jawbones, left and right inferior alveolar canals, maxillary sinuses, pharynx, upper and lower teeth, bridges, crowns, and implants. For evaluation, we obtained access to the private test set, collected by a different centre on different acquisition machines.

8 Models

We trained a range of model architectures to build a diverse model set for transfer comparison. All of the models trained by us are trained on the training set of a single 'source' dataset. We additionally investigated several publicly available models taken from the ToothFairy2 challenge [12], the Bioimage Model Zoo [63] and large generalist models taken from popular repositories (Cellpose-SAM [65], μSAM [3], TotalSegmentator [93]). Information listing the full set of models considered can be found in Sec. 4.2 in the main paper. In order to fully test the performance of consistency ranking we set up experiments that range from comparing identical models trained on different datasets, heterogenous multi-architecture sets and large generalists vs specialist models. We only tested ranking models where the source and target modalities matched, as shifting between modalities results in very large domain gaps that make successful direct application very unlikely. Hence, we organised our experiments by dataset group, below is a summary of the exact model sets used in each experiment.

Semantic Segmentation:

- *Mitochondria segmentation (Tab. 1b and Tab. 3)* – $|\mathcal{M}| = 15$; *architectures:* U-Net, Residual-U-Net, UNETR; *Source Datasets:* EPFL, Hmito, Rmito, VNC
- *Nuclei segmentation (Tab. 1b)* – $|\mathcal{M}| = 7$; *architectures:* U-Net; *Source Datasets:* BBBC039, Go-Nuclear, HeLaCytoNuc, Hoechst, S-BIAD895, SELMA3D, S-BIAD1410
- *Human Jaw (Tab. 2)* – $|\mathcal{M}| = 8$; *architectures:* 2D/3D nnU-Net, Swin-UNETR, TotalSegmentator, UMamba, VMamba;
  Source Datasets: ToothFairy2

Instance Segmentation:

- *Nuclei segmentation (Tab. 4)* – $|\mathcal{M}| = 5$; *architectures:* U-Net;
  Source Datasets: BBBC039, HeLaCytoNuc, Hoechst, S-BIAD895, S-BIAD1410

- *Cell segmentation (Tab. 4)* – $|\mathcal{M}| = 8$; *architectures:* U-Net, Residual-U-Net, UNETR; *Source Datasets:* FlyWing, Ovules, PNAS
- *Cell Segmentation Generalist (Fig. 4)* – $|\mathcal{M}| = 5$; *architectures:* U-Net, μSAM, Cellpose-SAM; *Source Datasets:* N/A

In the following paragraph we detail the model training setup. All the self-trained networks were trained until convergence on the source dataset and performed highly on the held out source test set. The classification networks were trained with a Binary Cross entropy (BCE) loss, while the segmentation networks used a combination of Dice and CE losses. For models trained for the ToothFairy2 challenge dataset, the nnUNet planner was used for preprocessing and hyperparameters selection. The models were trained from scratch on a single 48GB A40 Nvidia GPU using CUDA 11.8 and PyTorch 2.1.2. All other models were trained from scratch on a set of 8 12GB NVIDIA GeForce RTX 3080 Ti GPUs using CUDA12.1 and PyTorch 2.5.1.

8.1 Instance Segmentation Training

For the instance segmentation models trained by us we follow a popular instantiation process [67] of predicting cell boundaries and then using seeded watershed and graph partitioning algorithm GASP [7] to obtain instances.

9 Model Perturbation

Input Perturbations: We tested a range of different input perturbations by applying several popular image transformations at test-time: additive Gaussian noise, gamma correction, and changes in brightness and contrast. The transformations are defined as follows,

$$\text{AdditiveGaussianNoise:} \quad x' = x + \mathcal{N}(0, \sigma), \tag{6}$$

$$\text{RandomBrightness:} \quad x' = x + \theta_B, \tag{7}$$

$$\text{RandomContrast:} \quad x' = \mu(x) + \theta_C(x - \mu(x)), \tag{8}$$

$$\text{RandomGammaCorrection:} \quad x' = x^{\gamma}, \tag{9}$$

where the strength of each perturbation can be controlled by σ, θ_B, θ_C and γ parameters respectively.

Feature Space Perturbations: We relied upon test-time application of spatial DropOut [83] to provide feature space perturbation. We tested several approaches for applying DropOut to our models. Firstly, we tested applying Spatial DropOut only to the bottleneck layer of U-Net style architectures. Our experiments for bottleneck only DropOut are shown on the nuclei datasets both for

semantic and instance segmentation (Tabs. 11a, 11b and 21). Secondly, in addition to the bottleneck layer we also investigated adding DropOut to all layers that connect the encoder to the decoder, namely skip connections in U-Net style networks. Our experiments for bottleneck + skip connection DropOut are shown on the ToothFairy2 dataset for multiclass semantic segmentation (Tab. 14). Lastly, we investigate applying DropOut uniformly to every layer of the network, this represents the most general and architecture agnostic approach considered. Our experiments for all layer DropOut are shown on the mitochondria semantic segmentation datasets along with the cell instance segmentation datasets (Tabs. 6a, 6b and 19).

We found that all the considered feature space perturbation approaches were able to perform effectively, as shown in the tables in the following sections. However, it is worth noting that the first approach of applying DropOut to the bottleneck layer only can be sensitive to residual connections. The second approach of applying DropOut to both the bottleneck layer and skip connections works very well on the model sets tested, but does assume that the models have skip connections in the architecture, otherwise it reduces to the first case (bottleneck only) and could suffer the same sensitivity to residual connections. The last approach of applying DropOut to all layers is the most architecture agnostic and resulted in very strong correlation between the CMR metric and performance score.

10 Unsupervised Domain Adaptation (UDA) Approaches

When considering the reuse of pre-trained models users can either directly apply a model with frozen weights, in what we refer to as zero-shot reuse, or after applying some unsupervised domain adaptation (UDA) method to models in order to adapt them to the target data. Application of UDA has the potential to improve model performance on the target data, but requires users to have the resources and know-how to properly make use of existing methods. Furthermore, UDA methods can often be unstable and can even lead to degradation of model performance on the target data [92]. Hence, even if UDA is applied the ranking task still persists, and now the best adapted model must be selected.

In order to investigate ranking of post-UDA models we utilised two common UDA approaches to adapt our pre-trained source networks to target data. Firstly, we used a Mean Teacher [81] approach that relies on a 'student' 'teacher' paired network setting. The Mean Teacher framework for self-supervised training initialises both the student and teacher networks using the pre-trained source model weights and then relies on the 'teacher' branch to provide pseudo labels via its predictions. The student branch is trained via gradient descent using a consistency loss to the pseudo labels. Finally, the teacher weights are then updated via an exponential moving average of the student weights. We did not employ any additional pseudo-label selection steps as we found 'no-selection' was a sufficient strategy in many cases for successful UDA. In addition, in our experiments we are not directly interested in improving the performance of UDA, but

rather our real question is: can we rank the final post-UDA models regardless of whether UDA successfully improved target performance or indeed led to model degradation?

The second UDA approach we employed was Adaptive Batch Normalization [48] (AdaBN). AdaBN provides a simple and parameter-free approach for UDA and is based on updating the statistics of all Batch Normalization layers in a network to match the target data. This simple approach proved surprisingly effective at improving post-UDA model performance.

For both the Mean Teacher and AdaBN approaches we were able to successfully rank the post-UDA model performance, outcompeting the state-of-the-art baseline Transfer Score [98]. In Tab. 3, we showed the results for single CMR perturbation strengths, in Tab. 16 we show the CMR correlation to post-UDA F1-score across a range of perturbation strengths.

11 Semantic Segmentation

In the following section we investigate ranking of semantic segmentation models, each of the tables Tabs. 5a, 5b, 6a, 6b, 7a, 7b, 11a and 11b shows the correlation scores between our CMR metrics (CMR-EI and CMR-NHD) and the performance F1-score for sets of models applied directly to a single target dataset, across a range of perturbation strengths. In the main paper, due to space constraints, we show the correlation scores for a single perturbation strength averaged over the set of target datasets.

We investigate the dependency of CMR on the strength of perturbation applied for both input and feature space perturbations. The tables show that CMR's performance is stable across a wide range of perturbations.

11.1 Mitochondria Perturbation Sweep

For mitochondria semantic segmentation we investigated ranking a model set with $|\mathcal{M}| = 15$. The models were trained separately using all four Mitochondria datasets (EPFL [54], Hmito [25], Rmito [25] and VNC [68]) as different sources. The model set comprised of a range of model architectures: 2D U-Net (trained both with and without augmentations), Residual U-Net and UNETR. Hence we built a diverse set of domain-specialist models which we apply to each target dataset in turn and rank the corresponding performance via CMR.

To investigate input space perturbations we applied additive Gaussian noise to the inputs between the strengths $\sigma = 0.01 - 0.2$. We show the results for both CMR-EI (Tab. 5a) and CMR-NHD (Tab. 5b).

For feature space perturbation we investigated applying spatial DropOut equally to all layers of a network, where the strength of the perturbation is controlled by the proportion of feature maps dropped at each layer p_d. We investigated DropOut proportions in the range $p_d = 0.001 - 0.1$. We show results for both CMR-EI (Tab. 6a) and CMR-NHD (Tab. 6b). The tables show that CMR's performance is stable across a wide range of perturbations.

Table 5: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Mitochondria [25, 54, 68] across a range of Gaussian Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 15$. σ was randomly sampled from the noted range. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-EI	Kτ	0.73	(**)	0.66	(**)	0.77	(**)	0.46	(0.06)
Gauss, σ	Sρ	0.88	(**)	0.79	(**)	0.90	(**)	0.61	(0.06)
[0.01, 0.03]	Pr	0.82	(**)	0.75	(**)	0.87	(**)	0.63	(*)
CMR-EI	Kτ	0.77	(**)	0.73	(**)	0.79	(**)	0.52	(*)
Gauss, σ	Sρ	0.90	(**)	0.86	(**)	0.90	(**)	0.63	(*)
[0.03, 0.05]	Pr	0.86	(**)	0.82	(**)	0.84	(**)	0.67	(*)
CMR-EI	Kτ	0.81	(**)	0.73	(**)	0.82	(**)	0.61	(**)
Gauss, σ	Sρ	0.93	(**)	0.86	(**)	0.92	(**)	0.66	(*)
[0.05, 0.07]	Pr	0.88	(**)	0.83	(**)	0.82	(**)	0.71	(*)
CMR-EI	Kτ	0.83	(**)	0.75	(**)	0.90	(**)	0.55	(*)
Gauss, σ	Sρ	0.93	(**)	0.87	(**)	0.95	(**)	0.63	(*)
[0.07, 0.1]	Pr	0.91	(**)	0.82	(**)	0.88	(**)	0.73	(**)
CMR-EI	Kτ	0.87	(**)	0.73	(**)	0.88	(**)	0.58	(**)
Gauss, σ	Sρ	0.95	(**)	0.87	(**)	0.95	(**)	0.65	(*)
[0.1, 0.12]	Pr	0.91	(**)	0.80	(**)	0.85	(**)	0.75	(**)
CMR-EI	Kτ	0.85	(**)	0.73	(**)	0.84	(**)	0.58	(**)
Gauss, σ	Sρ	0.94	(**)	0.87	(**)	0.94	(**)	0.67	(*)
[0.12, 0.15]	Pr	0.90	(**)	0.81	(**)	0.84	(**)	0.78	(**)
CMR-EI	Kτ	0.77	(**)	0.71	(**)	0.84	(**)	0.55	(*)
Gauss, σ	Sρ	0.92	(**)	0.86	(**)	0.94	(**)	0.62	(*)
[0.15, 0.2]	Pr	0.86	(**)	0.84	(**)	0.89	(**)	0.78	(**)

(b) CMR-NHD Correlation scores

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-NHD	Kτ	0.73	(**)	0.66	(**)	0.71	(**)	0.52	(*)
Gauss, σ	Sρ	0.89	(**)	0.82	(**)	0.90	(**)	0.63	(0.05)
[0.01, 0.03]	Pr	0.83	(**)	0.75	(**)	0.54	(*)	0.72	(**)
CMR-NHD	Kτ	0.77	(**)	0.67	(**)	0.73	(**)	0.48	(*)
Gauss, σ	Sρ	0.93	(**)	0.83	(**)	0.91	(**)	0.59	(0.05)
[0.03, 0.05]	Pr	0.83	(**)	0.74	(**)	0.61	(*)	0.74	(**)
CMR-NHD	Kτ	0.79	(**)	0.68	(**)	0.71	(**)	0.48	(*)
Gauss, σ	Sρ	0.94	(**)	0.83	(**)	0.90	(**)	0.59	(0.05)
[0.05, 0.07]	Pr	0.84	(**)	0.76	(**)	0.59	(*)	0.69	(*)
CMR-NHD	Kτ	0.73	(**)	0.69	(**)	0.75	(**)	0.55	(*)
Gauss, σ	Sρ	0.91	(**)	0.85	(**)	0.92	(**)	0.64	(*)
[0.07, 0.1]	Pr	0.90	(**)	0.74	(**)	0.68	(**)	0.72	(**)
CMR-NHD	Kτ	0.77	(**)	0.71	(**)	0.77	(**)	0.55	(*)
Gauss, σ	Sρ	0.92	(**)	0.85	(**)	0.92	(**)	0.64	(*)
[0.1, 0.12]	Pr	0.92	(**)	0.73	(**)	0.69	(**)	0.75	(**)
CMR-NHD	Kτ	0.81	(**)	0.69	(**)	0.77	(**)	0.55	(*)
Gauss, σ	Sρ	0.93	(**)	0.82	(**)	0.92	(**)	0.64	(*)
[0.12, 0.15]	Pr	0.91	(**)	0.75	(**)	0.71	(**)	0.77	(**)
CMR-NHD	Kτ	0.75	(**)	0.69	(**)	0.75	(**)	0.55	(*)
Gauss, σ	Sρ	0.91	(**)	0.82	(**)	0.91	(**)	0.65	(*)
[0.15, 0.2]	Pr	0.88	(**)	0.79	(**)	0.78	(**)	0.77	(**)

Table 6: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Mitochondria [25, 54, 68] across a range of DropOut Perturbation Strengths. DropOut was applied equally to all layers of the networks. For each ranking experiment $|\mathcal{M}| = 15$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-EI	Kτ	0.75	(**)	0.56	(**)	0.73	(**)	0.48	(*)
DropOut	Sρ	0.93	(**)	0.78	(**)	0.90	(**)	0.60	(*)
$p_d = 0.001$	Pr	0.91	(**)	0.76	(**)	0.90	(**)	0.78	(**)
CMR-EI	Kτ	0.75	(**)	0.58	(**)	0.77	(**)	0.48	(*)
DropOut	Sρ	0.93	(**)	0.79	(**)	0.91	(**)	0.64	(*)
$p_d = 0.005$	Pr	0.91	(**)	0.76	(**)	0.92	(**)	0.79	(**)
CMR-EI	Kτ	0.83	(**)	0.58	(**)	0.75	(**)	0.55	(*)
DropOut	Sρ	0.95	(**)	0.78	(**)	0.90	(**)	0.68	(*)
$p_d = 0.01$	Pr	0.94	(**)	0.77	(**)	0.92	(**)	0.79	(**)
CMR-EI	Kτ	0.87	(**)	0.68	(**)	0.77	(**)	0.52	(*)
DropOut	Sρ	0.96	(**)	0.83	(**)	0.90	(**)	0.66	(*)
$p_d = 0.02$	Pr	0.95	(**)	0.79	(**)	0.94	(**)	0.79	(**)
CMR-EI	Kτ	0.85	(**)	0.68	(**)	0.77	(**)	0.61	(**)
DropOut	Sρ	0.96	(**)	0.83	(**)	0.90	(**)	0.76	(**)
$p_d = 0.03$	Pr	0.95	(**)	0.79	(**)	0.94	(**)	0.78	(**)
CMR-EI	Kτ	0.83	(**)	0.66	(**)	0.79	(**)	0.52	(*)
DropOut	Sρ	0.95	(**)	0.83	(**)	0.91	(**)	0.64	(*)
$p_d = 0.04$	Pr	0.94	(**)	0.77	(**)	0.91	(**)	0.73	(**)
CMR-EI	Kτ	0.85	(**)	0.66	(**)	0.79	(**)	0.61	(**)
DropOut	Sρ	0.96	(**)	0.83	(**)	0.91	(**)	0.73	(*)
$p_d = 0.05$	Pr	0.95	(**)	0.78	(**)	0.90	(**)	0.71	(**)
CMR-EI	Kτ	0.75	(**)	0.49	(*)	0.57	(**)	0.52	(*)
DropOut	Sρ	0.90	(**)	0.65	(*)	0.71	(**)	0.69	(**)
$p_d = 0.1$	Pr	0.87	(**)	0.68	(**)	0.66	(**)	0.65	(*)

(b) CMR-NHD Correlation scores

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-NHD	Kτ	0.69	(**)	0.58	(**)	0.61	(**)	0.58	(**)
DropOut	Sρ	0.89	(**)	0.76	(**)	0.80	(**)	0.72	(*)
$p_d = 0.001$	Pr	0.90	(**)	0.69	(**)	0.78	(**)	0.80	(**)
CMR-NHD	Kτ	0.71	(**)	0.62	(**)	0.63	(**)	0.61	(**)
DropOut	Sρ	0.90	(**)	0.80	(**)	0.81	(**)	0.76	(**)
$p_d = 0.005$	Pr	0.90	(**)	0.70	(**)	0.81	(**)	0.82	(**)
CMR-NHD	Kτ	0.77	(**)	0.66	(**)	0.69	(**)	0.64	(**)
DropOut	Sρ	0.93	(**)	0.82	(**)	0.88	(**)	0.76	(*)
$p_d = 0.01$	Pr	0.94	(**)	0.72	(**)	0.83	(**)	0.81	(**)
CMR-NHD	Kτ	0.81	(**)	0.66	(**)	0.69	(**)	0.61	(*)
DropOut	Sρ	0.94	(**)	0.81	(**)	0.88	(**)	0.77	(**)
$p_d = 0.02$	Pr	0.97	(**)	0.73	(**)	0.86	(**)	0.81	(**)
CMR-NHD	Kτ	0.81	(**)	0.62	(**)	0.69	(**)	0.64	(**)
DropOut	Sρ	0.94	(**)	0.78	(**)	0.88	(**)	0.82	(**)
$p_d = 0.03$	Pr	0.96	(**)	0.74	(**)	0.87	(**)	0.79	(**)
CMR-NHD	Kτ	0.81	(**)	0.60	(**)	0.67	(**)	0.48	(*)
DropOut	Sρ	0.94	(**)	0.76	(**)	0.83	(**)	0.61	(0.06)
$p_d = 0.04$	Pr	0.95	(**)	0.73	(**)	0.85	(**)	0.73	(**)
CMR-NHD	Kτ	0.87	(**)	0.62	(**)	0.71	(**)	0.52	(*)
DropOut	Sρ	0.96	(**)	0.76	(**)	0.84	(**)	0.68	(*)
$p_d = 0.05$	Pr	0.96	(**)	0.74	(**)	0.85	(**)	0.71	(*)
CMR-NHD	Kτ	0.73	(**)	0.47	(*)	0.52	(**)	0.52	(*)
DropOut	Sρ	0.89	(**)	0.65	(**)	0.65	(**)	0.71	(*)
$p_d = 0.1$	Pr	0.88	(**)	0.65	(**)	0.64	(*)	0.64	(*)

11.2 Nuclei Perturbation Sweep

For nuclei semantic segmentation we investigated ranking a model set with $|\mathcal{M}| = 7$. We built the model set by training a set of specialist 2D U-Net models on seven distinct source datasets: BBBC039 [52], Go-Nuclear [86], HeLaCytoNuc [19], Hoechst [5], S_BIAD895/SB-895 [16], SELMA3D 2024 challenge [26] and S_BIAD1410 [31].

We then investigated applying a range of input augmentations; additive Gaussian noise, Gamma correction, Brightness adjustment and Contrast adjustment (as defined in Sec. 9). We investigated additive Gaussian noise with strengths in the range $\sigma = 0.0 - 0.2$, Gamma correction with strengths in the range $\gamma = 0.8 - 1.2$, where $\gamma = 1.0$ equals no adjustment, Brightness adjustment with strengths in the range $\theta_B = 0.0 - 0.2$ and Contrast adjustment with strengths in the range $\theta_C = 0.8 - 1.2$, where $\theta_C = 1.0$ equals no adjustment. We show the results for both CMR-EI (Tabs. 7a, 8a, 9a and 10a) and CMR-NHD (Tabs. 7b, 8b, 9b and 10b).

For feature space perturbation we investigated applying spatial DropOut only to the bottleneck layer of networks, where the strength of the perturbation is controlled by the proportion of feature maps dropped at each layer p_d. We investigated DropOut proportions in the range $p_d = 0.05 - 0.5$. We show results for for both CMR-EI (Tab. 11a) and CMR-NHD (Tab. 11b). Again over a wide range of perturbation strengths across all considered input and feature space perturbations the correlation scores achieved by our CMR metrics remain largely stable. Although, as discussed in the main paper, Sec. 3.3, care should be taken to ensure perturbations remain 'tolerable' in the extreme perturbation case, too weak perturbation can lead to a loss of correlation between consistency based metrics and transfer performance. For example, the first row of Tab. 7a for the DSB2018 target dataset shows the weakest level of Gaussian input noise (σ sampled from $[0.0, 0.05]$) can cause the CMR-NHD metric to become unstable on DSB2018.

Table 7: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52] across a range of Gaussian Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 7$. σ is randomly sampled from the noted range. *pval.* $<$ 0.05 (*), *pval.* $<$ 0.01 (**).

(a) CMR-EI Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-EI	$K\tau$	0.62	(0.08)	0.52	(0.12)	0.62	(0.06)	0.60	(0.19)
Gauss, σ	$S\rho$	0.71	(0.09)	0.68	(0.14)	0.75	(0.07)	0.77	(0.11)
[0.0, 0.05]	Pr	0.99	(**)	0.98	(**)	0.95	(**)	0.98	(**)
CMR-EI	$K\tau$	0.62	(0.07)	0.52	(0.16)	0.52	(0.14)	0.60	(0.15)
Gauss, σ	$S\rho$	0.79	(*)	0.68	(0.11)	0.71	(0.10)	0.77	(0.10)
[0.05, 0.1]	Pr	1.00	(**)	0.99	(**)	0.94	(**)	0.98	(**)
CMR-EI	$K\tau$	0.71	(*)	0.52	(0.16)	0.52	(0.13)	0.73	(0.06)
Gauss, σ	$S\rho$	0.86	(*)	0.68	(0.10)	0.71	(0.11)	0.83	(0.09)
[0.1, 0.2]	Pr	1.00	(**)	0.99	(**)	0.93	(**)	0.99	(**)

(b) CMR-NHD Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-NHD	$K\tau$	0.43	(0.21)	0.14	(0.77)	0.71	(*)	0.73	(0.07)
Gauss, σ	$S\rho$	0.64	(0.12)	0.18	(0.69)	0.86	(*)	0.83	(0.06)
[0.0, 0.05]	Pr	1.00	(**)	0.98	(**)	0.99	(**)	0.99	(**)
CMR-NHD	$K\tau$	0.52	(0.13)	0.71	(*)	0.71	(*)	0.47	(0.27)
Gauss, σ	$S\rho$	0.71	(0.08)	0.79	(*)	0.86	(*)	0.54	(0.30)
[0.05, 0.1]	Pr	1.00	(**)	1.00	(**)	0.93	(**)	0.98	(**)
CMR-NHD	$K\tau$	0.71	(*)	0.81	(*)	0.62	(0.07)	0.60	(0.14)
Gauss, σ	$S\rho$	0.86	(*)	0.89	(*)	0.82	(*)	0.71	(0.12)
[0.1, 0.2]	Pr	1.00	(**)	1.00	(**)	0.92	(**)	0.98	(**)

Table 8: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52] across a range of Brightness Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 7$. θ_B is randomly sampled from the noted range. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-EI	Kτ	0.71	(*)	0.43	(0.25)	0.52	(0.16)	0.60	(0.12)
Brt, θ_B	Sρ	0.86	(*)	0.64	(0.15)	0.71	(0.08)	0.77	(0.08)
[0.0, 0.05]	Pr	0.99	(**)	0.91	(**)	0.94	(**)	0.96	(**)
CMR-EI	Kτ	0.81	(*)	0.62	(0.08)	0.81	(*)	0.73	(0.07)
Brt, θ_B	Sρ	0.93	(*)	0.75	(0.07)	0.93	(**)	0.83	(0.05)
[0.05, 0.1]	Pr	0.99	(**)	0.99	(**)	0.94	(**)	0.99	(**)
CMR-EI	Kτ	0.71	(*)	0.62	(0.06)	0.81	(*)	0.73	(0.06)
Brt, θ_B	Sρ	0.86	(*)	0.75	(0.07)	0.93	(*)	0.83	(0.07)
[0.1, 0.2]	Pr	0.99	(**)	0.99	(**)	0.94	(**)	0.99	(**)

(b) CMR-NHD Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-NHD	Kτ	0.52	(0.12)	0.71	(*)	0.71	(*)	0.47	(0.25)
Brt, θ_B	Sρ	0.71	(0.08)	0.79	(0.05)	0.86	(*)	0.54	(0.29)
[0.0, 0.05]	Pr	1.00	(**)	0.82	(*)	0.92	(**)	0.82	(0.05)
CMR-NHD	Kτ	0.71	(*)	0.62	(0.07)	0.71	(*)	0.73	(0.05)
Brt, θ_B	Sρ	0.86	(*)	0.71	(0.08)	0.86	(*)	0.83	(0.06)
[0.05, 0.1]	Pr	1.00	(**)	0.98	(**)	0.90	(*)	0.99	(**)
CMR-NHD	Kτ	0.71	(*)	0.71	(*)	0.90	(*)	0.73	(0.07)
Brt, θ_B	Sρ	0.86	(*)	0.82	(*)	0.96	(**)	0.83	(0.07)
[0.1, 0.2]	Pr	1.00	(**)	0.99	(**)	0.93	(**)	0.99	(**)

Table 9: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52] across a range of Contrast Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 7$. θ_C is randomly sampled from the noted range. *pval.* $<$ 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-EI	Kτ	0.81	(*)	0.33	(0.35)	0.52	(0.16)	0.47	(0.32)
Ctr, θ_C	Sρ	0.93	(*)	0.50	(0.26)	0.68	(0.12)	0.71	(0.11)
[0.8, 0.9]	Pr	0.95	(**)	0.91	(**)	0.94	(**)	0.98	(**)
CMR-EI	Kτ	0.71	(*)	0.33	(0.36)	0.52	(0.15)	0.47	(0.26)
Ctr, θ_C	Sρ	0.86	(*)	0.50	(0.27)	0.68	(0.12)	0.71	(0.12)
[0.9, 0.95]	Pr	0.98	(**)	0.95	(**)	0.95	(**)	0.98	(**)
CMR-EI	Kτ	0.62	(0.09)	0.43	(0.28)	0.62	(0.08)	0.47	(0.27)
Ctr, θ_C	Sρ	0.71	(0.10)	0.64	(0.13)	0.75	(0.09)	0.71	(0.13)
[1.05, 1.1]	Pr	0.98	(**)	0.97	(**)	0.94	(**)	0.98	(**)
CMR-EI	Kτ	0.62	(0.07)	0.43	(0.26)	0.62	(0.09)	0.47	(0.29)
Ctr, θ_C	Sρ	0.71	(0.11)	0.64	(0.15)	0.75	(0.05)	0.71	(0.15)
[1.1, 1.2]	Pr	0.99	(**)	0.97	(**)	0.95	(**)	0.98	(**)

(b) CMR-NHD Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-NHD	Kτ	0.81	(*)	0.43	(0.23)	0.52	(0.12)	0.73	(0.05)
Ctr, θ_C	Sρ	0.93	(*)	0.54	(0.20)	0.71	(0.09)	0.83	(0.05)
[0.8, 0.9]	Pr	0.94	(**)	0.91	(**)	0.92	(**)	0.99	(**)
CMR-NHD	Kτ	0.81	(*)	0.43	(0.25)	0.81	(*)	0.60	(0.12)
Ctr, θ_C	Sρ	0.93	(*)	0.54	(0.24)	0.93	(*)	0.77	(0.09)
[0.9, 0.95]	Pr	0.98	(**)	0.96	(**)	0.97	(**)	0.98	(**)
CMR-NHD	Kτ	0.81	(*)	0.43	(0.27)	0.71	(*)	0.73	(0.07)
Ctr, θ_C	Sρ	0.93	(**)	0.54	(0.21)	0.89	(*)	0.83	(0.07)
[1.05, 1.1]	Pr	0.99	(**)	0.97	(**)	0.92	(**)	0.99	(**)
CMR-NHD	Kτ	0.81	(*)	0.43	(0.24)	0.71	(0.05)	0.60	(0.14)
Ctr, θ_C	Sρ	0.93	(*)	0.64	(0.12)	0.89	(*)	0.77	(0.08)
[1.1, 1.2]	Pr	0.99	(**)	0.99	(**)	0.91	(**)	0.99	(**)

Table 10: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52] across a range of Gamma Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 7$. γ is randomly sampled from the noted range. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-EI	Kτ	0.71	(*)	0.52	(0.12)	0.81	(**)	0.60	(0.13)
Gamma, γ	Sρ	0.86	(*)	0.68	(0.13)	0.89	(*)	0.77	(0.09)
[0.8, 0.9]	Pr	0.98	(**)	0.96	(**)	0.98	(**)	0.97	(**)
CMR-EI	Kτ	0.71	(*)	0.52	(0.14)	0.71	(*)	0.47	(0.26)
Gamma, γ	Sρ	0.86	(*)	0.68	(0.13)	0.82	(*)	0.71	(0.16)
[0.9, 0.95]	Pr	0.99	(**)	0.96	(**)	0.98	(**)	0.96	(**)
CMR-EI	Kτ	0.71	(*)	0.52	(0.13)	0.62	(0.06)	0.47	(0.26)
Gamma, γ	Sρ	0.86	(*)	0.68	(0.12)	0.75	(0.08)	0.71	(0.14)
[1.05, 1.1]	Pr	0.98	(**)	0.94	(**)	0.97	(**)	0.97	(**)
CMR-EI	Kτ	0.81	(*)	0.52	(0.16)	0.81	(*)	0.73	(0.06)
Gamma, γ	Sρ	0.93	(**)	0.68	(0.14)	0.89	(*)	0.83	(0.07)
[1.1, 1.2]	Pr	0.96	(**)	0.94	(**)	0.98	(**)	0.96	(**)

(b) CMR-NHD Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-NHD	Kτ	0.71	(*)	0.62	(0.05)	0.90	(*)	0.87	(*)
Gamma, γ	Sρ	0.86	(*)	0.75	(0.06)	0.96	(**)	0.94	(*)
[0.8, 0.9]	Pr	0.97	(**)	0.91	(**)	0.99	(**)	0.95	(**)
CMR-NHD	Kτ	0.62	(0.07)	0.52	(0.14)	0.81	(*)	0.87	(*)
Gamma, γ	Sρ	0.75	(0.06)	0.68	(0.13)	0.93	(*)	0.94	(*)
[0.9, 0.95]	Pr	0.97	(**)	0.92	(**)	1.00	(**)	0.95	(**)
CMR-NHD	Kτ	0.62	(0.08)	0.43	(0.26)	1.00	(**)	0.73	(0.05)
Gamma, γ	Sρ	0.82	(0.05)	0.57	(0.19)	1.00	(**)	0.83	(0.08)
[1.05, 1.1]	Pr	0.93	(**)	0.88	(*)	1.00	(**)	0.83	(*)
CMR-NHD	Kτ	0.71	(*)	0.62	(0.06)	0.81	(*)	1.00	(**)
Gamma, γ	Sρ	0.79	(*)	0.75	(0.06)	0.93	(*)	1.00	(*)
[1.1, 1.2]	Pr	0.84	(*)	0.89	(*)	1.00	(**)	0.80	(0.06)

Table 11: Per-dataset correlation to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52] across a range of DropOut Perturbation Strengths. DropOut was applied only to the bottleneck layer of networks. For each ranking experiment $|\mathcal{M}| = 7$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) CMR-EI Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-EI	Kτ	0.71	(*)	0.43	(0.24)	0.52	(0.18)	0.60	(0.13)
DropOut	Sρ	0.86	(*)	0.64	(0.15)	0.71	(0.09)	0.77	(0.11)
$p_d = 0.05$	Pr	0.99	(**)	0.92	(**)	0.90	(*)	0.97	(**)
CMR-EI	Kτ	0.71	(*)	0.43	(0.25)	0.52	(0.15)	0.60	(0.14)
DropOut	Sρ	0.86	(*)	0.64	(0.14)	0.71	(0.12)	0.77	(0.12)
$p_d = 0.1$	Pr	0.99	(**)	0.86	(*)	0.86	(*)	0.96	(**)
CMR-EI	Kτ	0.71	(*)	0.43	(0.24)	0.52	(0.12)	0.60	(0.12)
DropOut	Sρ	0.86	(*)	0.64	(0.13)	0.71	(0.09)	0.77	(0.08)
$p_d = 0.2$	Pr	0.99	(**)	0.86	(*)	0.85	(*)	0.96	(**)
CMR-EI	Kτ	0.71	(*)	0.52	(0.12)	0.52	(0.14)	0.60	(0.15)
DropOut	Sρ	0.86	(*)	0.68	(0.13)	0.71	(0.11)	0.77	(0.13)
$p_d = 0.3$	Pr	0.99	(**)	0.79	(*)	0.85	(*)	0.96	(**)
CMR-EI	Kτ	0.81	(*)	0.71	(*)	0.52	(0.15)	0.73	(0.06)
DropOut	Sρ	0.93	(**)	0.82	(*)	0.71	(0.09)	0.83	(0.06)
$p_d = 0.4$	Pr	0.98	(**)	0.83	(*)	0.80	(*)	0.95	(**)
CMR-EI	Kτ	0.90	(*)	0.81	(*)	0.52	(0.13)	0.73	(0.05)
DropOut	Sρ	0.96	(*)	0.89	(*)	0.71	(0.07)	0.83	(0.05)
$p_d = 0.5$	Pr	0.98	(**)	0.77	(*)	0.89	(*)	0.95	(**)

(b) CMR-NHD Correlation scores

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	SB-895	*pval.*
CMR-NHD	Kτ	0.71	(*)	0.43	(0.23)	0.52	(0.15)	0.87	(*)
DropOut	Sρ	0.86	(*)	0.64	(0.16)	0.68	(0.11)	0.94	(*)
$p_d = 0.05$	Pr	0.97	(**)	0.01	(0.98)	0.65	(0.11)	0.81	(0.05)
CMR-NHD	Kτ	0.71	(*)	0.71	(0.05)	0.62	(0.06)	0.73	(0.05)
DropOut	Sρ	0.86	(*)	0.86	(*)	0.82	(*)	0.83	(0.07)
$p_d = 0.1$	Pr	0.97	(**)	-0.03	(0.95)	0.61	(0.15)	0.68	(0.13)
CMR-NHD	Kτ	0.90	(*)	0.52	(0.12)	0.71	(*)	0.73	(0.07)
DropOut	Sρ	0.96	(*)	0.68	(0.13)	0.89	(*)	0.83	(0.05)
$p_d = 0.2$	Pr	0.97	(**)	0.10	(0.82)	0.67	(0.10)	0.74	(0.10)
CMR-NHD	Kτ	0.90	(*)	0.62	(0.07)	0.81	(*)	0.73	(0.05)
DropOut	Sρ	0.96	(**)	0.71	(0.10)	0.93	(*)	0.83	(0.06)
$p_d = 0.3$	Pr	0.97	(**)	-0.01	(0.99)	0.63	(0.13)	0.74	(0.09)
CMR-NHD	Kτ	0.81	(*)	0.71	(0.05)	0.71	(*)	0.60	(0.13)
DropOut	Sρ	0.89	(*)	0.89	(*)	0.89	(*)	0.71	(0.14)
$p_d = 0.4$	Pr	0.97	(**)	0.18	(0.70)	0.63	(0.13)	0.71	(0.11)
CMR-NHD	Kτ	0.81	(*)	0.71	(*)	0.52	(0.13)	0.73	(0.05)
DropOut	Sρ	0.89	(*)	0.89	(*)	0.68	(0.13)	0.83	(0.05)
$p_d = 0.5$	Pr	0.96	(**)	0.11	(0.81)	0.84	(*)	0.68	(0.14)

11.3 Baseline Transferability Metrics

We also show the correlation scores per dataset for all the baseline transferability metrics. Tab. 12 shows the results for the four mitochondria datasets (EPFL [54], Hmito [25], Rmito [25] and VNC [68]) and Tab. 13 shows the results for the four nuclei datasets (BBBC039 [52], DSB2018 [14], Hoechst [5], S_BIAD895 [16]). In the main paper due to space constraints we showed the correlation scores averaged within the dataset groups.

Table 12: Per-dataset baseline correlations to F1-score ranking for Semantic Segmentation of Mitochondria. For each ranking experiment $|\mathcal{M}| = 15$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CCFV	Kτ	0.03	(0.89)	-0.09	(0.70)	-0.21	(0.30)	-0.18	(0.45)
	Sρ	0.02	(0.96)	-0.15	(0.56)	-0.29	(0.30)	-0.22	(0.51)
	Pr	-0.03	(0.92)	-0.17	(0.54)	-0.23	(0.40)	-0.29	(0.36)
GBC	Kτ	0.45	(*)	0.43	(*)	0.34	(0.09)	0.33	(0.15)
	Sρ	0.60	(*)	0.54	(*)	0.51	(0.06)	0.41	(0.17)
	Pr	0.65	(*)	0.54	(*)	0.47	(0.08)	0.23	(0.48)
LEEP	Kτ	0.79	(**)	0.90	(**)	0.94	(**)	0.91	(**)
	Sρ	0.89	(**)	0.98	(**)	0.98	(**)	0.98	(**)
	Pr	0.96	(**)	0.96	(**)	0.98	(**)	0.98	(**)
NLEEP	Kτ	0.26	(0.21)	0.77	(**)	0.65	(**)	-0.06	(0.86)
	Sρ	0.32	(0.23)	0.90	(**)	0.82	(**)	-0.06	(0.92)
	Pr	0.26	(0.34)	0.66	(*)	0.63	(*)	-0.01	(0.97)
Hscore	Kτ	-0.12	(0.58)	0.33	(0.10)	0.36	(0.06)	0.36	(0.11)
	Sρ	-0.21	(0.46)	0.50	(0.05)	0.53	(*)	0.50	(0.09)
	Pr	0.00	(1.00)	0.49	(0.07)	0.46	(0.08)	0.54	(0.07)
RegHscore	Kτ	-0.12	(0.52)	0.31	(0.12)	0.40	(*)	0.36	(0.10)
	Sρ	-0.21	(0.44)	0.49	(0.06)	0.58	(*)	0.50	(0.11)
	Pr	0.00	(1.00)	0.48	(0.07)	0.46	(0.08)	0.54	(0.07)
LogME	Kτ	-0.24	(0.23)	0.10	(0.61)	0.25	(0.22)	0.21	(0.39)
	Sρ	-0.33	(0.23)	0.17	(0.56)	0.32	(0.24)	0.25	(0.43)
	Pr	-0.30	(0.27)	0.05	(0.85)	0.09	(0.75)	0.41	(0.18)
NCTI	Kτ	-0.05	(0.85)	0.30	(0.14)	0.27	(0.16)	0.33	(0.13)
	Sρ	-0.07	(0.75)	0.42	(0.12)	0.35	(0.23)	0.44	(0.17)
	Pr	0.07	(0.81)	0.42	(0.12)	0.42	(0.12)	0.58	(0.05)
TS	Kτ	0.54	(*)	0.28	(0.14)	0.08	(0.77)	0.09	(0.74)
	Sρ	0.72	(*)	0.33	(0.23)	0.02	(0.94)	0.11	(0.70)
	Pr	0.60	(*)	0.19	(0.50)	0.14	(0.62)	0.00	(0.99)
NuNo	Kτ	0.49	(*)	0.2	(0.28)	0.06	(0.85)	-0.06	(0.87)
	Sρ	0.68	(*)	0.20	(0.51)	-0.03	(0.92)	-0.04	(0.90)
	Pr	0.52	(*)	0.10	(0.72)	-0.05	(0.86)	-0.04	(0.90)
Dispersion	Kτ	0.03	(0.94)	-0.07	(0.78)	-0.06	(0.73)	0.00	(1.00)
	Sρ	0.05	(0.87)	-0.13	(0.66)	-0.12	(0.66)	-0.08	(0.81)
	Pr	-0.07	(0.81)	-0.25	(0.36)	-0.20	(0.49)	-0.17	(0.60)

Table 13: Per-dataset baseline correlations to F1-score ranking for Semantic Segmentation of Nuclei [5, 14, 16, 52]. For each ranking experiment $|\mathcal{M}| = 12$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		BBBC039	*pval.*	DSB2018	*pval.*	Hoechst	*pval.*	S_BIAD895	*pval.*
CCFV	Kτ	0.05	(1.00)	0.05	(1.00)	0.24	(0.54)	-0.47	(0.29)
	Sρ	0.14	(0.77)	0.04	(0.98)	0.21	(0.61)	-0.60	(0.25)
	Pr	-0.31	(0.51)	-0.01	(0.99)	0.45	(0.31)	-0.78	(0.07)
GBC	Kτ	-0.14	(0.80)	-0.14	(0.82)	0.24	(0.54)	-0.47	(0.29)
	Sρ	-0.21	(0.67)	-0.21	(0.65)	0.36	(0.45)	-0.60	(0.21)
	Pr	-0.16	(0.72)	0.11	(0.81)	0.32	(0.48)	-0.59	(0.22)
LEEP	Kτ	0.81	(*)	0.71	(*)	0.14	(0.77)	1.00	(**)
	Sρ	0.89	(*)	0.86	(*)	0.32	(0.51)	1.00	(*)
	Pr	0.95	(**)	0.92	(**)	0.94	(**)	0.94	(*)
NLEEP	Kτ	0.14	(0.77)	-0.05	(1.00)	0.14	(0.77)	0.20	(0.79)
	Sρ	0.18	(0.72)	-0.04	(1.00)	0.14	(0.76)	0.20	(0.70)
	Pr	0.62	(0.14)	0.30	(0.52)	0.09	(0.85)	0.26	(0.62)
Hscore	Kτ	0.33	(0.33)	-0.14	(0.78)	-0.05	(1.00)	1.00	(*)
	Sρ	0.36	(0.46)	-0.36	(0.45)	0.00	(1.00)	1.00	(*)
	Pr	0.36	(0.43)	0.13	(0.77)	-0.04	(0.94)	0.78	(0.07)
RegHscore	Kτ	0.33	(0.41)	0.05	(0.97)	0.14	(0.72)	0.87	(*)
	Sρ	0.36	(0.45)	-0.07	(0.92)	0.21	(0.66)	0.94	(*)
	Pr	0.29	(0.52)	0.18	(0.70)	0.07	(0.88)	0.84	(*)
LogME	Kτ	0.14	(0.82)	-0.14	(0.78)	-0.33	(0.41)	-0.47	(0.25)
	Sρ	0.18	(0.67)	-0.21	(0.68)	-0.50	(0.24)	-0.60	(0.23)
	Pr	-0.36	(0.43)	-0.54	(0.21)	-0.12	(0.80)	-0.94	(*)
NCTI	Kτ	0.33	(0.39)	-0.14	(0.76)	-0.05	(1.00)	1.00	(**)
	Sρ	0.36	(0.45)	-0.36	(0.45)	0.00	(1.00)	1.00	(*)
	Pr	0.37	(0.41)	0.14	(0.76)	-0.02	(0.96)	0.72	(0.10)
TS	Kτ	0.14	(0.84)	-0.33	(0.41)	-0.05	(1.00)	0.33	(0.45)
	Sρ	0.25	(0.64)	-0.43	(0.35)	0.11	(0.77)	0.49	(0.35)
	Pr	-0.02	(0.97)	-0.56	(0.20)	-0.13	(0.79)	0.37	(0.48)
NuNo	Kτ	0.33	(0.36)	-0.43	(0.23)	-0.14	(0.82)	0.60	(0.13)
	Sρ	0.57	(0.16)	-0.43	(0.37)	-0.11	(0.85)	0.66	(0.19)
	Pr	0.15	(0.75)	-0.48	(0.28)	-0.04	(0.93)	0.68	(0.14)
Dispersion	Kτ	0.24	(0.54)	-0.24	(0.56)	-0.24	(0.53)	-0.47	(0.29)
	Sρ	0.25	(0.58)	-0.43	(0.36)	-0.39	(0.44)	-0.60	(0.26)
	Pr	0.43	(0.33)	-0.02	(0.97)	-0.62	(0.14)	-0.56	(0.25)

11.4 Multiclass ToothFairy2 Perturbation Sweep

For multiclass semantic segmentation we built a diverse set of models with $|\mathcal{M}| = 8$. All but one of the models were trained on the ToothFairy2 datasets taken from the MICCAI2024 challenge [12, 13], the exception being the generalist TotalSegmentator [93] model. The model set comprised of a wide range of architectures: 2D nnU-Net, residual-encoder 3D nnU-Net, SwinUNETR, TotalSegmentator, UMamba and VMamba.

We investigated applying both additive Gaussian noise (controlled by σ) and Gamma Correction (controlled by γ) as input perturbations. For feature space perturbation we investigated applying spatial DropOut to the bottleneck and skip connection layers of the network, where the strength of the perturbation is controlled by the proportion of feature maps dropped at each layer p_d. We report the performance of both CMR-EI (Tab. 14) and CMR-NHD (Tab. 15).

Table 14: CMR-EI Correlation scores for multiclass Semantic Segmentation of ToothFairy2 human jaw dataset across a range of Perturbations. $|\mathcal{M}| = 8$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		IACs		Teeth		Mand.		Sinus.		Overall	
			pval.		*pval.*		*pval.*		*pval.*		*pval.*
CMR-EI	Kτ	0.90	(**)	0.81	(**)	0.62	(*)	0.43	(0.3)	0.71	(*)
Gauss	Sρ	0.96	(**)	0.89	(*)	0.82	(*)	0.50	(0.3)	0.86	(*)
(σ = 2.0x)	Pr	0.99	(**)	0.95	(**)	0.73	(0.1)	0.38	(0.4)	0.93	(**)
CMR-EI	Kτ	0.93	(**)	0.86	(*)	0.65	(*)	0.40	(0.2)	0.71	(*)
Gauss	Sρ	0.97	(**)	0.85	(0.1)	0.85	(**)	0.44	(0.1)	0.84	(*)
(σ = 2.5x)	Pr	0.96	(*)	0.94	(**)	0.73	(*)	0.39	(0.3)	0.90	(*)
CMR-EI	Kτ	0.90	(**)	0.84	(**)	0.65	(*)	0.43	(0.2)	0.71	(*)
Gauss	Sρ	0.97	(**)	0.89	(**)	0.85	(**)	0.44	(0.2)	0.86	(*)
(σ = 3.0x)	Pr	0.96	(*)	0.92	(**)	0.59	(0.2)	0.42	(0.2)	0.90	(*)
CMR-EI	Kτ	0.93	(**)	0.77	(*)	0.65	(0.1)	0.57	(0.1)	0.78	(*)
Gamma	Sρ	0.96	(**)	0.88	(**)	0.74	(0.2)	0.61	(0.1)	0.88	(0.1)
(γ = 0.2)	Pr	0.95	(*)	0.71	(*)	0.66	(0.1)	0.50	(0.1)	0.70	(0.2)
CMR-EI	Kτ	0.97	(**)	0.77	(**)	0.68	(0.1)	0.48	(0.2)	0.81	(**)
Gamma	Sρ	0.98	(**)	0.90	(**)	0.80	(*)	0.51	(0.2)	0.89	(**)
(γ = 0.5)	Pr	0.94	(*)	0.83	(**)	0.75	(*)	0.42	(0.3)	0.89	(**)
CMR-EI	Kτ	1.00	(**)	0.71	(*)	0.43	(0.3)	0.43	(0.2)	0.81	(**)
Gamma	Sρ	1.00	(**)	0.86	(*)	0.54	(0.2)	0.57	(0.2)	0.89	(*)
(γ = 2.0)	Pr	0.99	(**)	0.89	(**)	0.74	(0.1)	0.38	(0.4)	0.90	(**)
CMR-EI	Kτ	1.00	(**)	0.60	(0.2)	0.83	(**)	0.44	(0.2)	0.58	(0.2)
DropOut	Sρ	1.00	(**)	0.77	(0.1)	0.82	(*)	0.52	(0.2)	0.72	(0.1)
(p_d = 0.3)	Pr	0.95	(**)	0.94	(*)	0.87	(*)	0.49	(0.1)	0.86	(*)
CMR-EI	Kτ	0.87	(*)	0.67	(0.1)	0.87	(*)	0.87	(*)	0.93	(*)
DropOut	Sρ	0.94	(**)	0.77	(0.1)	0.87	(*)	0.84	(*)	0.95	(*)
(p_d = 0.5)	Pr	0.83	(*)	0.85	(**)	0.89	(**)	0.88	(**)	0.93	(**)
CMR-EI	Kτ	0.65	(*)	0.58	(0.2)	0.68	(0.1)	0.59	(0.1)	0.60	(0.2)
DropOut	Sρ	0.79	(*)	0.61	(0.2)	0.75	(0.1)	0.58	(0.2)	0.57	(0.4)
(p_d = 0.7)	Pr	0.70	(*)	0.82	(*)	0.86	(0.1)	0.55	(0.2)	0.79	(**)

Table 15: CMR-NHD Correlation scores for multiclass Semantic Segmentation of ToothFairy2 human jaw dataset across a range of Perturbations. $|\mathcal{M}| = 8$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		IACs		Teeth		Mand.		Sinus.		Overall	
			pval.		*pval.*		*pval.*		*pval.*		*pval.*
CMR-NHD	Kτ	0.81	(**)	1.00	(**)	0.52	(0.2)	0.81	(*)	0.90	(**)
Gauss	Sρ	0.93	(*)	1.00	(**)	0.68	(0.1)	0.89	(*)	0.96	(**)
(σ = 2.0x)	Pr	0.98	(**)	0.98	(**)	0.65	(0.1)	0.73	(0.1)	0.98	(**)
CMR-NHD	Kτ	0.81	(**)	1.00	(**)	0.43	(0.2)	0.81	(**)	0.90	(**)
Gauss	Sρ	0.93	(**)	1.00	(**)	0.57	(0.2)	0.89	(*)	0.96	(**)
(σ = 2.5x)	Pr	0.99	(**)	0.96	(**)	0.60	(0.2)	0.67	(**)	0.96	(**)
CMR-NHD	Kτ	0.81	(**)	1.00	(**)	0.52	(0.2)	0.81	(**)	0.90	(**)
Gauss	Sρ	0.93	(**)	1.00	(**)	0.68	(0.1)	0.89	(*)	0.96	(**)
(σ = 3.0x)	Pr	0.99	(**)	0.94	(**)	0.50	(0.3)	0.64	(0.1)	0.94	(**)
CMR-NHD	Kτ	0.81	(*)	0.71	(*)	0.43	(0.3)	0.62	(0.1)	0.81	(**)
Gamma	Sρ	0.93	(**)	0.86	(*)	0.49	(0.1)	0.75	(0.1)	0.89	(**)
(γ = 0.2)	Pr	0.99	(**)	0.67	(0.1)	0.42	(0.2)	0.64	(0.1)	0.72	(0.1)
CMR-NHD	Kτ	0.90	(**)	0.81	(*)	0.90	(**)	0.81	(*)	0.81	(**)
Gamma	Sρ	0.96	(**)	0.93	(**)	0.96	(**)	0.89	(**)	0.89	(**)
(γ = 0.5)	Pr	0.96	(**)	0.94	(**)	0.79	(*)	0.97	(**)	0.94	(**)
CMR-NHD	Kτ	0.90	(*)	0.90	(**)	0.81	(**)	0.90	(**)	1.00	(**)
Gamma	Sρ	0.96	(**)	0.96	(**)	0.89	(**)	0.96	(**)	1.00	(**)
(γ = 2.0)	Pr	0.98	(**)	0.95	(**)	0.80	(*)	0.98	(**)	0.96	(**)
CMR-NHD	Kτ	0.90	(**)	0.71	(*)	0.81	(**)	0.43	(0.2)	0.81	(**)
DropOut	Sρ	0.96	(**)	0.86	(*)	0.89	(*)	0.43	(0.4)	0.89	(**)
(p_d = 0.3)	Pr	0.86	(**)	0.95	(**)	0.83	(*)	0.77	(*)	0.95	(**)
CMR-NHD	Kτ	0.87	(*)	0.75	(0.1)	0.87	(**)	0.83	(*)	0.73	(0.1)
DropOut	Sρ	0.94	(**)	0.77	(0.1)	0.94	(*)	0.94	(*)	0.83	(0.1)
(p_d = 0.5)	Pr	0.80	(0.1)	0.94	(**)	0.96	(**)	0.90	(**)	0.93	(**)
CMR-NHD	Kτ	0.62	(0.1)	0.43	(0.2)	0.81	(*)	0.44	(0.2)	0.53	(0.3)
DropOut	Sρ	0.71	(0.1)	0.50	(0.3)	0.93	(**)	0.43	(0.3)	0.49	(0.2)
(p_d = 0.7)	Pr	0.69	(0.1)	0.85	(*)	0.86	(**)	0.57	(0.2)	0.71	(*)

11.5 Post-UDA Perturbation Sweep

In this section we investigate the performance of the CMR metric for predicting the model target performance ranking after the application of two unsupervised domain adaptation (UDA) approaches, Mean Teacher [81] (Tab. 16a) and adaptive batch normalisation (AdaBN) [48] (Tab. 16b). The tables show a sweep over a range of perturbation strengths and report the correlation scores per target dataset. In the main paper due to space limitations we reported the correlation scores averaged over the target datasets for a single perturbation strength. We investigated applying additive Gaussian noise to the input of the models with strengths varying in the range $\sigma = 0.01 - 0.2$.

Table 16: Per-dataset post-UDA CMR-EI Correlation scores to F1-score ranking for Semantic Segmentation of Mitochondria across a range of Gaussian Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 12$. σ is randomly sampled from the noted range. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) Mean Teacher

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-EI	Kτ	0.71	(**)	0.64	(**)	0.56	(*)	0.15	(0.57)
Gauss, σ	Sρ	0.85	(**)	0.78	(*)	0.75	(*)	0.22	(0.53)
[0.01, 0.03]	Pr	0.93	(**)	0.71	(*)	0.52	(0.10)	0.11	(0.73)
CMR-EI	Kτ	0.82	(**)	0.67	(**)	0.60	(*)	0.09	(0.71)
Gauss, σ	Sρ	0.90	(**)	0.79	(**)	0.78	(*)	0.15	(0.65)
[0.03, 0.05]	Pr	0.95	(**)	0.75	(**)	0.59	(0.05)	0.16	(0.61)
CMR-EI	Kτ	0.82	(**)	0.71	(**)	0.64	(**)	0.09	(0.75)
Gauss, σ	Sρ	0.90	(**)	0.81	(**)	0.80	(*)	0.17	(0.61)
[0.05, 0.07]	Pr	0.96	(**)	0.76	(**)	0.64	(*)	0.19	(0.55)
CMR-EI	Kτ	0.78	(**)	0.67	(**)	0.67	(**)	0.12	(0.65)
Gauss, σ	Sρ	0.89	(**)	0.78	(*)	0.85	(**)	0.16	(0.63)
[0.07, 0.1]	Pr	0.96	(**)	0.77	(**)	0.67	(*)	0.22	(0.49)
CMR-EI	Kτ	0.82	(**)	0.75	(**)	0.64	(*)	0.12	(0.60)
Gauss, σ	Sρ	0.92	(**)	0.83	(**)	0.81	(**)	0.16	(0.65)
[0.1, 0.12]	Pr	0.96	(**)	0.77	(**)	0.69	(*)	0.27	(0.39)
CMR-EI	Kτ	0.78	(**)	0.75	(**)	0.64	(*)	0.21	(0.38)
Gauss, σ	Sρ	0.91	(**)	0.83	(**)	0.81	(**)	0.24	(0.45)
[0.12, 0.15]	Pr	0.93	(**)	0.76	(**)	0.70	(*)	0.28	(0.38)
CMR-EI	Kτ	0.67	(**)	0.71	(**)	0.64	(*)	0.18	(0.45)
Gauss, σ	Sρ	0.85	(**)	0.82	(**)	0.81	(**)	0.22	(0.47)
[0.15, 0.2]	Pr	0.89	(**)	0.72	(*)	0.71	(*)	0.28	(0.37)

(b) AdaBN

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CMR-EI	Kτ	0.67	(**)	0.73	(**)	0.55	(*)	0.33	(0.30)
Gauss, σ	Sρ	0.80	(**)	0.85	(**)	0.71	(*)	0.45	(0.25)
[0.01, 0.03]	Pr	0.93	(**)	0.93	(**)	0.92	(**)	0.57	(0.11)
CMR-EI	Kτ	0.70	(**)	0.73	(**)	0.61	(**)	0.33	(0.21)
Gauss, σ	Sρ	0.87	(**)	0.88	(**)	0.76	(*)	0.42	(0.29)
[0.03, 0.05]	Pr	0.89	(**)	0.92	(**)	0.89	(**)	0.62	(0.07)
CMR-EI	Kτ	0.73	(**)	0.73	(**)	0.58	(*)	0.33	(0.25)
Gauss, σ	Sρ	0.88	(**)	0.88	(**)	0.74	(**)	0.42	(0.26)
[0.05, 0.07]	Pr	0.85	(**)	0.91	(**)	0.88	(**)	0.71	(*)
CMR-EI	Kτ	0.70	(**)	0.73	(**)	0.55	(*)	0.39	(0.16)
Gauss, σ	Sρ	0.86	(**)	0.88	(**)	0.73	(*)	0.47	(0.21)
[0.07, 0.1]	Pr	0.76	(**)	0.91	(**)	0.87	(**)	0.71	(*)
CMR-EI	Kτ	0.67	(**)	0.73	(**)	0.58	(**)	0.33	(0.28)
Gauss, σ	Sρ	0.83	(**)	0.88	(**)	0.76	(**)	0.42	(0.27)
[0.1, 0.12]	Pr	0.59	(*)	0.90	(**)	0.86	(**)	0.70	(*)
CMR-EI	Kτ	0.67	(**)	0.73	(**)	0.61	(**)	0.39	(0.16)
Gauss, σ	Sρ	0.81	(**)	0.90	(**)	0.77	(**)	0.52	(0.15)
[0.12, 0.15]	Pr	0.44	(0.15)	0.89	(**)	0.85	(**)	0.71	(*)
CMR-EI	Kτ	0.64	(**)	0.67	(**)	0.58	(*)	0.39	(0.17)
Gauss, σ	Sρ	0.76	(*)	0.85	(**)	0.76	(**)	0.52	(0.15)
[0.15, 0.2]	Pr	0.38	(0.22)	0.86	(**)	0.82	(**)	0.74	(*)

Tab. 16 shows that the correlation score of our CMR approach remains very stable over a wide range of perturbation strengths. As noted in the main paper CMR ranking strongly correlates to F1-score ranking across three out of the four Mitochondria datasets, but struggles with the VNC dataset (but still beats the Transfer Score baseline see Tab. 3). We wanted to investigate this apparent underperformance on VNC further.

Firstly, it is important to note that VNC is by far the smallest dataset thus reducing its statistical robustness. The entire VNC dataset only 20 slices of 589×589 pixels. As noted in Sec. 7.1, due to the small dataset size in direct zero-shot application we purposely disregarded the source-to-source VNC model

transfer and instead used the full 20 slices as test data for models trained on other sources. However, in post-UDA ranking analysis this is no longer possible as we must split the data into train/test splits so that models can be adapted on the target training set and then tested on the test set. Hence, for the post-UDA ranking the models transferred to VNC are assessed on only 2 589×589 slices.

However, when we look into the errors made by the CMR ranking for post-UDA transfer to the VNC dataset we see that models trained on the EPFL dataset seem to be outliers with unexpectedly high consistency scores despite low performance. When investigating this further it seemed there was class confusion, see Fig. 5, in the predictions of the EPFL models, where 'none mitochondria' objects are consistently but erroneously segmented by the transferred EPFL source models. This results in low performance F1-score, but high consistency score.

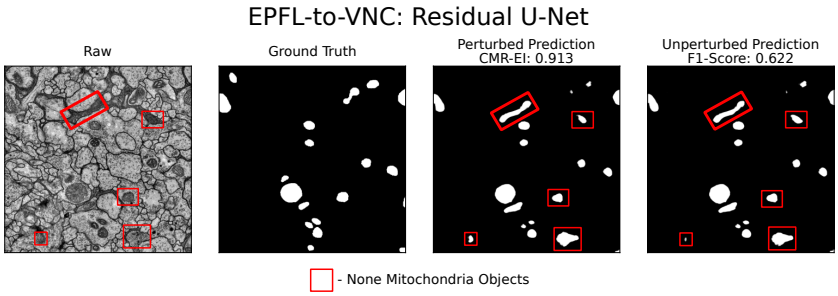


Fig. 5: EPFL-to-VNC Mean Teacher UDA Residual U-Net class confusion.

As noted in the main paper consistency regularisation in Mean Teacher training likely reinforces this class confusion between visually similar, but distinct structures present in EPFL and VNC, violating the assumption that source and target models address the same task. Hence, it is not surprising that CMR ranking after AdaBN UDA, which lacks a consistency regularisation component, performs better on the VNC target dataset, as shown in Tab. 16b.

11.6 Transferability Metrics: Final-Finetuned Correlation

For the supervised baseline methods compared against [8,36,47,61,66,91,99,100] the goal of the transferability metrics is to predict the final supervised finetuned transfer performance of a model based on its initial pre-trained state. In the main paper we investigate the real world scenario of unsupervised model transfer, hence we are limited to assessing direct zero-shot application of pre-trained models to target data or assessing models after application of UDA approaches. We showed already in Tab. 1 that, even with supervision, existing transfer metrics fail at direct zero-shot target performance ranking in the semantic setting. How about in the setting of predicting the final supervised finetuned performance of models?

Table 17: Transferability Metric correlation scores to final-finetuned F1-score performance ranking, Semantic Segmentation of Mitochondria with a range of Gaussian Input Perturbation Strengths. $|\mathcal{M}| = 11$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		EPFL	*pval.*	Hmito	*pval.*	Rmito	*pval.*	VNC	*pval.*
CCFV	Kτ	-0.16	(0.57)	-0.42	(0.10)	-0.53	(*)	-0.27	(0.25)
	Sρ	-0.24	(0.48)	-0.57	(0.05)	-0.69	(*)	-0.37	(0.24)
	Pr	-0.25	(0.46)	-0.52	(0.10)	-0.50	(0.12)	-0.40	(0.19)
GBC	Kτ	0.31	(0.21)	0.20	(0.47)	0.45	(0.07)	0.24	(0.34)
	Sρ	0.47	(0.14)	0.25	(0.45)	0.48	(0.16)	0.28	(0.36)
	Pr	0.24	(0.47)	0.53	(0.10)	0.41	(0.22)	0.45	(0.14)
LEEP	Kτ	0.56	(*)	-0.09	(0.77)	0.16	(0.55)	0.27	(0.28)
	Sρ	0.72	(*)	-0.04	(0.92)	0.36	(0.31)	0.41	(0.20)
	Pr	0.44	(0.17)	0.10	(0.78)	0.39	(0.23)	0.37	(0.23)
NLEEP	Kτ	0.09	(0.78)	-0.20	(0.41)	0.38	(0.09)	0.03	(0.92)
	Sρ	0.15	(0.66)	-0.21	(0.52)	0.45	(0.17)	-0.03	(0.92)
	Pr	0.11	(0.75)	0.10	(0.76)	0.45	(0.17)	-0.02	(0.95)
Hscore	Kτ	-0.02	(1.00)	0.27	(0.26)	0.45	(0.06)	0.15	(0.49)
	Sρ	0.09	(0.81)	0.42	(0.24)	0.53	(0.12)	0.28	(0.35)
	Pr	0.20	(0.56)	0.21	(0.54)	0.41	(0.20)	0.17	(0.60)
RegHscore	Kτ	-0.02	(1.00)	0.24	(0.37)	0.45	(0.06)	0.15	(0.58)
	Sρ	0.09	(0.78)	0.40	(0.22)	0.53	(0.09)	0.28	(0.40)
	Pr	0.20	(0.56)	0.21	(0.54)	0.41	(0.21)	0.17	(0.60)
LogME	Kτ	-0.13	(0.59)	-0.02	(0.98)	0.35	(0.17)	0.00	(1.00)
	Sρ	0.05	(0.84)	-0.01	(0.97)	0.41	(0.22)	0.13	(0.66)
	Pr	0.16	(0.64)	-0.08	(0.82)	0.03	(0.94)	0.29	(0.36)
NCTI	Kτ	-0.16	(0.54)	-0.09	(0.75)	0.16	(0.57)	0.00	(1.00)
	Sρ	-0.21	(0.54)	-0.24	(0.51)	0.26	(0.45)	-0.01	(0.99)
	Pr	0.08	(0.82)	-0.24	(0.49)	0.37	(0.26)	0.04	(0.89)

To investigate this we further trained a set of source models on the supervised training set of each of the mitochondria datasets in turn and then used the available ground truth to evaluate the performance of the finetuned models on the test set. Then using the pre-finetuning weights of the models we calculated all the base line transferability metrics and measured the correlation scores between pre-finetuned transferability metric and post-finetuned performance score. Tab. 17 shows the correlation scores split by target dataset, we can see that all the metrics fail to meaningfully correlate with final-finetuned performance. Further highlighting that existing transferability metrics cannot be reliably used for predicting the transfer of semantic segmentation models applied to biomedical data.

12 Instance Segmentation

In the following section we investigate ranking of instance segmentation models, each of the tables Tabs. 18, 19, 20a to 20d, 21 and 22 shows the correlation scores between our CMR-ARS metric and instance segmentation performance metrics for sets of models applied directly to a single target dataset, across a range of perturbation strengths. In the main paper, due to space constraints, we show the correlation scores for a single perturbation strength averaged over the set of target datasets. We investigate the dependency of CMR on the strength of perturbation applied for both input and feature space perturbations. The tables show that CMR's performance is stable across a wide range of perturbations.

12.1 Cells Perturbation Sweep

For cell instance segmentation we investigated ranking a diverse set of domain specialist models with $|\mathcal{M}| = 8$. The models were trained on three different source datasets FlyWing [26], Ovules [96] and PNAS [95] and include 2D U-Nets (trained both with and without augmentations) and Residual 2D U-Nets.

We investigated applying additive Gaussian noise as an input perturbation between the strengths $\sigma = 0.01 - 0.2$. We show the results for CMR-ARS in Tab. 18.

For feature space perturbation we investigated applying spatial Test-time DropOut (TTD) equally to all layers of a network, where the strength of the perturbation is controlled by the proportion of feature maps dropped at each layer p_d. We investigated DropOut proportions in the range $p_d = 0.001 - 0.1$. We show results for for both CMR-ARS in Tab. 19.

Table 18: Per-dataset CMR-ARS correlation scores to mAP@[0.5:0.95] ranking for Instance Segmentation of Cells across a range of Gaussian Input Perturbation Strengths. $|\mathcal{M}| = 8$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		FlyWing	*pval.*	Ovules	*pval.*	PNAS	*pval.*
CMR-ARS	Kτ	0.64	(*)	0.43	(0.16)	0.93	(**)
Gauss, σ	Sρ	0.79	(*)	0.60	(0.16)	0.98	(**)
[0.01, 0.05]	Pr	0.79	(*)	0.95	(**)	0.83	(*)
CMR-ARS	Kτ	0.57	(0.07)	0.43	(0.21)	0.71	(*)
Gauss, σ	Sρ	0.76	(*)	0.57	(0.15)	0.83	(*)
[0.05, 0.1]	Pr	0.91	(**)	0.87	(**)	0.91	(**)
CMR-ARS	Kτ	0.57	(0.07)	0.64	(*)	0.57	(0.07)
Gauss, σ	Sρ	0.76	(*)	0.79	(*)	0.76	(*)
[0.1, 0.15]	Pr	0.94	(**)	0.88	(**)	0.89	(**)
CMR-ARS	Kτ	0.50	(0.11)	0.57	(0.06)	0.57	(0.06)
Gauss, σ	Sρ	0.74	(0.05)	0.74	(0.05)	0.76	(*)
[0.15, 0.2]	Pr	0.94	(**)	0.85	(*)	0.90	(**)
CMR-ARS	Kτ	0.57	(0.07)	0.64	(*)	0.86	(**)
Gauss, σ	Sρ	0.76	(*)	0.79	(*)	0.93	(**)
[0.2, 0.25]	Pr	0.94	(**)	0.88	(**)	0.87	(**)

12.2 Nuclei Perturbation Sweep

For nuclei instance segmentation we trained a set of domain specialist models with $|\mathcal{M}| = 5$. The models all have a 2D U-Net architecture and were trained on five datasets; BBBC039 [52], HeLaCytoNuc [19], Hoechst [5], S_BIAD895/SB-895 [16] and SBIAD1410 [31]. The source models were then transferred to the test sets of four target datasets BBBC039 [52], Hoechst [5], S_BIAD895 [16] and S_BIAD634/SB-634 [43].

We investigated applying a range of input augmentations; additive Gaussian noise, Gamma correction, Brightness adjustment and Contrast adjustment (as defined in Sec. 9). We investigated additive Gaussian noise with strengths in the range $\sigma = 0.0 - 0.2$, Gamma correction with strengths in the range $\gamma = 0.8 - 1.2$, where $\gamma = 1.0$ equals no adjustment, Brightness adjustment with strengths in

Table 19: CMR-ARS Correlation scores for Instance Segmentation of Cells across a range of DropOut Perturbation Strengths. $|\mathcal{M}| = 8$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		FlyWing	*pval.*	Ovules	*pval.*	PNAS	*pval.*
CMR-ARS	Kτ	0.64	(*)	0.50	(0.11)	0.71	(*)
DropOut	Sρ	0.81	(*)	0.74	(0.06)	0.83	(*)
(p_d = 0.001)	Pr	0.93	(**)	0.82	(*)	0.81	(*)
CMR-ARS	Kτ	0.71	(*)	0.50	(0.14)	0.71	(*)
DropOut	Sρ	0.83	(*)	0.74	(*)	0.83	(*)
(p_d = 0.005)	Pr	0.93	(**)	0.83	(*)	0.86	(*)
CMR-ARS	Kτ	0.57	(0.07)	0.50	(0.11)	0.71	(*)
DropOut	Sρ	0.76	(*)	0.74	(0.06)	0.83	(*)
(p_d = 0.01)	Pr	0.94	(**)	0.84	(*)	0.88	(**)
CMR-ARS	Kτ	0.64	(*)	0.64	(*)	0.79	(*)
DropOut	Sρ	0.79	(*)	0.79	(*)	0.90	(*)
(p_d = 0.02)	Pr	0.94	(**)	0.87	(**)	0.92	(**)
CMR-ARS	Kτ	0.50	(0.11)	0.79	(*)	0.64	(*)
DropOut	Sρ	0.74	(0.07)	0.90	(*)	0.81	(*)
(p_d = 0.03)	Pr	0.95	(**)	0.89	(**)	0.91	(**)
CMR-ARS	Kτ	0.50	(0.12)	0.57	(0.07)	0.64	(*)
DropOut	Sρ	0.74	(*)	0.76	(*)	0.81	(*)
(p_d = 0.04)	Pr	0.95	(**)	0.79	(*)	0.86	(*)
CMR-ARS	Kτ	0.50	(0.10)	0.57	(0.05)	0.71	(*)
DropOut	Sρ	0.74	(0.06)	0.71	(0.05)	0.83	(*)
(p_d = 0.05)	Pr	0.95	(**)	0.83	(*)	0.88	(**)
CMR-ARS	Kτ	0.57	(0.06)	0.43	(0.17)	0.57	(*)
DropOut	Sρ	0.76	(*)	0.60	(0.14)	0.71	(0.07)
(p_d = 0.1)	Pr	0.95	(**)	0.78	(*)	0.88	(**)

the range $\theta_B = 0.0 - 0.2$ and Contrast adjustment with strengths in the range $\theta_C = 0.8 - 1.2$, where $\theta_C = 1.0$ equals no adjustment. We show the results for both CMR-ARS (Tabs. 20a to 20d)

For feature space perturbation we investigated applying spatial DropOut only to the bottleneck layer of networks, where the strength of the perturbation is controlled by the proportion of feature maps dropped at each layer p_d. We investigated DropOut proportions in the range $p_d = 0.05 - 0.5$. We show results for CMR-ARS in Tab. 21.

We again observe stable correlation performance across a wide range of perturbation strengths. Although, as discussed in the main paper, Sec. 3.3, care should be taken to ensure perturbations remain 'tolerable' in the extreme perturbation case, too strong of a perturbation can lead to a loss of correlation between consistency based metrics and transfer performance. For example, the last row of Tab. 21 shows the strongest level of DropOut ($p_d = 0.5$) applied can cause the CMR-ARS metric to become unstable on S_BIAD634.

Table 20: CMR-ARS Correlation scores to mAP@[0.5:0.95] ranking for Instance Segmentation of Nuclei across a range of Input Perturbation Strengths. For each ranking experiment $|\mathcal{M}| = 5$. Perturbation strength are controlled by sampling σ, θ_B, θ_C and γ from the noted ranges. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

(a) Gaussian Input Perturbation

Metric		BBBC039	*pval.*	Hoechst	*pval.*	SB-895	*pval.*	SB-634	*pval.*
CMR-ARS	Kτ	0.80	(0.08)	0.80	(0.11)	0.33	(0.75)	0.80	(0.10)
Gauss, σ	Sρ	0.90	(0.10)	0.90	(0.08)	0.60	(0.42)	0.90	(0.08)
[0.0, 0.05]	Pr	0.99	(**)	0.85	(0.07)	0.47	(0.53)	0.93	(*)
CMR-ARS	Kτ	0.20	(0.80)	0.80	(0.09)	0.33	(0.75)	1.00	(*)
Gauss, σ	Sρ	0.50	(0.42)	0.90	(0.08)	0.40	(0.75)	1.00	(*)
[0.05, 0.1]	Pr	0.72	(0.17)	0.89	(*)	0.99	(*)	0.80	(0.10)
CMR-ARS	Kτ	0.40	(0.54)	1.00	(*)	0.67	(0.33)	0.80	(0.09)
Gauss, σ	Sρ	0.60	(0.36)	1.00	(*)	0.80	(0.33)	0.90	(0.08)
[0.1, 0.2]	Pr	0.50	(0.39)	0.93	(*)	0.94	(0.06)	0.79	(0.11)

(b) Brightness Input Perturbation

Metric		BBBC039	*pval.*	Hoechst	*pval.*	SB-895	*pval.*	SB-634	*pval.*
CMR-ARS	Kτ	0.60	(0.22)	0.60	(0.26)	0.00	(1.00)	0.80	(0.10)
Brt, θ_B	Sρ	0.70	(0.25)	0.70	(0.21)	0.20	(0.92)	0.90	(0.09)
[0.0, 0.05]	Pr	0.80	(0.10)	0.83	(0.08)	0.77	(0.23)	0.72	(0.17)
CMR-ARS	Kτ	0.40	(0.46)	0.80	(0.11)	0.33	(0.75)	0.80	(0.07)
Brt, θ_B	Sρ	0.60	(0.32)	0.90	(0.10)	0.40	(0.75)	0.90	(0.10)
[0.05, 0.1]	Pr	0.69	(0.19)	0.92	(*)	0.80	(0.20)	0.84	(0.08)
CMR-ARS	Kτ	0.60	(0.23)	0.80	(0.09)	0.67	(0.33)	0.40	(0.49)
Brt, θ_B	Sρ	0.70	(0.23)	0.90	(0.10)	0.80	(0.33)	0.50	(0.40)
[0.1, 0.2]	Pr	0.81	(0.10)	0.85	(0.07)	0.99	(*)	0.43	(0.47)

(c) Contrast Input Perturbation

Metric		BBBC039	*pval.*	Hoechst	*pval.*	SB-895	*pval.*	SB-634	*pval.*
CMR-ARS	Kτ	0.80	(0.09)	1.00	(*)	0.00	(1.00)	0.80	(0.08)
Ctr, θ_C	Sρ	0.90	(0.09)	1.00	(*)	-0.20	(0.92)	0.90	(0.07)
[0.8, 0.9]	Pr	0.85	(0.07)	0.94	(*)	0.18	(0.82)	0.83	(0.08)
CMR-ARS	Kτ	0.80	(0.09)	1.00	(*)	0.67	(0.33)	0.80	(0.10)
Ctr, θ_C	Sρ	0.90	(0.09)	1.00	(*)	0.80	(0.33)	0.90	(0.09)
[0.9, 0.95]	Pr	0.96	(*)	0.99	(**)	0.52	(0.48)	0.87	(0.05)
CMR-ARS	Kτ	0.74	(0.13)	1.00	(*)	0.67	(0.33)	0.60	(0.25)
Ctr, θ_C	Sρ	0.82	(0.13)	1.00	(*)	0.80	(0.33)	0.80	(0.13)
[1.05, 1.1]	Pr	0.99	(**)	0.95	(*)	0.58	(0.42)	0.80	(0.10)
CMR-ARS	Kτ	0.80	(0.11)	0.60	(0.23)	0.33	(0.75)	0.80	(0.07)
Ctr, θ_C	Sρ	0.90	(0.10)	0.70	(0.20)	0.40	(0.75)	0.90	(0.09)
[1.1, 1.2]	Pr	0.83	(0.08)	0.81	(0.10)	0.38	(0.62)	0.90	(*)

(d) Gamma Input Perturbation

Metric		BBBC039	*pval.*	Hoechst	*pval.*	SB-895	*pval.*	SB-634	*pval.*
CMR-ARS	Kτ	0.40	(0.42)	0.80	(0.08)	-0.33	(0.75)	0.40	(0.49)
Gamma, γ	Sρ	0.60	(0.37)	0.90	(0.10)	-0.40	(0.75)	0.60	(0.36)
[0.8, 0.9]	Pr	0.62	(0.26)	0.82	(0.09)	0.06	(0.94)	0.64	(0.25)
CMR-ARS	Kτ	0.40	(0.50)	0.80	(0.08)	0.33	(0.75)	0.80	(0.08)
Gamma, γ	Sρ	0.60	(0.37)	0.90	(0.10)	0.40	(0.75)	0.90	(0.07)
[0.9, 0.95]	Pr	0.54	(0.35)	0.89	(*)	0.58	(0.42)	0.76	(0.14)
CMR-ARS	Kτ	0.60	(0.22)	0.80	(0.09)	0.00	(1.00)	0.80	(0.08)
Gamma, γ	Sρ	0.70	(0.22)	0.90	(0.08)	-0.20	(0.92)	0.90	(0.09)
[1.05, 1.1]	Pr	0.94	(*)	0.84	(0.07)	0.44	(0.56)	0.82	(0.09)
CMR-ARS	Kτ	0.40	(0.53)	1.00	(*)	0.33	(0.75)	0.40	(0.45)
Gamma, γ	Sρ	0.60	(0.37)	1.00	(*)	0.40	(0.75)	0.60	(0.34)
[1.1, 1.2]	Pr	0.71	(0.18)	0.87	(0.05)	0.68	(0.32)	0.76	(0.14)

Table 21: CMR-ARS Correlation scores to mAP@[0.5:0.95] ranking for Instance Segmentation of Nuclei across a range of DropOut Perturbation Strengths, controlled by p_d. For each ranking experiment $|\mathcal{M}| = 5$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		BBBC039	*pval.*	Hoechst	*pval.*	S_BIAD895	*pval.*	S_BIAD634	*pval.*
CMR-ARS	Kτ	0.40	(0.43)	0.80	(0.08)	1.00	(0.08)	0.80	(0.12)
DropOut	Sρ	0.60	(0.32)	0.90	(0.08)	1.00	(0.08)	0.90	(0.08)
(p_d = 0.05)	Pr	0.83	(0.08)	0.90	(*)	0.93	(0.07)	0.88	(0.05)
CMR-ARS	Kτ	0.40	(0.51)	1.00	(*)	1.00	(0.08)	0.80	(0.08)
DropOut	Sρ	0.50	(0.44)	1.00	(*)	1.00	(0.08)	0.90	(0.11)
(p_d = 0.1)	Pr	0.87	(0.05)	0.86	(0.06)	0.93	(0.07)	0.91	(*)
CMR-ARS	Kτ	0.40	(0.54)	1.00	(*)	1.00	(0.08)	0.80	(0.09)
DropOut	Sρ	0.50	(0.45)	1.00	(*)	1.00	(0.08)	0.90	(0.07)
(p_d = 0.2)	Pr	0.95	(*)	0.93	(*)	0.72	(0.28)	0.78	(0.12)
CMR-ARS	Kτ	0.60	(0.25)	1.00	(*)	1.00	(0.08)	0.80	(0.08)
DropOut	Sρ	0.70	(0.23)	1.00	(*)	1.00	(0.08)	0.90	(0.10)
(p_d = 0.3)	Pr	0.87	(0.05)	0.89	(*)	0.76	(0.24)	0.82	(0.09)
CMR-ARS	Kτ	0.60	(0.22)	1.00	(*)	0.67	(0.33)	1.00	(*)
DropOut	Sρ	0.70	(0.24)	1.00	(*)	0.80	(0.33)	1.00	(*)
(p_d = 0.4)	Pr	0.95	(*)	0.91	(*)	0.91	(0.09)	0.96	(*)
CMR-ARS	Kτ	0.60	(0.22)	0.80	(0.10)	0.67	(0.33)	0.00	(1.00)
DropOut	Sρ	0.70	(0.25)	0.90	(0.07)	0.80	(0.33)	0.00	(1.00)
(p_d = 0.5)	Pr	0.98	(**)	0.80	(0.10)	0.89	(0.11)	0.49	(0.40)

12.3 SEG Hyperparameters

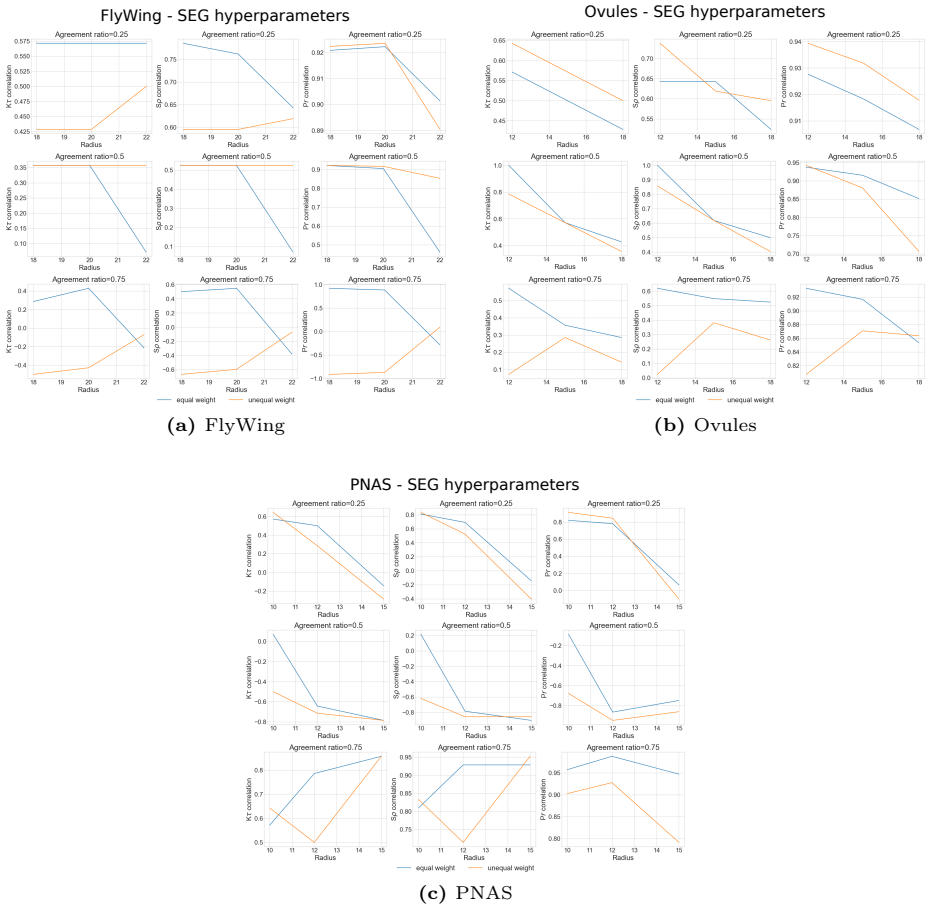
As described in Sec. 2 the SEG [78] method is the only other approach that aims to address ranking instance segmentation models. SEG has two hyperparameters: the agreement ratio a_r defining the ensemble consensus for pseudo ground truth and the centroid dilation radius r. In the main paper results Tab. 4 we followed the recommendation of the paper by Sims et al. [78] and set $a_r = 0.75$ and estimated r as half the mean instance size from a small random sample. However, for further analysis of the method we also investigated alternative hyperparameter settings.

Fig. 6 and Fig. 7 show the resulting correlation scores over sweeps of SEG hyperparameters. In each subplot we tested agreement ratios of 0.25, 0.5 and 0.75 and sweeped over a range of radii, r, spanning from below half the mean instance size to above half the mean instance size for each dataset (found using ground truth labels). Each subplot reports the corresponding Kendall Tau (K_τ), Spearman's Rho (S_ρ) and Pearson r (P_r) correlation coefficients. The SEG paper is based on an ensemble approach and discusses whether all the model votes should be equally weighted or if an adaptive weighting should be applied to automatically define the importance of each model. For completeness, we report both the equal weighted ensemble (blue line) and the adaptive weighted ensemble (orange line).

We found that SEG was very unstable and the model ranking correlation scores sensitive to hyperparameter selection. Furthermore, there is no suitable unsupervised approach for tuning these hyperparameters. Without labels estimating the appropriate radius to select is challenging and a variation of 2 pixels can substantially effect the correlation score. The same is true for agreement ratio, without knowing beforehand how many models in the candidate set are performing well (which we can't know before ranking) setting the ratio becomes very challenging. Hence, we found that selection of suitable hyperparameters for the SEG method was a major hurdle to its practical utility.

12.4 Covid-IF External Models Perturbation Sweep

In the main paper in Fig. 4 we also ranked the predicted target performance of a set of publicly available models [3,65,67] for instance cell segmentation on the Covid-IF [67] dataset. The models all have very different instantiation processes, but through calculating the ranking metric in output segmentation space we are able to directly compare the varied set of models. The Micro-SAM models [3] perform Automatic Instance Segmentation (AIS), which utilises an additionally trained decoder to predict a three channel output: the distance to the object centre, the distance to the object boundary and foreground probabilities. The predictions are then instantiated using seeded watershed. Cellpose-SAM [65] use a custom decoder to directly predict vector flow fields, which can then be converted to instance segmentations. 'Powerful-chipmunk' is a specialist U-Net model taken from the Bioimage Model Zoo [63], the model is trained specifically



(a) FlyWing **(b)** Ovules

(c) PNAS

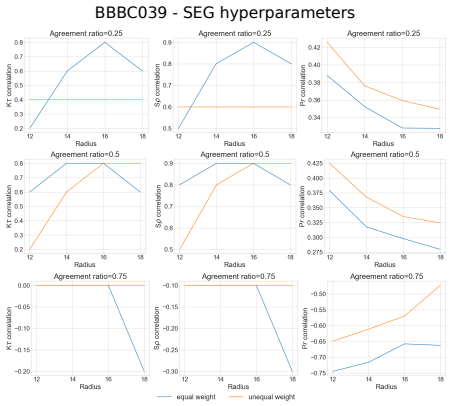
Fig. 6: Sweep of SEG [78] hyperparameters for ranking Cell Instance segmentation models.

on the Covid-IF data and predicts object foreground and boundary predictions, which are then converted to instances using watershed and GASP.

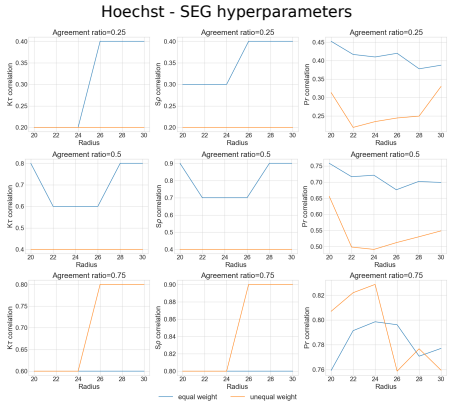
In Fig. 4 from the main paper we show the correlation between CMR-ARS and mSA performance for a single additive Gaussian noise perturbation strength. In Tab. 22 we investigate a range of Gaussian noise strengths $\sigma = 0.01 - 0.05$.

12.5 Generalist Models: Cells vs Nuclei Prediction

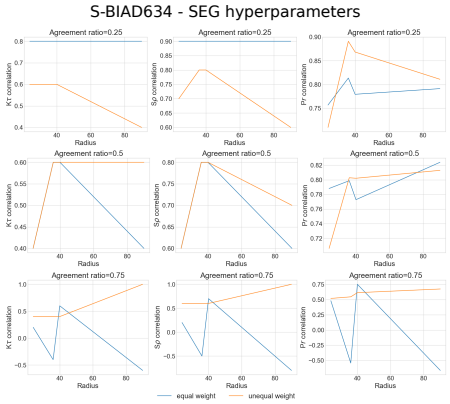
A key assumption of the proposed consistency based approach is that the source and target tasks align, i.e., if a model was trained for nuclei segmentation then it should only be evaluated for nuclei segmentation. For generalist models like μSAM [3] and Cellpose-SAM [65], this assumption may not fully hold in all cases. The models are finetuned for both cell and nuclei and have no automated



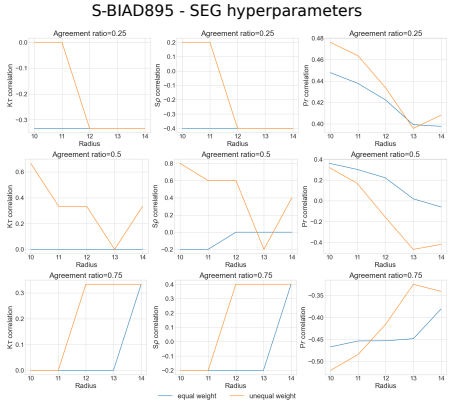
(a) BBBC039



(b) Hoechst



(c) S-BIAD634



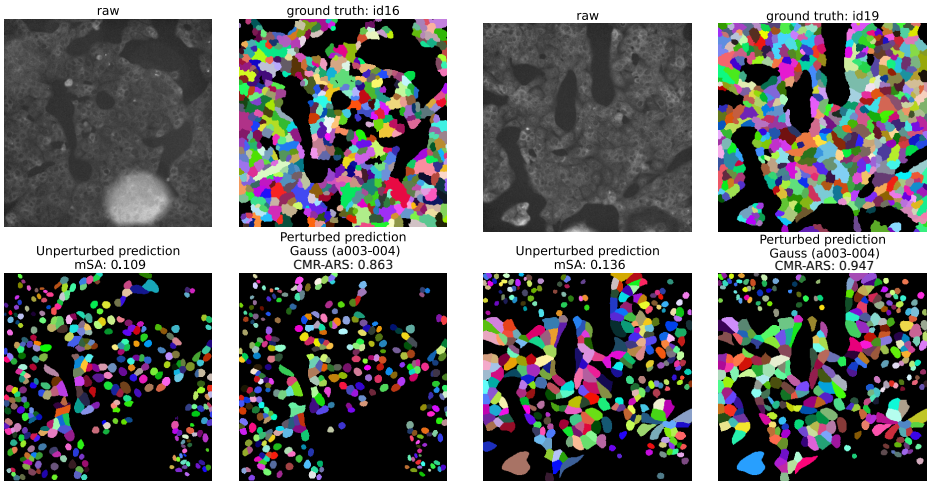
(d) S-BIAD895

Fig. 7: Sweep of SEG [78] hyperparameters for ranking Nuclei Instance segmentation models.

mechanism to control which structure is segmented. In datasets like Covid-IF, a channel can implicitly contain information about both classes. We observed that Cellpose-SAM and μSAM occasionally switch to segmenting nuclei instead of cells (Figs. 8a and 8b) particularly in 'challenging' sections of the image. It is hard to quantify is this always due to a 'mode' switch to nuclei segmentation or just poor cell segmentation performance, however it seems clear that at least in some cases the models segment small objects (much smaller than cells) that clearly correspond to the nuclei information present in the serum channel. This 'mode' switch reduces cell segmentation performance, but not consistency scores, as predictions remain internally stable. However, despite some examples of 'off-task' segmentation in general Cellpose-SAM and μSAM are strong performing cell segmentation models and in the majority of cases are able to stay on task given a good raw input signal. Hence, we can see in Fig. 4 and Tab. 22 that

Table 22: CMR-ARS Correlation scores to mSA for Instance Segmentation of Covid-IF across a range of Gaussian Input Perturbation Strengths, controlled by σ sampled from the noted range. For each ranking experiment $|\mathcal{M}| = 5$. *pval.* < 0.05 (*), *pval.* < 0.01 (**).

Metric		Covid-IF	*pval.*
CMR-ARS	Kτ	0.80	(0.09)
Gauss, σ	Sρ	0.90	(0.08)
$[0.01, 0.02]$	Pr	0.86	(0.06)
CMR-ARS	Kτ	1.00	(*)
Gauss, σ	Sρ	1.00	(*)
$[0.02, 0.03]$	Pr	0.93	(*)
CMR-ARS	Kτ	1.00	(*)
Gauss, σ	Sρ	1.00	(*)
$[0.03, 0.04]$	Pr	0.91	(*)
CMR-ARS	Kτ	1.00	(*)
Gauss, σ	Sρ	1.00	(*)
$[0.04, 0.05]$	Pr	0.93	(*)



(a) ID 16 **(b)** ID 19

Fig. 8: Cells vs Nuclei Prediction of Cellpose-SAM.

our CMR based ranking of these large generalist models and specialist BMZ model [63] strongly correlates with the true target performance ranking.

An important note is that the ranking in Fig. 4 does not show the general superiority of any one model, but is a target dataset specific ranking of automatic segmentation performance. One factor to keep in mind is that the Covid-IF dataset that we tested ranking on was explicitly not included in the training data of μSAM, while inclusion for Cellpose-SAM is uncertain. However, this represents realistic ranking scenarios where models will vary their performance across different target datasets due to differences in source-to-target distribution and model architecture. Hence the need for target specific model ranking as provided by our method.