Tree-based Semantic Losses: Application to Sparsely-supervised Large Multi-class Hyperspectral Segmentation

Junwen Wang[1], Oscar Maccormac[1,2], William Rochford[1,2], Aaron Kujawa[1], Jonathan Shapey[1,2], and Tom Vercauteren[1]

[1] King's College London, London, UK
junwen.wang@kcl.ac.uk
[2] King's College Hospital, London, UK

Abstract. Hyperspectral imaging (HSI) shows great promise for surgical applications, offering detailed insights into biological tissue differences beyond what the naked eye can perceive. Refined labelling efforts are underway to train vision systems to distinguish large numbers of subtly varying classes. However, commonly used learning methods for biomedical segmentation tasks penalise all errors equivalently and thus fail to exploit any inter-class semantics in the label space. In this work, we introduce two tree-based semantic loss functions which take advantage of a hierarchical organisation of the labels. We further incorporate our losses in a recently proposed approach for training with sparse, background-free annotations. Extensive experiments demonstrate that our proposed method reaches state-of-the-art performance on a sparsely annotated HSI dataset comprising 107 classes organised in a clinically-defined semantic tree structure. Furthermore, our method enables effective detection of out-of-distribution (OOD) pixels without compromising segmentation performance on in-distribution (ID) pixels.

Keywords: Hyperspectral imaging · Semantic segmentation · Out-of-distribution detection · Weakly supervised learning.

1 Introduction

Hyperspectral imaging (HSI) is an emerging optical imaging technique that collects and processes spectral data across a range of wavelengths [25]. By splitting light into multiple narrow bands beyond the capability of human vision, HSI captures additional details that are imperceptible to the naked eye. This technique can provide diagnostic information about tissue properties, enabling objective characterisation of tissues without the use of external contrast agents. Surgical image segmentation is an important application of HSI. The use of deep learning for biomedical segmentation with spectral imaging data is increasing [13], alongside the growing availability of medical HSI datasets [4, 6, 12, 26].

Developing an effective segmentation model for HSI data remains challenging. One reason is that most existing HSI datasets are sparsely annotated in order

to reduce the substantial workload [4, 12, 26]. In the sparse annotation setting, only a few representative areas of foreground objects are labelled. Examples of this labelling strategy are illustrated in the second column of Fig. 3. In a recent study, Wang et al. [28] introduced a medical image segmentation framework that enables pixel-level out-of-distribution (OOD) detection using sparsely annotated data containing only positive classes. Their method uses these sparse annotations to learn robust feature representations, which subsequently facilitate the reliable detection of OOD pixels via conventional OOD detection techniques originally developed for classification tasks. This approach achieves state-of-the-art OOD detection performance while preserving classification accuracy for in-distribution (ID) classes.

Another challenge relates to the granularity at which the data is annotated and that at which segmentation models should operate. Recent works have examined the impact of training models with various levels of labelling granularity – such as pixels, patches, and entire images [9, 24]. Other studies have provided comparative analyses of different pixel-level algorithms for brain tissue differentiation [22]. Underpinning these questions is the drive for a holistic and refined understanding of the surgical scenes. Annotation efforts are ongoing to provide training data across large number of potentially subtly varying classes. Combined with sparse annotations processes, many of these classes may only have small amounts of training samples. It is thus important to take advantage of the semantics of the labels and realise that some types of errors are more acceptable than others. However, no previous work in the field of surgical imaging has leveraged the structure of the label space as a source of information. The importance of label semantics is getting recognised in the general field of computer vision [2, 8, 15]. In the task of brain tumour segmentation, Fidon et al. [7] proposed a variant of the Dice score for multi-class segmentation based on the Wasserstein distance in the probabilistic label space. However, this method was only demonstrated in the context of a small number of labels from the BraTS challenge [23], and it does not generalise to sparse, background-free annotations.

In this work, we propose leveraging prior knowledge of label structures through two tree-based semantic losses for sparsely supervised segmentation. By incorporating a comprehensive label hierarchy curated by domain experts, the model achieves superior performance compared with the standard cross-entropy baseline. Our proposed tree-based hierarchical weighting scheme establishes a new state-of-the-art multi-class hyperspectral image segmentation method. Furthermore, we integrate these losses into the OOD detection framework of [28] to enable OOD detection without compromising performance on ID data.

2 Material and methods

Surgical imaging dataset The data was obtained from patients undergoing microscopic cranial neurosurgery as part of an ethically approved single-centre, prospective clinical observational investigation employing a prototype hyperspectral imaging system (NeuroHSI study: REC reference 22/LO/0046, ClinicalTri-

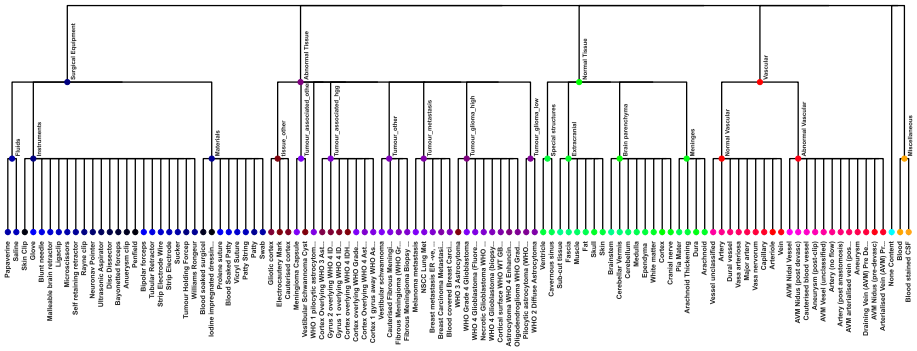

Fig. 1. Full tree-based label hierarchy of the surgical HSI dataset. From top to bottom, the hierarchy progresses from coarse object categories to specific classes. The colour coding matches the ground-truth mask at each level.

als.gov ID NCT05294185). Informed consent was obtained from all participants. The primary objective was to evaluate the intraoperative utility of a 4×4, 16-band visible-range snapshot mosaic camera (IMEC CMV2K-SSM4X4-VIS).

The dataset comprises 22,829 annotated frames derived from 45 distinct surgical cases, encompassing both neuro-oncological and neurovascular pathologies. Multiple videos were acquired for each case, with each recording representing a specific surgical phase intended to capture relevant intraoperative details. Individual phases included varying visual perspectives to account for changes in camera positioning during video acquisition. Training snapshot data are first processed with a demosaicking pipeline [17, 18], resulting in frames comprising $16 \times 1088 \times 2048$ pixels enabling reconstruction of synthetic sRGB visualisation from the HSI data.

The sparse, background-free annotations encompass 107 subclasses organised into a hierarchical structure defined by neurosurgeons. Fig. 1 presents the complete label hierarchy, including four primary categories at the top level of the tree: surgical equipment, abnormal tissue, normal tissue, and vascular structures with corresponding colour references displayed at each node in the label hierarchy. A subset of frames from each video was manually annotated by an experienced neurosurgeon, and where feasible, these annotations were then propagated across subsequent frames algorithmically using the registration-based propagation feature of the ImFusion Labels software. Each propagated annotation was verified and corrected (when needed) by a neurosurgeon prior to final submission.

In addition to human-labelled categories, an additional label was generated by a content area estimation algorithm [3]. This algorithm provides a robust estimation of non-informative areas near the edge of the image. By treating these areas as part of our label hierarchy, the model can more effectively discriminate content regions. Fig. 3 presents example sRGB image-annotation overlays.

Wasserstein distance in label space Let $\mathbf{L}$ be the label space (e.g. as in Fig. 1) with C leaf nodes, where $\mathbf{L} = \{1, 2, ..., C\}$. Let $p, q \in P(\mathbf{L})$ be probability vectors on $\mathbf{L}$. The Wasserstein distance between p and q is the minimal cost to transform p into q given the ground distances $M_{l,l'} \in \mathbb{R}^+$ between any two labels l and l'. The ground distance is represented as a matrix M and the associated Wasserstein distance $W^M(p, q)$ is defined through an optimal transport problem:

$$\begin{aligned} W^M(p,q) &= \min_{T_{l,l'}} \sum_{l,l' \in \mathbf{L}} T_{l,l'} M_{l,l'} \\ \text{subject to } \forall l \in \mathbf{L}, \sum_{l' \in \mathbf{L}} T_{l,l'} = p_l \quad &\text{and} \quad \forall l' \in \mathbf{L}, \sum_{l \in \mathbf{L}} T_{l,l'} = q_{l'} \end{aligned} \tag{1}$$

By leveraging the distance matrix M on $\mathbf{L}$, the Wasserstein distance yields a semantically-meaningful way of comparing two label probability vectors. Given a tree structure as in Fig. 1 with weights associated to the edges, a semantic ground distance can be induced by the path lengths between the leaf nodes. If $q = g$ is a *crisp* ground truth, a closed-form expression of (1) is given in [7]:

$$W^M(p, g) = \sum_{l,l' \in \mathbf{L}} M_{l,l'} p_l g_{l'} = p^T M g \tag{2}$$

Supervised training of segmentation models has been optimised for pixel-level cross-entropy (CE) which does not account for any label semantics:

$$CE(p, g) = -\sum_{l} g_l \log(p_l) \tag{3}$$

Several works have shown benefits in combining CE with task-specific losses [21]. We thus propose a simple combination of cross-entropy and Wasserstein distance which we refer to as *Wasserstein+CE* loss:

$$\mathcal{L}^M_{\text{wce}}(p, g) = \alpha CE + \beta W^M \tag{4}$$

where we set $\alpha = \beta = 0.5$ across all experiments. In the case of $\alpha = 0$ and $\beta = 1$, equation (4) reduces to the vanilla Wasserstein distance $\mathcal{L}^M_w$. In the case of $\alpha = 1$ and $\beta = 0$, equation (4) reduces to standard CE loss $\mathcal{L}_{ce}$.

Tree-based semantic cross-entropy loss We propose a novel semantic loss function which takes advantage of the tree structure, computes the (aggregated) probabilities across the nodes (not just the leaf nodes), and evaluates an extended CE loss on all node probabilities. Let the label tree $\mathcal{T}$ be composed of K levels with 0 corresponding to the deepest level (leaf nodes). Let A be the adjacency matrix associated with $\mathcal{T}$. Let $\tilde{p}$ be a zero-padding of p to initially associate non-leaf nodes with a zero mass, and $p^\dagger$ be the vector collecting all the probabilities:

$$p^\dagger = (\sum_{k \geq 0} A^k)\tilde{p} \tag{5}$$

where $A^k = 0$ for $k > K$. We express the extended weighted CE as:

$$CE^{\mathcal{T}}(p, g) = -\sum_{v} w_v g_v^{\dagger} \log(p_v^{\dagger}) \tag{6}$$

where w_v is the weight of the edge associated with v as a child node. We note that if $w_v = 1$ for all leaf nodes and $w_v = 0$ otherwise, equation (6) reduces to (3). We refer to our semantically-informed variant of the CE loss as *tree-based semantic cross-entropy* loss $\mathcal{L}_{\text{tce}}^{M}$.

Out-of-distribution segmentation We build on the recent work [28] to segment OOD samples from sparse multi-class positive-only annotations. Given a 2D image x, each annotated spatial location (i, j) from x has a corresponding annotation y_{ij}, where $y_{ij} \in \{c\} = \{1, 2, \ldots, C\}$ and C is the number of classes that marked as ID. $c = 0$ is retained to denote OOD pixels. Our proposed framework employs a confidence threshold τ to detect OOD samples at the pixel level:

$$\hat{y}_{ij} = \begin{cases} \arg\max_c \boldsymbol{S}_{ij}^c, & \text{if } \max_c \boldsymbol{S}_{ij}^c > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\boldsymbol{S}_{ij}^c$ is a scoring function that captures the probability of pixel (i, j) belonging to the ID class c, while acknowledging the possibility of it being OOD in the input image. $\boldsymbol{S}_{ij}^c$ can be selected from various OOD detection methods used in image classification tasks [10, 11, 16, 19]. For simplicity, we let $\boldsymbol{S}$ be the network output probabilities aggregated at level k i.e. $\boldsymbol{S}_{ij} = p_k^{\dagger}$. The collected probabilities $p^{\dagger}$ are computed using (5) based on the leaf node's softmax probability $p = \text{softmax}\big(f(x)\big)_{ij}$, where $f(x)$ is the segmentation network with input x, trained with positive-only sparse annotation.

3 Experimental results

Implementation details For baseline (cross-entropy) training, we adopted a similar training pipeline as described in [28]. Specifically, we used a U-Net architecture with an EfficientNet-b5 encoder [27], pre-trained[3] on ImageNet [5] for all experiments. We conducted model training using the softmax output and a one-hot encoded sparse annotation mask for pixels marked as ID. We employed the Adam optimiser [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, together with an exponential learning rate scheme ($\gamma = 0.999$). We set the initial learning rate to 0.0001, used a mini-batch size of 5, and trained for a total of 50 epochs. For a fair comparison, we applied the same hyperparameters across all methods. For data augmentation, we adopted a similar setup to that reported in [28]: random rotation (rotation angle limit: 45°), random flipping, random scaling (scaling factor limit: 0.1), and random shifting (shift factor limit: 0.0625). All transformations

[3] https://github.com/qubvel/segmentation_models.pytorch

Table 1. Cross-validation results on top-level classes. For each method and metric, performance is reported at thresholds $\tau_0 = 0$ and τ_m. τ_m is chosen from the optimal threshold with highest performance in the validation set. The best performance among all losses is highlighted in bold. Rows shaded in grey represent the baseline results, which are equivalent to the standard CE or Wasserstein+CE training on leaf node classes only (i.e. without class semantics). The asterisk "$*$" indicates a strong baseline result which is equivalent to standard CE training on top-level nodes only.

Loss		**TPR** ↑		**BACC** ↑		**F1** ↑	
		$\boldsymbol{\tau_0 = 0}$	$\boldsymbol{\tau_m}$	$\boldsymbol{\tau_0 = 0}$	$\boldsymbol{\tau_m}$	$\boldsymbol{\tau_0 = 0}$	$\boldsymbol{\tau_m}$
$\mathcal{L}_{\mathrm{w}}$	M_t	$0.51_{\pm 0.03}$	$0.51_{\pm 0.03}$	$0.74_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.47_{\pm 0.05}$	$0.47_{\pm 0.05}$
$\mathcal{L}_{\mathrm{tce}}$	M_ℓ	$0.61_{\pm 0.04}$	$0.65_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.82_{\pm 0.02}$	$0.60_{\pm 0.02}$	$0.65_{\pm 0.03}$
	M_e	$0.64_{\pm 0.04}$	$0.76_{\pm 0.03}$	$0.80_{\pm 0.02}$	$0.88_{\pm 0.01}$	$0.62_{\pm 0.05}$	$0.74_{\pm 0.05}$
	$^*M_t$	$0.65_{\pm 0.05}$	$0.73_{\pm 0.12}$	$0.81_{\pm 0.02}$	$0.86_{\pm 0.06}$	$0.63_{\pm 0.04}$	$0.72_{\pm 0.10}$
	M_h	$0.65_{\pm 0.03}$	$0.75_{\pm 0.03}$	$0.81_{\pm 0.02}$	$0.87_{\pm 0.02}$	$0.64_{\pm 0.04}$	$0.76_{\pm 0.05}$
$\mathcal{L}_{\mathrm{wce}}$	M_ℓ	$0.61_{\pm 0.06}$	$0.70_{\pm 0.05}$	$0.79_{\pm 0.03}$	$0.85_{\pm 0.03}$	$0.62_{\pm 0.04}$	$0.72_{\pm 0.04}$
	M_e	$0.65_{\pm 0.04}$	$0.72_{\pm 0.09}$	$0.81_{\pm 0.02}$	$0.85_{\pm 0.04}$	$0.64_{\pm 0.01}$	$0.73_{\pm 0.07}$
	M_t	$0.66_{\pm 0.04}$	$0.77_{\pm 0.07}$	$0.81_{\pm 0.02}$	$0.88_{\pm 0.04}$	$0.63_{\pm 0.03}$	$0.78_{\pm 0.09}$
	M_h	$\mathbf{0.68}_{\pm 0.02}$	$\mathbf{0.80}_{\pm 0.03}$	$\mathbf{0.83}_{\pm 0.01}$	$\mathbf{0.90}_{\pm 0.02}$	$\mathbf{0.66}_{\pm 0.02}$	$\mathbf{0.82}_{\pm 0.06}$

were applied with a probability of 0.5. In addition, we apply ℓ^1-normalisation at each spatial location ij to account for the non-uniform illumination of the tissue surface. This is routinely applied in HSI because of the dependency of the signal on the distance between the camera and the tissue [1, 26]. For Wasserstein-based training, we used the same hyperparameters as those used in the baseline approach to ensure a fair comparison. Experiments were run on an NVidia DGX cluster with V100 (32GB) and A100 (40GB) GPUs.

Choice of tree weights / ground distances Because edge weights exist at different hierarchical levels, resulting in distinct matrices M that affect overall performance, we investigated the impact of various edge weight strategies for M. Specifically, M_t (respectively M_ℓ) represent a configuration in which the edge weight is set to 1 only at the top level (respectively leaf node) with 0 elsewhere. M_e and M_h denote configurations in which edge weights are assigned at all levels, with different ratios. In particular, M_e sets all edge weights to 1, while M_h assigns weights in a 100–10–1 ratio to the top, middle, and bottom level edges, respectively. The latter encodes expert intuition that errors high in the hierarchy are clinically more severe. For clarity, we note that the proposed $\mathcal{L}_{\mathrm{tce}}$ with M_ℓ is equivalent to standard CE loss $\mathcal{L}_{\mathrm{ce}}$. Beyond handcrafted choice, a more systematic approach would tune the edge weights automatically, such as with hyperparameter optimisation to search over candidate configurations [20].

Cross-validation results Table 1 presents the cross-validation results on top-level classes by selecting the output probability at the top level only (i.e., $p_K^\dagger$).

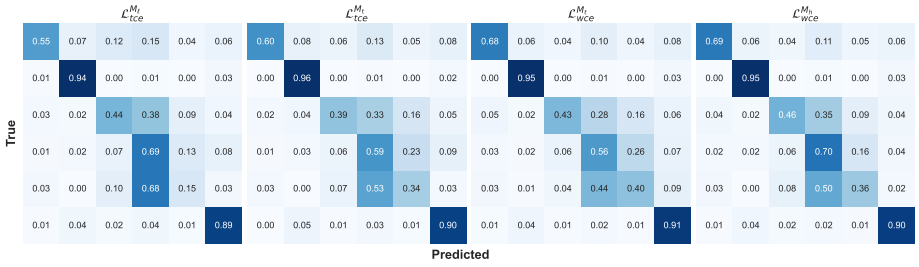

Fig. 2. Confusion matrices on top-level nodes. Results are averaged across all cross-validation folds. Class names from top-left to bottom-right: Other, Out-of-focus Area, Vascular, Normal Tissue, Abnormal Tissue and Surgical Equipment.

We report the results with different confidence thresholds τ. Where $\tau_0 = 0$ represents no OOD detection and τ_m represents the threshold which maximises ID score. Potentially τ_m can be also computed by incorporating held-out OOD classes in the two-level cross-validation [28]. However, this is not presented here due to time constraints in running cross-validation with held-out OOD class. For all loss functions we evaluate the *True Positive Rate (TPR)*, *Balanced Accuracy (BACC)*, and *F1 scores*. Results are reported by averaging across classes under one-vs-rest strategy, where the positives are pixels of such class, and the negative are the pixels of all the other classes. For statistical evaluation, we conducted paired-sample t-test on the test image scores. To limit the number of paired comparisons, we focus on F1-score and comparison between the final tree-based results (M_h configuration for $\mathcal{L}_{tce}$ or $\mathcal{L}_{wce}$) and their corresponding baselines (M_ℓ configuration). Both comparisons were found statistically significant ($p < 0.0001$). For model performance on leaf node classes, we report F1 scores at τ_m for both losses. For $\mathcal{L}_{\text{wce}}$, the mean of F1 scores at τ_m based on M_ℓ and M_h are 0.069 and 0.073, respectively. For $\mathcal{L}_{\text{tce}}$, the results are 0.068 and 0.037, respectively.

While both loss outperforms the baseline, our results shows that M_t performs similarly to M_e, suggesting that top-level edge weights have the greatest impact on performance when evaluating accumulated probabilities on corresponding nodes. For $\mathcal{L}_{\text{wce}}$, employing M_h yields the best performance, even surpassing the strong baseline M_t, demonstrating that an optimal M can achieve state-of-the-art results on both top-level and leaf nodes. For OOD segmentation, our findings exhibit trends similar to those reported in previous work on three medical image datasets [28]. By removing pixels considered outliers, all methods gain further improvements.

Error analysis Evaluation relying solely on overall segmentation performance is insufficient to capture how effectively a model infers relationships among different tissue types. It neglects the possibility that the model may choose semantically coherent labels, even if these do not precisely align with the ground truth. To

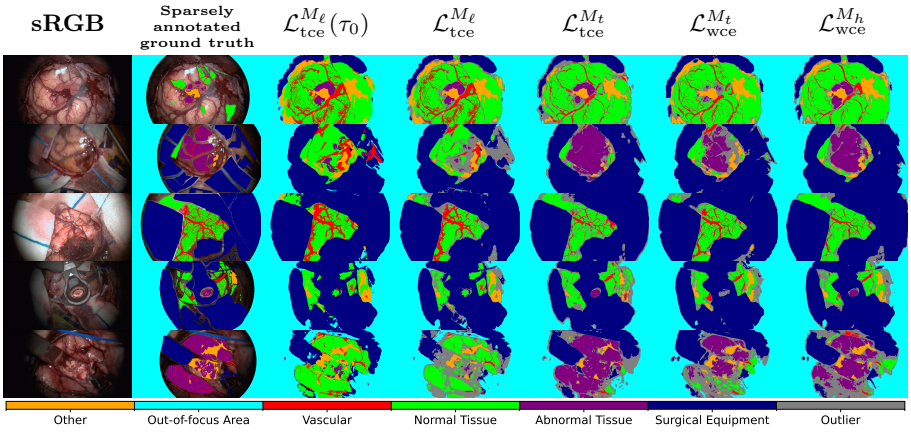

Fig. 3. Qualitative result on top-level classes. We show the result of same image using different methods at confidence threshold τ_m. Baseline results at $\tau_0 = 0$ are added to represent result without outlier detection.

explore these aspects, we plot a multi-class confusion matrix for the top-level nodes $p_K^\dagger$, as shown in Fig. 2. Because class distributions vary across folds, we average the results of each cross-validation fold to ensure a fair comparison.

Under the cross-entropy baseline, the model struggles to distinguish normal from abnormal tissues, which constitute a semantically important distinction. By contrast, both tree-based semantic losses (excluding M_ℓ cases) exhibit a more meaningful confusion pattern between normal and abnormal tissues. These findings suggest that the proposed method successfully exploits hierarchical relationships within the label space.

Qualitative results Fig. 3 presents qualitative results comparing different loss functions on some challenging cases. Although all models have the capacity to differentiate leaf node classes, for simplicity, we visualise predictions at the top-level nodes using the hierarchy in Fig. 1. We present results at $\tau_0 = 0$ to illustrate the outcome without OOD segmentation for the cross-entropy baseline. For the Wasserstein-based loss, we include results using ground distance matrices M_t and M_h for comparison. Comparing against M_ℓ baseline, $\mathcal{L}_{\text{wce}}$ and $\mathcal{L}_{\text{tce}}$ show qualitative results that are more semantically plausible in terms of differentiating normal and abnormal tissues. For other classes such as vascular ones, they also show improved segmentation performance by reducing false positive prediction.

4 Conclusion

We propose two semantically driven loss functions for sparsely supervised segmentation, relying on a tree-structured label space defined by domain experts.

Both Wasserstein+CE and tree-based semantic CE losses leverage prior knowledge of inter-class relationships: the former represents these relationships through a distance matrix in label space, while the latter extends the standard cross-entropy loss to incorporate weighted probabilities aggregated at each node in the tree. Additionally, we integrate these loss functions into an OOD segmentation framework, thereby enabling pixel-level OOD detection without compromising performance on ID data.

Regarding the optimal weighting of hierarchical levels, our experiments on four distance matrices reveal that top-level weights exert the greatest influence on performance when the evaluation is conducted at the corresponding level. Furthermore, we found that the hierarchical weighting scheme further enhances performance, achieving state-of-the-art results on the dataset for both top-level and leaf node labels. Moreover, error analysis and qualitative evaluations demonstrate that these approaches offer improved differentiation between normal and abnormal tissues compared with standard cross-entropy baselines.

Disclosure of Interests. TV and JS are co-founders and shareholders of Hypervision Surgical Ltd, London, UK. The authors have no other relevant interests to declare. This project received funding by the National Institute for Health and Care Research (NIHR) under its Invention for Innovation (i4i) Programme [NIHR202114]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work was supported by core funding from the Wellcome/EPSRC [WT203148/Z/16/Z; NS/A000049/1]. OM is funded by the EPSRC DTP [EP/T517963/1].

References

1. Bahl, A., Horgan, C.C., Janatka, M., MacCormac, O.J., Noonan, P., Xie, Y., *et al.*: Synthetic white balancing for intra-operative hyperspectral imaging. Journal of Medical Imaging **10**, 046001 (2023)
2. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12503–12512, Seattle, WA, USA (2020)
3. Budd, C., Garcia-Peraza Herrera, L.C., Huber, M., Ourselin, S., Vercauteren, T.: Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**, 1215–1224 (2023)
4. Carstens, M., Rinner, F.M., Bodenstedt, S., Jenke, A.C., Weitz, J., Distler, M., *et al.*: The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science. Scientific Data **10**, 3 (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
6. Fabelo, H., Ortega, S., Kabwama, S., Callico, G.M., Bulters, D., Szolna, A., *et al.*: HELICoiD project: a new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations. In: Hyperspectral Imaging Sensors: Innovative Applications and Sensor Standards 2016, p. 986002 (2016)

7. Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., *et al.*: Generalised Wasserstein Dice Score for Imbalanced Multi-class Segmentation Using Holistic Convolutional Networks. In: 9th International MICCAI Brainlesion Workshop, pp. 64–76, Cham (2018)
8. Frogner, C., Zhang, C., Mobahi, H., Araya, M., Poggio, T.A.: Learning with a Wasserstein Loss. In: Advances in Neural Information Processing Systems (2015)
9. Garcia Peraza Herrera, L.C., Horgan, C., Ourselin, S., Ebner, M., Vercauteren, T.: Hyperspectral image segmentation: a preliminary study on the Oral and Dental Spectral Image Database (ODSI-DB). Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**, 1290–1298 (2023)
10. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: International Conference on Learning Representations (2017)
11. Hsu, Y.-C., Shen, Y., Jin, H., Kira, Z.: Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10948–10957 (2020)
12. Hyttinen, J., Fält, P., Jäsberg, H., Kullaa, A., Hauta-Kasari, M.: Oral and Dental Spectral Image Database—ODSI-DB. Applied Sciences **10**, 7246 (2020)
13. Khan, U., Paheding, S., Elkin, C.P., Devabhaktuni, V.K.: Trends in Deep Learning for Medical Hyperspectral Image Analysis. IEEE Access **9**, 79534–79548 (2021)
14. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization, (2017).
15. Le, T., Yamada, M., Fukumizu, K., Cuturi, M.: Tree-Sliced Variants of Wasserstein Distances. In: Advances in Neural Information Processing Systems (2019)
16. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: Advances in Neural Information Processing Systems (2018)
17. Li, P., Ebner, M., Noonan, P., Horgan, C., Bahl, A., Ourselin, S., *et al.*: Deep learning approach for hyperspectral image demosaicking, spectral correction and high-resolution RGB reconstruction. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**, 409–417 (2022)
18. Li, P., MacCormac, O., Shapey, J., Vercauteren, T.: A self-supervised and adversarial approach to hyperspectral demosaicking and RGB reconstruction in surgical imaging. In: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024 (2024)
19. Liang, S., Li, Y., Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In: International Conference on Learning Representations (2018)
20. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: a research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118 (2018)
21. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., *et al.*: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021)
22. Martín-Pérez, A., Martinez-Vega, B., Villa, M., Leon, R., Martinez de Ternero, A., Fabelo, H., *et al.*: Machine Learning Performance Trends: A Comparative Study of Independent Hyperspectral Human Brain Cancer Databases, SSRN Scholarly Paper (2024).
23. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., *et al.*: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**, 1993–2024 (2015)

24. Seidlitz, S., Sellner, J., Odenthal, J., Özdemir, B., Studier-Fischer, A., Knödler, S., *et al.*: Robust deep learning-based semantic organ segmentation in hyperspectral images. Medical Image Analysis **80**, 102488 (2022)
25. Shapey, J., Xie, Y., Nabavi, E., Bradford, R., Saeed, S.R., Ourselin, S., *et al.*: Intraoperative multispectral and hyperspectral label-free imaging: A systematic review of in vivo clinical studies. Journal of Biophotonics **12**, e201800455 (2019)
26. Studier-Fischer, A., Seidlitz, S., Sellner, J., Bressan, M., Özdemir, B., Ayala, L., *et al.*: HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs. Scientific Data **10**, 414 (2023)
27. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114 (2019)
28. Wang, J., Wang, Z., MacCormac, O., Shapey, J., Vercauteren, T.: OOD-SEG: Out-Of-Distribution detection for image SEGmentation with sparse multi-class positive-only annotations. arXiv preprint arXiv:2411.09553 (2024)