HyperSORT: Self-Organising Robust Training with hyper-networks

Samuel Joutard[1⋆], Marijn Stollenga[1⋆], Marc Balle Sanchez[1,2], Mohammad Farid Azampour[2,3], and Raphael Prevost[1]

[1] ImFusion, Munich, Germany
[2] Chair for Computer Aided Medical Procedures (CAMP), Technical University of Munich, Germany
[3] Munich Center for Machine Learning (MCML), Munich, Germany
joutard@imfusion.com

Abstract. Medical imaging datasets often contain heterogeneous biases ranging from erroneous labels to inconsistent labeling styles. Such biases can negatively impact deep segmentation networks performance. Yet, the identification and characterization of such biases is a particularly tedious and challenging task. In this paper, we introduce HyperSORT, a framework using a *hyper-network* predicting UNets' parameters from latent vectors representing both the image and annotation variability. The hyper-network parameters and the latent vector collection corresponding to each data sample from the training set are jointly learned. Hence, instead of optimizing a single neural network to fit a dataset, HyperSORT learns a complex distribution of UNet parameters where low density areas can capture noise-specific patterns while larger modes robustly segment organs in differentiated but meaningful manners. We validate our method on two 3D abdominal CT public datasets: first a synthetically perturbed version of the AMOS dataset, and TotalSegmentator, a large scale dataset containing real unknown biases and errors. Our experiments show that HyperSORT creates a structured mapping of the dataset allowing the identification of relevant systematic biases and erroneous samples. Latent space clusters yield UNet parameters performing the segmentation task in accordance with the underlying "learned" systematic bias. The code and our analysis of the TotalSegmentator dataset are made available: https://github.com/ImFusionGmbH/HyperSORT

Keywords: Hyper Networks · Robust Training · Self-Organising.

1 Introduction

The development of deep learning solutions for medical image analysis requires a thorough review of the training data and its annotation [24]. Indeed, data irregularities such as wrong annotations or acquisition errors can perturb the training process and ultimately degrade the final algorithm capabilities [20].

⋆ These authors contributed equally to this work.

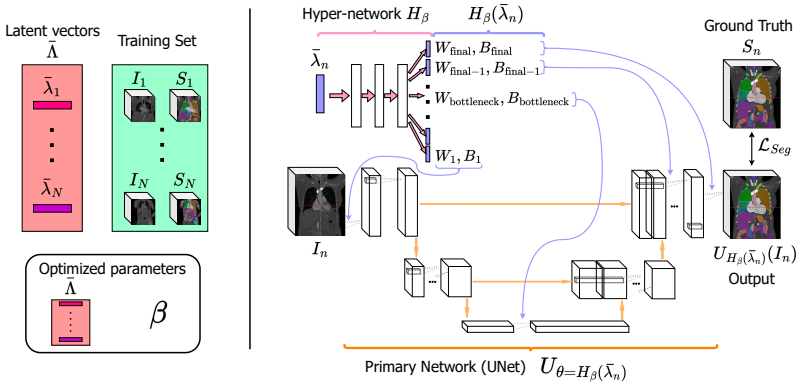

Fig. 1. Overview of HyperSORT. The hyper-network H_β generates the UNet parameters θ from a conditioning vector $\bar{\lambda}_n$ specific to a training sample I_n, S_n.

Medical data curation still heavily relies on human analysis [8], making it a particularly lengthy and error-prone step.

HyperSORT tackles this problem by modeling the annotation process with an additional hidden variable. As such, this hidden variable can parameterize differences between raters or annotation errors. A hyper-network [10] conditions the segmentation UNet [19] behavior on this variable. During training, HyperSORT jointly learns the parameters of the hyper-network and the empirical distribution of the annotation conditioning hidden variable. An overview of the proposed method is shown in Figure 1. HyperSORT provides both robustly trained versions of the segmentation UNet and a meaningful mapping of the training set which can be used to curate the training set and identify systematic biases.

We demonstrate the performance and usability of HyperSORT on two large 3D datasets. As a first proof of concept where the main mode of annotation variability is known and controlled, we injected synthetic perturbations into the AMOS dataset [13]. Second, as a real use case, we used the TotalSegmentator [22] training set. Indeed, this widely recognized dataset has been largely improved and corrected from V1 to V2, offering a form of pseudo ground truth for abnormal cases. In these experiments, we show that HyperSORT generates performing segmentation UNets while providing a meaningful map of the training set which can be interpreted and used to detect erroneous labels.

2 Related works

Dataset quality control Segmentation data curation is challenging as, unlike for classification tasks where an annotation is either right or wrong, segmentation masks can be partially right and wrong. While several methods have been developed for regression/classification data curation (e.g. [5]), the literature on

data curation for segmentation tasks is scarcer. A first typical approach is to rely on repeated cross-validations and use validation metrics such as the Dice score as a proxy for annotation quality [15]. Alternatively, one can rely on a pretrained quality control regressor such as the recently developed Quality Sentinel [6]. While these methods can flag some erroneous cases, HyperSORT pushes the analysis beyond by providing a meaningful mapping for the whole dataset.

Learning from noisy labels To circumvent the challenge of achieving gold standard annotations, methodologies improving models' robustness have been developed. When a rater stratification is available, disentanglement [25] or sampling reweighing [17] can be used. In the general case, probabilistic modeling allows to predict a segmentation distribution [2]. Alternatively, losses [9,26], architecture choices [21,12], or specific training strategies [7] have been shown to improve models' robustness to erroneous labels. For more details on noisy labels detection and robustness, we refer the interested reader to [23] for a more comprehensive review. HyperSORT combines enhanced quality control and robust learning by generating performing networks from potentially noisy labels alongside a mapping of the training set that can be used to discover erroneous cases and systematic biases.

Hypernetworks Hypernetworks have been used as a way to condition the behavior of a primary neural network with respect to a user-provided variable. In the context of medical imaging, it was first used to dynamically tune the regularization strength of deep deformable registration networks [11,18]. More recently, hyper-networks were used to condition a 3D segmentation network on the input image spacing resolution [14]. Hypernetworks can also enable synergistic learning from datasets with heterogeneous annotations by conditioning the network on the structure to segment [3]. All these approaches make use of explicit conditioning variables, either hand-crafted or coming from meta-data. The new paradigm introduced here instead leverages hyper-networks by learning and discovering relevant implicit conditioning within the training set.

3 Method

Supervised segmentation learning often assumes a data distribution $\mathcal{D}$ from which input/segmentation pairs are sampled $(I, S) \sim \mathcal{D}$. A segmentation network, typically a UNet U_θ [19] with parameters θ, is then optimized to minimize an error measure $\mathcal{L}_{Seg}(U_\theta(I), S)$ under the data distribution. However, this assumes that the error in the data annotation is independent and identically distributed (iid) and centered around the actual "ground truth" label [1]. These assumptions do not always hold in the medical domain. Indeed, the scarcity of available training data and the complex annotation process (often relying on bootstrapped, semi-automatic approaches [4] and show-casing high inter-rater variability [2]) require a refined formulation.

Modelization of the Labeling Process Instead, we model the data distribution more precisely by considering the labeling process: $\Omega(I, \lambda) \rightarrow S$, where Ω is an unknown deterministic *oracle* function, and $\lambda \in \mathcal{R}^n$ a latent vector that parameterizes the oracle annotation behavior. Our data distribution explicitly models the label generation process: $\mathcal{D} = \{I, \Omega(I, \lambda) | I \sim \mathcal{I}; \lambda \sim \Lambda\}$, where $\mathcal{I}$ and Λ are the distribution of images I and latent vectors λ respectively. The λ vectors model the labeling process and can, for instance, represent *erroneous labels* or a specific *labeling style* from an annotator, as we will show more concretely in Section 4. Our modelization splits the annotation error between a systematic component modeled by λ and a centered iid additive noise [1], relaxing our learning assumptions.

HyperSORT Our model approximates the Oracle function Ω and the set of annotation style λ on an existing training set. Firstly, we associate a trainable latent vector $\bar{\lambda}_n$ to each training sample $(I_n, S_n) \in \bar{\mathcal{D}}$ where $\bar{\mathcal{D}}$ is the empirical data distribution, i.e. the training set. We consider the well established UNet [19] architecture U_θ parameterized by θ. Instead of directly optimizing θ, we introduce a hyper-network H_β, parameterized by β [10], which predicts the UNet parameters θ from a latent vector $\bar{\lambda}$. The hyper-network parameters β and the set of proxy latent vectors $\bar{\Lambda} = \{\bar{\lambda}_n\}_{n \leq |\bar{\mathcal{D}}|}$ are jointly optimized as:

$$\min_{\beta, \bar{\Lambda}} \sum_{n=1}^{|\bar{\mathcal{D}}|} \mathcal{L}_{Seg}(U_{H_\beta(\bar{\lambda}_n)}(X_n), S_n) + \mathcal{L}_{reg}(\bar{\lambda}_n) \tag{1}$$

where $\mathcal{L}_{Seg}$ is the Dice + CrossEntropy loss and $\mathcal{L}_{reg}$ is the L1-norm regularization term on the latent vectors. This regularization term pushes the latent vectors towards the origin of the latent space, thus making the main annotation mode located around the zero vector $\overrightarrow{\mathbf{0}}$. Consequently, the most unusual cases typically end up isolated further away from the origin and can be identified. Upon convergence, the hyper-network H_β mimics the Oracle Ω and the learned distribution of proxy latent vectors $\bar{\Lambda}$ estimates the annotation style distribution Λ. Using a hyper-network to parameterize the oracle has two important advantages. First, it allows a low-dimensional latent parameterization of annotation styles which creates an interpretable map of the training set in the latent space. Second, as opposed to learning multiple unrelated sets of UNet parameters, using a hyper-network has been shown to enable synergistic learning [3] allowing to make the different annotation styles benefit from each other.

Inference When segmenting a new image, we choose the latent variable that the hyper-network H_β will use to adjust the UNets weights. We typically consider the centroids of the latent vector clusters that have formed during training, thus corresponding to different annotation styles. Given the regularization term $\mathcal{L}_{reg}$, a canonical choice is to use $\lambda = \overrightarrow{\mathbf{0}}$, as a representative of the main annotation style in the training set. Alternatively, a choice can be given to the user to dynamically select the most relevant annotation style for the case to segment.

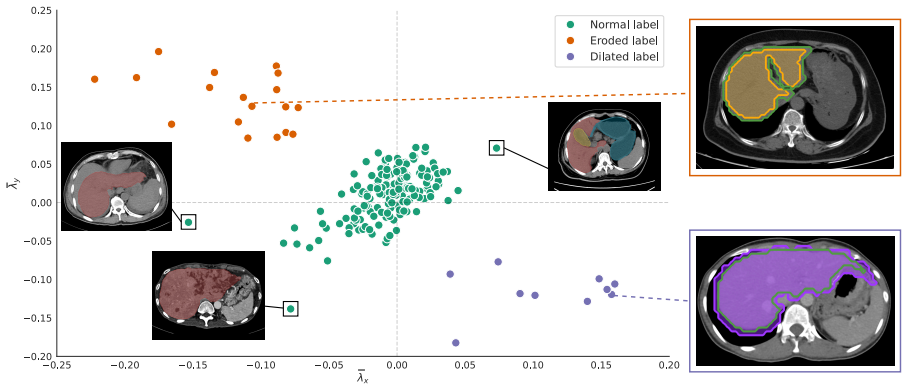

Fig. 2. (Left) Obtained AMOS latent space. The most eccentric cases of the $\overrightarrow{\mathbf{0}}$ cluster are challenging cases (from left to right): incomplete liver segmentation, liver tissue heterogeneity, and abnormal abdominal anatomy where the gallbladder and stomach are shown for reference. (Right) Inference using the $\overrightarrow{\mathbf{0}}$ latent on images from the erosion ■ and dilation ■ clusters.

In addition, given a set of preferred annotation styles selected by an annotator, HyperSORT provides the corresponding UNet parameters which can be used to correct erroneous labels and generate better pseudo-labels.

4 Experiments and Results

HyperSORT is agnostic to the choice of network architecture. Since we target segmentation applications, we use a standard 3D UNet architecture with 3 downsampling stages, 16 channels at the highest resolution and 3 convolutional layers at every resolution stage (ReLU activation, followed by instance normalization). We used 2-dimensional latent vectors to facilitate the visualization and analysis of our results. The hyper-network simply consists of a fully connected network with 3 hidden layers of size 50 each (ReLU activation). The final UNet parameters prediction is followed by a custom activation $x :\longrightarrow tanh(x) * 5$ capping the norm of predicted parameters. This experimentally stabilizes the training. All parameters are trained using the Adam optimizer with an initial learning rate of 10^{-4} until convergence. Additional details can be found in our public repository[1].

4.1 Proof-of-concept using synthetic label perturbations

As a first proof of concept, we create a rough approximation of a multi-rater scenario, with some being more conservative than others regarding organ boundaries. We derived a liver segmentation dataset with 200 CT scans from the

[1] https://github.com/ImFusionGmbH/HyperSORT

AMOS training dataset [13]. We perturb ∼15% of the dataset by performing 3, 4 or 5 iterations of *erosion* on ∼7.5% of the scans, and *dilation* on another ∼7.5%, leaving 85% of labels unperturbed. The learned latent vector distribution $\bar{\Lambda} \subset \mathbb{R}^2$ is shown in Figure 2. We observe that the $[1.0, -1.0]$ direction captures the tightness of liver boundary. Moreover, the synthetic labeling styles (normal, eroded and dilated) are clearly separated and ordered in a meaningful way in the latent space. The eroded and the dilated clusters are also more spread along that direction as they contain variability regarding the number of times each morphological operation was applied. We also see that the central cluster makes use of the other, orthogonal direction to capture some additional secondary variability. For instance, the three most distant cases from that cluster are all challenging, as illustrated in Figure 2. Finally, as shown on the right-hand side of Figure 2, UNet parameters from preferred clusters can be used to correct erroneous annotations from the training set, making of HyperSORT a particularly convenient tool for bootstrapping scenarios.

4.2 Application to the TotalSegmentator dataset

The TotalSegmentator (TS) dataset [22] is a comprehensive collection of 1204 annotated CT scans from different institutions, scanners, and protocols. The corresponding label maps contain more than 100 anatomical structures. Its impressive size has made it very popular in the research community and the basis of numerous papers. However, due to its annotation process (iterative learning, via manual refinement of the predictions of existing models), the label maps may sometimes contain artifacts and over/under-segmentations. Therefore, a second iteration (TS-V2) of the dataset has recently been published and aims to fix some of these errors. This makes it a suitable test-bed for our approach, since we can use TS-V1 as a "flawed" dataset and TS-V2 as the proxy ground truth. From V1 to V2, $\sim 50\%$ of liver annotations were corrected and $\sim 20\%$ required significant adjustments ($>$10000 voxels changed). This experiment allows us to demonstrate our two main claims:

Clusters capture annotation "styles" and generate robust networks On this real use case, the meaning of the five clusters illustrated in Figure 3 is more subtle. Yet, we show here that they all produce usable UNets with significant "annotation style" variation. We consider the five UNet parameter sets obtained from the centroid of each of these clusters. In comparison, we train five randomly initialized UNets with the same architecture as our primary network. We evaluate these models on the CT-1K dataset subtask 2 [16], another large dataset containing 361 diverse abdominal CT scans which do not overlap with TS. Performances are reported in Table 1. In addition to these 2×5 models, we also report the performances of the "best out of 5 models", simulating a "human in the loop" inference scenario. We observe that all UNets generated from HyperSORT yield competitive performances on a large test set. We also note that even smaller clusters containing a limited amount of samples achieve good

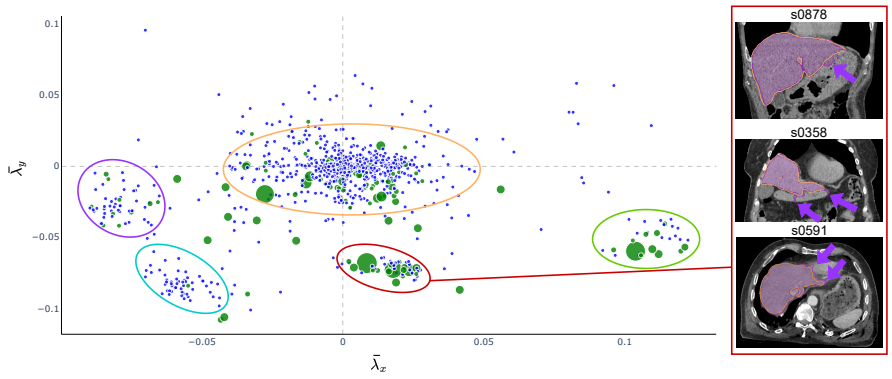

Fig. 3. (Left) Obtained TS latent space. Blue and green dots respectively correspond to non-corrected and corrected cases from V1 to V2. Green dots' radii are proportional to the number of voxels modified. 5 visual clusters are highlighted with colored ellipses. (Right) Slices from 3 cases belonging to the ■ cluster that were not modified between V1 and V2. Purple arrows highlight erroneous liver labeling. Corrective predictions from the ■ cluster UNet are shown.

generalization performances thanks to, we hypothesize, the synergistic learning capabilities of hyper-networks highlighted in [3].

Most importantly, we note that HyperSORT allows a better exploration of the solution space by providing five UNets which segment the liver in five different ways. Indeed, the Dice standard deviation per case between UNet predictions is on average two times larger within HyperSORT UNets (0.5) than within the five randomly initialized UNets (0.2) ($p_{value} \leq 10^{-5}$). This fine exploration of the solution space provides systematically better predictions in the "human in the loop" inference scenario ($p_{value} \leq 10^{-5}$). As the obtained solutions are better differentiated while remaining meaningful, the annotation styles on the left-hand side of the HyperSORT learned latent space (e.g. ■ cluster) must better correspond to the annotation style of the CT-1k dataset labels.

Beyond this quantitative evaluation, we observe in Figure 3 that the red cluster contains several cases that were corrected from V1 to V2. This figure also highlights several samples from that cluster that were not corrected while showcasing erroneous annotations. This suggests that this cluster captures a specific form of systematic annotation error and explains the poorer performance of that cluster's UNet on the CT-1K dataset. This also confirms the practical value of HyperSORT as a tool providing a rich and meaningful analysis of a dataset alongside robust and diverse UNet parameters.

The latent map can be used to identify erroneous cases To evaluate the ability of HyperSORT to detect erroneous cases, we use the changes applied to the liver label from V1 to V2 as pseudo ground truth. Only cases that had at least 1 voxel changed were considered for the pseudo ground truth as

Table 1. Models performances on the CT-1k dataset. Colors correspond to UNets obtained from HyperSORT latent clusters centroid.

UNet seed	1	2	3	4	5	Best
Dice (std)	96.4 (1.8)	96.1 (1.5)	96.4 (1.3)	96.5 (1.4)	96.4 (1.5)	96.6 (1.2)
HyperSort UNet						Best
Dice (std)	96.4 (1.3)	96.7 (1.5)	97.1 (1.5)	95.9 (1.4)	96.4 (1.5)	97.2 (1.4)

we are sure that these cases were checked from V1 to V2. We compare four different predictors for this experiment. First, Quality Sentinel [6] as a recently released annotation quality regressor for segmentation. Then, we train a UNet (same architecture as HyperSORT's primary network) only on the cases that were not modified between V1 and V2. As explained in [15], we can then use this model's Dice loss on the remaining of the training set as a proxy for label quality. We refer to this baseline as "Test-Dice". Test-Dice has an edge over the other predictors as the training/test set split is done leveraging the pseudo ground-truth. Note that, this UNet's generalization capabilities are on par with UNets trained on the whole dataset (96.4 test Dice score on the CT-1k dataset). We consider two possible predictors derived from HyperSORT's mapping of the training set. Following our assumption that the zero cluster captures the normative behavior, we use the proxy latent vector norms $\{||\bar{\lambda}_n||_2\}_{n\leq|\bar{\mathcal{D}}|}$. In addition, as a measure of "isolation" for training cases, we consider the mean distance to all cases $\{\frac{1}{|\bar{\mathcal{D}}|}\sum_m ||\bar{\lambda}_m - \bar{\lambda}_n||_2\}_{n\leq|\bar{\mathcal{D}}|}$. These two HyperSORT derived predictors achieve a respective Spearman correlation with the amount of voxels modified between V1 and V2 of 0.2166 and 0.1723. On the other hand, Quality Sentinel negative scores have an unexpected negative correlation with the amount of changes (-0.0499). Test-Dice, despite its edge, also achieves a lower correlation score of 0.1150. Hence, both HyperSORT-derived features better correlate with the amount of modifications applied from TS V1 to V2. In addition, we stress that the obtained latent vector map $\{\bar{\lambda}_n\}_{n\leq|\bar{\mathcal{D}}|}$ characterizes the training set beyond wrong label detection as shown before. This highlights HyperSORT's ability to identify candidates for label improvement.

4.3 Discussion

We showed that HyperSORT can capture outliers and variations in the dataset that could affect the model quality. However, a remaining problem is to differentiate 'bad labels' from 'challenging correct labels', which can both be associated with large latent vectors, preventing them from being represented in the main mode of the model distribution. On the other hand, this can also indicate under-sampled cohorts of the data distribution that would otherwise be ignored, and can help reveal biases in existing datasets. Regarding the choice of a 2-dimensional latent vector, it facilitates visual inspection and was sufficient in our experiments to capture meaningful variations. Higher dimensional

latent vectors could allow a more homogeneous relationship between the latent space Euclidean norm and annotation style variations, facilitating cluster interpretation. Evaluating the necessity of higher dimensional latent space for other datasets is left for future work. Finally, while we focused here for the sake of conciseness on liver segmentation from CT, HyperSORT can be applied on any segmentation task including challenging structures such as the intestine or multiclass problems. Such complex tasks are more likely to exhibit superposed systematic biases within data samples, making their identification with vanilla clustering methods more challenging. Our public repository makes the extension of HyperSORT to any architecture particularly straightforward and displays the obtained latent space on a series of well known public datasets. We hope that this will help further curate and improve these datasets.

5 Conclusion

In this paper, we introduced HyperSORT which leverages hyper-networks in a novel way to finely stratify the training set and help identifying both erroneous cases and systematic biases while producing performing robustly trained networks. As shown in our experiments, HyperSORT simultaneously acts as a structured curation and corrective tool that could be used systematically when training new models on large datasets.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bach, F.: Learning Theory from First Principles. Adaptive Computation and Machine Learning series, MIT Press (2024)
2. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 119–127. Springer International Publishing, Cham (2019)
3. Billot, B., Dey, N., Turk, E.A., Grant, E., Golland, P.: Network conditioning for synergistic learning on partial annotations. In: Medical Imaging with Deep Learning (2024), https://openreview.net/forum?id=sfjgmuvLS7
4. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis **71**, 102062 (2021). https://doi.org/https://doi.org/10.1016/j.media.2021.102062, https://www.sciencedirect.com/science/article/pii/S1361841521001080
5. Chen, J., Ramanathan, V., Xu, T., Martel, A.L.: Detecting noisy labels with repeated cross-validations **LNCS 15010** (October 2024)
6. Chen, Y., Zhou, Z., Yuille, A.L.: Quality sentinel: Estimating label quality and errors in medical segmentation datasets. CoRR **abs/2406.00327** (2024), https://doi.org/10.48550/arXiv.2406.00327

7. Dong, W., Du, B., Xu, Y.: Shape-intensity knowledge distillation for robust medical image segmentation. Frontiers of Computer Science **19**(9), 199705 (Jan 2025). https://doi.org/10.1007/s11704-024-40462-2, https://doi.org/10.1007/s11704-024-40462-2
8. Galbusera, F., Cina, A.: Image annotation and curation in radiology: an overview for machine learning practitioners. European Radiology Experimental **8**(1), 11 (Feb 2024). https://doi.org/10.1186/s41747-023-00408-y, https://doi.org/10.1186/s41747-023-00408-y
9. Gonzalez-Jimenez, A., Lionetti, S., Gottfrois, P., Gröger, F., Pouly, M., Navarini, A.A.: Robust t-loss for medical image segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 714–724. Springer Nature Switzerland, Cham (2023)
10. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=rkpACe1lx
11. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.: Learning the effect of registration hyperparameters with hypermorph. Machine Learning for Biomedical Imaging **1**, 1–30 (2022). https://doi.org/10.59275/j.melba.2022-74f1
12. Iqbal, S., Khan, T.M., Naqvi, S.S., Naveed, A., Meijering, E.: Tbconvl-net: A hybrid deep learning architecture for robust medical image segmentation. Pattern Recognition **158**, 111028 (2025). https://doi.org/https://doi.org/10.1016/j.patcog.2024.111028, https://www.sciencedirect.com/science/article/pii/S0031320324007799
13. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
14. Joutard, S., Pietsch, M., Prevost, R.: HyperSpace: Hypernetworks for spacing-adaptive image segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15009. Springer Nature Switzerland (October 2024)
15. Lad, V., Mueller, J.: Estimating label quality and errors in semantic segmentation data via any model. arXiv preprint arXiv:2307.05080 (2023)
16. Ma, J., Zhang, Y., Gu, S., Zhang, Y., Zhu, C., Wang, Q., Liu, X., An, X., Ge, C., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., Wang, C., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 6695–6714 (2020), https://api.semanticscholar.org/CorpusID:225094385
17. Mirikharaji, Z., Yan, Y., Hamarneh, G.: Learning to segment skin lesions from noisy annotations. CoRR **abs/1906.03815** (2019), http://arxiv.org/abs/1906.03815
18. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. pp. 35–45. Springer International Publishing, Cham (2021)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015 (2015)
20. Sylolypavan, A., Sleeman, D., Wu, H., Sim, M.: The impact of inconsistent human annotations on ai driven clinical decision making. npj Digital Medicine **6**(1), 26 (Feb 2023). https://doi.org/10.1038/s41746-023-00773-3, https://doi.org/10.1038/s41746-023-00773-3

21. Ș. Vădineanu, Pelt, D., Dzyubachyk, O., Batenburg, J.: An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. In: Medical Imaging with Deep Learning (2022), https://openreview.net/forum?id=C4B46ZS7MSB
22. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024, https://doi.org/10.1148/ryai.230024
23. Wei, Y., Deng, Y., Sun, C., Lin, M., Jiang, H., Peng, Y.: Deep learning with noisy labels in medical prediction problems: a scoping review. Journal of the American Medical Informatics Association **31**(7), 1596–1607 (05 2024). https://doi.org/10.1093/jamia/ocae108, https://doi.org/10.1093/jamia/ocae108
24. Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. Radiology **295**(1), 4–15 (2020). https://doi.org/10.1148/radiol.2020192224, https://doi.org/10.1148/radiol.2020192224, pMID: 32068507
25. Zhang, L., Tanno, R., Xu, M.C., Jacob, J., Ciccarelli, O., Barkhof, F., C. Alexander, D.: Disentangling human error from the ground truth in segmentation of medical images. NeurIPS (2020)
26. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 8792–8802. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)