Eurographics Workshop on Visual Computing for Biology and Medicine (2022)
T. Schultz and H.-Y. Wu (Editors)

Short Paper

Visual Analytics to Assess Deep Learning Models for Cross-Modal Brain Tumor Segmentation

Caroline Magg and Renata Georgia Raidou

TU Wien, Austria

Abstract

Accurate delineations of anatomically relevant structures are required for cancer treatment planning. Despite its accuracy, manual labeling is time-consuming and tedious—hence, the potential of automatic approaches, such as deep learning models, is being investigated. A promising trend in deep learning tumor segmentation is cross-modal domain adaptation, where knowledge learned on one source distribution (e.g., one modality) is transferred to another distribution. Yet, artificial intelligence (AI) engineers developing such models, need to thoroughly assess the robustness of their approaches, which demands a deep understanding of the model(s) behavior. In this paper, we propose a web-based visual analytics application that supports the visual assessment of the predictive performance of deep learning-based models built for cross-modal brain tumor segmentation. Our application supports the multi-level comparison of multiple models drilling from entire cohorts of patients down to individual slices, facilitates the analysis of the relationship between image-derived features and model performance, and enables the comparative exploration of the predictive outcomes of the models. All this is realized in an interactive interface with multiple linked views. We present three use cases, analyzing differences in deep learning segmentation approaches, the influence of the tumor size, and the relationship of other data set characteristics to the performance. From these scenarios, we discovered that the tumor size, i.e., both volumetric in 3D data and pixel count in 2D data, highly affects the model performance, as samples with small tumors often yield poorer results. Our approach is able to reveal the best algorithms and their optimal configurations to support AI engineers in obtaining more insights for the development of their segmentation models.

CCS Concepts

• ***Human-centered computing*** → ***Visual Analytics;*** • ***Applied computing*** → ***Life and medical sciences;***

1. Introduction

Cancer treatment planning requires the accurate segmentation of tumor(s) and of the surrounding structures. In clinical practice, this is often done manually and is used as input to, e.g., treatment planning systems for radiation therapy (RT). Yet, the manual delineation of anatomically relevant structures is very time-consuming and cumbersome. At the same time, missing, biased or inaccurate annotations—especially if treatment protocols involve multimodal images with visually preferred modalities based on contrast and resolution—are further burdening the segmentation process [ZD21]. As a solution to this, deep learning (DL) models, such as neural networks, have emerged and their applicability is being investigated for the automatic segmentation of tumors [NBM*18]. However, DL models are currently far from being integrated to clinical practice, mainly due to robustness reasons and lack of trust on behalf of the clinical experts [GSG*21].

To overcome this, artificial intelligence (AI) engineers, who develop DL models, need to provide an assessment of the robustness of their approaches upon development. Currently, standard evaluation, i.e., with performance measures such as the Dice Similarity Coefficient (DSC) and Averaged Symmetric Surface Distance (ASSD), reduces the complex behavior of an entire approach to a numerical value [WSL*19]. Although this provides a good high-level indication of the performance of the designed approach, it does not provide any further insights into, e.g., indications of consistent errors, correlations to the underlying data, model behavior with regard to special cases, or model trends and limitations. Patterns at a patient-, slice-, or feature-level cannot be detected and analyzed. Furthermore, the prediction results of the models cannot be cross-investigated or correlated with image-derived features to provide additional understanding of the model behavior with regard to the underlying imaging information. All this can help AI engineers to design networks that are more robust, trustable, and generalizable. Visualization and visual analytics are playing a significant role in establishing methods for explainable AI (XAI) [AS22].

In this work, we particularly focus on the visual assessment of DL models for tumor segmentation, stemming from the field of *cross-modal domain adaptation* [WSL*19]. Domain adaptation is a sub-field of transfer learning, which can be particularly useful in the clinical domain, as modalities and scanning protocols often change.

DOI: 10.2312/vcbm.20221193

It allows knowledge learned on one source distribution (e.g., one modality) to be transferred to another distribution. However, as in many DL methodologies, the underlying working mechanisms are not entirely understood, while the performance and the predictive outcomes need to be assessed using a multi-level approach that provides insights at a patient-level, slice-level, or feature-level.

The *contribution* of this work is the design and development of a visual analytics approach—called *crossMoVA*—that supports AI engineers in the visual exploration of the predictions of their cross-modal model(s) and the respective performances, as well as in the investigation of the model(s) behavior with regard to image-derived features. Our framework supports: **(1)** the *comparative visualization* of multiple cross-modal domain adaptation models at different levels of detail; and **(2)** the *correlation* of performance measures or model behaviors with radiomic features, at varying definitions of a tumor's Region of Interest (ROI).

2. Related Work

Several solutions have been proposed in the past for the visual analysis of automatic segmentation outcomes. Landesberger et al. [LBB16] propose visual analytics solutions for the visualization of statistical shape model results. Their approaches cover the analysis of the entire segmentation workflow and the analysis of systematically occurring error for single models. Raidou et al. [RMB*16] published a method to explore and visually assess segmentation errors of single shape models. The tool was extended by Reiter et al. [RBGR18] to incorporate the influence of tumor shape and size variability on the segmentation results. Although all these approaches deal with understanding segmentation outcomes and the impact of the algorithm configuration, they do not deal particularly with DL algorithms. Understanding the complex behavior of DL methods for (tumor) segmentation is a significant topic in the field of XAI [AS22], but the assessment of cross-modal approaches through visual analytics has not been yet investigated.

Another important aspect in assessing cross-modal domain adaptation solution is the investigation of potential relationships or patterns between radiomic features and model performance. Although visual analytics approaches for radiomics is a well-trodden topic (e.g., [MWLH*20]), none of the previous works has established a link to model performances. At the same time, the goal of our work is to not only understand the model performance, but also to support a multi-level analysis drilling down from cohort data to patient data, and further down to slice data, by employing flexible comparative techniques. The correlation between performance measures and radiomics features is supported for various ROIs, which is anticipated to reveal information for the robust design of DL solutions.

3. Data and Task Analysis

Data: The data set used for this work contains brain MRI scans from 250 patients diagnosed with Vestibular Schwannoma and is publicly available at The Cancer Imaging Archive (TCIA) [SKD*21]. The data set contains two types of scans, namely high-resolution T2-weighted (hrT2) and contrast-enhanced T1-weighted (ceT1), and ground truth (GT) labels for the tumor structure (see TCIA repository for data set details [SKD*21]). Eight patients were incomplete and, therefore, excluded. From the remaining, we used 194 for training the deep learning algorithms and 48 for testing.

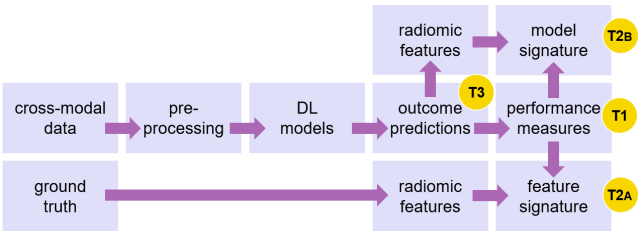


Figure 1: *The workflow employed for crossMoVA, including linking to the tasks (**T1–3**) that the application fulfills.*

Task Analysis: Given the specific requirements of AI engineers developing cross-modal domain adaptation models for the segmentation of brain tumors, we target the following tasks:

T1: Multi-level model performance comparison, i.e., cohort-based, per-patient and per-slice comparison of multiple segmentation models. This is required for the visual assessment of the model(s) robustness.

T2: Discovery of relationship between model performance and image-derived features. This is anticipated to provide insights about the behavior of cross-modal models with regard to underlying imaging features, and comprises two sub-tasks:

- **$T2_A$:** Correlation of error metrics with image-derived features of the GT tumor mask.
- **$T2_B$:** Correlation of model clusters with image-derived features of the tumor masks, as predicted by the models.

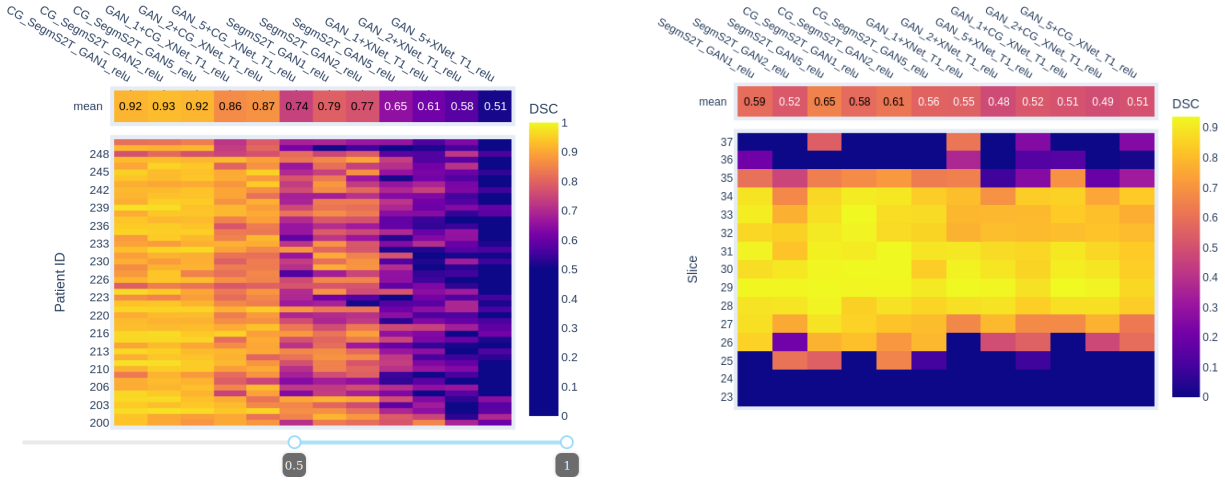
T3: Linking back to the anatomical space and the initial imaging information, to identify if the model predictions have specific behaviors with regard to the tumor characteristics and/or patient anatomy.

T1 represents the three different levels of detail of 3D cohort image data, i.e., entire cohort, individual patient, and single 2D image scan. The connection between performance metrics or model clusters and radiomic features is drawn by **T2**. **T3** represents the visual representation of model comparison in the anatomical space. To address the tasks discussed above, we design and develop a web-based application called *crossMoVA*. The implementation uses TensorFlow, Dash and plotly, and is publicly available on GitHub.

4. Design and Development of *crossMoVA*

The workflow followed for the design and development of *crossMoVA* is depicted schematically in Figure 1. A detailed description can be found in the thesis, from which this paper stems [Mag21].

Pre-processing: The image values are clipped to the 1st and 99th percentile to remove extreme outliers. After a volume-based z-score normalization and min–max normalization to the range [0,1] for the entire volume, each 2D slice is normalized to have a pixel value range of [0,1] and resized to the shape 256×256 (height×width). The former three processing steps are volume-based to preserve the volume statistics. The latter two are needed to apply in-house devel-



(a) *Cohort performance heatmap with individual patients.* **(b)** *Patient performance heatmap with individual slices.*

Figure 2: *A heatmap encodes performance measures at three levels of detail, i.e., overall in the cohort, per-patient, and per-slice, to support **T1** for comparing 12 models (one per column). The first row (mean) provides a summary with averaged performance values per model.*

oped 2D segmentation methods to predict the tumor segmentation mask (**T3**). All these steps are performed offline.

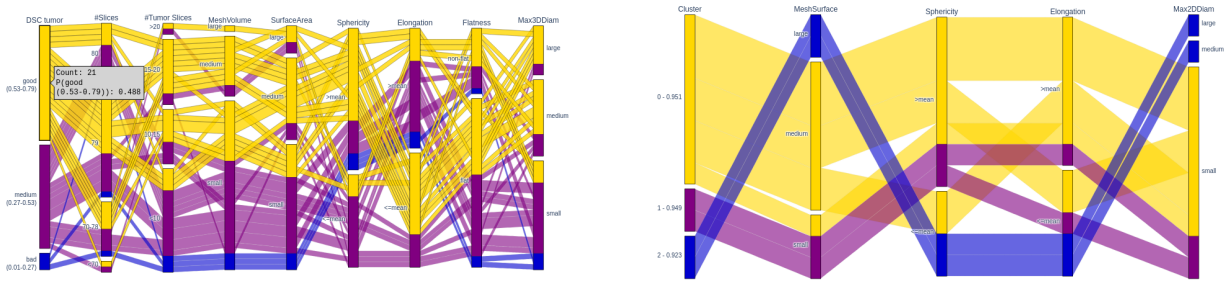
Development of segmentation models: 20 DL-based segmentation pipelines are developed and compared. Baseline networks and domain adaptation segmentation approaches using image synthesis with CycleGAN [ZPIE17] and an additional UNet-based segmentation network [RFB15] are implemented using two different strategies. Either we generate synthetic ceT1 images from real hrT2 scans and apply a T1-specific segmentation network (i.e., `GAN_X+XNet_T1`, with `X` describing at which steps the CycleGAN discriminator is updated), or we generate synthetic hrT2 images and use T1-transferred labels for a supervised training of a T2-dedicated segmentation network (i.e., `SegmS2T`). Both strategies have a standard and an enhanced version (i.e., naming in the upcoming sections contains `CG`). The standard version uses only semantic segmentation for training, whereas the enhanced version employs a classification-guided module to include classification of tumor presence into the training. For more details about the network architectures, we refer the reader to our GitHub repository. We would like to note here that the algorithms mentioned are only examples and could theoretically be exchanged for others.

Model clustering based on performance measures: *Performance measure*s, i.e., Dice Similarity Coefficient (DSC), Averaged Symmetric Surface Distance (ASSD), Accuracy (ACC), True Positive Rate (TPR), and True Negative Rate (TNR) [TH15] are calculated by comparing the predictions of the previously developed models to the GT tumor mask. The metrics are computed in three ways: they are either *averaged* over the entire data set, computed *per 3D patient* sample, or *per image slice*. This is done to support the multi-level comparison of the model robustness (**T1**), but also generates feature vectors for the subsequent clustering of the models according to their performance. This will be further employed for **T2**. The models are, to this end, grouped into 2 to 5 clusters based on the linkage distance of the performance vectors, using the *agglomerative clustering algorithm with Ward linkage* [War63]. This approach is preferred, as it minimizes the variance of the merged groups in a bottom-up approach. The resulting clusters are ranked in descending order based on their averaged cluster performance, i.e., cluster 1 has the highest averaged performance metric.

Radiomic features extraction and grouping: For discovering potential relationships of the model performance to imaging-derived features (**T2**), *radiomic features of shape and first order* are extracted using the image data and a pre-defined ROI [LRVL*12]. This ROI is either the binary GT mask data (for $\mathbf{T2_A}$) or the binary predicted segmentation mask (for $\mathbf{T2_B}$). The features are supplemented by task-specific features, i.e., total number of slices, tumor presence, and tumor size. Subsequently, the values of a feature are *grouped into classes*. Most are divided into three equal classes (i.e., low–mid–high values), but exceptions for divisions are features with statistically meaningful value limits, such as skewness (positive vs. negative) and kurtosis (below vs. above 3).

Creation of feature and model signatures: The combination of all features, i.e., performance measures, radiomic features, and task-specific characteristics, creates an n-dimensional vector per data sample, i.e., per patient or per slice. We call this vector a *feature signature* and this will be used in $\mathbf{T2_A}$. The combination of model cluster membership (as resulting from the clustering above) and radiomic feature extraction generates a vector per slice and is called a *model signature*, which will be further employed in $\mathbf{T2_B}$.

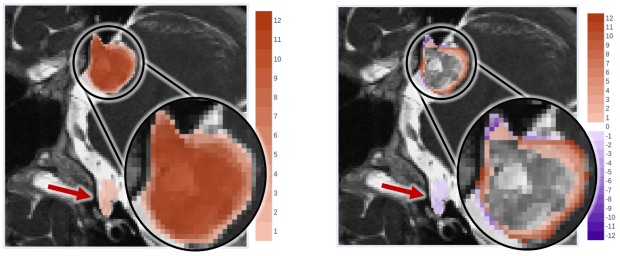
Multi-level model performance comparison (T1): For each segmentation approach, the performance measures need to be investigated at three different levels of detail, i.e., for the entire data set, per patient and per image slice. To this end, we employ a so-called *performance heatmap*, as shown in Figure 2. The performance metrics of each segmentation approach are represented within a two-dimensional matrix, where the columns indicate individual models and the rows individual patients or slices, depending on whether the performance is assessed at a cohort level (Figure 2a) or at a patient



(a) *Feature signature per data sample.*

(b) *Model signature per model prediction.*

Figure 3: *Parallel set diagram to encode (a) radiomic shape feature signatures per data sample, i.e., patient or slice data (**T2$_A$**), or (b) model signatures for slice-wise model prediction (**T2$_B$**). Good performance is shown with yellow, intermediate with purple, and bad with blue.*



(a) *Total sum encoding*

(b) *Subtraction encoding*

Figure 4: *The predicted segmentation masks are visualized with two explicit encoding rules to support **T3**: (a) The total sum visualizes the agreement among N algorithms ($N = 12$); (b) The subtraction compares these segmentation masks to the GT. The red arrow marks a prediction component mirrored around the x-axis.*

level (Figure 2b). The numerical values of the metrics are then colorcoded with a perceptually uniform sequential plasma color map, where yellow indicates good and blue bad performance values. An additional slider scales the color map, i.e., compresses the color map to a range of values. The performance summary per model (i.e., averaged over all patients or over all slices) is given as an annotated heatmap above the first row of the matrix. Here, the numerical values are used as an additional annotation to the colored cells. This representation allows an *n*-by-*n* comparison of all models.

Relationship between model performance and image-derived features (T2): To unveil potential relationships between performance and underlying imaging features, we need to visualize the feature signature for **T2$_A$** or the model signature for **T2$_B$**. We employ *parallel set diagrams* (PSD) [KBH06] to visualize the multivariate, categorical space of feature and model signatures, as shown in Figures 3a and 3b, respectively. The classes of a feature, as resulting from the feature grouping discussed above, are visualized as blocks on the respective parallel vertical axis. The polylines of data samples with similar signatures are merged to a band that spans along the parallel vertical axes. These bands are colorcoded with a binned plasma color map to indicate bad, medium and good performance with blue, purple, and yellow, respectively.

Linking back to the anatomical and imaging space (T3): To visually compare multiple predictions, the predicted segmentation masks need to be compared to each other and to the GT on the anatomical and imaging space. We employ a scalable visualization technique that requires an explicit encoding with two encoding rules. First, we use the *total sum of all masks*, as shown in Figure 4a. It is the equivalent of plotting the cumulative segmentation mask for a single algorithm. A pixel value of 0 means that no algorithm produced a positive prediction for this pixel. Second, we use a *subtraction rule*, as shown in Figure 4b. The total sum subtracted from the GT mask, where the tumor pixel value is the number of algorithms to compare, results in a map with positive and negative pixel values. To encode this, we use a divergent colormap with 0 being white and transparent. On the positive side of the scale, we indicate under-segmentation in red, and on the negative side over-segmentation in blue. A value of 0 corresponds to perfect prediction i.e., background or tumor has been predicted correctly by all algorithms. For both rules, the order of the masks does not matter.

5. Use Cases

We present three use cases to demonstrate the usability of our approach. The cases were conducted by the first author, who is also a trained AI developer with 3 years of experience.

Case 1: ***What is the difference between a number of segmentation approaches?*** Four different segmentation approaches with three different settings each (i.e., in total 12 approaches) are compared. The cohort performance heatmap in Figure 2a reflects the four approaches in the performance analysis. We can identify the overall best pipeline (i.e., `CG_SegmS2T_GAN2_relu` with an average DSC of 0.93 in the first row of the heatmap). The extended versions (i.e., marked with `CG`) show a similar performance for slices with tumor information, but are superior to the standard counterparts when considering the entire data set, as indicated by the higher values of the performance heatmap. The reason is shown in the slice performance and prediction heatmaps, as less false positive predictions (i.e., over-segmentation) occurs. This reduces the error averaged over a 3D volume. Looking at individual slices in Figure 2b, we observe the trend that bad clusters often have extreme values for mesh surface, i.e., either high or low values, reflecting the over- or under-segmentation (see also blue band in Figure 3b). Since the model signature does not show any correlation to the GT, no conclusion about the quality of high or low values can be made without visual inspection. The clusters do not show any trend with regard to a grouping of specific settings or variants.

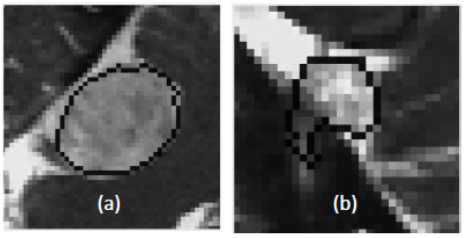


Figure 5: *Example of (a) a homogeneous vs. (b) a heterogeneous ROI indicated with a black contour for the GT.*

Case 2: ***How does the tumor size influence the segmentation performance?*** Subjects with small tumor volumes and slices with a low tumor pixel count show poor results. This can be observed when highlighting the correlation between features for tumor size (`MeshVolume`, `SurfaceArea` and `Max3DDiam`) and performance (`DSC tumor`) in Figure 3a. In the slice performance heatmap of Figure 2b, edge slices at the top and bottom of the tumor volume show worse performance (blue top and bottom rows). Larger tumor structures are more likely to be accurately predicted (yellow middle rows), and this behavior is consistent for all models.

Case 3: ***Are there other data set characteristics related to the model performance?*** During the visual inspection of samples with poor performance, we observed that low DSC and high ASSD values sometimes correspond to predictions mirrored around the x-axis, i.e., where the mirror axis runs through the image center parallel to the x-axis. An example is shown in Figure 4, which depicts a prediction with a correct (zoomed in) and a mirrored component (red arrow). This behavior can not be assigned to a particular approach and is a recurring error. As Figure 4b shows, the tumor center is predicted with high accuracy and reliability, while the tumor margin shows variability and inconsistency (indicated with red or purple). Using the link between performance measures and feature signatures, the visual inspection of prediction heatmaps shows a weak link between homogeneous ROIs with distinct borders and a good performance (Figure 5a), and heterogenous ROIs with more fuzzy borders and a bad performance (Figure 5b).

6. Conclusions and Future Work

In this work, we designed and developed *crossMoVA*, a visual analytics tool to support the comparison of the behavior of multiple cross-modal brain tumor segmentation algorithms at different levels of detail. Future research in radiomics may reveal new definitions that require visual investigation and may yield relationships to performance measures. This topic opens interesting directions with regard to the scalability of our design. Finally, a thorough user study to provide valuable feedback about current limitations and potential extensions is required. *crossMoVA* is an initial step towards gaining a deeper insight into the performance and results of deep neural segmentation networks and the correlation to image-derived features of medical structures.

References

[AS22] ALICIOGLU G., SUN B.: A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics 102* (2022), 502–520. 1, 2

[GSG*21] GILLMANN C., SMIT N. N., GRÖLLER M. E., PREIM B., VILANOVA A., WISCHGOLL T.: Ten challenges in medical visualization. *IEEE Computer Graphics & Applications 41*, 5 (2021), 7–15. 1

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568. 4

[LBB16] LANDESBERGER T. V., BASGIER D., BECKER M.: Comparative Local Quality Assessment of 3D Medical Image Segmentations with Focus on Statistical Shape Model-Based Algorithms. *IEEE Transactions on Visualization and Computer Graphics 22*, 12 (2016), 2537–2549. 2

[LRVL*12] LAMBIN P., RIOS-VELAZQUEZ E., LEIJENAAR R., CARVALHO S., VAN STIPHOUT R. G., GRANTON P., ZEGERS C. M., GILLIES R., BOELLARD R., DEKKER A., ET AL.: Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer 48*, 4 (2012), 441–446. 3

[Mag21] MAGG C.: Development and visual assessment of a multimodality brain segmentation pipeline. *M.Sc. Thesis, TU Wien* (2021). URL: https://doi.org/10.34726/hss.2021.92822. 2

[MWLH*20] MÖRTH E., WAGNER-LARSEN K., HODNELAND E., KRAKSTAD C., HALDORSEN I. S., BRUCKNER S., SMIT N. N.: RadEx: Integrated Visual Exploration of Multiparametric Studies for Radiomic Tumor Profiling. *Computer Graphics Forum 39* (2020). 2

[NBM*18] NIKOLOV S., BLACKWELL S., MENDES R., FAUW J. D., ET AL.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *CoRR abs/1809.04430* (2018). URL: http://arxiv.org/abs/1809.04430, arXiv:1809.04430. 1

[RBGR18] REITER O., BREEUWER M., GRÖLLER E., RAIDOU R. G.: Comparative Visual Analysis of Pelvic Organ Segmentations. In *EuroVis 2018—Short Papers* (2018), pp. 37–41. 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), Springer, pp. 234–241. 3

[RMB*16] RAIDOU R. G., MARCELIS F. J. J., BREEUWER M., GRÖLLER E., VILANOVA A., WETERING H. M. M. V. D.: Visual Analytics for the Exploration and Assessment of Segmentation Errors. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (2016). 2

[SKD*21] SHAPEY J., KUJAWA A., DORENT R., WANG G., BISDAS S., ET AL.: Segmentation of vestibular schwannoma from magnetic resonance imaging: An open annotated dataset and baseline algorithm [data set]. Available from: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70229053, 2021. (accessed 05.07.2021). 2

[TH15] TAHA A. A., HANBURY A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging 15*, 1 (2015), 1–28. 3

[War63] WARD J. H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 301 (1963), 236–244. 3

[WSL*19] WANG G., SHAPEY J., LI W., DORENT R., DIMITRIADIS A., BISDAS S., ET AL.: Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 264–272. 1

[ZD21] ZOU X., DOU Q.: Domain Knowledge Driven Multi-modal Segmentation of Anatomical Brain Barriers to Cancer Spread. In *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data* (2021), Springer, pp. 16–26. 1

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2242–2251. 3