International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Exploiting Architectural and Loss Complementarity: A Diversity-Driven Ensemble for Brain Tumor Segmentation on BraTS 2020

S. U. Ravi Kumar Chavali[1], P. V. Kumar[2]

[1]Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India
[2]School of Engineering, Anurag University, Hyderabad, Telangana, India

Abstract
Deep-learning brain-tumor segmentation on BraTS is dominated by U-Net-family models, yet a prior matched-condition study established that no single architecture or loss dominates and that different configurations excel on different tumor sub-regions. It is asked whether this complementarity is an exploitable resource. Using sixteen previously trained U-Net-family configurations (four decoders × four losses, sharing a ResNet34 encoder), softmax-averaging ensembles are formed and evaluated on 2,242 BraTS 2020 test slices, with test-time augmentation applied uniformly so that gains reflect ensembling alone. A sixteen-model ensemble raises mean Dice from 0.831 to 0.842 (Δ = +0.011, 95% CI [+0.007, +0.015], $p < 0.001$), with the largest gain in the tumor-core region. A three-model complementary subset recovers most of the benefit, and both architectural and loss diversity contribute significantly. Ensembling converts matched-comparison complementarity into a small but statistically robust accuracy gain at modest additional cost.

Keywords: Brain Tumor Segmentation, BraTS 2020, Deep Ensemble, Model Complementarity, Test-Time Augmentation

1. Introduction

1.1 Clinical Context and Motivation

Automated segmentation of gliomas from multi-modal magnetic resonance imaging (MRI) is a central task in neuro-oncology, supporting diagnosis, surgical planning, radiotherapy target definition, and longitudinal monitoring. Following the Brain Tumor Segmentation (BraTS) challenge convention, performance is reported on three hierarchical tumor sub-regions: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET). Over the past decade, deep convolutional networks built on the U-Net encoder-decoder family have come to dominate this task, and a large literature proposes architectural and loss-function refinements, each typically reporting a small Dice improvement over a baseline.

In a prior matched-condition study [1], four U-Net-family decoders (U-Net, UNet++, MAnet, LinkNet) crossed with four segmentation losses (Dice, Focal, Tversky, Focal-Tversky) were compared on BraTS 2020, holding the encoder, preprocessing, augmentation, optimizer, and evaluation protocol identical across all sixteen configurations. That study reached a null result at the aggregate level: under matched conditions, the top configurations were statistically indistinguishable in mean Dice. Crucially, however, it also observed that the best configuration differed by sub-region — no single model was best everywhere,

and different architecture-loss pairings excelled on WT, TC, and ET respectively. This residual, structured disagreement between otherwise comparable models is precisely the condition under which ensembling is expected to help.

1.2 From Matched Comparison to Complementarity

The observation that models of similar aggregate accuracy make systematically different errors is the classical precondition for ensemble learning [12, 13]. If the sixteen matched configurations of [1] are individually near-equivalent but disagree on which voxels they misclassify, then combining their predictions should reduce variance and recover accuracy that no single model attains. This paper tests that hypothesis directly. Because all sixteen models were already trained under identical conditions in [1], ensembling can be studied in a uniquely controlled setting: any gain measured is attributable to prediction combination alone, not to differences in preprocessing, augmentation, or training schedule.

1.3 Contributions

C1 — Ensembling yields a significant aggregate gain. It is shown that softmax-averaging ensembles of the matched U-Net-family configurations improve mean per-slice Dice over the best single model by a small but statistically significant margin (paired bootstrap, $p < 0.001$), with the gain concentrated in the tumor-core region.

C2 — A three-model subset recovers most of the benefit. A compact complementary ensemble of three models, selected on validation performance, recovers the large majority of the full sixteen-model gain, giving a favorable accuracy-versus-cost trade-off for deployment.

C3 — Both architectural and loss diversity contribute. A controlled ablation isolating architecture diversity from loss diversity shows that each axis produces a significant gain on its own, and that mixing both yields the largest improvement.

C4 — A rigorous, reproducible protocol. All results use only the pre-trained checkpoints of [1] (no new training); test-time augmentation is applied uniformly to every configuration including the single-model baseline, so reported gains reflect ensembling rather than augmentation; and ensemble membership is selected on a held-out validation set with no access to the test data.

2. Related Work

2.1 Ensemble Learning

Ensemble methods combine multiple predictors to obtain accuracy and robustness beyond that of any individual member. Their theoretical basis is well established: an ensemble improves on its members when those members are individually accurate and make uncorrelated errors [12]. Bagging [13] reduces variance by averaging models trained on resampled data, while deep ensembles [14] have shown that independently trained neural networks, differing only in random initialization, yield both improved accuracy and better-calibrated uncertainty. In the present setting, diversity arises not from resampling or reinitialization but from deliberate variation of decoder architecture and training loss across otherwise identical models.

2.2 Ensembles in Brain Tumor Segmentation

Ensembling is a recurring theme among top-performing BraTS entries. The EMMA approach of Kamnitsas et al. [15] combined multiple models and architectures to win the BraTS 2017 challenge, arguing explicitly that heterogeneity across models confers robustness to the biases of any single configuration. The nnU-Net framework [10, 11], which won multiple BraTS editions, likewise relies on cross-validation ensembling as a standard component of its pipeline. These results establish that ensembling helps on BraTS; what has been less examined is how much of the benefit is attributable to

architectural versus loss-function diversity when all other training conditions are held fixed — the question this paper isolates.

2.3 Test-Time Augmentation

Test-time augmentation (TTA) averages predictions over label-preserving input transformations, and has been shown to improve both accuracy and uncertainty estimation in medical image segmentation [16]. Because TTA and model ensembling both operate by averaging predictions, a fair evaluation of ensembling must control for TTA. The same flip-based TTA is therefore applied to every configuration, including the single-model baseline, so that the measured ensemble gain is not confounded by augmentation.

2.4 Positioning of the Present Work

This study differs from prior BraTS ensembling work in its controlled design. Rather than assembling a heterogeneous collection of independently tuned models, sixteen configurations that were trained under strictly matched conditions [1] are ensembled. This makes it possible to attribute any accuracy gain to prediction combination alone, quantify how the gain decomposes into architectural and loss diversity, and identify a minimal complementary subset that captures most of the benefit. The work thus completes an arc: a matched comparison first showed that no single configuration dominates, and here that same complementarity is converted into a concrete, if modest, accuracy gain.

3. Research Method

3.1 Base Configurations

The sixteen trained models from [1] are reused without modification. Each model pairs one of four U-Net-family decoders — U-Net, UNet++, MAnet, and LinkNet — with one of four segmentation losses: Dice [17], Focal [18], Tversky [19] ($\alpha = 0.3$, $\beta = 0.7$), and Focal-Tversky [20] ($\gamma = 1.333$). All sixteen share an identical ResNet34 encoder pre-trained on ImageNet [6], identical preprocessing (per-modality z-score normalization, central cropping, resizing to 128 × 128, four-channel input), identical flip-only augmentation, and an identical Adam optimizer schedule. Models were implemented with the Segmentation Models PyTorch library [21]. Following [1], all evaluation is performed per slice on the held-out test set of 2,242 axial tumor-bearing slices; four-class labels denote background, necrotic/non-enhancing core (NCR), peritumoral edema, and enhancing tumor (ET), and the WT, TC, and ET regions are derived from these labels.

3.2 Ensemble Construction

Let a configuration produce, for input slice x, a per-class softmax probability map. For an ensemble of M member configurations, the member probability maps are averaged and the per-pixel class with maximum averaged probability is taken, as in Equations (1) and (2).

$$\bar{p}_c(x) = \frac{1}{M} \sum_{m=1}^{M} p_c^{(m)}(x) \quad (1)$$

$$\hat{y}(x) = \arg\max_c \; \bar{p}_c(x) \quad (2)$$

Probability averaging (a soft vote) is used throughout in preference to hard majority voting, as it preserves the confidence of each model and is standard for softmax classifiers. All member probabilities are computed at 32-bit precision and combined on the CPU to avoid precision-related artifacts.

3.3 Test-Time Augmentation

The probability map of every configuration is itself computed as an average over a set T of four label-preserving flip transforms (identity, horizontal flip, vertical flip, and both), each applied to the input and inverted on the output before averaging, as in Equation (3).

$$p_{TTA}(x) = \frac{1}{|T|} \sum_{t \in T} t^{-1}(f(t(x))) \quad (3)$$

Because the same TTA is applied identically to the single-model baseline and to every ensemble member, the improvement attributed to ensembling is measured over an already TTA-enhanced baseline and therefore reflects prediction combination alone, not augmentation.

3.4 Ensemble Families and Configuration Selection

To avoid selecting ensemble members on the test set, all membership decisions use validation-set performance recorded during training in [1]. Six models are evaluated. The baseline is the single configuration with the highest validation mean Dice (UNet++ + Dice). Five ensembles are then formed: the full All-16 ensemble; a Top-5 ensemble of the five highest-validation configurations; a Complementary ensemble formed by taking the union of the validation-best configuration overall and the validation-best configuration for each of WT, TC, and ET (three distinct models: UNet++ + Dice, UNet++ + Focal, and U-Net + Dice); and two diversity-controlled ensembles — an Arch-diverse ensemble that fixes the loss to Dice and varies the architecture across all four decoders, and a Loss-diverse ensemble that fixes the architecture to UNet++ and varies the loss across all four formulations. The two diversity-controlled ensembles are designed to isolate the contribution of architectural diversity from that of loss diversity.

3.5 Evaluation and Statistical Testing

For each region r in {WT, TC, ET}, per-slice Dice is computed between the predicted region mask P and ground-truth mask G as in Equation (4), with the convention that a slice containing no ground-truth voxels of region r and no predicted voxels of region r scores 1.

$$\text{Dice}_r = \frac{2\,|P_r \cap G_r|}{|P_r| + |G_r|} \quad (4)$$

The overall score for a slice is the mean of its WT, TC, and ET Dice. To test whether an ensemble improves on the best single model, a paired bootstrap over slices is used: for each of 10,000 resamples (fixed seed 42), slices are drawn with replacement and the mean per-slice Dice difference between ensemble and baseline is computed. The observed mean difference Δ, its 95% confidence interval from the bootstrap distribution, and a two-tailed p-value are reported; a difference is called significant when the 95% interval excludes zero. It is noted that a slice-level bootstrap treats slices as independent and is therefore mildly anti-conservative, since slices from the same patient are correlated; this point is revisited in the limitations.

4. Results

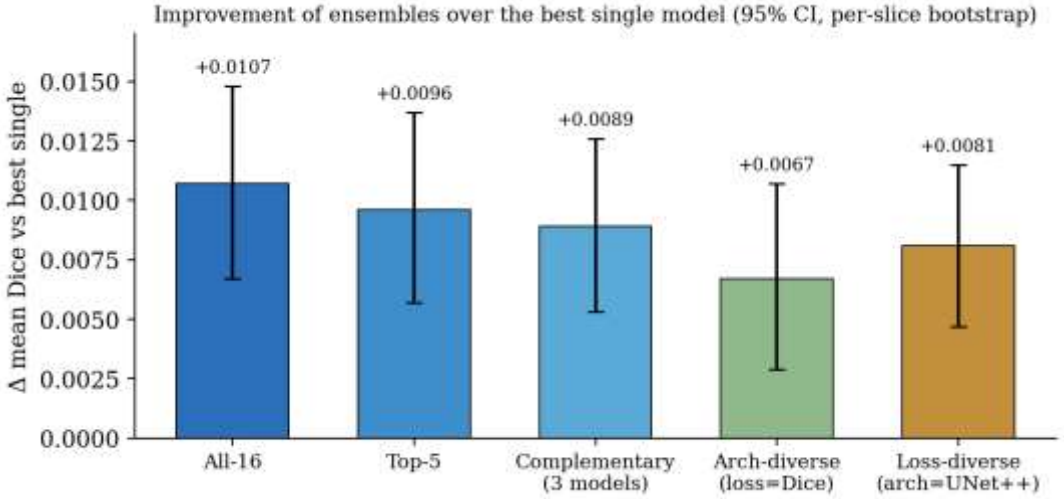
4.1 Ensembles Improve on the Best Single Model

Table 1 reports per-region and mean Dice for the best single configuration and for each of the five ensembles, together with the paired-bootstrap comparison of each ensemble against the single model. Every ensemble improves on the best single model, and in every case the 95% confidence interval on the mean-Dice difference excludes zero. The full All-16 ensemble achieves the highest mean Dice, raising it from 0.8308 to 0.8416, a difference of +0.0107 (95% CI [+0.0067, +0.0148], $p < 0.001$).

Table 1: Per-Slice Test Performance (n = 2,242 Slices) of the Best Single Configuration and Five Ensembles on BraTS 2020. M Is the Number of Member Models. Δ Is the Mean Per-Slice Dice Difference Versus the Best Single Model, With Its 95% Bootstrap Confidence Interval and Two-Tailed p-Value.

Model	M	WT	TC	ET	Mean	Δ vs single	95% CI	p
Best single (UNet++ + Dice)	1	0.858	0.807	0.828	0.8308	—	—	—
All-16 ensemble	16	0.867	0.827	0.831	0.8416	+0.0107	[+0.0067, +0.0148]	< 0.001
Top-5 ensemble	5	0.868	0.823	0.830	0.8405	+0.0096	[+0.0057, +0.0137]	< 0.001
Complementary ensemble	3	0.863	0.823	0.833	0.8397	+0.0089	[+0.0053, +0.0126]	< 0.001
Arch-diverse (loss = Dice)	4	0.861	0.821	0.830	0.8375	+0.0067	[+0.0029, +0.0107]	< 0.001
Loss-diverse (arch = UNet++)	4	0.866	0.820	0.830	0.8389	+0.0081	[+0.0047, +0.0115]	< 0.001

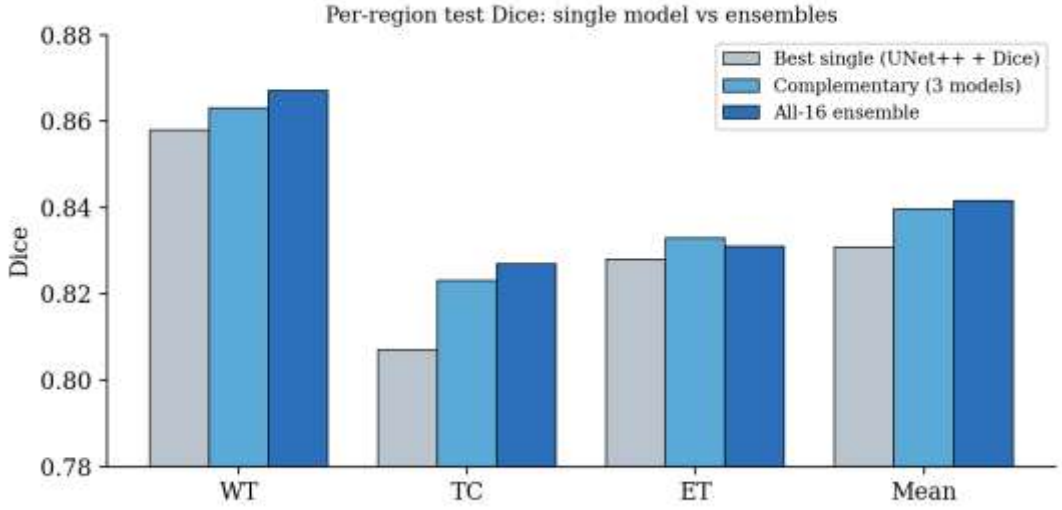
Figure 1: Improvement in Mean Per-Slice Dice of Each Ensemble Over the Best Single Model. Bars Show the Observed Mean Difference Δ; Error Bars Show the 95% Bootstrap Confidence Interval (10,000 Resamples). All Five Intervals Lie Entirely Above Zero.



4.2 The Gain Is Concentrated in the Tumor Core

Breaking the improvement down by sub-region (Figure 2) shows that the ensemble benefit is not uniform. Relative to the best single model, the All-16 ensemble improves tumor-core Dice most strongly, from 0.807 to 0.827 (+0.020), while whole-tumor Dice rises modestly (0.858 to 0.867) and enhancing-tumor Dice is essentially unchanged (0.828 to 0.831). This pattern is consistent with the matched comparison of [1], in which tumor core was the region of greatest inter-model disagreement; ensembling recovers accuracy precisely where the individual models were least consistent. No region degrades under any ensemble.

Figure 2: Per-Region Test Dice (WT, TC, ET) and Overall Mean for the Best Single Model, the Three-Model Complementary Ensemble, and the Full Sixteen-Model Ensemble. The Largest Ensemble Gain Is in the Tumor-Core Region; No Region Degrades.



4.3 A Compact Complementary Ensemble Captures Most of the Gain

Practically, the full sixteen-model ensemble is expensive at inference. The three-model complementary ensemble — UNet++ + Dice, UNet++ + Focal, and U-Net + Dice — attains a mean Dice of 0.8397, a gain of +0.0089 over the best single model. This recovers about 83% of the +0.0107 full-ensemble gain using roughly one-fifth of the models, and it achieves the highest enhancing-tumor Dice of any configuration (0.833). For inference-constrained deployment, this compact ensemble offers most of the accuracy benefit at a small fraction of the cost, echoing the efficiency emphasis of the matched comparison in [1].

4.4 Both Architectural and Loss Diversity Contribute

The two diversity-controlled ensembles isolate the source of the benefit. The Arch-diverse ensemble (four decoders, loss fixed to Dice) improves mean Dice by +0.0067 (95% CI [+0.0029, +0.0107]), and the Loss-diverse ensemble (four losses, architecture fixed to UNet++) improves it by +0.0081 (95% CI [+0.0047, +0.0115]); both are statistically significant. Loss diversity yields a marginally larger gain than architecture diversity, but the two intervals overlap substantially, so the honest reading is that both axes contribute meaningfully. That the full ensemble, which mixes both forms of diversity, exceeds either controlled ensemble indicates the two sources of diversity are complementary rather than redundant.

5. Discussion

5.1 Interpreting the Gain

The central finding is positive but modest: under strictly matched conditions, ensembling U-Net-family configurations produces a small, statistically robust improvement in mean Dice over the best single model, concentrated in the tumor core. This should be read together with the null result of [1]. There, varying architecture or loss in isolation did not move aggregate accuracy; here, combining those same varied models does. The two findings are consistent: individual configurations are near-equivalent in aggregate yet disagree in structured ways, and it is exactly that disagreement that ensembling exploits. The magnitude of the gain (about one Dice point) is commensurate with the small inter-model spread reported in [1] and should not be oversold, but it is real and reproducible.

5.2 Practical Implications

For practitioners, the useful message is that most of the ensemble benefit is available cheaply. A three-model complementary ensemble recovers the large majority of the full-ensemble gain, and its members can be selected on validation performance without test-set access. Where inference budget permits only a single forward pass, the matched comparison already identified efficient single-model choices; where a modest ensemble is affordable, the three-model configuration here is a practical operating point. Test-time augmentation contributes to all of these results and is essentially free at inference relative to the cost of additional models.

5.3 Limitations

Three limitations bound these conclusions. First, evaluation and significance testing are performed at the slice level; because slices from the same patient are correlated, the slice-level bootstrap is mildly anti-conservative, and patient-level aggregation would yield wider intervals. Given the size and consistency of the observed effect the direction of the result is not expected to change, but a patient-level confirmation is a direct and worthwhile extension. Second, all experiments use a single dataset (BraTS 2020) and a 2D slice-based pipeline inherited from [1]; generalization to 3D pipelines and other cohorts is untested. Third, the absolute gain is small, as expected for an ensemble of matched models; ensembling is a reliable way to extract the last fraction of a Dice point from an existing model family, not a substitute for a fundamentally stronger model. Addressing the specific failure modes identified in [1] through targeted architectural or loss design remains complementary future work.

6. Conclusion

Starting from a matched-condition comparison that found no single U-Net-family configuration dominant on BraTS 2020, it was asked whether the residual complementarity among those configurations could be turned into an accuracy gain. It can. Softmax-averaging ensembles of the sixteen matched models improve mean per-slice Dice over the best single model by a small but statistically significant margin, with the improvement concentrated in the tumor-core region and no region degrading. A three-model complementary ensemble recovers most of the benefit at a fraction of the inference cost, and a controlled ablation shows that architectural and loss diversity each contribute and are complementary. The study completes a coherent arc: matched rigor first showed that the easy levers of architecture and loss do not move aggregate accuracy in isolation, and here that same rigor shows how their complementarity, combined, yields a reliable if modest gain. Patient-level confirmation and extension to 3D pipelines are natural next steps.

7. Acknowledgement

The authors thank the organizers of the Multimodal Brain Tumor Segmentation (BraTS) challenge for making the BraTS 2020 dataset publicly available, and acknowledge the use of publicly available open-source segmentation libraries. This work reuses the trained models and evaluation infrastructure of the companion study [1]. No new model training was required for the ensembling experiments reported here. This work received no specific funding from any agency.

8. References

1. S. U. Ravi Kumar Chavali and P. V. Kumar, "A Controlled Comparison of U-Net Decoder Architectures and Loss Functions for Brain Tumor Segmentation on BraTS 2020," companion manuscript, 2024 (unpublished).
2. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, pp. 234–241, 2015.
3. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in Deep Learning in Medical Image Analysis (DLMIA), LNCS 11045, Springer, pp. 3–11, 2018.
4. T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation," IEEE Access, vol. 8, pp. 179656–179665, 2020.
5. A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," in Proc. IEEE Visual Communications and Image Processing (VCIP), pp. 1–4, 2017.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, pp. 770–778, 2016.
7. B. H. Menze, A. Jakab, S. Bauer, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.
8. S. Bakas, H. Akbari, A. Sotiras, et al., "Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features," Scientific Data, vol. 4, article 170117, 2017.
9. U. Baid, S. Ghodasara, S. Mohan, et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," arXiv:2107.02314, 2021.
10. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.
11. F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for Brain Tumor Segmentation," in Brainlesion (BraTS 2020), arXiv:2011.00848, 2020.
12. T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, LNCS 1857, Springer, pp. 1–15, 2000.
13. L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
14. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," in Proc. NeurIPS, pp. 6402–6413, 2017.
15. K. Kamnitsas, W. Bai, E. Ferrante, et al., "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation," in Brainlesion (BrainLes 2017), LNCS 10670, Springer, pp. 450–462, 2018.

16. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks," Neurocomputing, vol. 338, pp. 34–45, 2019.
17. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in Proc. 4th International Conference on 3D Vision (3DV), pp. 565–571, 2016.
18. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. IEEE ICCV, pp. 2980–2988, 2017.
19. S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," in Machine Learning in Medical Imaging (MLMI), LNCS 10541, Springer, pp. 379–387, 2017.
20. N. Abraham and N. M. Khan, "A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 683–687, 2019.
21. P. Iakubovskii, "Segmentation Models PyTorch," GitHub repository, 2019. https://github.com/qubvel/segmentation_models.pytorch