ATVNet: Adaptive Transformer-Enhanced V-Net for Automated msTBI Lesion Segmentation

Balázs Zavadil[1 [0009−0009−5211−0539]], Bálint Szabó[1 [0000−0003−0350−6499]], Ákos Szlávecz[1 [0009−0000−9019−175X]], and Balázs Benyó[1 [0000−0003−2770−9127]]

Department of Control Engineering and Information Technology,
Faculty of Electrical Engineering and Informatics,
Budapest University of Technology and Economics, 1111 Műegyetem rkp. 3.,
Budapest, Hungary
zavadilb@edu.bme.hu
{bszabo,szlavecz,bbenyo}@iit.bme.hu
https://www.iit.bme.hu/

Abstract. In this paper, we propose Adaptive TransVNet (ATVNet), a transformer-enhanced VNet-based framework for automated moderate-to-severe traumatic brain injury (msTBI) lesion segmentation from T1-weighted MRI images. Unlike better-characterized neurological conditions such as stroke, multiple sclerosis, or brain tumors, msTBI lesions exhibit substantial variability in size, location, and appearance, making automated segmentation a challenging task. The proposed architecture extends the conventional V-Net framework by incorporating transformer modules at multiple low-resolution stages and introducing learnable residual gating to adaptively regulate the contribution of transformer features. The gating parameters are initialized to preserve the original convolutional network behavior and gradually learn the optimal integration of local and global representations during training. The model was developed and evaluated within the AIMS-TBI26 challenge, using extensive three-dimensional data augmentation and a combined BCE-Dice-Focal loss function. Preliminary experiments demonstrated improved performance over both V-Net and TransVNet variants, with ATVNet achieving a mean DSC of 56.52% on the benchmark train-validation split. A 5-fold ensemble strategy was also evaluated but did not improve performance for either task. In the final challenge evaluation on a hidden test dataset, the submitted model achieved a DSC of 52.25%, an ASSD of 9.12 mm, and an HD95 of 32.02 mm for lesion segmentation. For lesion detection, the proposed approach achieved a balanced accuracy of 82.89%, with 75.00% sensitivity and 90.79% specificity. These results demonstrate the potential of the proposed approach within the AIMS-TBI26 challenge, while further evaluation on larger, more diverse datasets and additional performance metrics is needed to fully assess its general applicability.

Keywords: msTBI, MRI, Lesion segmentation, TransVNet, Adaptive transformer gating

1 Introduction

Moderate-to-severe traumatic brain injury (msTBI) can result from significant external forces such as head impacts. It causes highly heterogeneous patterns of structural brain damage, including hemorrhages and contusions that can lead to life-threatening complications. These lesions vary in size, location, morphology, and tissue involvement [1], creating significant challenges for decision support pipelines. Current automated approaches for lesion diagnosis are limited, as manual segmentation is labor-intensive, and many automated methods require imaging modalities that are not consistently available across large multi-site datasets.

To address these problems, the AIMS-TBI26 challenge aims to advance automated lesion detection and segmentation methods using routinely acquired MRI data. The challenge comprises two tasks: (1) lesion detection, which aims to identify the presence of msTBI-related lesions from T1-weighted MRI data, and (2) lesion segmentation, which aims to generate binary lesion masks for the identified abnormalities.

In this work, we propose Adaptive TransVNet (ATVNet), a transformer-enhanced framework for automated msTBI lesion segmentation. The method extends our previously proposed VNet-based segmentation model [2] by incorporating recent advances in hybrid convolutional-transformer architectures. In particular, TransVNet, originally proposed for seismic data segmentation [3], demonstrated the effectiveness of embedding transformer modules within a V-Net framework. Building on this idea, our architecture introduces transformer pathways at multiple low-resolution stages and employs learnable residual gating to adaptively fuse transformer and convolutional features. The proposed method is evaluated on both the lesion detection and lesion segmentation tasks of the AIMS-TBI26 challenge.

2 Methods and Data

2.1 Data

Dataset The training dataset provided by the challenge organizers consisted of 553 annotated T1-weighted MRI images and an additional 100 scans for validation. Each image was accompanied by a corresponding lesion mask containing lesion annotations when lesions were present; otherwise, the mask was empty, indicating the absence of visible lesions.

The final performance of the proposed method was assessed on a separate test dataset that was withheld from challenge participants and accessible only through the evaluation system on the grand-challenge.org platform. Leaderboard rankings were determined automatically, using Dice Similarity Coefficient (DSC), 95th percentile Hausdorff distance (HD95), and Average Symmetric Surface Distance (ASSD) for the segmentation task, while balanced accuracy was used to evaluate detection performance.

2.2 Data Preprocessing

Image preprocessing was performed to standardize the input data and ensure compatibility with the proposed neural networks, which require fixed-size inputs. The preprocessing pipeline consisted of cropping each image to its non-zero bounding box, resampling it to an isotropic voxel spacing of $(1, 1, 1)$ mm, reorienting it to the Right-Anterior-Superior (RAS) coordinate system, and, when necessary, applying zero-padding to achieve the required input dimensions.

2.3 Methods

Network Architecture CNN-based encoder-decoder architectures, such as U-Net [4] and V-Net [5], are widely used for medical image segmentation due to their strong spatial feature extraction capabilities. However, their limited receptive fields motivate the integration of transformer-based modules, which enable modeling of long-range dependencies through self-attention mechanisms [6,7].

Our initial baseline model was based on V-Net, which uses an encoder-decoder structure with skip connections to preserve spatial information during segmentation.

To address the CNN's limited receptive field, transformer blocks were initially introduced at the bottleneck, forming the TransVNet architecture. In subsequent experiments, transformer modules were additionally incorporated into the two lowest-resolution skip connections to provide global context at multiple network stages, while remaining within the available GPU memory constraints.

The transformer module consists of token embedding, transformer encoding, and feature reconstruction stages. Four transformer layers, each containing multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks, were used to model global dependencies among spatial features. The MSA employs 16 attention heads with a hidden dimension of 768, followed by an MLP layer with a dimension of 3072.

The final architecture, ATVNet, shown in Fig. 1, extends this design through learnable residual gating. Scaling parameters (α) regulate the contribution of transformer features within the bottleneck and skip pathways. All α parameters are initialized to zero, allowing the model to begin as a standard V-Net and gradually learn the optimal contribution of transformer features during training. Due to GPU memory limitations introduced by the transformer modules and the 3D input volumes, training was performed with a batch size of 1.

Data Augmentation and Sampling For the segmentation task, models were trained exclusively on lesion-positive images, as the hidden test set for this phase consisted only of scans containing lesions. This decision also reduced the number of training samples, leading to shorter training times and enabling more extensive hyperparameter optimization before the challenge deadline.

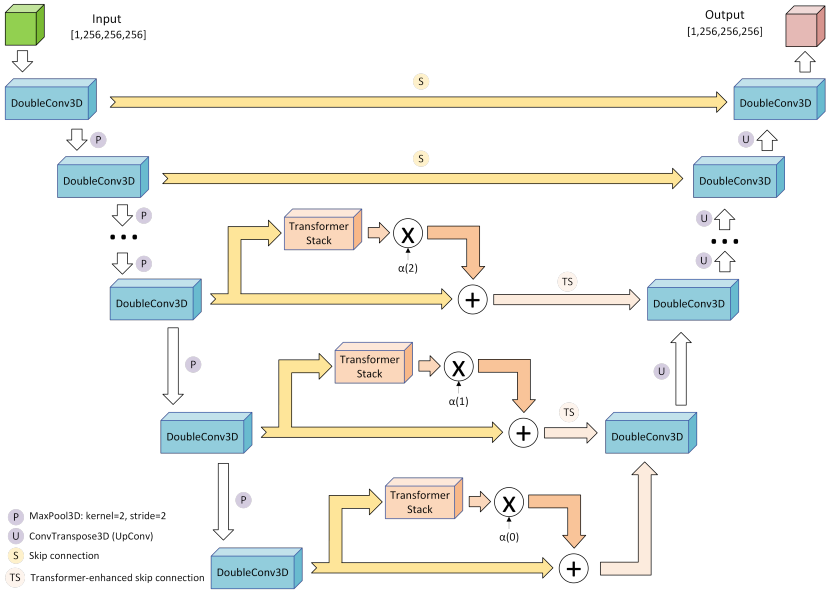

Fig. 1. Overview of the proposed ATVNet architecture. Transformer-enhanced skip connections with adaptive gating coefficients $\alpha(0)$, $\alpha(1)$, and $\alpha(2)$ integrate transformer features into the decoder while preserving spatial information from encoder features.

For the detection task, balanced sampling was used, providing the model with 50% lesion-positive and 50% lesion-negative scans during training to address class imbalance, as in this case, detecting true negatives is equally important.

The data augmentation strategy for both tasks was based on the augmentation pipeline used by the winning method of the AIMS-TBI25 challenge [8], the BLSegMamba [9] algorithm. This pipeline was selected due to its comprehensive set of augmentations for T1-weighted MRI images, which provided strong variability during training and improved the model's ability to generalize.

Spatial diversity was introduced through 3D transformations, including random rotations around all axes ($\pm 30°$) and isotropic scaling ($0.7 - 1.4$), applied with a probability of $p = 0.2$. To improve robustness to MRI noise and acquisition variability, intensity augmentations included Gaussian noise ($p = 0.1$), Gaussian blur ($\sigma = 0.5 - 1.0$; $p = 0.2$ per sample), brightness and contrast adjustments ($p = 0.15$ each), and gamma correction (range $0.7 - 1.5$; $p = 0.1$). Additionally, random mirror flipping along the axial, coronal, and sagittal axes was applied to exploit anatomical symmetry and to make the model robust to differently oriented scans.

Loss Function and Optimizer As a loss function, we used binary cross-entropy (BCE) for VNet-based models, as it yielded higher segmentation performance. However, the later models incorporating transformers at any level of

the network struggled to converge, and they tended to collapse into predicting empty masks for all inputs.

To address this, the combined loss proposed by BLSegMamba [9] was used. It is defined as a weighted combination of BCE loss, Dice loss, and Focal loss, as shown in Eq. (1):

$$\mathcal{L} = 0.3\,\mathcal{L}_{\text{BCE}} + 0.4\,\mathcal{L}_{\text{Dice}} + 0.3\,\mathcal{L}_{\text{Focal}} \tag{1}$$

As the optimizer, Adam [10] was used, the learning rate was initialized to 1.8×10^{-4} and reduced by a factor of 2 whenever the training loss plateaued for more than 25 epochs. The learning rate was decreased until it reached a minimum of 5×10^{-6}.

Early Stopping and Checkpointing Training was terminated early if validation loss did not improve for 150 consecutive epochs. After each epoch, the latest model and optimizer states were saved to resume interrupted training. For segmentation, the checkpoint with the highest DSC on lesion-positive validation cases was selected, while for detection, the checkpoint with the highest balanced accuracy on the complete validation set was saved.

5-Fold Ensemble For both segmentation and detection tasks, a 5-fold ensemble strategy was used to reduce model variance. During inference, the voxel-wise probability maps generated by the five fold-specific models were averaged after sigmoid activation, and the final segmentation mask was obtained by thresholding the ensemble probability map at $\tau = 0.5$. Averaging predictions across multiple models can reduce the influence of model-specific errors and improve the reliability of the final output.

Connected-Component Filtering for Detection For lesion detection, no dedicated classification head was used. Instead, segmentation probability maps were converted into volume-level predictions using connected-component (CC) filtering to reduce false-positive components and improve specificity. The minimum retained component size was optimized on validation data by sweeping thresholds in 5-voxel increments and selecting the threshold with the highest balanced accuracy. For the 5-fold ensemble, out-of-fold predictions yielded an optimal threshold of 105 voxels, while the single submitted model yielded 130 voxels on its validation set. A volume was classified as lesion-positive if at least one component remained after filtering.

Implementation The experiments were conducted on NVIDIA A100 GPUs using CUDA-enabled parallel computation. The models were implemented and trained using PyTorch as the primary deep learning framework. Data augmentation was performed using the batchgenerators library [11]. Training and evaluation of the proposed models were carried out on the GPU and AI partitions of the Komondor high-performance computing (HPC) system.

3 Results

To enable a direct comparison between models during the preliminary experiments, a fixed train-validation split was employed. The validation set corresponded to the dataset later released for the validation phase of the challenge, after teams submitted their containerized solutions for evaluation. Each model was trained and evaluated using this identical data partitioning scheme, and the resulting performance metrics are presented in Table 1.

Table 1. Performance comparison on lesion-positive scans of different models during preliminary testing using the fixed train-validation split. Metrics are computed on a per-volume basis and then averaged across all volumes.

Model	DSC (%)	ASSD (mm)	HD95 (mm)
V-Net	45.79%	11.36	27.74
TransVNet	49.05%	8.71	28.70
ATVNet	**56.52%**	**5.71**	**20.39**

The computational characteristics of the evaluated models are summarized in Table 2. While TransVNet and ATVNet have higher parameter counts and GPU memory requirements than V-Net, their inference times remain comparable.

Table 2. Computational characteristics of the evaluated segmentation models. GPU memory denotes peak memory usage. All metrics were measured on a single NVIDIA A100 GPU with a batch size of 1 and an input size of $256 \times 256 \times 256$. Inference time was averaged over 50 forward passes for each model.

Model	Parameters (M)	Train VRAM (GB)	Inference VRAM (GB)	Inference (s/vol.)
V-Net	129.66	19.50	7.88	0.216
TransVNet	187.94	20.23	8.60	0.217
ATVNet	258.22	26.22	9.56	0.243

Based on the results in Table 1, the ATVNet model was chosen, and we performed 5-fold cross-validation on the entire dataset. These results are shown in Table 3. The resulting five models were used to construct the final ensemble.

For the segmentation finals, both the best-performing ATVNet model from the preliminary train-validation split and its 5-fold ensemble were submitted. The achieved results according to the official challenge evaluation metrics are shown in Table 4.

Similarly, for the detection finals, a single-fold model and a 5-fold ensemble of the ATVNet model were submitted. Each model was trained on both lesion-positive and lesion-negative samples, rather than only on positive cases. The checkpoints with the highest validation balanced accuracy were selected for final testing. Segmentation outputs were converted into volume-level predictions using connected-component filtering. The results are shown in Table 5.

Table 3. Five-fold cross-validation results for the ATVNet model, trained and evaluated only on lesion-positive images. Metrics are computed on a per-volume basis and then averaged across all volumes. Mean ± standard deviation is reported in the final row.

Fold	DSC (%)	ASSD (mm)	HD95 (mm)
Fold 1	51.77%	9.93	27.09
Fold 2	52.04%	12.07	29.63
Fold 3	54.31%	8.38	26.65
Fold 4	57.54%	7.44	26.05
Fold 5	56.18%	5.07	17.81
Mean ± Std	54.37 ± 2.53	8.58 ± 2.63	25.45 ± 4.48

Table 4. The segmentation performance of the two algorithms (ATVNet and its 5-fold ensemble) submitted to the final test phase of the challenge, according to leaderboard metrics.

Model	DSC (%)	ASSD (mm)	HD95 (mm)
ATVNet	52.25%	9.12	32.02
ATVNet (ens.)	48.29%	10.10	30.28

Table 5. Detection performance of the two submitted algorithms (ATVNet and its 5-fold ensemble) submitted to the final test phase of the challenge, according to leaderboard metrics.

Model	Bal. Acc. (%)	Sensitivity (%)	Specificity (%)
ATVNet	82.89%	75.00%	90.79%
ATVNet (ens.)	80.08%	69.38%	90.79%

4 Discussion

Preliminary results on a single train-validation split indicated that ATVNet outperformed V-Net and TransVNet and was therefore selected for 5-fold cross-validation and final submission. However, this comparison is limited because cross-validation was not performed on the baseline models due to time constraints in the challenge. Performance differences may be affected by the selected split, training randomness, and differences in loss functions, motivating evaluation across multiple data splits.

The 5-fold cross-validation results on lesion-positive images demonstrate consistent segmentation performance across different data splits, with a mean DSC of 54.37% and relatively low variation between folds. The observed variability in ASSD and HD95 suggests some sensitivity to the specific lesion characteristics present in each validation subset, but the overall stability of the results indicates reasonable generalization of the ATVNet model.

Although an ensemble was expected to improve segmentation and detection performance, the submitted 5-fold ensembles did not outperform the best-performing single ATVNet models. For segmentation, the ensemble achieved a lower DSC (48.29% vs. 52.25%) but slightly improved HD95 (30.28 mm vs. 32.02

mm), suggesting smoother but less accurate lesion boundaries. For detection, balanced accuracy decreased from 82.89% to 80.08% due to reduced sensitivity, while specificity remained unchanged, increasing the number of missed lesions without reducing the number of false-positive detections.

This behavior may be related to equal weighting of fold models despite differences in validation performance, as well as probability thresholding and connected-component filtering optimized for individual models rather than the ensemble. These explanations cannot be verified because only aggregate leaderboard metrics were available for the final test dataset.

Future work will include 5-fold cross-validation of both V-Net and TransVNet, systematic ablation studies to isolate the contributions of transformer skip connections, adaptive gating, and the combined loss, as well as further analysis of the learned gating coefficients across training and network levels. Statistical significance testing or confidence-interval estimation will also be used to assess the robustness of the observed improvements.

When it comes to improving lesion segmentation and detection performance, different ensemble strategies and test-time augmentation should be investigated to improve robustness. For lesion detection, a dedicated detection head could be explored as an alternative to connected-component filtering. These approaches may provide more reliable predictions and improve both segmentation and detection performance.

5 Conclusion

In this work, we presented ATVNet, a transformer-enhanced V-Net architecture with adaptive residual gating for automated msTBI lesion segmentation and detection from T1-weighted MRI. Within the AIMS-TBI26 challenge setting, ATVNet demonstrated improved preliminary segmentation performance compared with both V-Net and TransVNet, supporting the potential benefit of integrating adaptive transformer-based feature modeling with convolutional representations. The final submitted single-model approach achieved a DSC of 52.25%, an ASSD of 9.12 mm, and an HD95 of 32.02 mm for lesion segmentation, while achieving a balanced accuracy of 82.89% for lesion detection. The proposed 5-fold ensemble did not improve final performance, highlighting the challenges of applying ensemble strategies with voxel-wise prediction averaging. Further evaluation on larger and more diverse datasets, along with additional ablation studies are required to fully assess the generalizability and clinical applicability of the proposed approach.

Acknowledgments

We acknowledge Digital Government Development and Project Management Ltd. for granting us access to the Komondor HPC facility in Hungary. The project was supported by the National Research, Development and Innovation Office under grant number OTKA K137995.

References

1. Natalie V. Covington and Melissa C. Duff. Heterogeneity is a hallmark of traumatic brain injury, not a limitation: A new perspective on study design in rehabilitation research. *American Journal of Speech-Language Pathology*, 30(2S):974–985, 2021.
2. Balázs Zavadil, András Lenkovics, Bálint Szabó, Ákos Szlávecz, and Balázs Benyó. Hybrid traumatic brain injury lesion segmentation using voxel-based v-net model and connected component filtering. In Spyridon Bakas, Emily Dennis, Mehdi Astaraki, Ujjwal Baid, Gian Marco Conte, Martha Foltyn-Dumitru, Zhifan Jiang, Dominic Labella, Marie-Christin Metz, Udunna Anazodo, Maria Correia de Verdier, Florian Kofler, Hongwei Bran Li, Marius George Linguraru, and Nazanin Maleki, editors, *Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries*, pages 311–321, Cham, 2026. Springer Nature Switzerland.
3. Yang Lei, Chenqiang Zhang, Wenjing Wu, Mingchun Chen, Xiaotao Wen, Xilei He, Chenggang Bai, Siping Qin, Ying Li, and Lijing Wang. 3d fault detection method using transvnet. *Frontiers in Earth Science*, Volume 13 - 2025, 2025.
4. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
5. Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016.
6. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021.
7. Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.
8. Aims-tbi25 grand challenge. Accessed: 2025-09-23.
9. Yueyue Zhu, Xiaoyu Bai, Haotian Jiang, and Geng Chen. Blsegmamba: An optimized segmamba framework fornbsp;mstbi lesion segmentation innbsp;mri. In *Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries: MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 23, 2025, Proceedings, Part II*, page 301–310, Berlin, Heidelberg, 2026. Springer-Verlag.
10. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
11. Fabian Isensee, Paul Jäger, Jakob Wasserthal, David Zimmerer, Jens Petersen, Simon Kohl, Justus Schock, Andre Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. batchgenerators - a python framework for data augmentation, January 2020.