BioMassters: A Benchmark Dataset for Forest Biomass Estimation using Multi-modal Satellite Time-series

1 Appendix A: Top-performing models details

1.1 U-TAE Model details

We adapted U-TAE Model [1]. We consider input as an image time sequence X, organized into a four-dimensional tensor of shape $T \times C \times H \times W$, with T the length of the sequence, C the number of channels, and $H \times W$ the spatial extent.

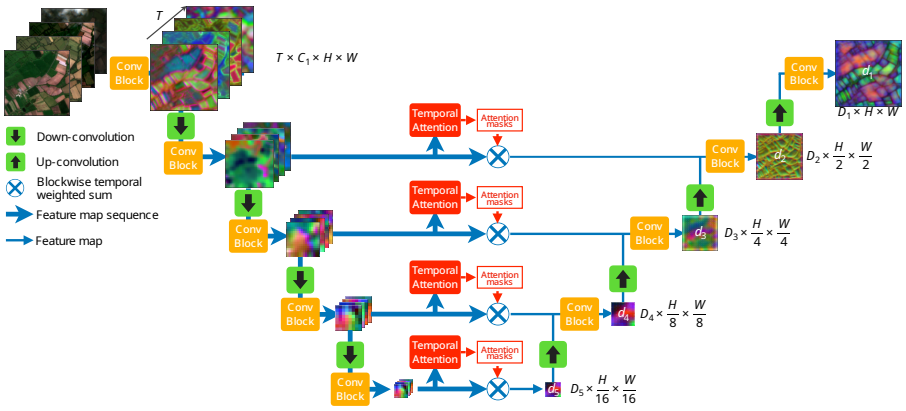

Figure 1: U-TAE Model Architecture (edited from [1])

Spatio-Temporal Encoding The model encodes a sequence X in three steps: (a) each image in the sequence is embedded simultaneously and independently by a shared multi-level spatial convolutional encoder, (b) a temporal attention encoder collapses the temporal dimension of the resulting sequence of feature maps into a single map for each level, (c) a spatial convolutional decoder produces a single feature map with the same resolution as the input images, see Figure. 1.

The adapted model has two major differences then the U-TAE Model [1]. (a) Spatial Encoding: Unlike [1] we do not use group normalization in encoder because we do not see any improvements over batch normalization. (b) Temporal Encoding : Unlike [1] we use a simplified attention-based scheme without grouping strategy which processes the temporal dimension at each feature map resolution. For each resolution map, we apply shared attention weights independently at each pixel of e^l, the feature map sequence at the level resolution l. This generates a temporal attention mask a^l for each pixel. The masks a^l at level l of the encoder are then used as weights to aggregate e^l on the temporal dimension resulting f^l map:

$$f^l = \sum_{t=1}^{T} a_t^l \odot e_t^l, \tag{1}$$

with $\odot$ term-wise multiplication with channel broadcasting.

Training details We take `tf_efficientnetv2_l_in21k`encoder from `timm` framework. The inputs to the encoder are 15-band ($C = 15$) images with a resolution of $W \times H = 256 \times 256$ from joint Sentinel-1 and Sentinel-2 satellite missions. The encoder is shared for all $T = 12$ months. We directly optimize RMSE loss for 900 epochs using AdamW optimizer with learning rate 10^{-3} and CosineAnnelingLR scheduler. We don't compute loss for high AGB values over 400. We use vertical flips, rotations, and random month dropout as augmentations. Month dropout removes images.

1.2 Swin UNETR Model details

Data Preprocessing The 0.1st and 99.9th percentiles are obtained as the lower and upper bound of inliers of each feature and AGB. Outliers are replaced by lower or upper limitations. The missing features of a specific month or modality (Sentinel-1 and Sentinel-2) are substituted by a zero array. Features are normalized using the Z-score method, and AGB is rescaled to range [0, 1]. After that, for each training sample, 4 Sentinel-1 features and 11 Sentinel-2 features of 12 months are concatenated to 4D tensor in shape [15, 12, 256, 256], AGB is in shape [1, 256, 256] as training target.

Model and Losses The adapted Swin UNEet TRansformer (Swin UNETR) is applied as the spatial-temporal regression model. The original Swin UNETR is designed for semantic segmentation of 3D medical images [2]. It has a UNet-based architecture with Swin Transformer V1 [3] as an encoder to extract multi-scale features using self-attention in an efficient shifted window partitioning scheme. We replace the attention layer of V1 block with the attention proposed in Swin Transformer V2 [4] to improve the training stability.

The adapted Swin UNETR contains a 4-stage encoder to learn multi-scale features from input, and then a 5-stage decoder upsamples feature maps to the same spatial-temporal size as input. Feature maps are averaged on the

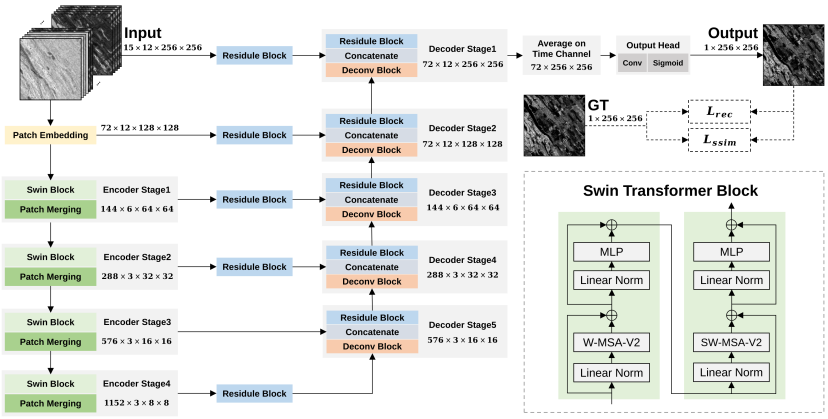

Figure 2: Swin UNETR model architecture

time channel and fed into the output head to generate AGB prediction. We apply mean absolute error as the reconstruction loss L_{rec} to measure content consistency between AGB ground truth I and prediction $\hat{I}_i$ as

$$L_{rec} = \frac{1}{n}\sum_{i=1}^{n}(I_i - \hat{I}_i) \tag{2}$$

where n is the number of pixels in AGBM.

In addition, structure similarity loss L_{ssim} is used in training to produce more visually pleasing AGBM predictions. Structure similarity comprehensively measures the differences between images in brightness, contrast, and structure, and it correlates more with human perception of image quality. L_{ssim} is calculated as

$$L_{ssim} = 1 - SSIM(I, \hat{I}) \tag{3}$$

where SSIM is implemented as [5].

The total loss L_{total} is the weighted combination of L_{rec} and L_ssim where λ_1 is the weight of L_{rec}, λ_2 is the weight of L_{ssim}.

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{ssim} \tag{4}$$

Training details The adapted Swin UNETR is trained for 100 epochs using AdamW optimizer with constant betas (0.9, 0.99) and weight decay 0.01. The learning rate increases linearly from 0.0 to 0.001 in the first 10 epochs, then it anneals in a cosine schedule to 0.0 in the last 90 epochs. The batch size is set to 4 to take full advantage of GPU A100-40G. In each training step, Volumentations-3D [6] carries out the 3D data augmentation on features and AGB targets, including vertical flipping, horizontal flipping and randomly rotating in 90 degrees with the probability of 0.1 for each operation. In the loss function, λ_1 is set to 1.0, and λ_2 is set to 0.2. The training samples are split

into 5 folds to train 5 models. For each testing sample, the average of 5 outputs is the final AGB prediction.

1.3 UNET++ Model details

Data Preprocessing Sentinel-1 (S1) and Sentinel-2 (S2) imagery were preprocessed into six cloud-free median composites (Table 1) to reduce data dimensionality while preserving the maximum amount of information. S1 imagery was reduced to seasonal median composites and then stacked. Similarly, S2 imagery was first cloud masked using a 50% threshold of the cloud probability layer, and then also reduced to seasonal median composites and stacked. For two composites (i.e. 2SI and 4SI) multiple vegetation (e.g. NDVI for S2) and spectral indices (e.g. VV/VH ratio for S1) were also generated.

Training Details The data consisting of 8692 stacked images were divided into the train (98.9%, n=8596) and validation (1.1%, n=96) datasets. This was done in a stratified manner by binning AGB values into four 25th-percentile bins. S1, S2, and AGBD images were also standardized using mean and standard deviation calculated on the train set. Then 15 models were trained using a UNet++ architecture in combination with various encoders and attention blocks (e.g. scse) and median composites (Figure 1). The pre-trained on imagenet dataset models were further trained with multiple augmentations (e.g. flips and rotations), batch size of 32, AdamW optimizer with 0.001 initial learning rate, weight decay of 0.0001, and a ReduceLROnPlateau scheduler. UNet++ models were optimized using a Huber loss to reduce the effect of outliers in the data for 200 epochs. To improve the performance of each UNet++ model they were further fine-tuned (after freezing pre-trained encoder weights and removing augmentations) for another 100 epochs. For each UNet++ model, the average of the two best predictions was used for further ensembling and evaluation using a root-mean-square error (RMSE). The ensemble of all 15 models using a weighted average was used for the final evaluation of the test set (n=2773). The training of the ensemble model took approximately 360 hours (i.e. 24 hours/model) on a single NVIDIA GeForce RTX 3090 (24 GB VRAM).

References

[1] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *ICCV*, 2021.

[2] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, January 2022. arXiv:2201.01266 [cs, eess].

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, October 2021. IEEE.

[4] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution.

[5] Loss Functions for Image Restoration With Neural Networks | IEEE Journals & Magazine | IEEE Xplore.

[6] Roman Solovyev, Alexandr A. Kalinin, and Tatiana Gabruseva. 3D convolutional neural networks for stalled brain capillary detection. *Computers in Biology and Medicine*, 141:105089, February 2022.