A multimodal species distribution model incorporating remote sensing images and environmental features

xiaojuan zhang
Chengdu University of Technology
Yongxiu Zhou
Chengdu University of Technology
Peihao Peng (peihaopeng1963@126.com)
Chengdu University of Technology
Guoyan Wang
Chengdu University of Technology

Research Article

Keywords: deep learning, feature fusion, high-resolution remote sensing images, multimodal, species distribution models, Transformer network structure

Posted Date: April 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1599783/v1

Abstract

Context

Species distribution models (SDMs) are critical in conservation decision-making and ecological or biogeographical inference. However, species distribution models are typically limited to using a single source of information as input to the model. This leaves us with a poor understanding of species distribution and landscape ecology processes.

Objectives

The aims are to explore the feasibility of multimodal models for biodistribution prediction and address this limitation by developing a multimodal species distribution model that can simultaneously input multiple information sources to investigate a solution to the species distribution model's lack of accuracy with a single information source.

Methods

The proposed “multimodal species distribution model” used ResNet50 and Transformer network structures as the backbone for multimodal data modeling. (i) The ResNet50 network structure was used to compare different input sources' effects on the model's accuracy (SDMM-Net v1 and SDMM-Net v2). (ii) The Transformer network structure was used to investigate the effect of different feature fusion structures and feature fusion methods on the accuracy of the model (PreFuzeMM-Swin, PostFuzeMM-Swin and MidFuzeMM-Swin).

Results

Our model's accuracy was validated using the GEOLIFE2020 dataset, and we achieved the state-of-the-art (SOTA). We discovered that the prediction accuracy of the multimodal species distribution model with multiple data sources of remote sensing images, environmental variables, latitude and longitude information as inputs (29.56%) was higher than that of the model with only remote sensing images or environmental variables as inputs (25.72% and 21.68%, respectively). We also discovered that using a Transformer network structure to fuse data from multiple sources can significantly improve the accuracy of multimodal models, and the PreFuzeMM-Swin network model accuracy is higher than that of the PostFuzeMM-Swin network and MidFuzeMM-Swin network.

Conclusions

We proposed the first multimodal model with multiple sources of information as input for species distribution prediction to advance the research progress of multimodal models in the field of ecology.

1 Introduction

Species distribution models (SDMs) have become a fundamental tool in ecology, biodiversity conservation, biogeography and natural resource management (Franklin, 2013; Guillera-Arroita et al., 2015; Guisan et al., 2013; Guisan & Thuiller, 2005; Newbold, 2010). Traditional SDMs typically correlate the presence (or presence/absence) of species at multiple sites with relevant environmental covariates (temperature, precipitation, altitude, land cover, soil type, etc.) to estimate habitat preferences or predict distributions; these outputs are commonly used to inform ecological and biogeographical theory as well as conservation decisions (Bekessy et al., 2009; Ferrier et al., 2002; Keith et al., 2014; Pearce & Lindenmayer, 1998; Re ~ it Ak ~ akaya et al., 1995). Researchers in the field of ecology and geology use remote sensing images for automatic classification, providing a convenient means of classifying forests and classifying land use types. Although with the continuous development of remote sensing technology and the improvement of remote sensing image accuracy, many researchers started to try to incorporate remote sensing images as variables in the species distribution models (Brown et al., 2014; Cerrejón et al., 2020; K. S. He et al., 2015; Sumsion et al., 2019; B. Zhang et al., 2020), they only extracted the point information corresponding to species in remote sensing images and did not use the image information, which is still essentially the same as using environmental covariates to model, for example, Cerrejón et al (2020) (Bhattarai et al., 2020) used remote sensing data combined with species distribution models to predict and map bryophyte communities and diversity patterns in a 28436 km^2 northern forest in Quebec (Canada). Bhattarai et al (2020) (Scholl et al., 2020) used Sentinel-1 synthetic aperture radar (SAR) and Sentinel-2 multispectral images in combination with a total of 191 covariates from northern New Brunswick, Canada several site variables and mapped the distribution and abundance of spruce glowworm (SBW) host species using a random forest.

Just as the proposal of the imagenet dataset (Deng et al., 2009) has greatly advanced the development of deep learning techniques, many researchers in the field of remote sensing image-based SDM research have started to produce and open source their own datasets for comparing the performance and accuracy of different models, while advancing the technology in the field. Marconi et al (2019) held a competition for tree species classification based on remote sensing images, which offered three tracks on canopy segmentation, tree alignment and species classification, contributing to the development of remote sensing data for ecological and biological methods (Marconi et al., 2019). Deneu et al (2020) organized a competition called GeoLifeCLEF to study the relationship between the environment and the possible occurrence of species, a dataset that collected 1.9 million observations with their corresponding remote sensing data, the largest remote sensing dataset to date open-sourced for studying species distribution (Deneu et al., 2020.).

As deep learning algorithms make breakthroughs in various unimodal tasks, researchers are beginning to focus on multimodal tasks that are closer to the real world. Currently, multimodal models are beginning to be widely used in areas such as image captioning, Text-to-image Generation, Visual Question Answering, and Visual Reasoning (C. Zhang et al., 2019). Jun et al (2020) proposed a multimodal deep attention network model based on a deep self-attention network applied to image captioning, which can input an image and output the text description corresponding to this image, and they won the MSCOCO Image

Captioning Challenge in the then They won first place in the MSCOCO Image Captioning Challenge (Yu et al., 2020). Tao et al (2018) proposed a deep attentional multimodal similarity model to train a graphical text generator, which can supplement missing details in images based on the input text descriptions, and improved the best score by 14.14% on the CUB datasets, while improving the best score by 170.25% on the COCO datasets (Xu et al., 2018). Ben-Younes et al (2017) proposed a multimodal model based on Tucker decomposition for the visual question and answer domain (Ben-Younes et al., 2017). Such models typically input a textual description of a question and a corresponding image, and output the answer to that question. Ham et al (2017) propose a model of Dual Attention Networks (DAN) that uses visual and textual attention mechanisms to capture the interactions between vision and language, which can be used for multimodal inference and matching (Nam et al., 2017).

Before the emergence of the Transformer model (Vaswani et al., 2017), the backbone of multimodal modeling was primarily convolutional neural networks, which were first proposed for natural language processing (nlp) with good results, and then the Transformer model and its variants were applied to the field of vision, also with good results, and now some researchers have began to apply the Transformer model to multimodal domains. Lu J et al (2019) proposed the ViLBERT model for visual question and answer tasks, which used the bert model as the main architecture and achieved the best results on all four visual response datasets tested (Lu et al., 2019). Chen Y et al (2019) proposed a joint image text representation network called UNITER, which used a multilayer Transformer mechanism and achieved the best results on six V + L datasets (Chen et al., 2019). Li C et al (2021) proposed a new pre-training method SemVLP using the Transformer network as a pre-training model, which is effective in being able to align cross-modal representations with different semantic granularities (Li et al., 2021). We need to note that most of the current multimodal models based on Transformer are used for tasks that deal with the combination of vision and language, while models for species distribution prediction have not been proposed yet.

Although in the direction of species distribution prediction research, multiple modal information including environmental variables and remote sensing images can be obtained for a single sample point, so far we have not found researchers using multimodal models to study species distribution. Therefore, we propose a multimodal model based on Transformer for species distribution prediction to explore whether the accuracy of a species distribution prediction model using remote sensing images and environmental variables as inputs is higher than that of a model using only remote sensing images or environmental variables as inputs? How much does the different structural fusion methods of the Transformer-based backbone species distribution model affect the model results?

2 Materials And Methods

2.1 Dataset

We used GeoLifeCLEF 2020 (Cole et al., 2020) as the dataset for training and testing, which collected data from a total of 1.9 million sample points in the USA and France, including a total of 31,435

categories of plant and animal information. Each of these sample points has a latitude and longitude information and corresponding high resolution remote sensing image (Fig. 1) and is 256x256 in size and have a spatial resolution of 1 meter per pixel. The dataset also provides 19 climate data and 8 soil type information for each sample point (Table 1). To validate the effectiveness of the different algorithms, the authors of the GeoLifeCLEF 2020 dataset randomly selected 2.5% of the data as test data and the rest of the data as training data.

Table 1
Summary of the low-resolution environmental variable rasters provided. The first 19 rows correspond to the bio-climatic variables from WorldClim (Hijmans et al., 2005). The last 8 rows correspond to the soil variables from SoilGrid (Hengl et al., 2017).

Name	Description	Resolution
bio_1	Annual Mean Temperature	30 arcsec
bio_2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	30 arcsec
bio_3	Isothermality (bio_2/bio_7) (* 100)	30 arcsec
bio_4	Temperature Seasonality (standard deviation *100)	30 arcsec
bio_5	Max Temperature of Warmest Month	30 arcsec
bio_6	Min Temperature of Coldest Month	30 arcsec
bio_7	Temperature Annual Range (bio_5-bio_6)	30 arcsec
bio_8	Mean Temperature of Wettest Quarter	30 arcsec
bio_9	Mean Temperature of Driest Quarter	30 arcsec
bio_10	Mean Temperature of Warmest Quarter	30 arcsec
bio_11	Mean Temperature of Coldest Quarter	30 arcsec
bio_12	Annual Precipitation	30 arcsec
bio_13	Precipitation of Wettest Month	30 arcsec
bio_14	Precipitation of Driest Month	30 arcsec
bio_15	Precipitation Seasonality (Coefficient of Variation)	30 arcsec
bio_16	Precipitation of Wettest Quarter	30 arcsec
bio_17	Precipitation of Driest Quarter	30 arcsec
bio_18	Precipitation of Warmest Quarter	30 arcsec
bio_19	Precipitation of Coldest Quarter	30 arcsec
orcdrc	Soil organic carbon content (g/kg at 15cm depth)	250 m
phihox	Ph × 10 in H20 (at 15cm depth)	250 m
cecsol	cation exchange capacity of soil in cmolc/kg 15cm depth	250 m
bdticm	Absolute depth to bedrock in cm	250 m
clyppt	Clay (0–2 micro meter) mass fraction at 15cm depth	250 m
sltppt	Silt mass fraction at 15cm depth	250 m
(GeoLifeCLEF 2020)		

Name	Description	Resolution
sndppt	Sand mass fraction at 15cm depth	250 m
bldfie	Bulk density in kg/m3 at 15cm depth	250 m
(GeoLifeCLEF 2020)		

2.2 Multimodal models

Species distribution models (SDMs) can be classified into three types based on the type of input data to the model: those based on environmental variable algorithms, those based on remote sensing image algorithms, and those based on multimodal data. Environmental variable algorithms-based models use temperature, precipitation, longitude, latitude, elevation, slope, slope direction, and soil type as input to achieve the final classification via machine learning. Remote sensing-based classification models take remote sensing images (optical, radar, etc.) as input and classify the data using machine learning. They are currently frequently employed in the disciplines of vegetation classification and geology. Multimodal data, on the other hand, has received little attention, and the work that has been reviewed thus far mostly extracts remote sensing information relating to species points (surface temperature, vegetation indices, elevation, and so on) for modeling to achieve categorization. Models that use only tabular data as input, as shown in Fig. 2, are typically selected from machine learning models such as SVM and Random Forest. Because such models only learn scalar data, such as environmental variables, the accuracy of trained models is the lowest of the three types of models. Deep learning models that are commonly used take remote sensing images as input. In terms of test accuracy, we can see from the results of the GeoLifeCLEF 2021 competition (Lorieul et al., 2021) that deep learning models with remote sensing images as input outperformed machine learning models with environmental variable data as input. By analogy with the human information processing system, we know that the more comprehensive the information about the sample points, the more accurate we can judge the classification results of the samples. As a result, we propose a multimodal model that uses both remote sensing data and environmental variable data as input for SDM modeling.

2.2.1 Multimodal model based on ResNet50

Although deep learning species distribution prediction models based on remote sensing images outperformed machine learning approaches that only use environmental variables for prediction (Lorieul et al., 2021), completely eliminating the use of environmental variables can result in a significant loss of useful information when making predictions. To allow the model to extract useful information from both remote sensing images and environmental variable information, we created a multimodal model capable of simultaneously inputting image information and environmental variable information, as well as extracting and fusing remote sensing image information and environmental variable information from sample points.

Our proposed multimodal model based on ResNet50 uses ResNet (K. He et al., 2016) as the main network structure for extracting remote sensing image features. By introducing a network structure of residuals, the ResNet network is used to solve the problem of vanishing gradient caused by too deep layers of a convolutional neural network. Since its inception, the ResNet network has been widely used for classification, detection, segmentation, and other tasks. It is one of the most common backbone networks. The ResNet50 network structure consists of five stages, where stage 0 serving as the pre-processing layer of the input image, containing a 7x7 convolution and a 3x3 max pool layer, and the last four stages as the feature extraction layer of the image, all of which consist of Bottleneck, and their structures are relatively similar, and each stage begins with each layer using Bottleneck down for feature sampling. Stages 1, 2, 3 and 4 consist of 3, 4, 6 and 3 Bottlenecks respectively. Finally, an average pool layer produces a 2048-dimensional feature vector.

The SDMM-Net network structure we propose has two input branches (Fig. 3). One of the branches takes an RGB remote sensing image as input to provide image information to the model for that sample point, and this input is passed through a ResNet network (K. He et al., 2016) to produce a 2048-dimensional vector. The other branch feeds information on the 27 environmental variables at that sample site into the model. To facilitate network processing of the data, each environmental variable feature was normalized before being entered into the network. We start with a 27-dimensional vector, then pass it through an FC layer to get a 2048-dimensional vector, and then concat (fuse) the two 2048-dimensional vectors from the two branches to obtain a 4096-dimensional feature vector. Finally, the 4096-dimensional vector is followed by an FC layer and a softmax layer to produce a 31435-dimensional result (a total of 31435 species classes in the training dataset).

2.2.2 Multimodal model based on Swin Transformer

The Swin Transformer (Liu et al., 2021) model is an attention mechanism-based model structure proposed by Microsoft Research Asia. Once proposed, the network achieved SOTA on different datasets and is the most effective attention mechanism model at this stage. There are four versions of the official proposed Swin Transformer model, Swin-T, Swin-S, Swin-B and Swin-L. Because the computational effort of the Swin-T model is comparable to that of ResNet50, we designed a multimodal attention mechanism model based on the Transformer model using the Swin-T model as the backbone network in order to facilitate comparison of the Transformer-based multimodal model with the ResNet50-based multimodal model.

The Swin-T model has a hierarchical design with four stages, each of which reduces the resolution of the input feature map and expands the receptive field layer by layer, similar to how a CNN does. A Patch Embedding was performed during the image pre-processing stage to cut the image into individual blocks and embed them into the Embedding. Each stage consists of a Patch Merging module and several Block modules, where the Patch Merging module primarily reduces the image resolution at the beginning of each stage, while the Block consists of LayerNorm, MLP, Window Attention and Shifted Window Attention, which are used to extract the feature values of the image.

Current multimodal data fusion methods can usually be divided into data fusion, feature fusion and model fusion. We designed three types of Transformer multimodal models based on the Swin Transformer model structure according to different fusion methods (Fig. 4–6).

Pre-fuze-multimodel

The Pre-fuze-multimodel uses the Swin-T backbone network to extract the features of the data after feature fusion. After reading the data, the model first encodes the image data into a feature vector using the Patch Partition module, then normalizes the input environmental variable features using the feature norm module, then fuses the image feature vector and environmental variable features using the feature fusion module (feature norm), followed by the data fused input to the Swin T network structure, and finally outputs the classification results (Fig. 4).

Post-fuze-multimodel

After reading the original data, the Post-fuze-multimodel extracts the features of the remote sensing image using the Swin-T backbone network; at the same time, it uses the feature norm layer to normalize the environmental variable features and uses the FC layer to up-dimension the feature vector; it then performs feature fusion on the image features and environmental variable features; finally, it classifies the fused feature values (Fig. 5).

Mid-fuze-multimodel

The Mid-fuze-multimodel will fuse image features and environment variable features at each stage of Swin-T (Fig. 6). Because the network structure of Swin-T performs a down-sampling-like operation on the image at each stage, the feature input dimensions of the four stages here are 3136, 784, 196, and 49 dimensions respectively. In order to enable the image features and the environment variable features to be fused, the environment variable features must be up-dimensioned to the same dimension using the FC layer prior to the fusion of each stage structure.

Feature fusion methods

Feature addition and feature concatenation are two commonly used feature fusion methods. The feature addition method necessitates unifying two feature vectors to the same dimension and then directly adding the two vectors. The lengths of the two feature values are typically unified by up-dimensioning the lower dimensional feature values through the FC layer to the same dimensional length as the higher dimensional vector. Typically, feature concatenation is accomplished by directly concatenating an N-dimensional vector with an M-dimensional vector, yielding an N + M-dimensional vector. and we experimented with Transformer-based networks using two feature fusion methods. respectively.

3 Results

We used the top-30 correctness rate as the model evaluation metric to facilitate cross-sectional comparisons with other methods (Cole et al., 2020). Top-30 accuracy refers to using the model to predict the 30 species with the highest probability of occurring at that sample point, and a prediction is considered correct if the correct result is included among the 30 predicted results. We assume that each sample point i has a unique label C_i and the K categories with the highest confidence in the model prediction are $c_{i1}, \ldots c_{iK}$. If there exists a certain prediction value $c_{ij} = C_i$ the model prediction for that sample, is considered to be correct. We assume that

$$d_{ij} = d\left(c_{ij}, C_i\right) = \begin{cases} 1, if c_{ij} \neq C_i \\ 0, othersise \end{cases}$$

The following is the formula for calculating Top-K accuracy:

Precision@K=1-\frac{1}{N}\sum _{i=1}^{N}\begin{array}{c}min\\ K\end{array}{d}_{iK}
where N is the total number of test samples and K is set to 30.

3.1 Test results

To investigate the variation in accuracy of SDMs with different data type inputs, as well as the variation in model accuracy caused by different ways of fusing model structures, we propose the ResNet50 network model (including SDMM-Net v1 and SDMM-Net v2) and the Swin-T network model (including PreFuzeMM-Swin, PostFuzeMM-Swin and MidFuzeMM-Swin). Among the various types of models, we list the test results for the model with the highest accuracy (Table 2).

The test codes for the Top-30 most present species and Random forest models were taken directly from the code repository provided by E. Cole et al (https://github.com/maximiliense/GLC) (Cole et al., 2020). The Top-30 most present species prediction method is used as a baseline method, which directly uses the 30 most frequent species in the training data as the prediction result for each test sample point, this method uses entirely statistical methods to make predictions, and the predicted results are only statistically significant. The Random forest model is a popular machine learning model that uses a statistical approach to condition on environmental variables, predicting sample points with similar environmental variables as likely to have the same species. The test results for the ss model (using remote sensing images as input) from the test results provided by S. Seneviratne (Liu et al., 2021), which is the model with the highest accuracy among the public methods, the model is trained using RGB remote sensing images as training data, using a deep learning contrast representation learning approach. The rest of the models from the test results of our model based on the public test set in the GEOLIFE dataset.

According to the test results, we can observe that (1) the highest accuracy of our proposed PreFuzeMM-Swin model is 29.37%, the accuracy of this model has exceeded the existing ss model with the highest accuracy by 3 percentage points, improving the highest accuracy by 11.6%. This indicates that a multimodal model with multiple data types as input can significantly improve the accuracy of the existing

SDMs. (2) The accuracy of the random forest model using only environmental variables as training data was 21.68%, and the accuracy of the ResNet50 model using only remote sensing images as training data was 25.72%, indicating that both environmental variables and remote sensing data can be used as valid training data for SDM. (3) The accuracy of the SDMM-Net v2 model, which uses remote sensing images, environmental features and latitude and longitude as input data, is higher than that of the SDMM-Net v1 model, which uses only remote sensing images and environmental variable features as input data, indicating that adding latitude and longitude as input data can effectively improve the accuracy of the model. (4) The accuracy of the PreFuzeMM-Swin model with Swin-T as the backbone network outperforms the SDMM-Net v2 model with ResNet50 as the backbone network (29.56%>26.4%), indicating that the accuracy of the multimodal model with Transformer as the backbone network is higher than that of the multimodal model with the traditional ResNet structure. (5) PreFuzeMM-Swin model has a higher model accuracy than the MidFuzeMM-Swin model and PostFuzeMM-Swin model.

Table 2
Comparison of test results

model	**input data**			**top-30 accuraccy**
	RGB images	**environmental variable**	**longitude and latitude**	
Top-30 most present species	□	□	□	4.36
Random forest	□	✓	□	21.68
ss model	✓	□	□	26.32
ResNet50	✓	□	□	25.72
SDMM-Net v1	✓	✓	□	25.8
SDMM-Net v2	✓	✓	✓	26.4
Swin-T	✓	□	□	25.13
PreFuzeMM-Swin	✓	✓	✓	**29.56**
PostFuzeMM-Swin	✓	✓	✓	26.07
MidFuzeMM-Swin	✓	✓	✓	29.37

3.2 Comparison of different feature fusion structures

To compare the effects of different feature fusion structures on the accuracy of the multimodal models, we compared the final test accuracies of three models with Swin-T as the backbone network, PreFuzeMM-Swin, PostFuzeMM-Swin and MidFuzeMM-Swin, all of which were fused by choosing the feature addition method (Table 2, Fig. 7). We came to the following conclusions after 20 rounds of

training our model using the feature addition approach to feature fusion. (1) Regardless of the fusion structure, the accuracy of the multimodal model was higher than that of the base model using only remote sensing images as input. (2) PreFuzeMM-Swin has the highest model accuracy, which is slightly higher than MidFuzeMM-Swin. (3) The model accuracy of PostFuzeMM-Swin is the lowest.

3.3 Comparison of different feature fusion methods

We selected the two most common feature fusion methods: feature concatenation and feature addition for testing. According to the test results (Table 3), we found that the fusion accuracy of feature concatenation is much lower than that of feature addition when the Transformer is the main network, and even lower than that of the network without feature fusion, which may be due to the unique attention mechanism of the Transformer network.

Table 3 Comparison of different feature fusion methods

Fuze architecture / Fuze pattern	Pre	Post	Mid	base
addition	29.558	26.066	29.369	25.133
concatenation	7.367	20.977	23.505	

4 Conclusion And Discussion

In this paper, we propose a multimodal network model for species distribution prediction that can extract both image and environmental features from sample points. We called the network species distribution multimodel net (PreFuzeMM-Swin), tested it on the GeoLifeCLEF 2020 dataset, and obtained SOTA results. Our experiments demonstrate that multimodal models that incorporate both image information and environmental variables predict species distributions better than models using only image information or single variable inputs using only environmental variables. Furthermore, using Transformer-based multimodal models for SDM can significantly improve the accuracy of existing models. While different model structure fusion methods have a greater impact on the model accuracy, we found that the model accuracy of the PreFuzeMM-Swin network is higher than that of the PostFuzeMM-Swin network and MidFuzeMM-Swin network, and the model accuracy of the feature addition is higher than that of the feature concatenation.

At the same time, we discovered that the current species distribution prediction model's overall accuracy is still very low. The main reason for this is in the experimental data, specifically the uneven distribution of categories in the datasets we used, and secondly, the datasets we used are all single category labels, whereas the reality is that a sample point can correspond to multiple biological species. In a subsequent

study, we may consider using novel methods to address the existing issues of unbalanced data categories and single labels. Furthermore, we can think about adding more information to the remote sensing images and environmental data. As a result, in future research, we can concentrate on creating better datasets and then validating the performance of existing algorithms on this basis.

Declarations

ACKNOWLEDGEMENTS

We thank the subject editor and anonymous reviewers for their valuable comments and feedback on this manuscript. This work was funded by the National Natural Science Foundation of P.R. China (41671432), National Natural Science Foundation of P.R. China (31860123), the Second Tibetan Plateau Scientific Expedition and Research Program of P. R. China (2019QZKK0301), and the Biodiversity Survey and Evaluation of the Ministry of Ecology and Environment of P. R. China (2019HJ2096001006)

AUTHORS' CONTRIBUTIONS

Xiaojuan Zhang conceived the idea; Yongxiu Zhou trained the model. Peihao Peng provided the foundation, All authors participated in the writing of multiple drafts.

COMPETING INTERESTS

All authors declare no competing interests.

References

1. Bekessy, S. A., Wintle, B. A., Gordon, A., Fox, J. C., Chisholm, R., Brown, B., Regan, T., Mooney, N., Read, S. M., & Burgman, M. A. (2009). Modelling human impacts on the Tasmanian wedge-tailed eagle (Aquila audax fleayi). Biological Conservation, *142*(11), 2438–2448. https://doi.org/10.1016/j.biocon.2009.05.010
2. Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). *MUTAN: Multimodal Tucker Fusion for Visual Question Answering*.
3. Bhattarai, R., Rahimzadeh-Bajgiran, P., Weiskittel, A., Meneghini, A., & Maclean, D. (2020). *Spruce budworm tree host species distribution and abundance mapping using multi-temporal Sentinel-1 and Sentinel-2 satellite imagery*.
4. Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & Gibb, H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. Methods in Ecology and Evolution, *5*(4), 344–352. https://doi.org/10.1111/2041-210X.12163
5. Cerrejón, C., Valeria, O., Mansuy, N., Barbé, M., & Fenton, N. J. (2020). Predictive mapping of bryophyte richness patterns in boreal forests using species distribution models and remote sensing data. *Ecological Indicators*, *119*. https://doi.org/10.1016/j.ecolind.2020.106826

6. Chen, Y.-C., Li, L., Yu, L., Kholy, A. el, Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2019). *UNITER: UNiversal Image-TExt Representation Learning*. http://arxiv.org/abs/1909.11740
7. Cole, E., Deneu, B., Lorieul, T., Servajean, M., Botella, C., Morris, D., Jojic, N., Bonnet, P., & Joly, A. (2020). *The GeoLifeCLEF 2020 Dataset*. http://arxiv.org/abs/2004.04192
8. Deneu, B., Lorieul, T., Cole, E., Servajean, M., Botella, C., Bonnet, P., & Joly, A. (2020). *Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF)*. http://osr-cesbio.ups-tlse.fr/
9. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
10. Ferrier, S., Drielsma, M., Manion, G., & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. In *Biodiversity and Conservation* (Vol. 11).
11. Franklin, J. (2013). Species distribution models in conservation biogeography: Developments and challenges. In *Diversity and Distributions* (Vol. 19, Issue 10, pp. 1217–1223). https://doi.org/10.1111/ddi.12125
12. Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., Mccarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography, *24*(3), 276–292. https://doi.org/10.1111/geb.12268
13. Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. In *Ecology Letters* (Vol. 8, Issue 9, pp. 993–1009). https://doi.org/10.1111/j.1461-0248.2005.00792.x
14. Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., Mcdonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. Ecology Letters, *16*(12), 1424–1435. https://doi.org/10.1111/ele.12189
15. He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M. N., Schmidtlein, S., Turner, W., Wegmann, M., & Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? Remote Sensing in Ecology and Conservation, *1*(1), 4–18. https://doi.org/10.1002/rse2.7
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
17. Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE, *12*(2). https://doi.org/10.1371/journal.pone.0169748
18. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, *25*(15),

1965–1978. https://doi.org/10.1002/joc.1276

19. Keith, D. A., Elith, J., & Simpson, C. C. (2014). Predicting distribution changes of a mire ecosystem under future climates. Diversity and Distributions, *20*(4), 440–454. https://doi.org/10.1111/ddi.12173
20. Li, C., Yan, M., Xu, H., Luo, F., Wang, W., Bi, B., & Huang, S. (2021). *SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels*. http://arxiv.org/abs/2103.07829
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. http://arxiv.org/abs/2103.14030
22. Lorieul, T., Cole, E., Deneu, B., Servajean, M., Bonnet, P., & Joly, A. (2021). *Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images*. http://ceur-ws.org
23. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*.
24. Marconi, S., Graves, S. J., Gong, D., Nia, M. S., le Bras, M., Dorr, B. J., Fontana, P., Gearhart, J., Greenberg, C., Harris, D. J., Kumar, S. A., Nishant, A., Prarabdh, J., Rege, S. U., Bohlman, S. A., White, E. P., & Wang, D. Z. (2019). A data science challenge for converting airborne remote sensing data into ecological information. PeerJ, *2019*(2). https://doi.org/10.7717/peerj.5843
25. Nam, H., Ha, J.-W., Corp, N., & Kim, J. (2017). *Dual Attention Networks for Multimodal Reasoning and Matching*.
26. Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography, *34*(1), 3–22. https://doi.org/10.1177/0309133309355630
27. Pearce, J., & Lindenmayer, D. (1998). *Bioclimatic Analysis to Enhance Reintroduction Biology of the Endangered Helmeted Honeyeater (Lichenostomus melanops cassidix) in Southeastern Australia* (Vol. 6, Issue 3).
28. Re ~ it Ak ~ akaya, H., Mccarthy, M. A., & Pearce, J. L. (1995). LINKING LANDSCAPE DATA WITH POPULATION VIABILITY ANALYSIS: MANAGEMENT OPTIONS FOR THE HELMETED HONEYEATER Lichenostomus melanops cassidix. In *Biological Conservation* (Vol. 73).
29. Scholl, V. M., Cattau, M. E., Joseph, M. B., & Balch, J. K. (2020). Integrating national ecological observatory network (NEON) airborne remote sensing and in-situ data for optimal tree species classification. Remote Sensing, *12*(9). https://doi.org/10.3390/RS12091414
30. Sumsion, G. R., Bradshaw, M. S., Hill, K. T., Pinto, L. D. G., & Piccolo, S. R. (2019). Remote sensing tree classification with a multilayer perceptron. PeerJ, *2019*(2). https://doi.org/10.7717/peerj.6101
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762
32. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). *AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks*. https://github.com/taoxugit/AttnGAN.

33. Yu, J., Li, J., Yu, Z., & Huang, Q. (2020). Multimodal Transformer with Multi-View Visual Representation for Image Captioning. IEEE Transactions on Circuits and Systems for Video Technology, *30*(12), 4467–4480. https://doi.org/10.1109/TCSVT.2019.2947482
34. Zhang, B., Zhao, L., & Zhang, X. (2020). Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. *Remote Sensing of Environment*, *247*. https://doi.org/10.1016/j.rse.2020.111938
35. Zhang, C., Yang, Z., He, X., & Deng, L. (2019). *Multimodal Intelligence: Representation Learning, Information Fusion, and Applications*. https://doi.org/10.1109/JSTSP.2020.2987728

Figures

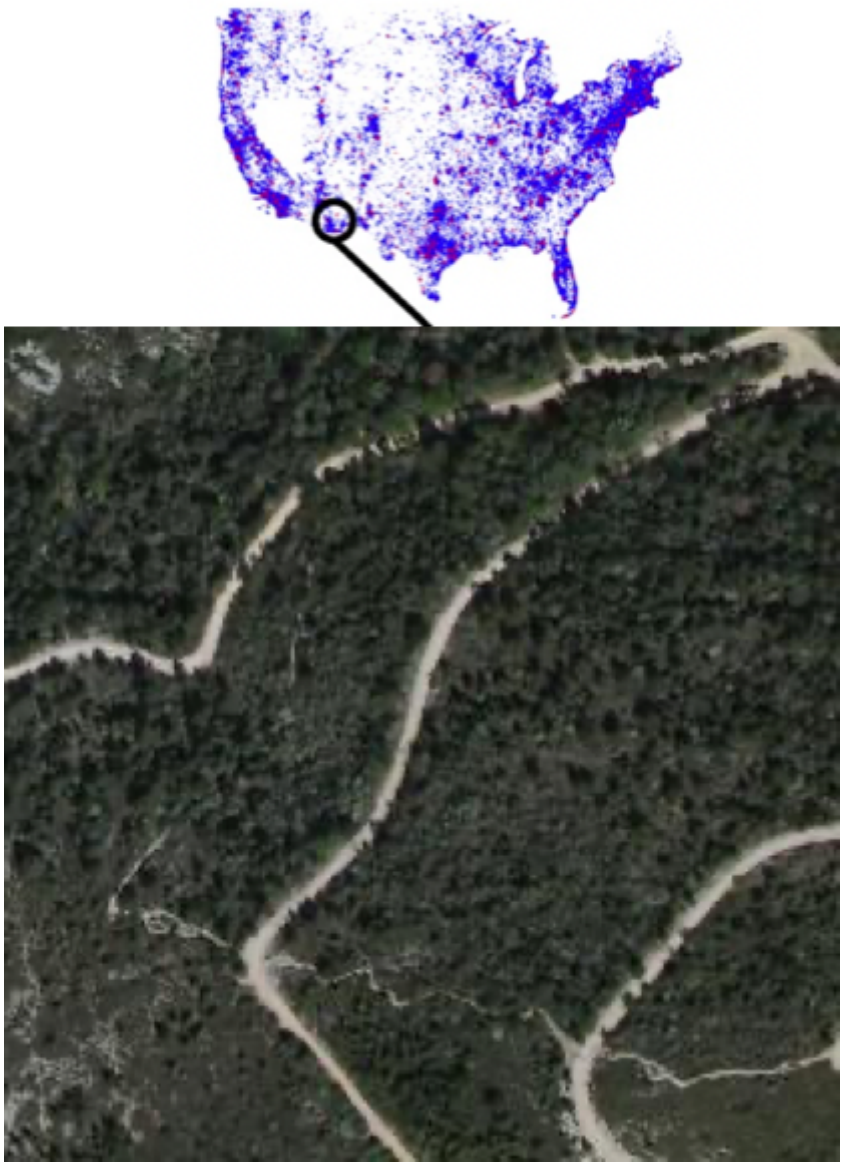

Figure 1

Each species observation is paired with high-resolution covariates (Lorieul et al., 2021).

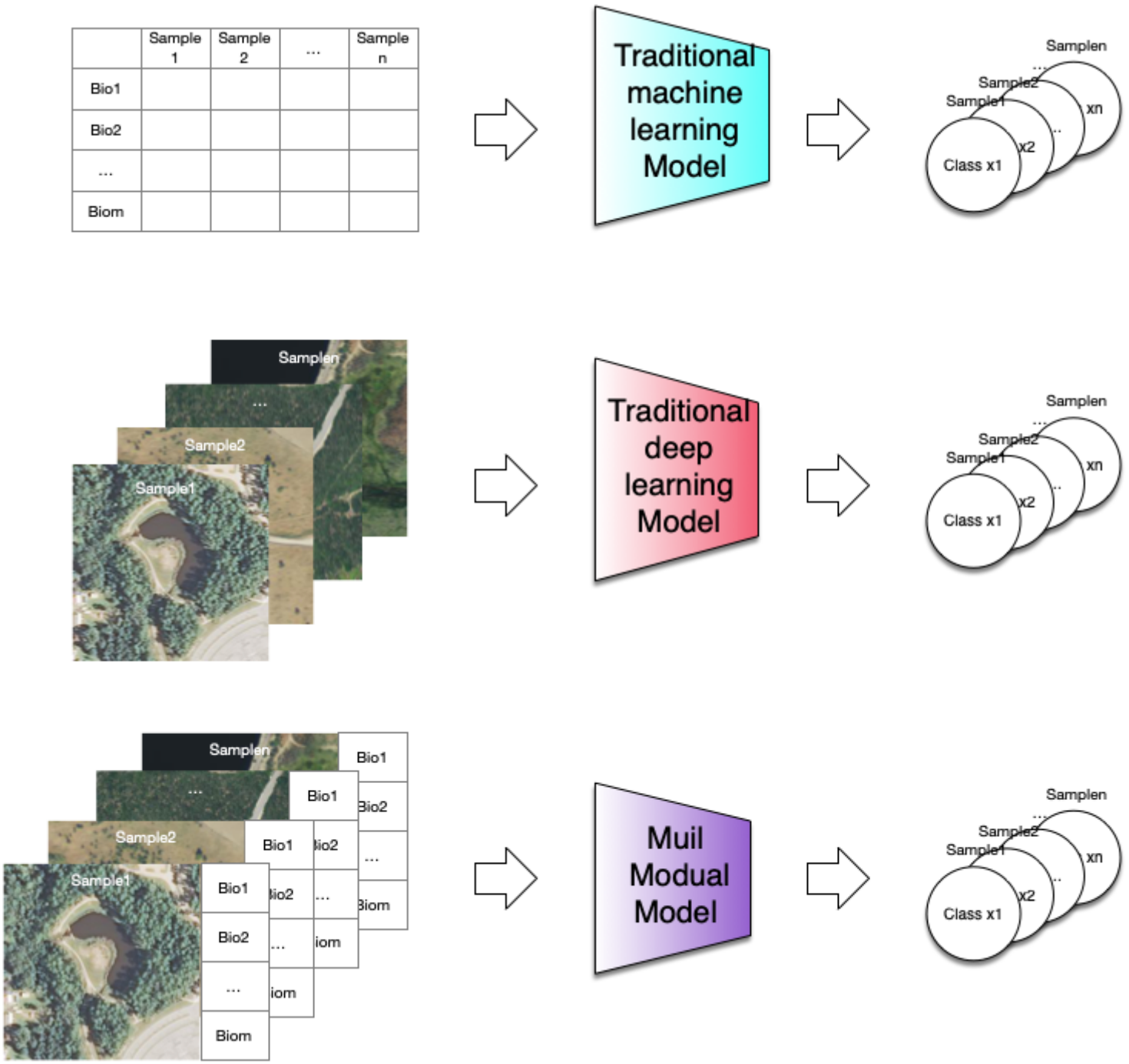

Figure 2

Classification of SDM models

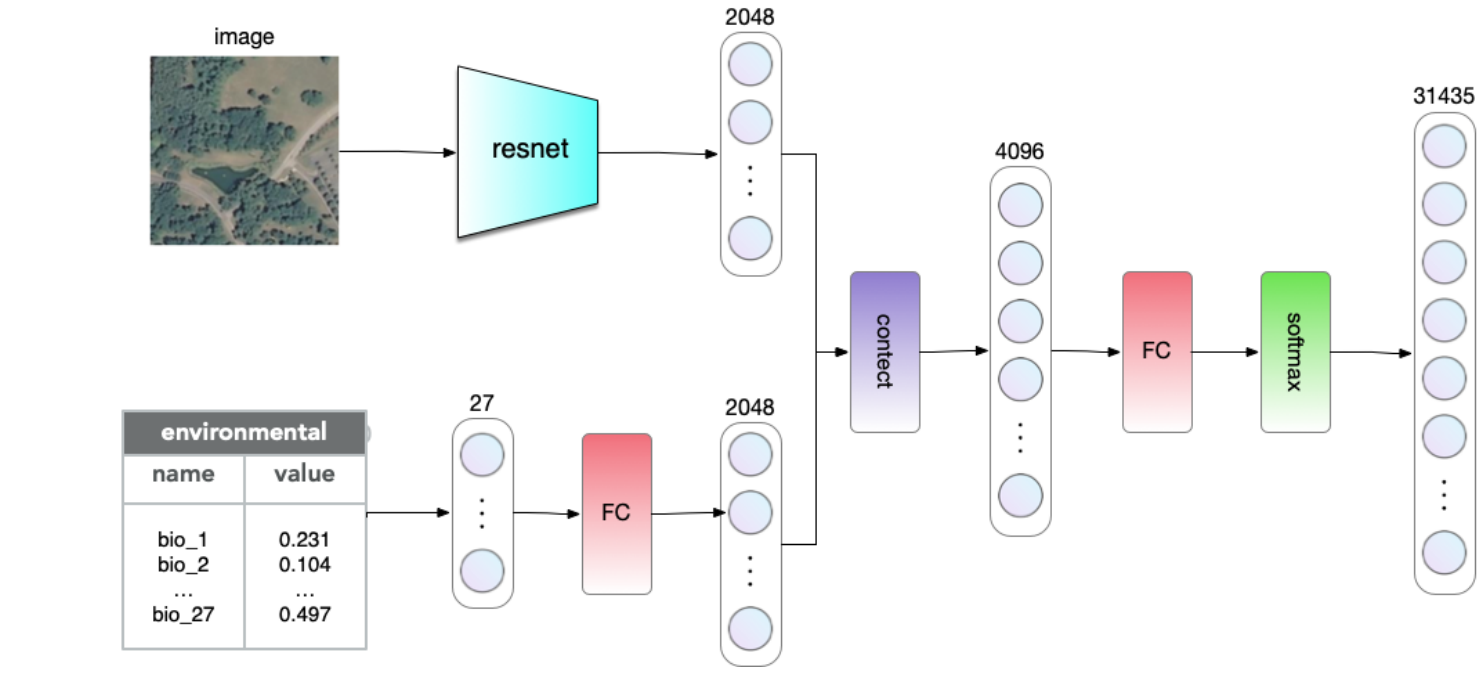

Figure 3

SDMM-Net network structure

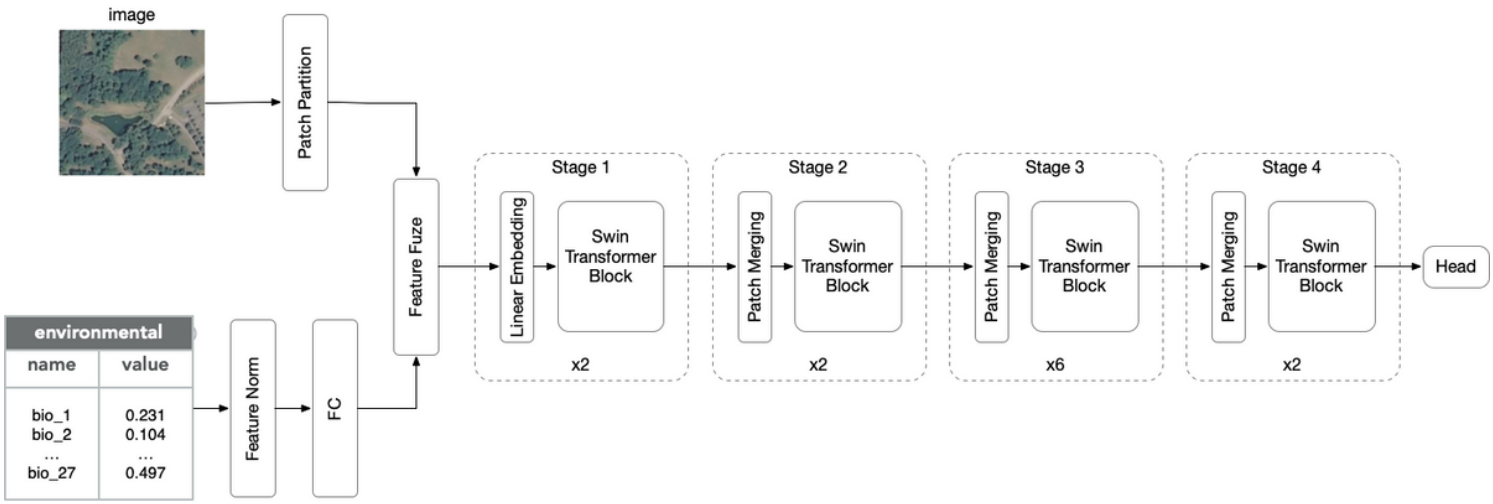

Figure 4

PreFuzemm-Swin network structure

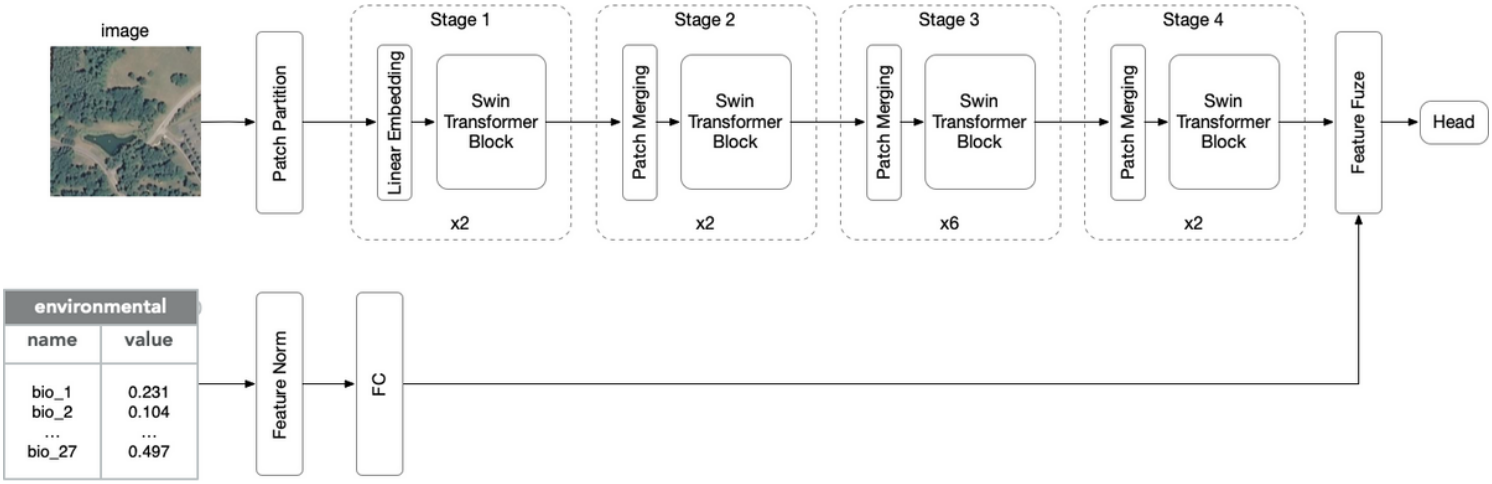

Figure 5

PostFuzemm-Swin network structure

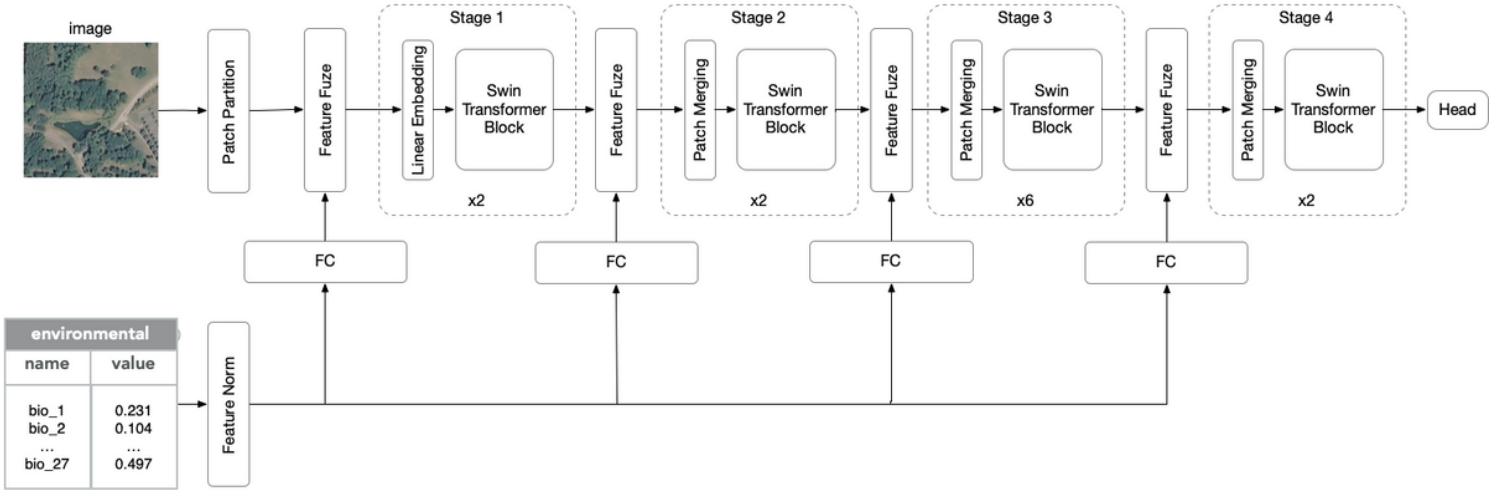

Figure 6

MidFuzemm-Swin network structure

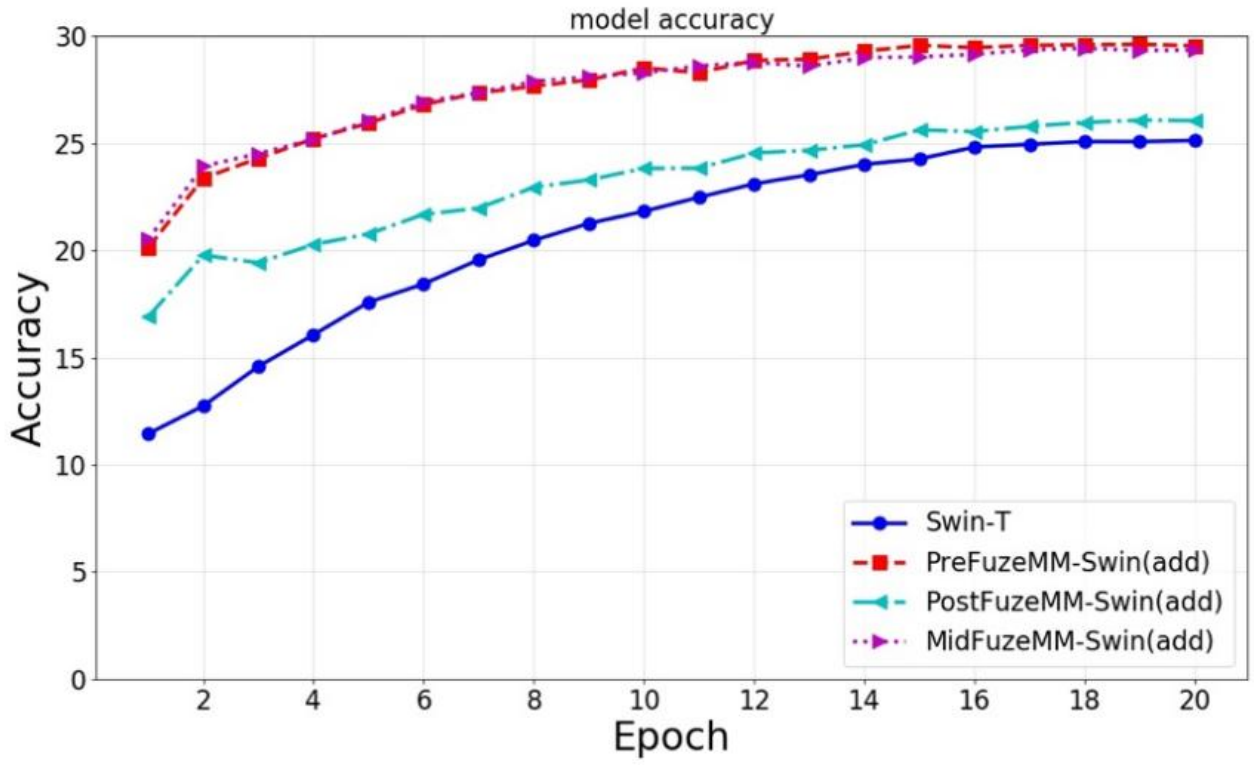

Figure 7

Comparison of different feature fusion structures on model accuracy