# Vision Transformer with Source-Target Attention from a Dilated Convolutional Structure for Remote Sensing

Tatsuki Shimura, Katsumi Tadamura and Toshikazu Samura,

***Abstract*—Attention-based Vision Transformers (ViTs) must be pre-trained on large datasets to give high performance. We propose a ViT with a dilated convolutional structure in the form of source-target attention (STA) and demonstrate that the proposed ViT performs better for remote sensing datasets and acquires attention much as the original ViTs pre-trained on a large dataset, even after pre-training on a small dataset. Our results suggest that the proposed structure efficiently acquires attention suitable for remote sensing from small datasets.**

***Keywords: Vision Transformer, Source-target attention, Dilated convolutional token, Attention acquisition, Remote Sensing***

## I. INTRODUCTION

A vision transformer (ViT) is a type of neural network that must be pre-trained on a large, labeled dataset to provide high-performance image recognition. ViTs are considered to be of limited applicability in remote sensing applications, which handle uncommon images that are difficult to label (e.g., synthetic aperture radar images). A ViT with a convolution input layer performs well while maintaining low training costs (e.g., [1]). However, network structures in which the convolutional process is incorporated into the ViT middle layer have rarely been investigated. We propose a network that introduces dilated convolution inputs in the form of source-target attention (STA) [2] to the ViT middle layer (ViT–STA).

## II. PROPOSED METHODS

The ViT [1] consists of a patch-token generator, self-attention (SA) encoders, and a multilayer perceptron (MLP) head (Fig. 1, excluding the shaded area). The patch-token generator divides an input image into sub-images as a patch (P) token. The SA encoders calculate attention using the query, key, and value calculated from a P token with a class token and position embedding. ViT predicts a class of input from the encoder outputs through the MLP head. We proposed ViT-STA, where an STA encoder replaces one SA. The STA encoder receives dilated-convolutional (DC) tokens and P tokens through different pathways (Fig. 1 shaded area). To calculate attention, a query is calculated from a DC token, and the key and value are calculated from a P token.

## III. FINDINGS

We prepared two pre-training datasets with different data sizes from down-scaled Image Net: a large dataset (900,000 data) and a small dataset (450,000 data). The original ViT with six SA encoders, and ViT–STA E*n*, where the *n*th SA encoder is replaced with an STA encoder, were pre-trained on the large or small dataset.

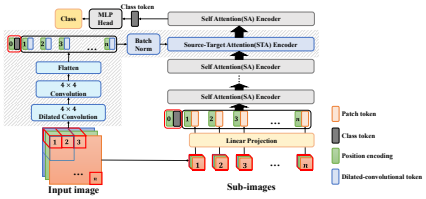


Figure 1. Structure of the Vision Transformer (ViT) with source–target attention (STA)

After that, they were fine-tuned on the EuroSAT remote sensing dataset [3]. In Fig. 2, the accuracies of the ViT-original were more than 3% lower during pre-training on the small dataset than during pre-training on the large dataset. Conversely, the accuracies of ViT-STA E3–E6 exceed those of ViT-original networks. Furthermore, the attention weights for test images in ViT-STA are similar to that of the ViT-original pre-trained on the large dataset.

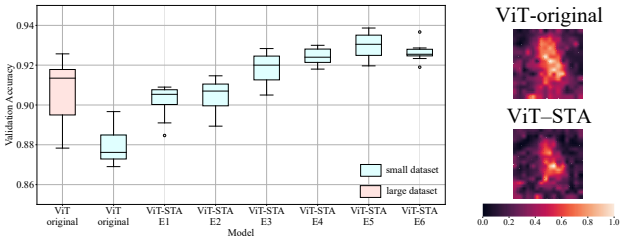


Figure 2. Network performances (left) and attention weights (right) acquired through fine-tuning on the EuroSAT remote sensing dataset.

## IV. CONCLUSION AND RECOMMENDATIONS

The proposed ViT-STAs provide better performance on the remote sensing dataset even when the pre-training data were halved from the large dataset through acquiring similar attention to those of a ViT pre-trained on the large dataset efficiently. These results suggest that the proposed structure enables ViT to acquire attention that improves classification for remote sensing efficiently from small datasets.

## REFERENCES

[1] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,". *in: International Conference on Learning Representations* 2021.

[2] A. Vaswani et al., "Attention is All you Need," *in: Conference on Neural Information Processing Systems (NIPS)* 2017.

[3] P. Helber et al., "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Top. Appl. Earth Obs. and Remote Sens.*, 2019, pp. 2217–2226.

*Research supported by JSPS KAKENHI Grant Number JP20H02417.

T. Shimura, K Tadamura are with Graduate School of Sciences and Technology for Innovation. T. Samura is with Organization for Research Initiatives, Yamaguchi University, 2-16-1 Tokiwadai Ube-shi, Yamaguchi, 755-8611, Japan ({c069vgw, tadamura, samura}@yamaguchi-u.ac.jp). T. Samura is also with Japan Aerospace Exploration Agency, 4-1-1 Asutopia, Ube-shi, Yamaguchi, 755-0195, Japan.