www.cad-journal.net
Computer-Aided Design
AND APPLICATIONS

UAV-based Anomaly Detection via a novel Spatial-Temporal Transformer for Precision Agriculture

Hebin Cheng[1], Heming Li[2] and Jian Lian[3]

[1] Shandong Management University, chenghebin@sdmu.edu.cn
[2] Shandong Management University, liheming@sdmu.edu.cn
[3] Shandong Management University, lianjianlianjian@163.com

Corresponding author: Heming Li, liheming@sdmu.edu.cn

Abstract. Low altitude security has gained widespread concern for its applications, e.g., cropland monitoring. In this work, an Internet of Drones architecture was firstly presented for low altitude scenarios. Within the presented framework, each drone can communicate with each other and the ground base stations while traveling along a fixed route. Since anomalous activities can significantly affect the normal operation of the Internet of Drones, it becomes critical to differentiate the normal surroundings from the abnormal surroundings around the drones. On the other hand, the vision transformer-based deep learning models have exhibited their superiority even over the convolutional neural network-based architectures due to the adoption of global receptive field. Bearing the above-mentioned issue in mind, we propose a novel deep learning pipeline by leveraging the vision transformer model for anomaly detection in the low altitude context. The sensitivity, specificity, accuracy, and F1 score of the proposed approach on a manually-collected dataset are 91.2%, 92.5%, 92.2%, and 93.7%, respectively. In addition, the experimental results demonstrate that the proposed deep learning pipeline can yield a promising outcome over the state-of-the-art algorithms.

Keywords: Transformer, Anomaly Detection, Low Altitude, Orchard, Machine Vision, Deep Learning.
DOI: https://doi.org/10.14733/cadaps.2024.S13.298-310

1 INTRODUCTION

The development of low-altitude security corridors (LASC) has attracted a great deal of attention. This architecture is built upon at least the following techniques, Internet of Things (IoT), Internet of Drones (IoD), modern communication, networking security, and artificial intelligence. By introducing machine learning algorithms, the low-altitude airspace can be fully exploited and the low-altitude capability of security management can be enhanced. Consequently, the machine vision community recently pays attention to the application of various learning methods in the IoD environment, such as automated identification of anomalous behaviors [13],[25] as shown in Fig.

1. These machine-learning approaches have shown promising performance in this field, but most of them still suffer from handling the vast amount of information within the district covered by IoD.

Figure 1: The diagram of object detection in LASC.

Yang et al. [32] studied the application of wireless sensor system in orchard management. Their presented method could be also used in orchard monitoring. To decrease the resource occupation for apple orchards, enhance the apple's quality, and provide ample information, the work of [11] proposed an automatic apple orchard monitoring system via the Internet of Things (IoT). The study of [3] presented a robotic platform to monitor the status of plants. Aiming at addressing the monitoring issues of apple orchard in China, the study of [19] presented the wireless sensor-based apple orchard monitoring pipeline. Other influential work includes [1],[4], [20],[26],[27],[35].

In recent years, the task of object detection has benefited from the deep learning backbones and their built-in detection-specific modules. Among them, the convolutional networks (ConvNet) with multi-scale and hierarchical architectures have significantly influenced the design for object detection [16]. However, the ConvNet may neglect the long-range relationship between its input samples since the performance of ConvNet primarily relies on local operator convolution.

Moreover, Vaswani et al. [28] presented the early work of transformer for NLP, and has become the de facto benchmark method in plenty of NLP applications. Due to the lack of inductive biases like locality, the transformer-based models usually need to be trained on large corpora and sequential mission, e.g., BERT [8]. It was designed for generating bi-directional embedding from the source text by leveraging the context of both directions and all layers. To note that the pre-trained Bert can then be further optimized to implement various tasks without substantial modifications made to the whole architecture. Self-attention plays a critical role in transformers. Straightforward application of self-attention to an image requires pair-wise operations between great amounts of pixels, especially in images with high resolution. To apply a transformer to image classification, the modification of self-attention needs to be taken into consideration in advance. For instance, the studies [15],[22],[33] exploited the local self-attention for each query pixel and the presented local multi-head dot-product blocks can take the place of convolution operators. Alternatively, self-attention for vision can be realized by leveraging it in different scales of images [30]. Plenty of attempts have also been made to combine CNN with self- attention. For instance, Bello et al. [2] augmented the capability of convolution modules by using self-attention mechanism.

Furthermore, contrary to the hierarchical transformers in the computer vision area, the vision transformer (ViT) [9] is a powerful yet non-hierarchical backbone for image classification. To be specific, the former transformers including, Swin [18], MViT [10], PVT [29], and PiT [14], followed the designs from ConvNet, e.g., convolution and pooling. On the other hand, plain backbones similar to ViT have been proposed with a single-scale strategy. For instance, UViT [6] introduced the width, depth of the network, and the resolution of input from ViT models while leveraging a progressive attention mechanism to deal with the images with high resolution. In the work of [5], Carion et al. presented a framework for object detection named after DETR, which uses a transformer-based model. Wu et al. [31] extracted the visual tokens to yield the representation by using ConvNet. Then they used transformers to operate on the extracted tokens to model the relationships between them.

Inspired by the work of ViT [9] and visual transformer [31], we propose a novel transformer-based framework composed of two channels, which focus on extracting both the spatially positional information and temporally sequential information, simultaneously. To be specific, we first split the images captured by the drones into image patches. Then the linear embeddings of the image patches are fed into the proposed transformer. The patches play the same role as the tokens in an NLP task. Accordingly, the mission of anomaly detection can be transferred as a task of image classification in a supervised fashion. To be specific, the anomalies in this study mainly refer to the anomalous objects shown in the captured images using UAVs. To avoid the global inductive bias, we trained the proposed model on a large dataset (ImageNet-ISLVRC) before the manually collected images [9]. Besides, we introduce a novel loss function to encourage the output to be accurate. To train this model, we collect 20,411 frames of the scenes in an LASC. The outcome from the experiments proved that the presented algorithm achieves superior outcomes over the current deep learning techniques.

In this work, the contributions include the followings:

- First of all, this is an early work of anomaly detection in LASC by IoD.
- We propose a transformer-based model to implement the above-mentioned task. Additionally, we leverage one novel loss function to guarantee an accurate outcome.
- Extensive evaluations demonstrate the superiority of this work over the state-of-the-art algorithms and shows robustness under various tasks.

2 METHODS

2.1 Dataset and Image Preprocessing

In this work, the proposed spatial-temporal transformer is firstly trained on the publicly available database ImageNet-ISLVRC [24], which is widely exploited for enhancing the performance of object detection and classification from 2010. It contains 50,000 images with labels of 1,000 categories as the training set. Then, we collected 6,803 frames from the photographs by leveraging 16 sets of drones (model: DJI JY03-4K; size: 31-40cm; channels: 4; material: plastic). During the capturing process, the sRGB color space is chosen using the light of white fluorescent. The collection is located at a university campus in Zibo, Shandong Province, China. In general, the drones were grouped into an integrated IoD, and each drone was arranged to follow a fixed route on the campus. To guarantee the cruising ability of the drones, each drone was charged every 15 minutes during the tour and collected the scene photos every 30 seconds. The resolution of the collected images is 8192*4096. After the collection process, each sample frame was manually labeled as anomalous or non-anomalous using a majority voting mechanism by three experts in the field of machine vision.

Moreover, the initially collected images were labeled using the annotation instrument LabelMe. And the annotated images were stored in MS COCO format [17]. Both the images and the corresponding JSON file were generated. Furthermore, to further enhance the diversity of the manually collected images, we performed a group of transformations on the images, including horizontal flip, vertical flip, and rotation. To note that each image and its transformed counterparts are labeled as the same category.

2.2 The Proposed Transformer-based Framework

In the following, we provide details of the proposed ST-Transformer architecture for anomaly detection in LASC. This model is built upon self-attention-based transformers. And the self-attention mechanism has been exploited in various works [10],[18],[29] along with or combined with the convolutional layers. On the other hand, in this study, we prove that transformer-based models can generate competitive outcomes over CNN and hybrid architectures. The proposed ST-Transformer is illustrated in Fig. 2, which is partially derived from the work of ViT [9] that adopts both the image patches and sequence of the patches as its input.

2.2.1 Input of the proposed model

Generally, both the spatial patches and the temporal patches are taken as input for the presented model, which are both captured from the collected images by the drones. To provide the spatial information, the position embeddings are also combined into the input of the transformer. As shown in Fig. 2, the presented ST-Transformer is composed of two separate paths (as shown in Fig. 3 and Fig. 4) without sharing the weight parameters.

In a real application, each UAV takes off after 30 seconds before capturing 1 second of video; During the trip, each UAV collects videos every 30 seconds. Note that the speed of the UAVs is less than 3m/s, and the UAVs can collect images from more than 100m away.

In each channel, following the vision transformer and different from the standard transformer, this model uses the 2D embeddings. To feed the proposed model, we resize the input image $x \in \mathbb{R}^{H \times W \times C}$ into image patches as $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, H, W respectively represent height and width of an image, C is the number of channels, P denotes the width (and height) of a patch. Then, the patches are transformed into a vector of length D. And the output of this process is the embedded patches.

Similar to the vision transformer, a learnable embedding is attached to form a sequence of embeddings ($\mathrm{z}_0^0 = x_{class}$), and its corresponding output from the transformer (z_L^0) is denoted as y. Furthermore, the position embedding is exploited to provide the positioning information in addition to the sequence of patches.

$$\mathrm{z}_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \tag{1}$$

Where $E \in \mathbb{R}^{P^2 \cdot C} \times D$ and $E_{pos} \in \mathbb{R}^{(N+1) \times D}$.

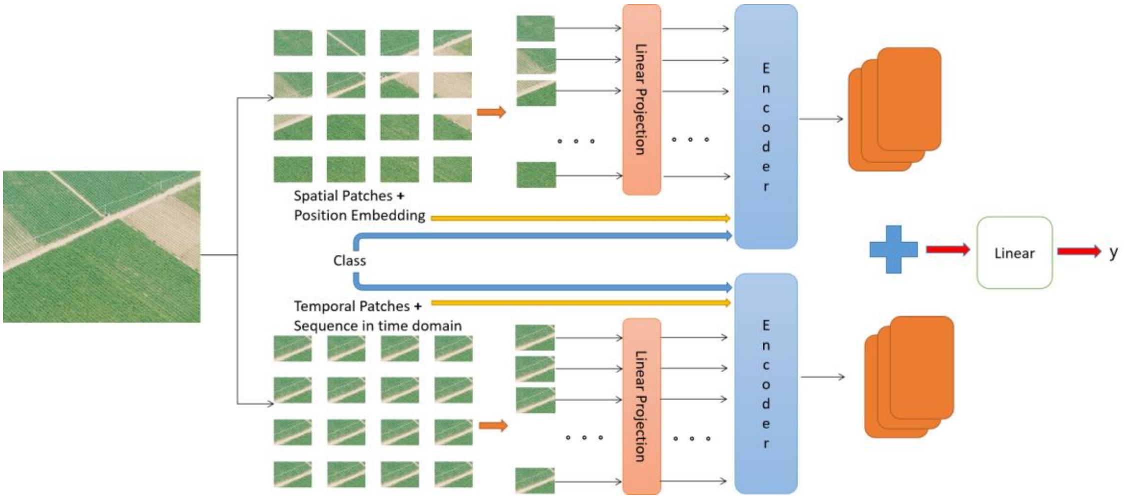

Figure 2: The proposed pipeline for anomaly detection in LASC.

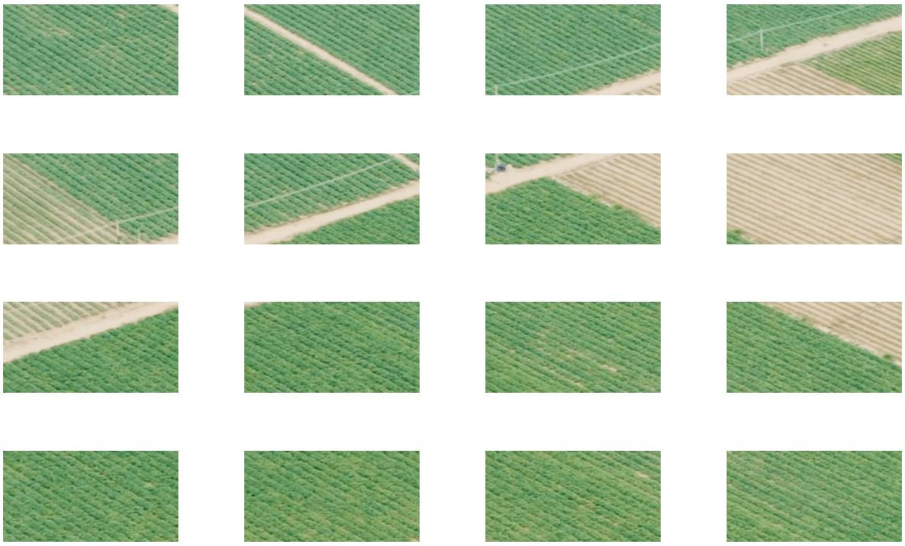

Figure 3: A sequence of spatial image patches.

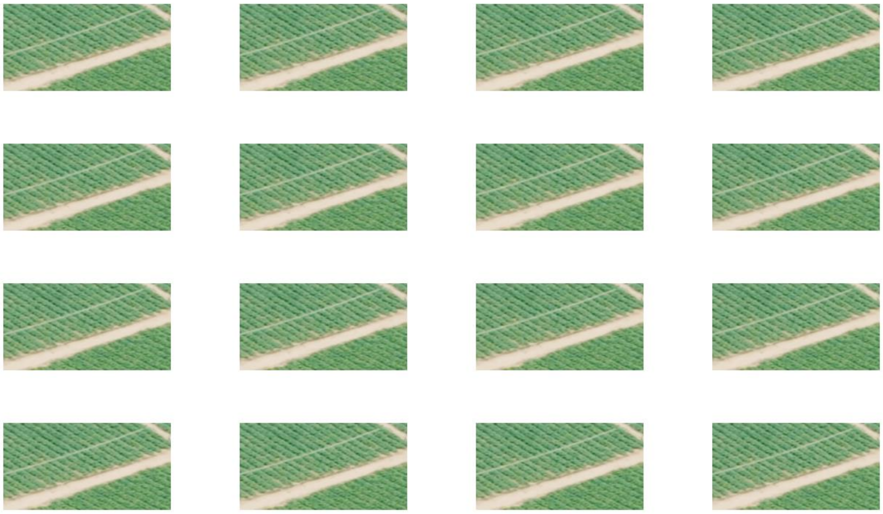

Figure 4: A sequence of temporal image patches.

2.2.2 *Encoder*

The above-mentioned z_0 is used as the input for the proposed transformer. Inspired by the work of [28], the presented model treats the input patches as tokens. Each encoder contains L layers, which includes a multi-head self-attention (MSA) layer and a multi-layer perception (MLP) layer.

Besides the MSA module, a module of layerNorm is leveraged before every block and the residual block is used after every block. Two layers MLP with a Gaussian error linear unit (GELU) as the classification head is attached to z_L^0 .

Note that the MSA module is built upon the conception of the self-attention (SA) mechanism. SA is used to measure the similarity of a query and the corresponding keys with weighting values. Therefore, the output can be obtained from the weighted sum of all of the values. To be specific, once there is an input $Z \in \mathbb{R}^{N \times D}$ composed of N vectors with the length of D.

$$[Q,K,V] = ZW_{QKV} \tag{2}$$

where W_{QKV} denotes the weight matrix that can be updated by training. All of the weights are computed into the probabilities P with the following function:

$$P = soft\max(\frac{QK^T}{\sqrt{D}}) \tag{3}$$

where D is the length of each vector in Q, K, and V. Finally, the output of the SA mechanism can be mathematically expressed as:

$$SA(Z) = PV \tag{4}$$

Moreover, the MSA mechanism equals adopts the SA mechanism several times in parallel, which contributes to extracting the information in the input for each head, respectively. The output of the MSA is the concatenated elements of all of the heads, which is expressed as:

$$MSA\ Z\ = \left[SA_1\ \ Z\ ;\ SA_2\ \ Z\ ;\ ...;\ SA_h\ \ Z\ \right]W_{MSA} \tag{5}$$

Where h denotes the number of heads in the MSA module.

Furthermore, at the end of the two-way structure, we add a linear layer to fuse the extracted feature maps from both channels.

$$y = Linear(LayeNorm[(Z_L^0)_{spatial} + (Z_L^0)_{temporal}]) \tag{6}$$

where Linear denotes a linear function, L =1 or 2, $(Z_L^0)_{spatial}$ and $(Z_L^0)_{temporal}$ represent the output of each channel, respectively.

Note that different from vision transformers, these bi-channels separately receive the temporal and spatial sequence of patches. To be specific, the position embeddings stand for the spatial and temporal sequence of the image patches in two channels, respectively.

3 RESULTS

3.1 Implementation Details

In general, we leveraged the images from ImageNet-ISLVRC [24] to perform the initial training of the proposed transformer. Furthermore, the primary settings include RMSprop as the optimizer, the learning rate 0.001 reduced by 0.5, and a batch of 8 images. This was realized by employing PyTorch [21] and 2 NVIDIA Telsa V100 GPUs (HBM2 32GB).

In the following experiments, firstly, we investigated the influence of 3 parameters, including layers (L), model dimension (D), and number of heads (h) on the proposed transformer by leveraging a fraction of the entire data samples. With the optimal parameter combination, we pre-trained the transformer on the ImageNet-ISLVRC []24. Moreover, we used the manually collected image samples to fine-tune the proposed transformer. Secondly, we conducted comparing experiments between the deep learning models and the proposed algorithm. The results show that the proposed transformer is superior over the state-of-the-arts in performance, including sensitivity, specificity, accuracy, F1 score, and ROC. At last, we performed a ablation study through evaluating the performance of a variety of the proposed model.

3.2 Loss function

The integration of spatial and temporal components is exploited as the loss function for the transformer-based pipeline.

$$Loss_{\mathrm{mod}\,el} = Loss_{Spatial} + Loss_{Temporal} \tag{7}$$

where $Loss_{Spatial}$ and $Loss_{Temporal}$ denote the spatial and temporal cross-entropy loss, respectively.

3.3 Evaluation Metrics

In this work, we chose the following metrics including sensitivity, specificity, accuracy, F1 score, and ROC in the experiments and could be formulated as the followings:

- Sensitivity. The ratio between true positives (TP) cases and (TP + FN), where FN denotes false negative.

$$Sensitivity=\frac{TP}{TP+FN} \tag{8}$$

- Specificity. The ratio between the true negatives (TN) and (TN + FP), where F P denotes false positives.

$$Specificity=\frac{TN}{TN+FP} \tag{9}$$

- Accuracy.

$$Accuracy=\frac{TP+TN}{TP+FN+TN+FP} \tag{10}$$

- F1 score.

$$F1=2\times\frac{Precision\cdot Recall}{Precision+Recall} \tag{11}$$

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

$$Recall=\frac{TP}{TP+FN} \tag{13}$$

To evaluate the presented algorithm robustly, a 10-fold cross-validation strategy was exploited for the experiments. To note that the average of these metrics over ten folds were taken as the experimental outcome derived from every round.

3.4 Influence of the Hyper-parameters on the Proposed Transformer

To obtain an optimal combination of the 3 parameters of the proposed model, we carried out the comparison experiments on a subset of the collected images with different combination of the parameters. And it is supposed to yield a better outcome of classification for the entire dataset.

Consequently, we evaluated the following combinations of these 3 parameters as shown in Table. 1. These combinations are named after Spatial Temporal Transformer with the actual values of the parameters.

Combination	*Layer(L)*	*Dimension(D)*	*Number of heads(h)*
STT_1_64_4	1	64	4
STT_1_64_8	1	64	8
STT_1_128_4	1	128	4
STT_1_128_8	1	128	8
STT_2_64_4	2	64	4

STT_2_64_8	2	64	8
STT_2_128_4	2	128	4
STT_2_128_8	2	128	8

Table 1: The combinations of the 3 parameters in the proposed transformer.

To note that only 3 parameters were chosen in this stage and more variants would be too many to follow up. With the comparison outcome shown in Fig. 5, we obtained transformer model with the combination of STT_2_64_4.

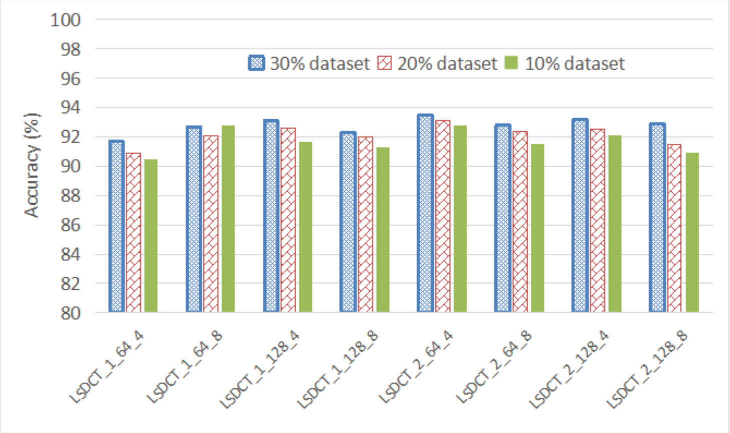

Figure 5: The influence of 3 parameters on the proposed transformer.

In addition, we conducted comparison experiments between the mean square error (MSE) loss, cross entropy (CE) loss, and the proposed loss function (as shown in Table. 2).

Loss function	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*	*F1 score(%)*
MSE	*89.4*	*88.2*	*89.3*	*88.7*
CE	*90.5*	*92.3*	*91.7*	91.5
Ours	*91.2*	*92.5*	*92.2*	93.7

Table 2: Influence of various loss functions on the proposed approach.

3.5 Comparison Between the State-of-the-art and the Proposed Transformer

As shown in Table. 3, the proposed method is superior in terms of sensitivity, specificity, accuracy, F1 score, and ROC which indicates that our approach can be more useful for anomaly detection than the state-of-the-arts.

Furthermore, we evaluated different initial weights of random initialization and trained on ImageNet- ISLVRC for the proposed transformer. After less than 15 epochs, the manner of non-random enters the state of convergence while the random set requires more than 30 epochs. Meanwhile, the transformer trained on Image-ISLVRC has a higher starting point and the loss differences share a uniform trend.

Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 score(%)
U-Net [23]	81.2	82.4	83.7	81.9
Mask R-CNN [12]	81.9	82.5	82.9	80.6
ExtremeNet [34]	81.7	82.3	83.6	82.3
TensorMask [7]	82.2	82.9	84.3	81.8
Visual Transformer [31]	89.5	85.6	87.3	86.2
ViT [9]	88.5	86.4	87.1	87.1
MViT [10]	87.9	87.5	88.1	87.9
PiT [14]	87.6	88.2	89.4	88.3
PVT [29]	89.3	89.2	90.1	89.5
UViT [6]	88.6	89.7	91.5	90.3
Swin Transformer [18]	89.1	87.2	88.0	87.4
Ours	91.2	92.5	92.2	93.7

Table 3: Performance comparison between state-of-the-art techniques and ours in terms of sensitivity, specificity, accuracy, and F1 score.

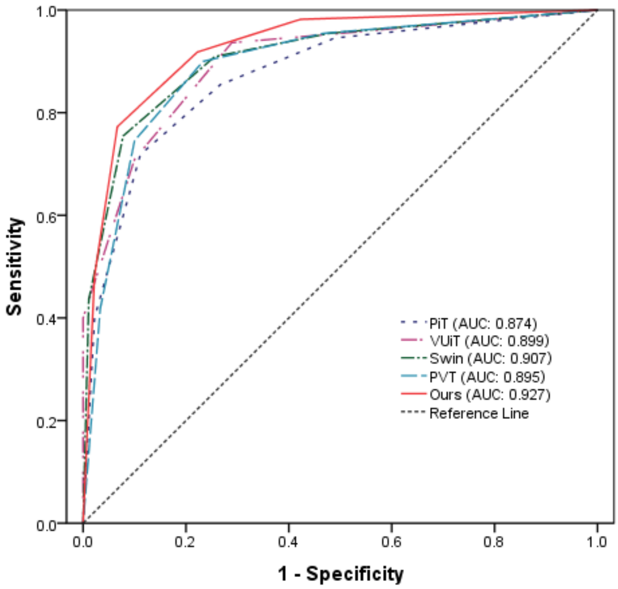

Figure 6: The ROCs of the competing algorithms.

3.6 Ablation Study

Since the proposed model can be considered as a hybrid architecture, we continued to measure the difference between the individual channels and the integrated model. Accordingly, we also computed the accuracy of discrimination leveraging the spatial channel and temporal channel, respectively. As shown in Fig. 7, we can observe that the hybrid model is superior to the individual ones.

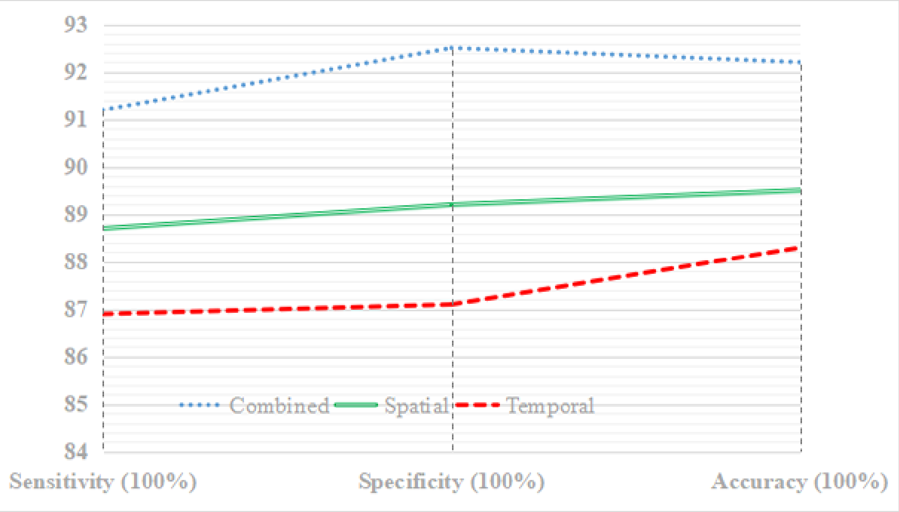

Figure 7: The comparison between the spatial, temporal, and combined architectures.

To fully exploit the intra-frame and inter-frame information, we proposed a spatial-temporal dual channel transformer. Meanwhile, to yield an accurate classification outcome, we leveraged the new loss function. From the results from both the comparison experiments and ablation study, we have proved the advantage of the two-channel structure.

4 DISCUSSION AND CONCLUSION

During the training of a deep learning model, a large number of samples are needed to implement the training of the model. Moreover, a certain number of samples could ensure that the deep learning model can solve the specific task. On one hand, too much data will lead to a waste of resources and may cause the problem of over-fitting. On the other hand, insufficient data might not produce a satisfactory performance and will directly affect the availability of the presented deep learning model. Accordingly, we proposed that the best solution to this type of task is to build a data-driven deep learning model with an appropriate size of data samples through regularized output.

In this work, a dual-channel model is presented for anomaly detection in low-altitude scenarios. This is also an application of vision transformer-based algorithm in low-altitude scenarios. According to the experimental results, the proposed transformer can yield accurate detection outcome by leveraging the hybrid architecture.

Notably, the proposed anomaly detection framework is built upon the model of the vision transformer. Unlike most of the anomaly detection techniques in the literature, our method can leverage both temporal and spatial information in the input image samples. Although the presented algorithm adopted an end-to-end learning strategy, it still needs to adapt to the needs of anomaly detection. By taking the attention mechanism commonly used by the transformer-based techniques, the associations between global pixels in a collected image using UAVs can be unveiled. Experimental results demonstrate that the presented approach could guarantee the performance of anomaly detection. Therefore, it is a valuable instrument for orchard monitoring in LASC.

In addition, this study also has several limitations. First of all, a publicly available database should be used rather than the leveraged private samples in the research. Secondly, the mechanism of transfer learning needs to be introduced.

Hebin Cheng, http://orcid.org/0009-0002-5829-2838
Heming Li, https://orcid.org/0000-0002-2579-7701
Jian Lian, http://orcid.org/0000-0003-0305-8454

REFERENCES

[1] Argov, Y.; Rössler, Y.; Voet, H.; Rosen, D.: Spatial dispersion and sampling of citrus whitefly, dialeurodes citri, for control decisions in a citrus orchard, Agricultural and Forest Entomology, 1(4), 1999, 305–318. https://doi.org/10.1046/j.1461-9563.1999.00041.x.

[2] Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q. V.: Attention augmented convolutional networks, CVF International Conference on Computer Vision (ICCV), IEEE, 2019, https://doi.org/10.1109/ICCV.2019.00338.

[3] Bietresato, M.; Carabin, G.; D'Auria, D.; Gallo, R.; Ristorto, G.; Mazzetto, F.; Vidoni, R.; Gasparetto, A.; Scalera, L.: A tracked mobile robotic lab for monitoring the plants volume and health, 12th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA), 2016, 1–6. https://doi.org/10.1109/MESA.2016.7587134.

[4] Bramlett, D. L.: Protection of pine seed orchards in the southeastern united states, Forest Ecology and Management, 19, 1987, 199–208. https://doi.org/10.1016/0378-1127(87)90028-4.

[5] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S.: End-to-end object detection with transformers, Computer Vision–ECCV 2020, 2020, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

[6] Chen, W.; Du, X.; Yang, F.; Beyer, L.; Zhai, X.; Lin, T.-Y.; Chen, H.; Li, J.; Song, X.; Wang, Z.; Zhou, D.: A simple single-scale vision transformer for object localization and instance segmentation, European Conference on Computer Vision, ECCV, 2022, 711–727. https://doi.org/10.1007/978-3-031-20080-9_41.

[7] Chen, X.; Girshick, R.; He, K.; Dollar, P.: Tensormask: A foundation for dense object segmentation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2061–2069. https://doi.org/10.1109/ICCV.2019.00215.

[8] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[9] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale, ICLR 2021. https://doi.org/10.48550/arXiv.2010.11929

[10] Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C.: Multiscale vision transformers, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 6804–6815. https://doi.org/10.1109/ICCV48922.2021.00675.

[11] Feng, C.; Wu, H. R.; Zhu, H. J.; Sun, X.: The design and realization of apple orchard intelligent monitoring system based on internet of things technology, Advanced Materials Research, 546, 2012, 898–902. https://doi.org/10.4028/www.scientific.net/AMR.546-547.898.

[12] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.: Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), 2980–2988. https://doi.org/10.1109/ICCV.2017.322.

[13] Henrio, J.; Nakashima, T.: Anomaly detection in videos recorded by drones in a surveillance context, 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2503-2508. https://doi.org/10.1109/SMC.2018.00429.

[14] Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S. J.: Rethinking spatial dimensions of vision transformers, CVF International Conference on Computer Vision (ICCV), IEEE, 2021, 11916–11925. https://doi.org/10.1109/ICCV48922.2021.01172.

[15] Hu, H.; Zhang, Z.; Xie, Z.; Lin, S.: Local relation networks for image recognition, International Conference on Computer Vision (ICCV), IEEE, 2019, 3463-3472. https://doi.org/10.1109/ICCV.2019.00356.

[16] Li, Y.; Mao, H.; Girshick, R.; He, K.: Exploring plain vision transformer backbones for object detection, European Conference on Computer Vision, 2022, 280-296. https://doi.org/10.1007/978-3-031-20077-9_17.

[17] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár P.; Zitnick C. L.: Microsoft coco: Common objects in context, Computer Vision–European Conference on Computer Vision, 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.

[18] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[19] Meng, X.; Cong, W.; Liang, H.; Li, J.: Design and implementation of apple orchard monitoring system based on wireless sensor network, 2018 IEEE International Conference on Mechatronics and Automation (ICMA) (IEEE), 200–204. https://doi.org/10.1109/ICMA.2018.8484350.

[20] Nagy, A.; Fórián, T.; Tamás, J.; Nyéki, J.: Szabó, Z.; Soltész, M.: Monitoring of water regime in an apple orchard, International Journal of Horticultural Science, 17(1-2), 2011, 29-32. https://doi.org/10.31421/IJHS/17/1-2./940.

[21] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Fimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; Devito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S.: Pytorch: An imperative style, high-performance deep learning library, NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, 8026–8037. https://doi.org/10.48550/arXiv.1912.01703.

[22] Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J.: Stand-alone self-attention in vision models, Proceedings of the 33rd International Conference on Neural Information Processing Systems, 7, 2019, 68–80. https://doi.org/10.48550/arXiv.1906.05909.

[23] Ronneberger, O.; Fischer, P.; Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[24] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Li, F.; ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV), 115(3), 2015, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[25] Sindhwani, V.; Sidahmed, H.; Choromanski, K.; Jones, B.: Unsupervised anomaly detection for self-flying delivery drones, 2020 IEEE International Conference on Robotics and Automation (ICRA), https://doi.org/10.1109/ICRA40945.2020.9197074.

[26] Tognetti, R.; Giovannelli, A.; d'Andria, R.; Fragnito, F.; Lavini, A.; Morelli, G.; Sebastiani, L.: Monitoring sap flow as indicator of transpiration and water status of an experimental olive tree orchard, Acta Horticulturae, 2012, 951, 167–174. https://doi.org/10.17660/ActaHortic.2012.951.20.

[27] Meng, X., Cong, W., Liang, H., & Li, J. (2018, August). Design and implementation of Apple Orchard Monitoring System based on wireless sensor network. In 2018 IEEE International Conference on Mechatronics and Automation (ICMA), 200-204, IEEE. https://doi.org/10.1109/ICMA.2018.8484350.

[28] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, E.; Polosukhin, I.: Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 6000-6010. https://doi.org/10.48550/ARXIV.1706.03762.

[29] Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 548–558. https://doi.org/10.1109/ICCV48922.2021.00061.

[30] Weissenborn, D.; Täckström, O.; Uszkoreit, J.: Scaling autoregressive video models, ICLR 2020, https://doi.org/10.48550/arXiv.1906.02634.

[31] Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; Vajda, P.; Labs, F. R.; AI, F.: Visual transformers: Token-based image representation and processing for computer vision, arXiv: Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2006.03677

[32] Yang, H.; Kuang, B.; Mouazen, A. M.: Wireless sensor network for orchard management, 2011 Third International Conference on Measuring Technology and Mechatronics Automation (IEEE), 3, 2011, 1162–1165. https://doi.org/10.1109/ICMTMA.2011.859.

[33] Zhao, H.; Jia, J.; Koltun, V.: Exploring self-attention for image recognition, CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, 10073–10082. https://doi.org/10.1109/CVPR42600.2020.01009.

[34] Zhou, X.; Zhuo, J.; Krähenbühl, P.: Bottom-up object detection by grouping extreme and center points, CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, 850–859. https://doi.org/10.1109/CVPR.2019.00094.

[35] Zhu, G.: The application of wireless sensor networks in management of orchard, Computer and Computing Technologies in Agriculture III: Third IFIP TC 12 International Conference, CCTA, 2009, 519–522. https://doi.org/10.1007/978-3-642-12220-0_75.