Vision Transformers in Real-time

**Abstract**

Vision Transformer (ViT) is often used in vision processing tasks such as image recognition. ViT was first introduced in 2017 in a research paper called “Attention is All You Need” and it became popular in the field of Natural Language Processing (NLP). Three years after that, in 2020, ViT was implemented and adapted into computer vision tasks; and in 2021, there were researches resulting that ViT’s model have better performance and efficiency than Convolutional Neural Network (CNN) with regards to image classification. In the more recent years, ViT have been put into test for real-time tasks. There are many research papers in reputable research papers databases which showed and proved that works have been done and still in progress when it comes to implementing and use ViT models for real-time tasks. This survey paper will give a high-level overview of the many researches done in ViT for real-time and also, provide some experiments, results, and findings stated in the research papers. This survey paper would be useful for anyone whom is interested in ViT in real-time tasks and for people whom wants a more higher-level for information and the works done in ViT for real-time. Ultimately, this survey paper would not go into extreme depths like what is presented in the referenced research papers but instead, this survey paper will be useful to give readers a glance and teaser on what is being researched in real-time ViT.

**Keywords**

ViT, Dataset, Object detection, Real-time, Transformer Network, ViT architecture, Transformer tracking

# Introduction

This survey paper covers several different and unique aspects of real-time examples and how it is being applied into ViT models and ultimately, presenting results whether or not the results are desirable or not. Also, it will cover some of the challenges and future plans of works that can still be done to improve the overall efficiency and performance of the respective ViT model.

In a bird’s eye view, the areas of real-time tasks which will be covered in this survey paper includes, but not only, detecting violence from real-time videos, applying ViT in real-time object detection in ultrasound, using transformer network to estimate earthquake’s location and magnitude from real-time seismic waveforms, implementing a 3-dimensional spatial transformer network for facial alignment and unconstrainted poses, make using of ViT models for real-time semantic communications, and adopting an efficient vision transformers for real-time style transfer.

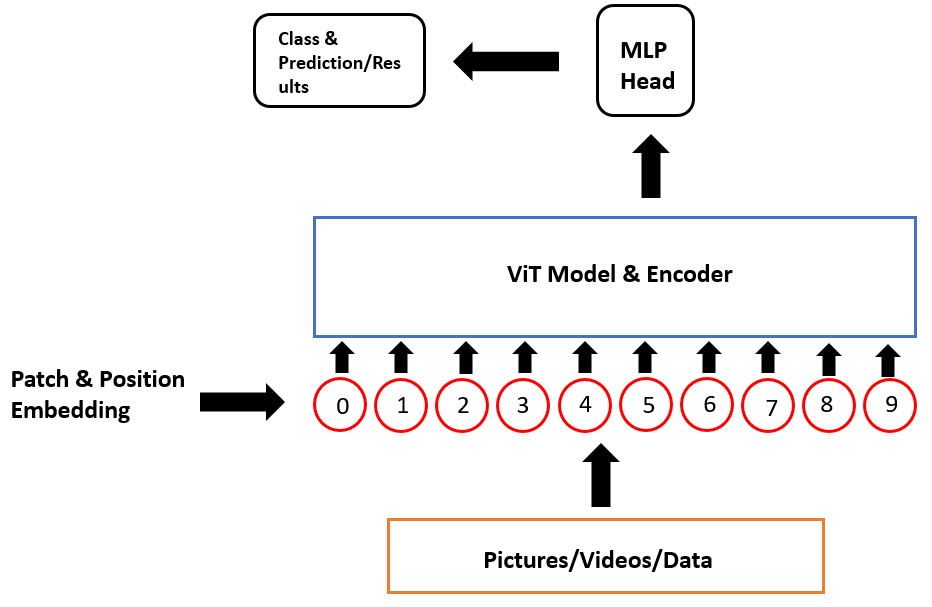
As the real-time tasks are of extreme different areas, this survey paper will present the most common features and aspects of what were being presented in most research papers. Having said that, this survey paper will present certain interesting findings and works specific to certain real-time tasks.

Lastly but not least, the structure of this survey paper will first give a summarized and high-level overview of how most of the ViT architectures and models look like, present some interesting experiments and works done in the referenced research papers, show some of the results and findings, in general, presented in the papers, conclusion, and lastly, acknowledgments and references.

# ViT in Real-time Model/Architecture

Although no two research papers use the same ViT model for their respective works, most research papers followed a relatively similar ViT architecture and workflow from passing their dataset to getting the results and predictions.

The figure below is a overview of how a general ViT architecture in real-time looks like based on the understanding of the different ViT architecture presented in the research papers citied.



**Figure 1. High-level overview of the real-time ViT architectures**

Based on the understanding of the different ViT models and architectures presented in the research papers, the above figure was made and its purpose is to give a general overview of how the different yet common architectures are in the papers.

Starting from the bottom, whatever dataset is being used such as images, videos, or any form of real-time data, the data will first go through a patching process. During this process, the data will be segmented into chunks of smaller data and then, each chunk of the smaller data will be flattened.

Then after, the next process would be called the position embedding process. During this process, each flattened chunk of data will be tagged with a location/position such as 0 to 9 in a sequence so that each position of the data is unique. This process is said to be extremely important; particularly for ViT models and architectures.

The next step after tagging the data with its embedded position/location, the data will be passed into the ViT model. At this phase, it is difficult to show the specifics in the figure because every model is different and is tailored to suit each real-time task. Thus, the figure shows a general box whereby, during this phase, the data will be in its training and validation phase.

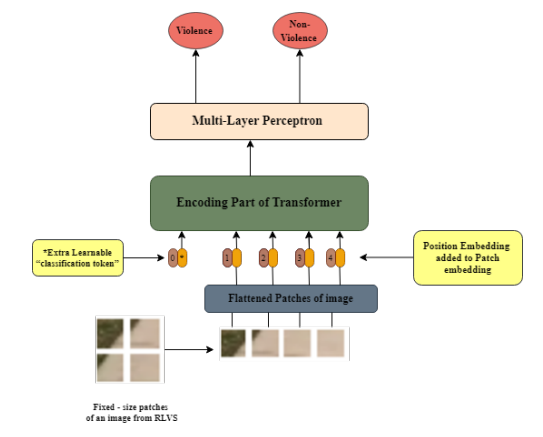
The second last phase of a general real-time ViT architecture is the multilayer perceptron (MLP) phase. In specific, for ViT model and architecture, it is called the MLP Head phase. During this phase, the outputs from the ViT model and encoder will be passed into a MLP for image classification. It is said that the MLP Head is able to capture the key features of images very well and hence, making the image classification work much simpler. Lastly, during this phase, the MLP Head is able to produce multiple outputs.

Last but not least, the final stage of a ViT in real-time architecture is to put the model and architecture into test and see it produce its results and predictions.

# Notable Experiments

## ViT for Violence Detection from Real-time Videos

In the research experiment for violence detection based on real-time videos using ViT, there were two forms of dataset which were based on the nature of the environment’s background of the video being taken. Also, three different models were used in the research; namely: ViT, ConvLSTM, and VGG16+LSTM. The main model used was the ViT model and the two LSTM models were used as baseline. However, all three models used different hyperparameters in order to see which model can achieve the bast accuracy and performance for detecting violence. Since this survey paper is about ViT, this paper will only present the in-depth architecture of the ViT (the below figure) that was used.

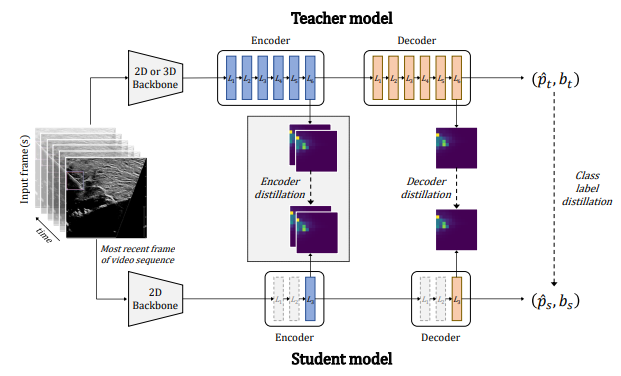


**Figure 2. Block diagram of the ViT used**

As seen in the figure above, the overall ViT architecture is very similar to what was explained in 2. ViT in Real-time Model/Architecture. Similarly, it goes through a flattening process, encoding process, MLP process, and finally, obtaining the results whether or not do the video contain any form of violence.

## Real-Time Video Object Detection in Ultrasound

This is the second notable and interest experiment that was refenced from one of the research papers. In a nutshell, the experiment was to make use of ultrasound and ViT model to detect objects in real-time videos. Below is the overview of the ViT model used for the experiment, specifically, attention distillation.



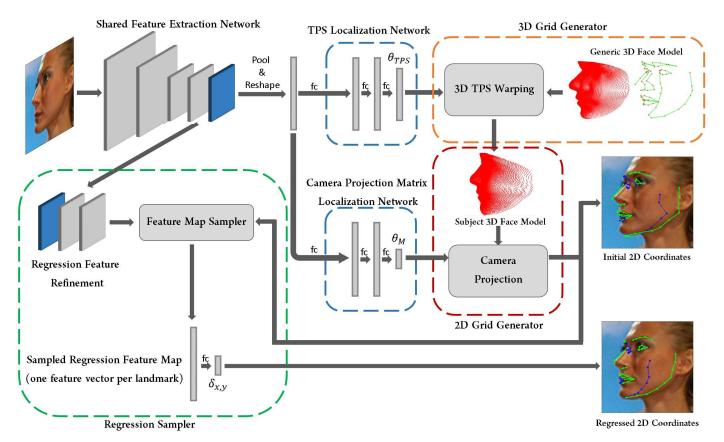
**Figure 3. Overview of model for detection transformers**

Self-attention matrices of large 2D and 3D teachers’ networks were used to break down information into smaller student network which have lesser decoder and encoder layers. Also, on top of attention distillation, the research also made used of DETR which is a cutting-edge object detection architecture.

## 3D Spatial Transformer Network for Facial Alignment and Unconstrained Poses

This research and experiment basically is used as part of face recognition and analysis technology and its purpose is to align faces which could affect facial recognition and analysis such as faces which are tilted, not facing the front, or not looking straight up. However, one of the constraints mentioned in the research paper was that the faces could not have any accessories such as glasses, masks, or hands coving part of the face.

There are many mathematical formulae and algorithms presented in detail throughout the research paper, however, this survey paper will not go into its depth but show a more general overview of the different aspects of work done in order to achieve face alignment. Below is the diagram presented in the research paper.



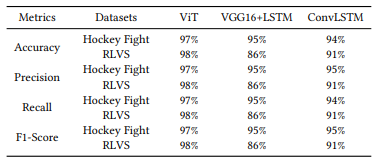
**Figure 4. Network design of the 3D TPS Spatial Transformer for facial alignment**

As a 3D model and a estimation of the position of the camera’s location can be determined by the network as an output, the visibility of the landmarks can be detected. Visible landmarks are denoted in green while the non-visible landmarks are denoted in blue (since in the figure above).

Some of techniques and computations used in the experiment includes camera projection transformers, 3D thin plate spline transformers, wrapped camera projection transformers, 2D landmark regression, and 3D model regression from 2D landmarks.

# Results

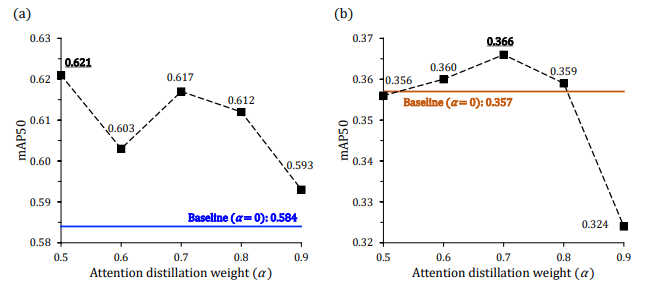
## ViT for Violence Detection from Real-time Videos



**Figure 5. Results comparison of ViT vs LSTM models**

As seen in the results table above, the proposed ViT model performed and produced better results compared to the two baseline LSTM models. The proposed ViT model have an overall accuracy of about 97 to 98% which is about 5% higher than the two baseline models.

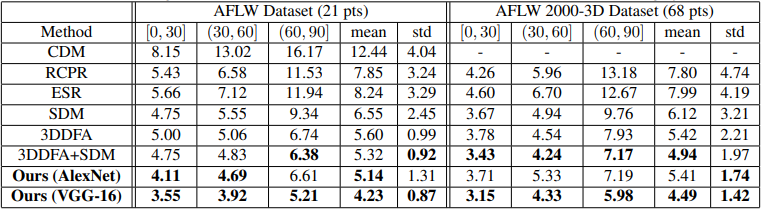
## Real-Time Video Object Detection in Ultrasound



**Figure 6. 3D attention distillation results from 0.5 to 0.9**

(a) shows the mAP results on test dataset and (b) shows the mAPshort on challenge dataset which have the short insertions only.

## 3D Spatial Transformer Network for Facial Alignment and Unconstrained Poses



**Figure 7. Normalized mean error (%) of face alignment results**

The best two numbers of each category are shown in bold. Thus, as seen from the figure above, the research paper’s proposed model which are AlexNet and VGG-16 generally performed better than the other existing models.

# Conclusion

As ViT is still a relatively new area of computer science, specifically, in the field of computer vision, object detection, and image classification, many more works need to be done to improve ViT as a whole. ViT, itself, is already very challenge but real-time ViT is taken to a whole new level and, indeed, there are works for improvements, said various research papers.

In this survey, three notable experiments out of the many have been shown and, although all three datasets and the application to ViT are vastly different, the general architecture of the ViT is still seemingly similar.

A common future work that was mentioned in most research papers was that it is agreeable that it is still extremely challenging to tackle real-time images, videos, or data and apply it into ViT as real-time have its inconsistency which have many possible such as the environment, terrain, lighting, and quality of the real-time data. Thus, ViT in real-time will still a challenge and, not only on the improvement on the ViT models and architectures, but also, on how can the ever-moving real-time data be more consistent so to have a more reliable model and trust-worthy results.

# Acknowledgements

My thanks to the 10 research papers which I read, understood, and summarized in order to write this survey paper.

# References

1. Arfin, S., Masood, S., Tarique Hesham, M., and Ahmed, A. 2021. *A Vision Transformer Model for Violence Detection from Real-Time Videos.* Research Paper.
2. Rubin, J., Erkamp, R., Srinivasa Naidu, R., Odungatta Thodiyil, A., and Chen, A. 2021. *Attention Distillation for Detection Transformers: Application to Real-Time Video Object Detection in Ultrasound.* Research Paper.
3. Munchmeyer, J., Bindi, D., Leser, U., and Tilmann, F. 2021. *EARTHQUAKE MAGNITUDE AND LOCATION ESTIMATION FROM REAL TIME SEISMIC WAVEFORMS WITH A TRANSFORMER NETWORK.* Research paper.
4. Bhagavatula, C., Zhu, C., Luu, K., and Savvides, M. 2021. *Faster Than Real-time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses.* Research Paper.
5. Yoo, H., Dai, L., Jung, T., Kim, S., and Chae, C. 2022. *Demo: Real-Time Semantic Communications with a Vision Transformer.* Research Paper.
6. Benmeziane, H., El Maghraoui, K., Ouarnoughi, H., and Niar, S. 2022. *Real-time Style Transfer with Efficient Vision Transformers.* Research Paper.
7. Chen, X., Yan, Bin., Zhu, J., Wang, D., Yang, X., and La, H. 2021. *Transformer Tracking.* Research Paper.
8. Zhao, H., Chen, Z., Guo, L., and Han, Z. 2021. *Video captioning based on vision transformer and reinforcement learning.* Research Paper.
9. Hussain, A., Hussain, T., Ullah, W., and Wook Baik, S. 2021. *Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos.* Research Paper.
10. Zhu, X., Jia, Y., Jian, Sun., Gu, L., and Fu, Zhang. 2021. *ViTT: Vision Transformer Tracker.* Research Paper.