Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2024.039436

ARTICLE

SwinVid: Enhancing Video Object Detection Using Swin Transformer

Abdelrahman Maharek[1,2,*], Amr Abozeid[2,3], Rasha Orban[1] and Kamal ElDahshan[2]

[1]Computer Science Department, Faculty of Artificial Intelligence and Informatics, Benha, Egypt

[2]Mathematics Department, Faculty of Sciences, Al-Azhar University, Cairo, Egypt

[3]Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Sakaka, Saudi Arabia

*Corresponding Author: Abdelrahman Maharek. Email: abdelrahman.mahareek@gmail.com

Received: 30 January 2023 Accepted: 11 May 2023 Published: 19 March 2024

ABSTRACT

What causes object detection in video to be less accurate than it is in still images? Because some video frames have degraded in appearance from fast movement, out-of-focus camera shots, and changes in posture. These reasons have made video object detection (VID) a growing area of research in recent years. Video object detection can be used for various healthcare applications, such as detecting and tracking tumors in medical imaging, monitoring the movement of patients in hospitals and long-term care facilities, and analyzing videos of surgeries to improve technique and training. Additionally, it can be used in telemedicine to help diagnose and monitor patients remotely. Existing VID techniques are based on recurrent neural networks or optical flow for feature aggregation to produce reliable features which can be used for detection. Some of those methods aggregate features on the full-sequence level or from nearby frames. To create feature maps, existing VID techniques frequently use Convolutional Neural Networks (CNNs) as the backbone network. On the other hand, Vision Transformers have outperformed CNNs in various vision tasks, including object detection in still images and image classification. We propose in this research to use Swin-Transformer, a state-of-the-art Vision Transformer, as an alternative to CNN-based backbone networks for object detection in videos. The proposed architecture enhances the accuracy of existing VID methods. The ImageNet VID and EPIC KITCHENS datasets are used to evaluate the suggested methodology. We have demonstrated that our proposed method is efficient by achieving 84.3% mean average precision (mAP) on ImageNet VID using less memory in comparison to other leading VID techniques. The source code is available on the website https://github.com/amaharek/SwinVid.

KEYWORDS

Video object detection; vision transformers; convolutional neural networks; deep learning

1 Introduction

Deep Convolutional Neural Networks (CNNs) have been employed to perform a variety of computer vision tasks, such as classifying and recognizing objects in still images and have significantly improved these techniques. These networks have also found applications in the field of medical image analysis, including the diagnosis of retinal disorders [1] and cancer [2]. In addition, CNNs have been used for finger vein recognition [3], which has potential applications in biometric identification

systems. The state-of-the-art object detection systems employ a two-stage approach known as the region proposal-based detection approach. The system initially produces regions of interest (RoI) using a Region Proposal Network (RPN), and afterwards classifies the objects found in those regions of Interest using a backbone network. The architectural advances of CNN-based backbone networks have led to a great performance improvement on the two-stage object detectors.

On the other hand, modeling in Natural Language Processing (NLP) has witnessed a great improvement led by using the Transformer [4]. The transformer architecture, which is optimized for sequence modeling, has made significant advancements in natural language processing because of its ability to use attention mechanisms to capture long-range dependencies. Given its outstanding performance in natural language processing, researchers are investigating how it may be adapted to computer vision tasks. Recently, the vision community has witnessed a modeling shift towards using transformer-based backbone architectures for classification and object detection. Vision Transformers [5,6] have shown a great improvement on the performance of classification and detection tasks of still images.

When compared to identifying objects in still images, video object detection (VID) is a challenging task. Although videos can be treated as a sequence of frames, the use of still-image object detectors on the frames of a video leads to poor object detection due to fast movements within the video [7]. Fast-moving objects can lead to problems such as blurring, out-of-focus shots, and altered postures that negatively impact the appearance of the objects, as illustrated in Fig. 1.

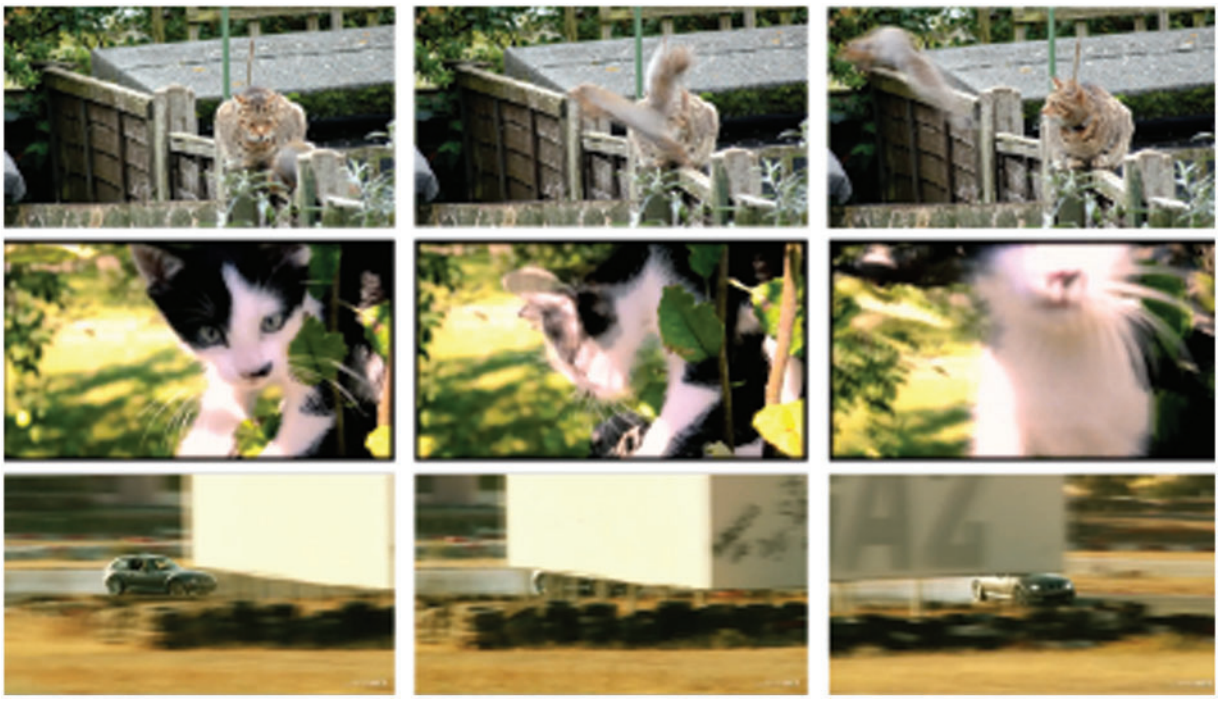

Figure 1: The degradation of appearance caused by fast-moving objects, which can include motion blur, changes in posture, and out-of-focus shots

Another limitation is the computational complexity of existing video object detection techniques. Some algorithms require extensive computing resources, making them unsuitable for real-time applications. Furthermore, as video resolutions and frame rates increase, the computational requirements

of these algorithms also increase, making them even more challenging to implement in real-world scenarios.

The lack of annotated data is another significant challenge in video object detection. It is often expensive and time-consuming to obtain high-quality annotated data for training these algorithms. As a result, many existing video object detection techniques rely on pre-trained models that may not generalize well to new and unseen data.

Despite these challenges, videos contain far richer temporal information about the same object from other nearby frames. Consequently, the primary challenge in enhancing detection performance is effectively utilizing information from other video frames to detect instances of the same object. An effective video detector should be able to recognize the semantically similar objects from notable views in order to enhance detection accuracy on degraded views.

In fact, many previous works are attempting to exploit the information from salient views to develop the accuracy of degraded views. SELSA [8] aggregates the high-level proposal features from multiple frames of the video to ensure that each proposal feature in the current frame includes information from other frames in the video. The final detection results are then obtained by the detection head using the aggregated features.

The aim of this study is to enhance the performance and efficiency of current video object detectors by proposing the use of Vision Transformer as the backbone network for video object detection. First, the feature maps are generated using the standard Swin-Transfomer block [6]. The feature maps produced by the backbone have the same resolution as those produced by ResNet [9]. The object proposals are then obtained using the Region Proposal Network (RPN). The RPN generates proposals from different video frames which are handed together with the feature map to the final classification layer. The ImageNet VID [10] and EPIC KITCHENS [11] datasets are used to evaluate the suggested methodology. We have demonstrated that our proposed method is efficient by achieving 84.3% mean average precision (mAP) on ImageNet VID using less memory in comparison to other state-of-the-art VID techniques. The proposed architecture enhances the accuracy of existing VID methods. Our experiments have demonstrated that using the suggested backbone network leads to a 1.2% increase in mean Average Precision (mAP) for the SELSA [8] method with less memory usage.

This paper is organized into six sections. Section 2 presents a related work on video object detection. Section 3 discusses the proposed SwinVid model. Section 4 presents the implementation details and discusses the experimental results on the ImageNet VID dataset. Additional experiments on the Epic Kitchens dataset and applying the proposed model to Video Instance Segmentation (VIS) are discussed in Section 5. Section 6 presents the paper's conclusion.

2 Related Work

2.1 Object Detection in Still Images

Object detection networks have made significant strides in the past decade thanks to the evolution of deep Convolutional Neural Networks (CNNs). The most effective object detection networks can be grouped into two main architectural frameworks:

(a) **Region proposal-based object detection networks (two-stage).** Initially, the RPN generates a collection of region proposals that are subsequently processed by feature extraction modules such as RoIPooling [12] and RoIAlign [13]. Subsequently, the detection head employs the extracted features to perform bounding box regression and classification in the second stage.

(b) **Regression/classification-based object detection networks (one-stage).** These networks approach the detection problem as a regression or classification problem by considering every position in the image as a potential object and then classifying each region of interest (RoI) to its corresponding class.

The typical two-stage network R-CNN [14] detects objects in still images using multiple stages: Initially, selective search [15] is devised to obtain the set of object proposals. Afterwards, the regional features are extracted using a CNN-based backbone network. Finally, the bounding boxes are classified and adjusted in a localization stage. Fast R-CNN [12] suggests extracting proposal features using an embedded Region of Interest (RoI) pooling layer between the final CONV layer and the first FC layer. Typically, region proposals are obtained by using selective search [15]. An extra Region Proposal Network (RPN) was developed by Faster R-CNN [9,16] framework. Instead of re-calculating the CONV features, the RPN operates nearly cost-freely because it shares the full image CONV features with the detection Network. To extract features for a proposal within a still image, these methods adopt RoI Pooling [12] or RoI Align [13]. As a result, the features obtained by RoI Align lack the temporal information in videos when applied to Video Object Detection (VID).

Unlike the two-stage object detectors, one-stage object detectors like the YOLO series [17–19], SSD [20], and RetinaNet [21], are also widely used for object detection tasks. One-stage detectors differ from two-stage detectors by making dense predictions on feature maps generated by a CNN, without the need for region proposals. In addition, they deliver both position and class probabilities. The one-stage method is faster as it can be optimized as a whole, rather than in individual stages. One-stage detectors, however, can hardly be extended for other vision tasks like key point detection, instance segmentation, or video object detection (VID). The two-stage approach is therefore adopted by the current proposed model.

2.2 Video Object Detection

Video object detection is the process of detecting and identifying objects within a video stream. This typically involves analyzing each frame of the video and using computer vision techniques to identify and track objects within the frame, such as people, cars, or animals. The aim of video object detection is to detect and classify objects within a video accurately and in real time, which has various applications like surveillance, self-driving cars, intelligent healthcare, and video analysis.

Deep Convolutional Neural Networks (CNNs) have played a vital role in the advancement of object detection and other computer vision fields. Many researchers have attempted to adapt CNN-based object detectors for use on video data. How to efficiently employ the temporal information from the video to enhance the performance of the detector on individual frames on individual frames is a significant challenge in object detection in videos [22].

One approach used to identify objects from a video stream is to apply post-processing methods [23,24] on the output of a still image detector that utilizes temporal information. Through the creation of object tubelets and the careful adjustment of the final classification and bounding box, that method aims to enhance the predictions of the still image detector over subsequent video frames.

Another approach to improve the keyframes involves using features from nearby reference frames to enhance the quality of the keyframes. This can be done through a variety of methods, including those based on optical flow, attention, and tracking. These methods seek to alleviate degradation in the keyframes by using information from nearby frames to fill in missing or degraded information. The three primary groups of these methods are optical flow-based, attention-based, and tracking-based.

Flow-based methods use the movement of pixels between frames, known as optical flow, to identify the transformations that have occurred between the frames. This information is used to improve object detection by providing additional context about the motion of objects in the scene. One advantage of optical flow-based methods is that they are computationally efficient, making them suitable for real-time VID applications. However, they may struggle to handle occlusions, and their accuracy can degrade in the presence of significant camera motion or scene clutter. Deep Feature Flow (DFF) [25] was the first approach to use fine-tuned optical flow computation within a network. It makes use of the optical flow calculated by FlowNet [26] to transfer and align the features of chosen keyframes to surrounding non-keyframes, therefore minimizing extra computations and boosting system performance. By aligning and aggregating features from keyframes using optical flow, the Flow-Guided Feature Aggregation (FGFA) [27] technique, an extension of the Deep Feature Flow (DFF) [25] method, aims to enhance the efficiency.

To reduce the high computational cost of aggregating features at the image level, several attention-based methods have been developed. SELSA [8] introduces a long-range feature aggregation method based on semantic similarity between region-level features. A memory-enhanced global-local aggregation module is used by MEGA [28] to more efficiently capture relationships between the instances of objects in different video frames. Temporal ROI [29] performs ROI alignment for fine-grained feature aggregation, and HVR-Net [30] incorporates intra-video and inter-video proposal relations to achieve additional improvement. Due to the high computational cost associated with video detectors, QueryProp [31] proposes a lightweight module to boost the efficiency of video object detection.

Besides attention-based techniques, D&T [32] solves video object detection by tracking the object using correlation maps of various frame features. Even though these methods can increase detection accuracy, they commonly rely on two-stage detectors, which can result in a decline in inference speed.

While various methods for video object detection exist, most prior works utilize a CNN-based backbone to extract the feature maps. In this study, we propose using the Vision Transformer as the backbone network for video object detectors. As per our experiments, incorporating the Vision Transformer leads to an improvement in the performance of existing VID methods with less memory used to detect objects from video frames.

2.3 Vision Transformer

Due to its capability to leverage attention methods to capture long-range dependencies, the Transformers architecture [4] was initially presented in 2017 for machine translation and sequence modeling. Although Transformers are considered the standard for Natural Language Processing (NLP), their applications to computer vision are still relatively limited.

With the increasing popularity of the Vision Transformer (ViT) [5] and Swin-Transformer [6], there has recently been a modeling shift in the architecture employed for computer vision from Convolutional Neural Networks (CNNs) to Vision Transformers [5,6]. Further research to expand the functionality of vision transformers has been inspired by this innovative work. The Vision Transformer (ViT) must, however, be trained on a large dataset like JFT-300M. DeiT [33] is a method that was developed to make Vision Transformer (ViT) more efficient on smaller datasets, such as ImageNet-1k, by incorporating various training strategies. Despite the promising results of Vision Transformer (ViT) in image classification tasks, its architecture makes it less suitable to be used as a general-purpose backbone for dense vision tasks or when the input image has a high resolution. This is because ViT has low-resolution feature maps and its complexity increases quadratically with the increase in image size. To address this limitation, the Swin-Transformer [6] was developed to incorporate inductive biases

such as locality, hierarchy, and translation invariance, making it a suitable general-purpose backbone for different image recognition tasks.

One of the main advantages of ViT over CNNs is the ability to capture long-range dependencies in an image. ViT applies self-attention mechanisms to capture these dependencies, allowing the network to understand the relationships between pixels in the image without relying on local information. This is especially useful for object detection tasks, where the objects of interest can be located in different parts of the image.

Another advantage of ViT is its ability to process input images of arbitrary size. In contrast, CNNs require input images of fixed size, which can be a limitation in some applications. With ViT, input images can be split into smaller patches, and each patch is processed independently. This allows ViT to handle images of varying sizes without requiring resizing or cropping.

In VID, the Swin-Transformer architecture has demonstrated improvements over traditional CNN-based approaches. Swin-Transformer applies a hierarchical architecture that processes images at multiple scales, allowing it to capture both local and global features. This architecture enables Swin-Transformer to handle objects at different scales and resolutions, making it suitable for VID tasks with varying object sizes.

Swin-Transformer also employs a window-based self-attention mechanism that allows it to capture long-range dependencies in an image. This feature enables Swin-Transformer to recognize complex object interactions and occlusions in a video, leading to improved object detection accuracy. Additionally, Swin-Transformer's ability to handle input images of arbitrary size makes it suitable for VID tasks that involve video sequences with varying resolutions and frame rates.

This work presents a transformer-based backbone for Video Object Detection (VID) and aims to shift the use of these types of architectures in VID models. Currently, the most commonly employed architecture for Video Object Detection (VID) is Convolutional Neural Network (CNN) and its variants. However, unified modeling between vision and Natural Language Processing (NLP) tasks may be possible with the use of Transformer-based backbones for vision tasks.

3 Method

In this section, the motivation for using a Transformer-based backbone network to improve video object detection is described in Section 3.1. The use of the Swin-Transformer with a state-of-the-art video object detector (VID) and the full architecture is then discussed in Section 3.2. Finally, in Section 3.3, the architecture variants are discussed.

3.1 Motivation

The major challenge to boosting object-detection accuracy and speed in videos is how to use the additional information offered by the video frames to improve the accuracy. Feature aggregation has been shown to be an effective method for addressing the degradation of appearance in video object detection (VID) [8]. Sequence-Level Semantics Aggregation (SELSA) [8] combines the high-level proposal features from multiple frames to ensure that every proposal feature in the current frame includes information from other frames. The final detection results are obtained by passing the aggregated features to the detection head. The detection head uses Faster-RCNN [16] and ResNet-101 [9] as the backbone network to produce feature maps. This study investigates the feasibility of using Transformer-based backbones to generate the feature map for Video Object Detection (VID) by

adapting the standard Swin-Transformer [6]. The backbone network produces feature maps using the same feature vector resolutions as standard convolutional networks but using less memory.

3.2 Network Architecture

Fig. 2 depicts the proposed method architecture. The version illustrated in Fig. 2, (SwinVid-T) is the tiny version of the suggested model. It first extracts the feature maps from the input frames using the standard tiny version of the Swin-Transformer block [6]. The feature maps produced by the backbone have the same resolution as those produced by ResNet [9]. The object proposals are then obtained using the Region Proposal Network (RPN).

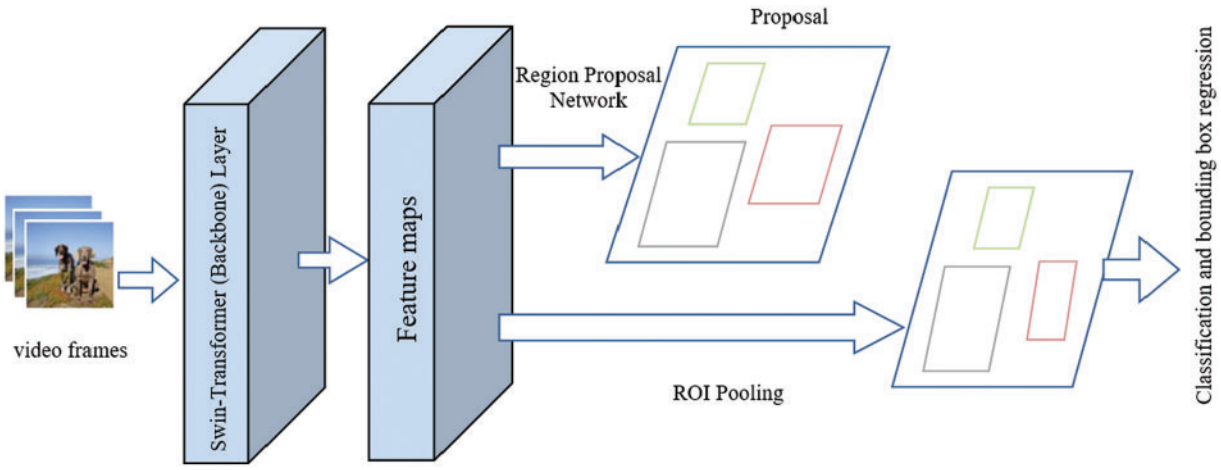

Figure 2: The architecture of the proposed model SwinVid. Feature maps are generated using the standard Swin-Transfomer block. Region Proposal Network (RPN) generates proposals from different video frames which are handed together with the feature map to the final classification layer

The architecture of standard Swin-Transformer architecture, specifically the tiny version (Swin-T) is shown in Fig. 3. Using a patch-splitting module similar to ViT [5], it starts by splitting input RGB video frames into non-overlapping patches. Each patch has a size of 4×4 and is considered as a "token", such that each patch has a feature dimension of $4 \times 4 \times 3 = 48$. A linear embedding layer is then used to project the raw-valued feature to the desired dimension (denoted as C).

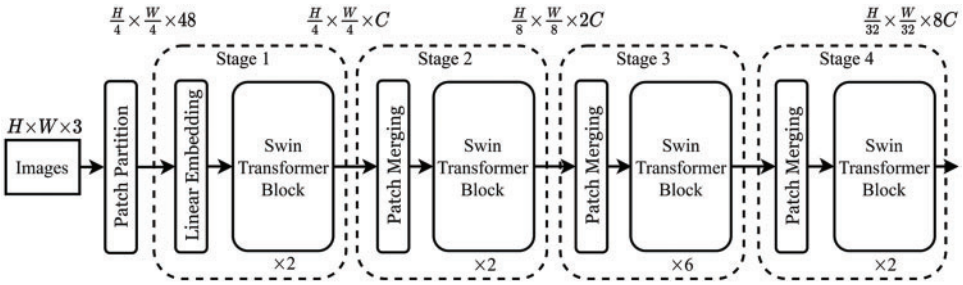

Figure 3: The structure of the tiny version of the Swin-Transformer [6] used as the backbone module

A group of modified self-attention Transformer blocks (Swin Transformer blocks) [6] are utilized on these patch tokens. The first stage (denoted as "Stage 1") incorporates both the linear embedding and Transformer blocks to retain the same number of tokens $\left(\frac{H}{4} \times \frac{W}{4}\right)$.

To decrease the tokens number as the network becomes deeper and generate a hierarchical representation, a patch-merging layer is employed. Nearby patches of the size 2×2 as treated as a "group". Each group is concatenated using the first patch merging layer, resulting attached concatenated feature of size 4C. This process reduces the number of tokens by a factor of 4 and sets the output dimension to 2C. Afterwards, feature transformation is performed using Swin-Transformer blocks, while maintaining a resolution of $\left(\frac{H}{8} \times \frac{W}{8}\right)$. The second stage, "Stage 2" consists of the initial patch merging and feature transformation. This initial patch merging and feature transformation block are called "Stage 2". The same transformation is repeated again for two additional stages, labeled "Stage 3" and "Stage 4". When combined, these phases result in a hierarchical representation with feature map resolutions comparable to common CNNs.

As a result, the backbone networks in current video object detection techniques and other vision tasks can be replaced by the Swin-Transformer backbone.

A feature aggregation module is used to reduce appearance degradation in video frames. We employ the standard SELSA [8] module which aggregates features from the semantic neighborhood based on semantic similarities.

Semantic Guidance: For each frame f, the Region Proposal Network (RPN) produces proposals set, denoted as $X^f = \left\{x_1^f, x_2^f, \ldots\right\}$. The semantic similarity between any two proposals $\left(x_i^k, x_j^l\right)$, in different frames is calculated using the generalized cosine similarity in Eq. (1).

$$w_{ij}^{kl} = \phi(x_i^k)^T \psi\left(x_j^l\right) \tag{1}$$

where ϕ (.) and ψ (.) are some global transformation functions. A higher degree of similarity means that two proposals are more likely to be in the same category.

Feature Aggregation: The semantic similarity between proposals has been determined and is being used to guide the creation of a reference proposal by combining features from multiple proposals. This new proposal feature is more comprehensive and resilient to changes in appearance such as pose changes, motion blur, and object deformation. Because the cosine similarity is calculated between proposals rather than the entire frame, the feature aggregation process is more reliable.

The similarities are normalized across all proposals using the SoftMax function to ensure that the magnitude of features is preserved after aggregation. Specifically, given a video with F randomly selected frames and N proposals generated from an individual video frame, the reference proposal's aggregated feature is calculated as in Eq. (2).

$$\overline{x}_i^k = \sum\nolimits_{l \in \Omega} \sum\nolimits_{j=1}^{N} \omega_{ij}^{kl} x_j^l \tag{2}$$

The SELSA module uses a set of frame indexes, designated as Ω, that are randomly chosen for the aggregation process. This module can be optimized using Stochastic Gradient Descent (SGD since its fully differentiable. Afterwards, the detecting head network is then fed with the improved proposal features. The diagram in Fig. 4 illustrates the functioning of the SELSA module.

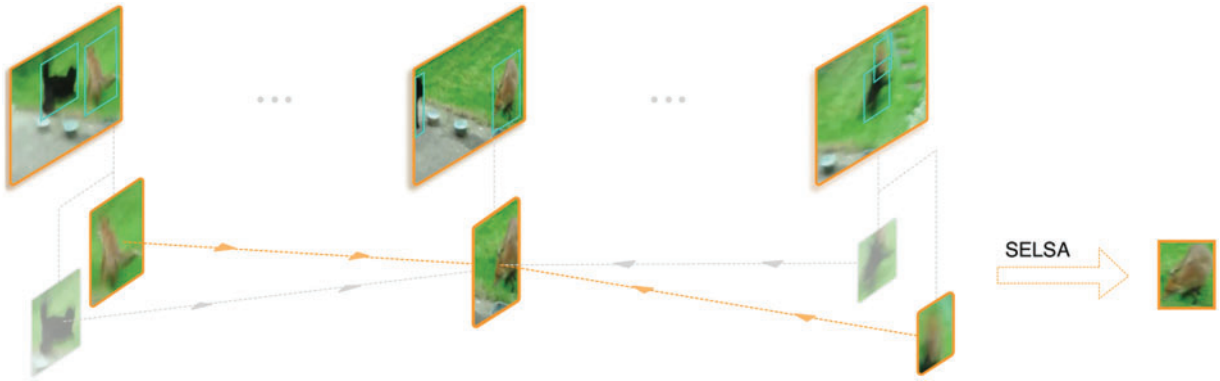

Figure 4: The architecture of the SELSA model [8]. Proposals are first extracted from different frames in a video. Semantic similarities are then calculated between these proposals. In order to develop robust features for object detection, the features from the proposals are finally aggregated based on these similarities. This procedure is employed to enhance the accuracy and reliability of object detection by considering multiple proposals and their relationships rather than relying on a single proposal or frame

3.3 Architecture Variants

We present two variations of the SwinVid model, known as SwinVid-T and SwinVid-S, which are based on the SELSA module with Swin-T and Swin-S, respectively. The default window size for both models is M = 7. The query dimension per head in the model is d = 32, and the expansion layer for each MLP is 4 for all experiments. The following are the hyperparameters for the architectural design of these model versions:

- Swin-T: C = (96), number of layers = {2, 2, 6, 2}
- Swin-S: C = (96), number of layers = {2, 2, 18, 2}

where C is the first stage's hidden layer's channel number.

4 ImageNet-VID Dataset Experiments

Initially, we will go through the datasets and evaluation metrics that were utilized for video object identification (VID) in Section 4.1. Afterward, we will present the implementation details of our method in Sections 4.2 and 4.3. Finally, the performance of our method will be compared against other related methods for video object detection on the ImageNet-VID dataset in the experimental Sections 4.4 and 4.5.

4.1 Dataset and Metrics of Evaluation

The suggested model (Swin-VID) is trained with a dataset consisting of ImageNet VID (30 object categories, 3862 video frames for training, and 555 video frames for validation) and DET datasets using the split specified in FGFA [27]. Each video is subsampled to a maximum of 15 frames, and the VID: DET balance is roughly 1:1. We test the performance of the proposed model on the ImageNet VID dataset [10]. The mean average precision (mAP) at the intersection of union (IoU) of 0.5 (mAP@IoU = 0.5) is reported.

The mean Average Precision (mAP) is a widely used metric to evaluate object detection algorithms, including video object detection (VID) algorithms. It is an extension of the AP metric, which computes the average precision across different object categories. In VID, mAP is computed by averaging the AP scores for each object category across all frames in a video sequence. mAP provides a comprehensive measure of the algorithm's performance by considering the detection accuracy across different object categories and frames. It is an essential metric for comparing the performance of different VID algorithms and selecting the best performing algorithm.

4.2 Implementation Details

Backbone Module: For the purpose of conducting ablation studies, the tiny version (Swin-T) [6] is used as the basic unit to create the feature map. The small version (Swin-S) [6] is also used to report the final results.

Detection Module: RPN is placed on the output of the feature map to generate proposals. The total number of anchors is 12, comprising of 4 scales $\{64^2, 128^2, 256^2, 512^2\}$, and 3 aspect ratios {1:2, 1:1, 2:1} producing a total of 300 proposals on each image. The RoI pooled features are first passed through two fully connected (FC) layers with 1024 units (also known as neurons or dimensions) each. These layers are used to extract important information from the features and reduce their dimensionality. After this, the output is passed through a classification layer that is used to predict the class of the object in the RoI, and a bounding box regression layer that is employed to predict the location of the object in the image. This pipeline is commonly used in object detection tasks using deep learning.

SELSA Module: This network design is incorporating two SELSA [8] modules after the Fully Connected (FC) layers. These modules are incorporated into the pipeline after the fully-connected layers, with the sequence being (FC $\rightarrow$ SELSA $\rightarrow$ FC $\rightarrow$ SELSA). This arrangement allows the SELSA modules to be integrated into the network architecture and contribute to the overall performance of the network.

4.3 Training and Testing Details

Pre-trained weights from the ImageNet VID dataset are utilized to initialize the backbone network. A batch size of 8 on 8 GPUs is used to train the networks using SGD for a total of 3 epochs. The learning rate is decreased by a factor of 10 at the 110 k and 165 k iterations from the original setting of 2.5×10^{-4}. A single frame and two additional randomly chosen frames from the same video are sampled during training (or the DET dataset's identical frames). During the inference phase, K frames are sampled from the same video as the inference frame, and all images are resized to have a shorter side of 600 pixels. This method is employed to make certain that the networks are effectively trained and capable of performing well in the task of identifying objects in videos.

4.4 The Effectiveness of SwinVid

SwinVid is proposed to improve the existing video object detectors by using a transformer-based backbone network. As shown in Table 1, the vision transformer-based backbone has improved the effectiveness of the standard SELSA module [8] on the ImageNet VID dataset [10] by 3.1% and 1.2% when replaced ResNet-50 and ResNetXt-101, respectively. Table 1 illustrates that the proposed method has enhanced the mAP of SELSA and achieved better results than Temporal ROI [29]. Fig. 5 shows the visual results of our method.

Table 1: The performance of different video object detectors and our detector on ImageNet-VID

Method	Backbone	mAP (%)
SELSA [1]	ResNet-50	78.4
Temporal ROI [2]	ResNet-50	79.8
SELSA [1]	ResNet-101	81.5
Temporal ROI	ResNet-101	82.6
SELSA	ResNeXt-101	83.1
Temporal ROI	ResNeXt-101	84.1
SwinVid (ours)	Swin-T	80.1
SwinVid (ours)	**Swin-S**	**84.3**

Figure 5: (Continued)

Figure 5: Visualizations of our method on YouTube VID dataset [34]

4.5 The Performance of SwinVid

SwinVid has shown a better mean Average Precision (mAP) than other Video Object Detectors. Additionally, the memory utilized is less than the memory utilized by Temporal ROI [29] and the standard SELSA module (ResNetXt101) as shown in Table 2.

Table 2: The memory usage of different video object detectors and our detector on ImageNet VID

Method	Backbone	Memory (GB)
SELSA [1]	ResNet-50	3.49
Temporal ROI [2]	ResNet-50	4.14
SELSA [1]	ResNet-101	5.18
Temporal ROI	ResNet-101	5.83
SELSA	ResNeXt-101	9.15
Temporal ROI	ResNeXt-101	9.74
SwinVid (ours)	Swin-T	4.12
SwinVid (ours)	**Swin-S**	**6.94**

As shown in Table 2, our method has also used a smaller amount of memory (6.94 GB) compared to Temporal ROI (9.74 GB) and the standard SELSA (9.15 GB) which shows the potential of replacing the CNN-based backbones with the vision transformer.

5 Additional Experiments

5.1 Experiments on Epic Kitchens

EPIC KITCHENS [11] is a large scale egocentric dataset, capturing daily activities happened in the kitchens. The EPIC KITCHENS dataset is more complicated and difficult, as each frame contains an average or maximum of 1.7 or 9 objects. The task of detecting objects in videos is challenging, as there are 454,255 object bounding boxes spanning 290 classes across 32 different kitchens. The dataset includes 272 video sequences captured in 28 kitchens for training and 106 sequences collected from

the same 28 kitchens (S1), as well as 54 sequences collected from four other kitchens that were not used during training (S2), for evaluation. The videos are annotated in one-second intervals.

We adopt the same network configuration as in the ImageNet VID dataset and do not use any data augmentation except for random horizontal flipping. We train the model using SGD for 600,000 iterations on four GPUs, starting with an initial learning rate of 2.5×10^{-4} that decreases by a factor of 10 after 300,000 iterations. During both training and testing, we select frames within a range of ± 10 s for the SELSA module.

Here we present some preliminary results on the EPIC KITCHENS dataset. As shown in Table 3, SwinVid improves over Temporal ROI [29] by 1.3/2.1 mAP for Seen/Unseen splits. Although the hyper parameters selection are far from being optimal, our method still achieves promising results. This shows that SwinVid is applicable to more complex video detection tasks.

Table 3: Performance comparison on EPIC KITCHENS validation set. S1 and S2 indicate seen and unseen splits

Method	mAP@0.5 (S1)	mAP@0.5 (S2)
SELSA [8]	38.0	34.8
SELSA + TROI [29]	42.2	39.6
SwinVid (ours)	**43.5**	**41.7**

5.2 Application to Video Instance Segmentation

Video instance segmentation is a computer vision task that involves identifying and segmenting each instance of an object in a video sequence. Unlike image segmentation, which only focuses on identifying and segmenting objects in a single image, video instance segmentation requires tracking and segmenting objects across multiple frames in a video.

Video instance segmentation is a challenging task because it requires not only accurate object detection and segmentation but also the ability to track objects over time, even when they move out of frame or are occluded by other objects.

We investigate SwinVid on Video Instance Segmentation (VIS). The dataset of VIS is YouTube-VIS [34] which contains 40 object categories. The dataset comprises of 2238 videos for training, 302 videos for validation and 343 videos for testing. The training process utilizes the training set, while the validation set is used for evaluation since the test set is not currently available.

MaskTrack R-CNN [34] is a variant of Mask R-CNN [13] that incorporates a track head to associate object instances across frames. MaskTrack R-CNN uses the ResNet-101 as the backbone network to extract the feature maps, and we replace ResNet-101 backbone with the Swin-Transformer [6] backbone. The results are shown in Table 4. We can see that the SwinVid consistently improves ResNet-50 and ResNeXt-101 baselines on all metrics involving AP, AP50 and AP75, which further demonstrates the flexibility of proposed SwinVid.

Table 4: Applying the Swin-Transformer to MaskTrack R-CNN in VIS. AP denotes mask AP which follows the COCO evaluation metric to use 10 IoU thresholds from 50% to 95% at step 5%

Method	Backbone	AP	AP_{50}	AP_{75}
MaskTrack R-CNN [34]	ResNet-50	30.3	51.1	32.6
MaskTrack R-CNN + TROI [29]	ResNet-50	33.5	57.0	36.6
MaskTrack R-CNN [34]	ResNeXt-101	34.9	58.8	36.5
MaskTrack R-CNN + TROI [29]	ResNeXt-101	38.0	63.3	40.3
SwinVid (ours)	Swin-T	36	60	40.3
SwinVid (ours)	**Swin-S**	**38.3**	**60.7**	**41.4**

6 Conclusion

Video object detection, as opposed to conventional object detection, aims to detect objects in video data rather than static images. Numerous applications, such as video surveillance, healthcare monitoring and autonomous driving, have played a significant role in the advancement of video object detection research.

This paper presents SwinViD, a technique that utilizes the standard Swin Transformer as the backbone module to enhance the efficiency of Video Object Detection (VID). The Swin-Transformer backbone utilized in SwinViD generates a hierarchical representation that shares the same feature vector resolutions as conventional CNNs, and it can be easily implemented to replace CNN-based backbones in existing video object detectors, resulting in enhanced performance and reduced memory consumption. We test our method on ImageNet-VID, EPIC KITCHENS and YouTube VIS datasets to demonstrate its effectiveness. The experimental results demonstrated that our proposed method is efficient by achieving 84.3% mean average precision (mAP) on ImageNet VID using less memory in comparison to other leading VID techniques. These results shows the potential of our proposed method and the potential of using the ViT as a backbone network for video object detectors.

Going forward, we plan to test the transformer-based backbone on other video detection techniques and aim to enhance the efficiency of video object detectors by fully utilizing the temporal dimension through an end-to-end transformer model. We plan to explore the potential of using SwinVid for real-time video object detection and tracking.

Acknowledgement: None

Funding Statement: The authors received no funding for this study.

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and design: A. Maharek, A. Abozeid, R. Orban, K. ElDahshan; data collection: A. Maharek, A. Abozeid, K. ElDahshan; experiments on various datasets: A. Maharek, A. Abozeid, K. ElDahshan; analysis and interpretation of results: A. Maharek, A. Abozeid, R. Orban, K. ElDahshan; draft manuscript preparation: A. Maharek, A. Abozeid, R. Orban, K. ElDahshan. All authors actively participated in discussions, provided critical insights, and reviewed the results. All authors have read and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository. The data that support the findings of this study are openly available in [ImageNet VID] at https://image-net.org/challenges/LSVRC/2017/ and [EPIC Kitchens] at https://epic-kitchens.github.io/2023.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Bilal, L. Zhu, A. Deng, H. Lu and N. Wu, "AI-based automatic detection and classification of diabetic retinopathy using U-Net and deep learning," *Symmetry*, vol. 14, no. 7, pp. 1427, 2022.

[2] A. Bilal, G. Sun, Y. Li, S. Mazhar and J. Latif, "Lung nodules detection using grey wolf optimization by weighted filters and classification using CNN," *Journal of the Chinese Institute of Engineers*, vol. 45, no. 2, pp. 175–186, 2022.

[3] A. Bilal, G. Sun and S. Mazhar, "Finger-vein recognition using a novel enhancement method with convolutional neural network," *Journal of the Chinese Institute of Engineers*, vol. 44, no. 5, pp. 407–417, 2021.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems*, California, USA, 2017.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. of Int. Conf. on Learning Representations*, Vienna, Austria, 2021.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei *et al.,* "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Montreal, Canada, pp. 10012–10022, 2021.

[7] C. Xu, Z. Gao, H. Zhang, S. Li and V. H. C. J. A. S. C. de Albuquerque, "Video salient object detection using dual-stream spatiotemporal attention," *Applied Soft Computing*, vol. 108, no. 12, pp. 107433, 2021.

[8] H. Wu, Y. Chen, N. Wang and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 9217–9225, 2019.

[9] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 770–778, 2016.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh *et al.,* "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[11] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari *et al.,* "Scaling egocentric vision: The epic-kitchens dataset," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 720–736, 2018.

[12] R. Girshick, "Fast R-CNN," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.

[13] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2961–2969, 2017.

[14] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, United States, pp. 580–587, 2014.

[15] J. R. Uijlings, K. E. van de Sande, T. Gevers and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[16] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[17] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 779–788, 2016.

[18] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 7263–7271, 2017.
[19] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "YOLOV4: Optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.
[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 21–37, 2016.
[21] H. Hu, J. Gu, Z. Zhang, J. Dai and Y. Wei, "Relation networks for object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 3588–3597, 2018.
[22] E. Mofreh, A. Abozeid, H. Farouk and K. A. El-Dahshan, "Multi-object semantic video detection and indexing using a 3D deep learning model," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 3, pp. 268–280, 2022.
[23] K. Kang, W. Ouyang, H. Li and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 817–825, 2016.
[24] A. Sabater, L. Montesano and A. C. Murillo, "Robust and efficient post-processing for video object detection," in *Proc. of 2020 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Las Vegas, Nevada, USA, pp. 10536–10542, 2020.
[25] X. Zhu, Y. Xiong, J. Dai, L. Yuan and Y. Wei, "Deep feature flow for video recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 2349–2358, 2017.
[26] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas *et al.,* "FlowNet: Learning optical flow with convolutional networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 2758–2766, 2015.
[27] X. Zhu, Y. Wang, J. Dai, L. Yuan and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 408–417, 2017.
[28] Y. Chen, Y. Cao, H. Hu and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, Washington, USA, pp. 10337–10346, 2020.
[29] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu *et al.,* "Temporal ROI align for video object recognition," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Virtual, pp. 1442–1450, 2021.
[30] M. Han, Y. Wang, X. Chang and Y. Qiao, "Mining inter-video proposal relations for video object detection," in *Proc. of European Conf. on Computer Vision*, Glasgow, Scotland, pp. 431–446, 2020.
[31] F. He, N. Gao, J. Jia, X. Zhao and K. Huang, "QueryProp: Object query propagation for high-performance video object detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Vancouver, BC, Canada, 2022.
[32] C. Feichtenhofer, A. Pinz and A. Zisserman, "Detect to track and track to detect," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 3038–3046, 2017.
[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles *et al.,* "Training data-efficient image transformers & distillation through attention," in *Proc. of Int. Conf. on Machine Learning*, Vienna, Austria, pp. 10347–10357, 2021.
[34] L. Yang, Y. Fan and N. Xu, "Video instance segmentation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 5188–5197, 2019.