PTSEFormer: Progressive Temporal-Spatial Enhanced TransFormer Towards Video Object Detection

Han Wang[1], Jun Tang[2], Xiaodong Liu[2], Shanyan Guan[3], Rong Xie[1], and Li Song[1,3⋆]

[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2] HIKVISION Inc.
[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

Abstract. Recent years have witnessed a trend of applying context frames to boost the performance of object detection as video object detection. Existing methods usually aggregate features at one stroke to enhance the feature. These methods, however, usually lack spatial information from neighboring frames and suffer from insufficient feature aggregation. To address the issues, we perform a progressive way to introduce both temporal information and spatial information for an integrated enhancement. The temporal information is introduced by the temporal feature aggregation model (TFAM), by conducting an attention mechanism between the context frames and the target frame (*i.e.*, the frame to be detected). Meanwhile, we employ a Spatial Transition Awareness Model (STAM) to convey the location transition information between each context frame and target frame. Built upon a transformer-based detector DETR, our PTSEFormer also follows an end-to-end fashion to avoid heavy post-processing procedures while achieving 88.1% mAP on the ImageNet VID dataset. Codes are available at https://github.com/Hon-Wong/PTSEFormer.

Keywords: video object detection, transformer

1 Introduction

Video Object Detection (VOD) [34,29,36,33,4] has emerged as a hot topic in computer vision. Given a target frame and its context frames, VOD aims to detect objects in the target frame, with the compensation of observation from context frames. By observing the same instance in different poses from context frames, many hard cases, such as blurry appearance and background occlusion, are possible to be tackled.

Previous works [4,13,11,33] usually aggregate features at one stroke, suffering from insufficient utilization of temporal information. In particular, they employ

⋆ Corresponding author.

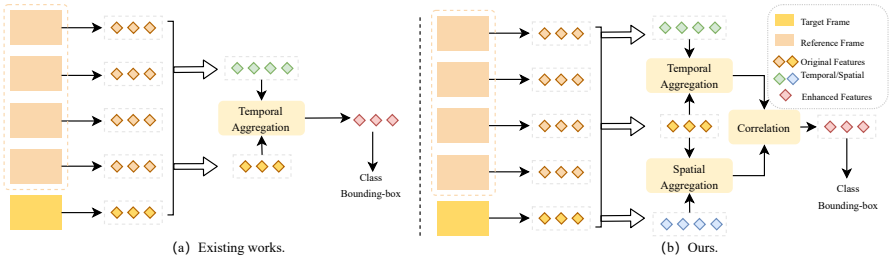

Fig. 1. The differences between existing works and ours. Previous works usually conduct temporal feature aggregation at one stroke, lacking in spatial information and suffering from insufficient feature aggregation. In contrast, our PTSEFormer utilizes both spatial and temporal information and performs feature aggregation in a progressive way.

isolated box-level associations [4,13,11] to enhance the instance feature of the target frame only using the extracted features of proposals, ignoring the spatial relations between frames. To diversify context frame features, those works put effort into how to excavate information from long-range context frames. However, as a common sense of human vision, information from a nearby time window is enough for detection in most scenarios. Specifically, when distinguishing a blurry object from the target frame, we often refer to the frame sliding near the target frame temporally, instead of observing the whole video. In this way, how to fully utilize the information from context frames, rather than enlarging the range of context frames, should be valued in the first place.

In this paper, we propose PTSEFormer to tackle the problems mentioned above. Motivated by DETR [2,35], PTSEFormer uses Transformer [27] as the basic structure to avoid complicated post-processing (*e.g.,* Seq-NMS [14], Tublet-Linking [17], Viterbi [10], Tublet-Rescore [3]). In contrast to aggregating features of the target frame and context frames at one stroke by attention layers [4,13,33] and conducting box-level associations upon extracted proposals [4,13,11], PTSEFormer conducts a progressive way to focus on both the temporal information and the spatial transition relations between frames. Specifically, **Temporal Feature Aggregation Module** is designed to introduce the temporal information to enhance the feature of the target frame with different perspectives towards the same objects in all the context frames. **Spatial Transition Awareness Module** is designed for estimating the position transition of the objects between the target frame and each context frame, enhancing the target feature with frame-to-frame spatial information. To build a balanced correlation model upon transformer decoder, we further propose the Gated Correlation model, which considers the imbalance caused by the residual connection layer and adds a gate to fix it.

Furthermore, as an important design of DETR, object queries contain inherent object position distribution learned from training data, and are fixed during inferring. We propose the Query Assembling Module(QAM) to regress object

queries directly from context frames. Due to the fact that it is more reasonable to infer position from adjacent context frames, rather than from fixed parameters decided by training data.

We conduct extensive experiments on ImageNet VID dataset [24] and achieve a **4.9%** absolute improvement on mAP compared to previous end-to-end state-of-the-art method [13] and **3.3%** absolute improvement on mAP compared to its variant with post-processing when applied on a ResNet-101 backbone, showing the effectiveness of our method.

2 Related Works

2.1 Vision Transformer

Recent years have witnessed great progress on vision transformers. ViT [8] first introduces a transformer architecture to the image classification and draws much attention. DETR [2,35] builds a transformer-based architecture for object detection, with delicately designed object queries to learn the position distribution of objects. After successful applications, transformers have achieved leading performance in many downstream tasks of computer vision. For instance, in visual object tracking (VOT), TrDimp/TrSiam [28] modifies the transformer decoder for correlation between features from images, as a replacement of classical correlation model (*i.e.*, depth-wise cross correlation [19]) in VOT. HiFT [1] also utilizes the transformer decoder for correlation on hierarchical features extracted from images via a CNN backbone. The multi-head attentions in the decoder seems naturally suitable for feature correlation. However, we cast doubt on the direct usage of the decoder as a feature fusion model for features in the same feature space.

2.2 Video Object Detection

Object detection suffers from image deterioration problems, such as motion blur, background occlusion, deformation, etc. To tackle this problem, many works [25,4,11,13] explored to use temporal context frames to provide compensation guidance (*i.e.*, the object at context frames with different viewpoints). Built upon a two-stage detector (*e.g.*, Faster-RCNN [22], R-FCN [6], FPN [20]), Early works [4,13,11,31] conduct box-level associations and achieve remarkable success. However, these methods highly rely on the features of proposals extracted by the two-stage detector, lacking spatial information. In recent years, the rapid progress of anchor-free object detectors obtain remarkable performance. We observe several attempts to introduce anchor-free methods to video object detection and boost the performance by spatial information. CHP [32] uses an anchor-free detector CenterNet [9] as a base detector and propagates its heat map by post-processing to deliver the spatial information. Apparently, it ignores the support from the temporal features. TransVOD [33] is the first to apply transformer architecture [27] into VOD and builds upon DETR. However, suffering from insufficient feature aggregation and lacking spatial information, its performance is

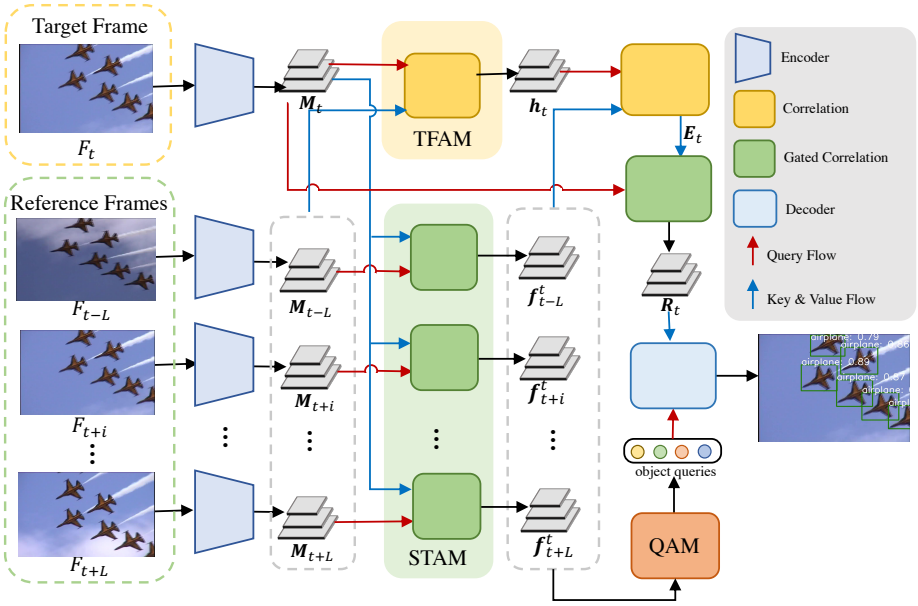

Fig. 2. Overview of the proposed PTSEFormer. First, image features $\boldsymbol{M}$ are extracted by a transformer-based encoder. The image features are further input to TFAM and STAM to obtain temporal feature $\boldsymbol{h}_t$ and spatial features $\{\boldsymbol{f}^i_t\}_{i=-L:L}$, and then are progressively aggregated. Finally, the aggregated feature, together with regressed object queries from QAM, is decoded for final detection result

inferior to those with box-level associations when applied on the same backbone. To address the limitations mentioned above, we propose an end-to-end framework with temporal-spatial feature aggregation design to better employ context frames information.

3 PTSEFormer

3.1 Overview

The overview of PTSEFormer is shown in Figure 2. Given a target frame F_t and its context frames $F^c_t = \{F_{t+i}\}_{i=-L:L}$, PTSEFormer detects the class and bounding-box of objects at F_t. To better explore the context information from F^c_t, PTSEFormer extracts both temporal features (representing the motion of objects) and spatial features (representing position and transformations of objects). Next, the temporal and spatial features are progressively aggregated. Then a decoder learns to infer the class and bounding boxes from the aggregated feature and the object query. Particularly, our object query is conditioned on F^c_t, and thus leads to more accurate object position distribution.

In Section 3.2, we introduce the details of encoding temporal and spatial memories, including feature extraction and progressive aggregation. Next, Section 3.3 introduces how to infer class and bounding-box of objects from the aggregated feature. Finally, the details of learning PTSEFormer are described in Section 3.4, including the total objective function and the network details.

3.2 Temporal and Spatial Encoding

We introduce how to extract temporal and spatial memories from the target frame F_t and its context frames F_t^c. First, a transformer-based encoder embeds F_t and F_t^c to latent feature maps respectively, termed as $\boldsymbol{M}_t$ and $\boldsymbol{M}_t^c = \{\boldsymbol{M}_{t+i}\}_{i=-L:L}$. Then our model obtains the temporal and spatial memories from $\boldsymbol{M}_t$ and $\boldsymbol{M}_t^c$ by two modules: Temporal Feature Aggregation Module (TFAM) and Spatial Transition Awareness Module (STAM). Finally, the temporal and spatial memories are progressively aggregated. We describe the details of each module below.

TFAM. As demonstrated in previous works [33,4,13,25], learning the temporal relation between F_t and F_t^c is beneficial for detecting objects with blurry appearance or distorted shape. Consequently, we propose TFAM to extract this temporal memory $\boldsymbol{h}_t$, which is formulated as:

$$\boldsymbol{h}_t = \mathcal{C}(\boldsymbol{M}_t, \boldsymbol{M}_t^c), \tag{1}$$

where $\mathcal{C}(\cdot,\cdot)$ is the correlation operator:

$$\mathcal{C}(Q, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V + Q, \tag{2}$$

where $Q \in \mathbb{R}^{N_Q \times d_k}$, $K, V \in \mathbb{R}^{N_V \times d_k}$, and '+' represents the residual connection.

STAM. STAM is proposed to learn relative positional transition of objects from a context frame F_{t+i} to the target frame F_t. Since the object identity annotation is unavailable in the VOD task, unsupervised learning of the relations of the objects at F_{t+i} and F_t is non-trivial.

A straightforward idea is to employ the correlation operator $\mathcal{C}(\cdot)$ to model the relative transitions between F_{t+i} and F_t. However, the imbalance weight on Q and V in Equation 2 makes it infeasible to match the objects at two frames. Specifically, the weights before Q and V are 1 and softmax($\frac{QK^T}{\sqrt{d_k}}$), respectively. The average value of softmax($\frac{QK^T}{\sqrt{d_k}}$) is decided by the size of Q and K. When the size goes large, the weight is far less than 1, leading to severer imbalance attention on Q and V. Commonly, this architecture is used for correlation between features from different space and dimensions, which naturally need biased attention. However, in some recent researches [28,33,1], it is also used for correlation between features in the same spaces without any modification. We believe the imbalanced attention could do harm to the performance.

To address the limitation mentioned above and inspired by the gate control design by GRU [5], we design a **Gated Correlation operation**, denoted as

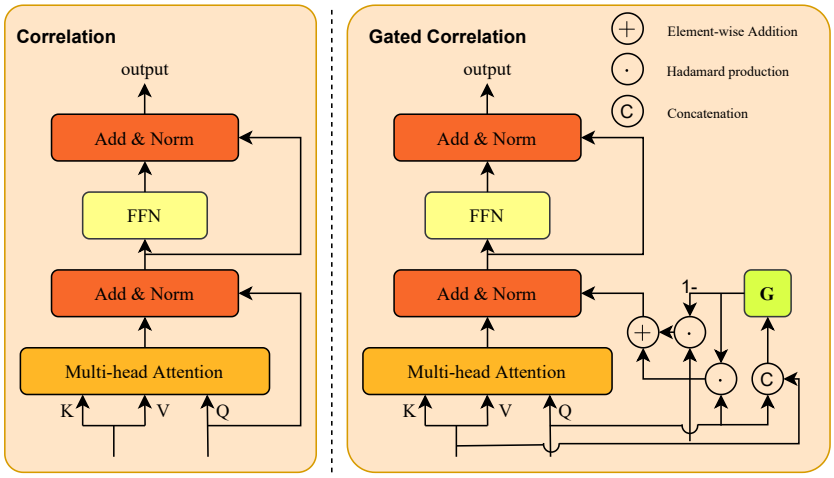

Fig. 3. Illustration of Correlation (left) and Gated Correlation (right).

$\mathcal{C}^g$. By adding a gate control to the residual connection of the decoder, we can change the weight before Q. Furthermore, to get the gate control awareness of the input Q and V, the control weight must be decided by Q and K. Thus, we pass Q and K through a fully connected gate layer for the weight. The process can be changed into:

$$\mathcal{C}^g = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V + M \odot Q + (1 - M) \odot V, \tag{3}$$

$$M = \sigma(\mathcal{G}([Q, V])), \tag{4}$$

where $\mathcal{G}(\cdot)$ refers to the gated function, consisting of a fully connected function. $\sigma(\cdot)$ is the Sigmoid function. $[\cdot, \cdot]$ is the concatenation operation, and $\odot$ refers to the Hadamard production. Note that Q, K, V and M must be of the same size. When initializing, the Sigmoid function in gate can project the output to (0, 1) with a primal value of 0.5, conducting fair attention on both Q and V.

The final STAM can be formulated as:

$$\boldsymbol{f}_i^t = \mathcal{C}^g(\boldsymbol{M}_t, \boldsymbol{M}_{t+i}), \tag{5}$$

where $i = -L : L$, and $\boldsymbol{f}_i^t$ is the extracted spatial memory.

Progressive Aggregation. We aggregate the $\boldsymbol{h}_t$ and $\boldsymbol{f}_i^t$ in a progressive way. First, $\boldsymbol{h}_t$ and $\boldsymbol{f}_i^t$ are combined with the Correlation operation $\mathcal{C}(\cdot)$ to generate a temporal-spatial memory $\boldsymbol{E}_t$. The formulation is written as:

$$\boldsymbol{E}_t = \mathcal{C}(\boldsymbol{h}_t, \{\boldsymbol{f}_i^t\}_{i=-L:L}). \tag{6}$$

By aggregating features from context frames, $\boldsymbol{E}_t$ contains both long-term temporal and spatial transition information. However, in some scenes, the context

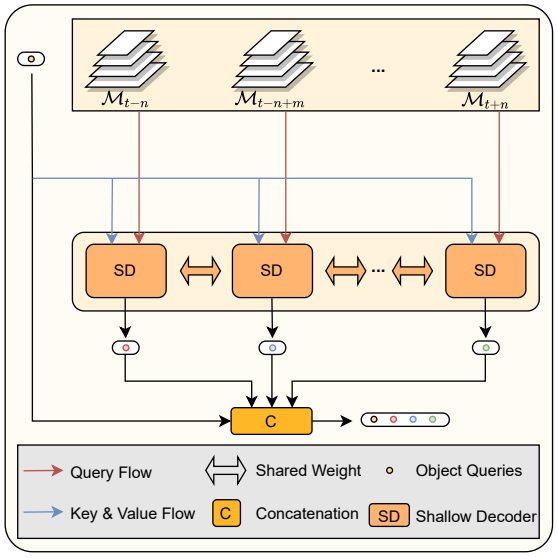

Fig. 4. Illustration of the Query Assembling Model (QAM). We apply a shared shallow decoder to combine the primal object queries and each context frame. All the output from shallow decoders are concatenated to form the final object queries.

frames are likely to be low-quality and the spatial-temporal memory may be useless and even misleading. In this situation, we should take more information of the current frame instead of context frame. Thus, we use a Gated Correlation between the feature of current frame and the temporal-spatial memory to obtain the final enhanced memory $\boldsymbol{R}_t$. The operation is denoted as Residual Gated Correlation, which can be written as:

$$\boldsymbol{R}_t = \mathcal{C}^g(\boldsymbol{E}_t, \boldsymbol{M}_t), \tag{7}$$

3.3 Enhanced Memory Decoding

In original DETR, a group of learned embeddings is designed to learn the position distribution of different objects. With each object query, the decoder decodes one bounding box and its class on the memory. Following the same protocols, we decode our enhanced memory $\boldsymbol{R}_t$ with a transformer decoder. However, there remains a question that the original object queries are fixed through time, cannot benefit from the context frames. Thus, we propose a Query Assembling Model to diversify the object query and convey the position distribution information through time.

Query Assembling Model. Query Assembling Model aims at propagating implicit position distribution information via object queries through time. As primal object queries in DETR are fixed embeddings in the inference stage and have no difference across frames, we apply a shallow correlation model to inherit

location information of the primal object queries and diversify information from features. The final object queries can be described as:

$$Q = [Q_p, \{\mathrm{SD}(Q_p, \boldsymbol{M}_{t+i}), i = -L : L\}], \tag{8}$$

where Q_p is the primal object query, and SD is a shallow transformer decoder with 2 layers. $[\cdot, \cdot]$ is the concatenation operation.

3.4 Learning PTSEFormer

Following DETR, we adopt a Hungarian algorithm [26] to calculate the matching cost between the ground truths and predictions. The objective function is formulated as follows:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box}, \tag{9}$$

$$\mathcal{L}_{box} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou}, \tag{10}$$

where $\mathcal{L}_{cls}$ is focal loss [21] for classification. $\mathcal{L}_{L1}$, $\mathcal{L}_{giou}$ represent the L1 loss and GIoU loss [23] for bounding box regression, respectively. λ_{cls}, λ_{box}, λ_{L1}, λ_{giou} are hyper-parameters to balance the multi-task losses.

Network Details. PTSEFormer is built upon the DETR with several modifications. The number of layers in the encoder and decoder is decreased to 2 for a trade-off between speed and precision. Notice that our method also adopts multi-scale features to boost the performance for detecting small objects. We adopt the ResNet models as our backbones. In particular, we adopt ResNet-101 [15] for a fair comparison with previous works. All the components (*i.e.*, TFAM, STAM, Correlation and Gated correlation) also have a two-layer structure. The number of heads in multi-head attention is fixed as 6 and the number of primal object queries is set to be 100, the same as the original DETR.

4 Experiments

4.1 Implement Details

Dataset and Metric. For a fair and convincing comparison, we conduct our experiments on ImageNet VID dataset [24] which is a large-scale public dataset for video object detection and contains more than 1M frames for training and more than 100k frames for validation. In particular, we train our model on the training split of ImageNet VID and DET dataset [24] following common protocols. Same as previous works [4,33], we adopt mean average precision (mAP) as our metric.

Training Details. We train our PTSEFormer on 8 GPUs of Tesla V100 with Adam [18], and each GPU holds one target frame and its reference frames. The whole training procedure lasts for 50 epochs, each taking almost 1.5 hours. The initial learning rate is 1e-4, with a drop in the 40th epoch to 1e-5. For each target frame, we randomly sample 2 frames from a sliding window with a length of 25

Methods	Base Detector	Stages	Backbone	mAP(%)
DFF [37]	R-FCN	2	ResNet-50	70.4
FGFA [36]	R-FCN	2	ResNet-50	74.0
RDN [7]	Faster-RCNN	2	ResNet-50	76.7
MEGA [4]	Faster-RCNN	2	ResNet-50	77.3
TransVOD [33]	Deformable DETR	1	ResNet-50	79.9
OURS	Deformable DETR	1	ResNet-50	**87.4**

Table 1. End-to-end methods comparisons (with ResNet-50 backbone).

Methods	Base Detector	Stages	Backbone	mAP(%)
LLTR [25]	FPN	2	ResNet-101	81.0
DFF [37]	R-FCN	2	ResNet-101	73.0
D&T [10]	R-FCN	2	ResNet-101	75.8
LSTS [16]	R-FCN	2	ResNet-101	77.2
FGFA [36]	R-FCN	2	ResNet-101	76.3
SELSA [31]	Faster-RCNN	2	ResNet-101	80.3
TROI [11] + SELSA [31]	Faster-RCNN	2	ResNet-101	82.0
MEGA [4]	Faster-RCNN	2	ResNet-101	82.9
HVRNet [13]	Faster-RCNN	2	ResNet-101	83.2
CHP [32]	CenterNet	1	ResNet-101	76.7
TransVOD [33]	Deformable DETR	1	ResNet-101	81.9
OURS	Deformable DETR	1	ResNet-101	**88.1**

Table 2. End-to-end methods comparisons (with ResNet-101 backbone).

as the reference frames. The input images are all resized to hold a shorter size of 800 pixels without any other extra data augmentation applied. All the networks including the single frame baseline are trained from the very beginning with a pre-trained backbone.

4.2 State-of-the-art Comparison

We first compare our PTSEFomer with several state-of-the-art methods in an end-to-end fashion. As shown in Table 1 and Table 2, we group these methods into two categories by their backbones. Previous end-to-end methods are also mostly built upon a two-stage detector without a post-processing procedure for VOD. The existing one-stage based VOD approaches, however, fall behind. Built upon a one-stage detector, we achieve much higher performance on mAP than existing methods with a magnificent margin. Reasonably, the larger backbone boosts the performance of all the methods, including ours. As illustrated in Table 1 and Table 2, Our PTSEFomer leads the performance with ResNet-50 and ResNet-101 [15].

We also compare our PTSEFormer with several state-of-the-art methods with post-processing procedures in Table 3. Post-processing proves useful in

Methods	Base Detector	Stages	Backbone	Post-processing	mAP(%)
PSLA [12]	R-FCN	2	ResNet-101	Seq-NMS	81.4
D&T [10]	R-FCN	2	ResNet-101	Viterbi	79.8
MANet [30]	R-FCN	2	ResNet-101	Seq-NMS	80.3
Scale-Time Lattice [3]	R-FCN	2	ResNet-101	Tublet-Rescore	79.6
FGFA [36]	R-FCN	2	ResNet-101	Seq-NMS	78.4
SELSA [31]	Faster-RCNN	2	ResNet-101	Seq-NMS	82.5
MEGA [4]	Faster-RCNN	2	ResNet-101	Seq-NMS	84.5
HVRNet [13]	Faster-RCNN	2	ResNet-101	Seq-NMS	84.8
CHP [31]	CenterNet	1	ResNet-101	Seq-NMS	78.4
TransVOD [33]	Deformable DETR	1	ResNet-101	-	81.9
OURS	Deformable DETR	1	ResNet-101	-	**88.1**

Table 3. State-of-the-art methods comparisons (with Post-processing).

Method	STAM	TFAM	mAP(%)
Single Frame Baseline [35]	✗	✗	81.2
PTSEFormer	✓	✗	84.5
PTSEFormer	✓	✓	**87.4**

Table 4. Ablation studies of STAM and TFAM.

many VOD methods, especially in those built upon an anchor-based detector. Indeed, most existing methods have their versions with post-processing to boost the performance. For instance, the most widely used post-processing, Seq-NMS, conducts an NMS operation through a sequence, boosting the mAP by 1%-2%. However, those post-processing procedures, though prove effective, demand extra computations. Thus, our PTSEFormer obtains an end-to-end structure. We declare that even we do not adopt post-processing, our method still obtains the best score on mAP.

4.3 Ablation Studies

Considering the speed, we adopt the ResNet-50 model as our backbone for ablation study. The effectiveness of each component of PTSEFormer is verified independently.

TFAM and STAM. To verify the effectiveness of the TFAM and STAM, we conduct ablation studies on both, respectively. As shown in Table 4, we add our STAM model and TFAM model step by step to verify the effectiveness of both. The use of STAM improves the mAP by 3.3%, performing spatial relations between the target frame and each reference frame and offering spatial transferring information. As mentioned above, TFAM conducts a temporal feature aggregation, providing the temporal memory of the target frame. The TFAM leads to an increase of 2.9% compared with only applying STAM.

QAM	GatedCorr	RGC	Multi-scale	mAP(%)
✗	✓	✓	✓	86.1
✓	✗	✓	✓	86.7
✓	✓	✗	✓	86.3
✓	✓	✓	✗	86.4
✓	✓	✓	✓	**87.4**

Table 5. Ablation studies on QAM, Gated Correlation, RGC and Multi-scale.

Query Assembling Model. Query Assembling Model carries the spatial information through time, offering implicit track information. The original object queries in DETR are fixed embeddings, expected to learn the position distribution of the objects in the dataset. We compare QAM with the original object queries in DETR in our experiment by replacing the QAM with original object queries. By comparing line 1 and line 5 in Table 5, results have shown the assistance from the specially designed QAM by an improvement of 1.3% on mAP.
Gated Correlation. To alleviate the imbalanced attention on Key and Value of the transformer decoder as a correlation model, we propose Gated Correlation to carry out a relation between temporal memory and spatial memories. To prove it useful, we replace it with the original transformer decoder. The results show a little drop in mAP which is illustrated in line 2 and line 5 in Table 5.
Residual Gated Correlation. The Residual Gated Correlation model is designed for gating out the memories from low-quality reference frames and boosts the performance of our method. We also investigate it in our experiment and the results from line 3 and line 5 in Table 5 show its positive influence on the performance. In particular, application of Residual Gated Correlation leads to a 1.1% increasement on mAP.
Multi-scale. Similar to the original DETR, the designs of our methods also benefit from the multi-scale features. We obtain 1% increment on mAP with a multi-scale architecture by comparing line 4 and line 5 in Table 5.

4.4 Visualization

Feature Visualization. We first visualize the feature maps of our network to figure out how our TFAM and STAM work. As depicted in Figure 5, we demonstrate three target frames and their corresponding reference frames and feature maps, respectively. The first column shows the original input frames (*i.e.*, target frame and its two reference frames, from top to bottom), and the second column shows the original memories after a shared backbone and encoder, referred to as $\boldsymbol{M}_t$ and $\boldsymbol{M}_{t+i}$. Obviously, it is hard to distinguish the object from the background on these feature maps. The third column shows the temporal memory $\boldsymbol{T}_t$ and the spatial memories $\boldsymbol{S}_{t+i}$. Compared with the original memory $\boldsymbol{M}_t$, it is clear that the $\boldsymbol{T}_t$ has much more attention on the target objects, which indicates that the temporal information does contribute to the distinguishing be-

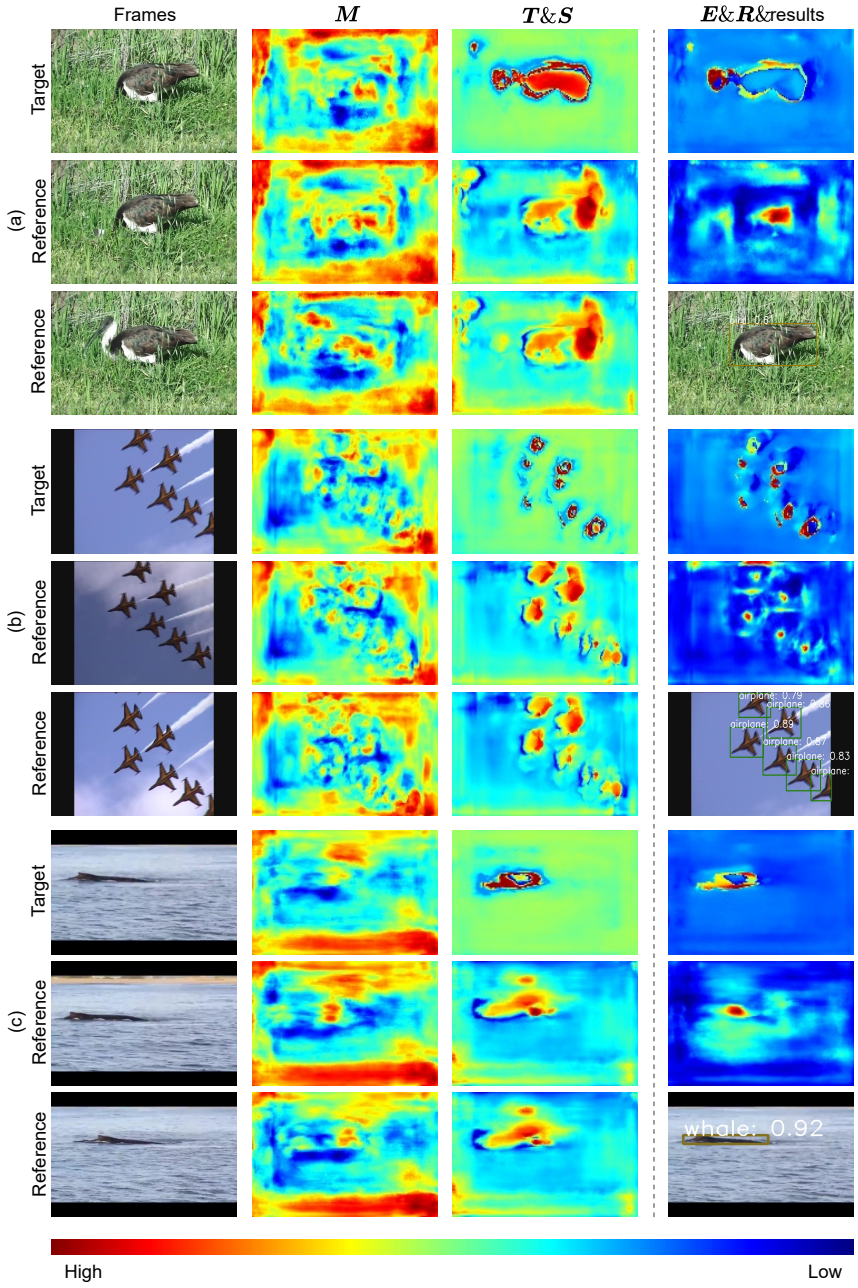

Fig. 5. The feature maps of our models. We select three target frames to figure out what the network learns.

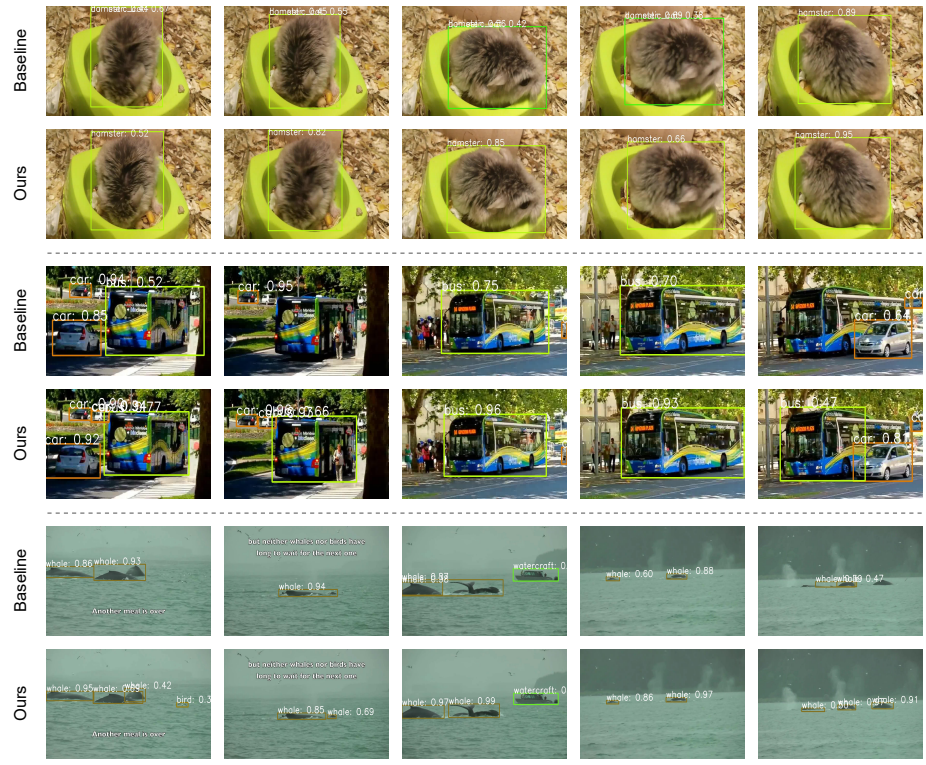

Fig. 6. Results Visualization. Our results are in the odd row, and single frame detector DETR results as baseline are in the even row. As shown in figure, our method is more robust against various image deterioration(*e.g.*,occlusion, deformation).

tween foreground and background. The last column shows the temporal-spatial memory $\boldsymbol{E}_t$, $\boldsymbol{R}_t$ after the Residual Gated Correlation and the detection results from top to bottom. Notice the color of the feature map indicates the value. Observing the the original memory $\boldsymbol{M}_t$ the temporal memory $\boldsymbol{T}_t$ and the final enhanced memory $\boldsymbol{R}_t$, it is easy to find a trend that the values of foreground and background become more easy to separate. The temporal information contributes to recognizing a object by introducing different poses of it. Furthermore, the spatial information helps our PTSEFormer to locate objects with higher confidence score by using spatial transition information. We declare that the reason of such excellent results is the contribution of temporal and spatial information from our TFAM and STAM.

Results Visualization. We present the results of both the single frame baseline method and our PTSEFormer in Figure 6. In particular, the detection results are exhibited in the time order. Compared with the single frame baseline method DETR, Our method shows the priority towards the image deterioration problems. By exploiting the temporal and spatial information, we get a higher confi-

dence score in normal situations and behave much better dealing with occlusion and posture deformation. For example, when the face of a hamster gets occluded by the background, the baseline single frame detector is confused about the category, and easily fooled to predict it as a domestic cat. However, its appearances in context frames are clear and easy to recognize, so our method succeeds in predicting the right category by introducing temporal information. In the second video, the detector is expected to detect several cars and a bus. Interfered by the background and occluded by a car, the baseline method fails at detection in some frames. In contrast, with the help of spatial information, our method can sense the motion of the bus and cars and produce the correct results. In the third video, when two whales get too close, it is hard for the baseline detector to recognize both, causing false detection. In this situation, our PTSEFormer behaves much better according to the temporal-spatial enhancement. It is necessary to introduce temporal-spatial information in this situation to better distinguish one object from another. Consequently, our PTSEFormer achieves much better performance than the single frame baseline method thanks to the temporal-spatial information.

5 Conclusion

In this work, we propose a progressive temporal-spatial enhanced transformer towards video object detection. Based on a one-stage object detector DETR, we boost the performance with proper design of introducing progressive feature aggregation. Temporal information and spatial information are proved useful to improve the robustness of detector against image deterioration. We also conduct extensive experiments on the public dataset ImageNet VID to verify the effectiveness of our method. We hope our work can shed light on the research on VOD applying anchor-free approaches.

Acknowledgements. This work was partly supported by MoE-China Mobile Research Fund Project (MCM20180702), the 111 Project (B07022 and Sheitc No. 150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions. And part of this work was done while Han Wang performed as an intern at HIKVISION.

References

1. Cao, Z., Fu, C., Ye, J., Li, B., Li, Y.: Hift: Hierarchical feature transformer for aerial tracking. In: ICCV. pp. 15457–15466 (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
3. Chen, K., Wang, J., Yang, S., Zhang, X., Xiong, Y., Loy, C.C., Lin, D.: Optimizing video object detection via a scale-time lattice. In: CVPR. pp. 7814–7823 (2018)
4. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: CVPR. pp. 10337–10346 (2020)
5. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
6. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems **29** (2016)
7. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: ICCV. pp. 7023–7032 (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: ICCV. pp. 6569–6578 (2019)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: ICCV. pp. 3038–3046 (2017)
11. Gong, T., Chen, K., Wang, X., Chu, Q., Zhu, F., Lin, D., Yu, N., Feng, H.: Temporal roi align for video object recognition. In: AAAI. pp. 1442–1450 (2021)
12. Guo, C., Fan, B., Gu, J., Zhang, Q., Xiang, S., Prinet, V., Pan, C.: Progressive sparse local attention for video object detection. In: ICCV. pp. 3909–3918 (2019)
13. Han, M., Wang, Y., Chang, X., Qiao, Y.: Mining inter-video proposal relations for video object detection. In: ECCV. pp. 431–446. Springer (2020)
14. Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S.: Seq-nms for video object detection. arXiv preprint arXiv:1602.08465 (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Jiang, Z., Liu, Y., Yang, C., Liu, J., Gao, P., Zhang, Q., Xiang, S., Pan, C.: Learning where to focus for efficient video object detection. In: ECCV. pp. 18–34. Springer (2020)
17. Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., et al.: T-cnn: Tubelets with convolutional neural networks for object detection from videos. TCSVT **28**(10), 2896–2907 (2017)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28** (2015)
23. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
25. Shvets, M., Liu, W., Berg, A.C.: Leveraging long-range temporal relationships between proposals for video object detection. In: ICCV. pp. 9756–9764 (2019)
26. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2325–2333 (2016)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
28. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: CVPR. pp. 1571–1580 (2021)
29. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: ECCV. pp. 542–557 (2018)
30. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: ECCV. pp. 542–557 (2018)
31. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: ICCV. pp. 9217–9225 (2019)
32. Xu, Z., Hrustic, E., Vivet, D.: Centernet heatmap propagation for real-time video object detection. In: ECCV. pp. 220–234. Springer (2020)
33. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: End-to-end video object detection with spatial-temporal transformers. arXiv preprint arXiv:2201.05047 (2022)
34. Zhu, H., Wei, H., Li, B., Yuan, X., Kehtarnavaz, N.: A review of video object detection: Datasets, metrics and methods. Applied Sciences **10**(21), 7834 (2020)
35. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
36. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: ICCV. pp. 408–417 (2017)
37. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: CVPR. pp. 2349–2358 (2017)