# D2FANet: Enhancing Video Object Detection with Dual-Domain Feature Aggregation Network

## Supplementary Material

This supplementary material offers additional details that support and extend the content presented in the main paper. It is organized into the following sections:

- Section A provides more related works.
- Section B provides more network architecture details.
- Section C provides more ablation studies.
- Section D provides more visualization results.

## A. More Related Works

Transformers [1, 4, 11, 12, 15, 16] have recently shown strong ability to model long-range dependencies in both spatial and temporal dimensions, which has driven substantial progress in computer vision tasks. Building on this foundation, several transformer-based methods [6, 8, 9] have been proposed for video object detection, each leveraging the transformer architecture to model spatiotemporal dependencies. For instance, TransVOD [13] uses a transformer decoder to aggregate multi-frame features, allowing object queries to attend to all frame features and capture temporal dynamics for consistent detection. PTSEFormer [7] employs a progressive temporal-spatial transformer to aggregate features across multiple frames, where object queries attend to all relevant frame features using multi-head attention to capture both spatial and temporal dependencies. FAIM [2] employs multi-head attention to selectively aggregate instance mask features and classification features across frames.

These methods generally exploit uniform attention or indiscriminate feature aggregation operations to model spatiotemporal information, but they may overlook the fact that different regions contribute unevenly to video object detection task. Recent works [3, 5] suggest that aggregating multiple tokens into a single representative token can enhance cross-frame spatiotemporal feature aggregation in transformer-based video models. However, such aggregation treats each token equally and ignores the inherent significance difference of different tokens. Therefore, it faces two main problems. First, information in salient regions may be lost or mixed with irrelevant content. Second, in minor or background regions, many tokens are redundant for simple semantics and incur excessive computational cost. In contrast, our D2FANet introduces an importance-guided mechanism that dynamically highlights the significance of different regions to guide feature aggregation.

## B. More Network Architecture Details

The schematic illustration of the vanilla spatiotemporal transformer decoder is illustrated in Figure 1. It consists of a multi-head self-attention (MHAttn) layer, a spatiotemporal deformable attention (STDAttn) layer, and a feed-forward network (FFN). We build upon the overall decoder framework of Deformable DETR [10], which integrates multi-scale features with deformable attention and feed-forward networks, but extend it from its original single-frame formulation to a spatiotemporal transformer decoder that performs attention across both spatial regions and temporal neighbors for effective cross-frame feature aggregation. Let $\mathbf{Z}$ denote the collection of all object queries from $M$ frames. To enable interaction across different frames, the decoder first applies a cross-frame multi-head self-attention (MHAttn) operation to process vanilla object queries:

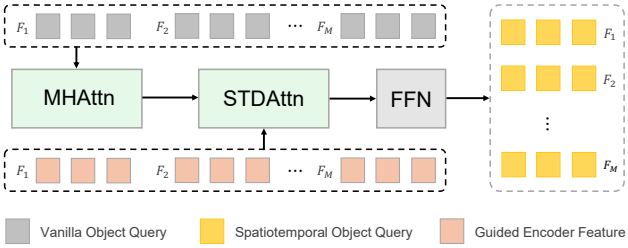


Figure 1. The schematic illustration of the vanilla spatiotemporal transformer decoder.

$$\mathbf{Z}' = \mathrm{MHAttn}(\mathbf{Z}), \tag{1}$$

where $\mathbf{Z}' \in \mathbb{R}^{M\times N\times D}$ represents the updated object queries after self-attention, and $M$, $N$, and $D$ denote the number of frames, the number of object queries per frame, and the feature dimension, respectively. This operation enables information exchange across the temporal dimension, allowing the object queries to model long-range inter-frame dependencies and refine their representations by incorporating context from other frames. Next, the updated object queries $\mathbf{Z}'$ are fed into the spatiotemporal deformable attention (STDAttn) layer together with the guided encoder features $\mathbf{F} = \{X_m\}_{m=1}^{M}$ extracted from all frames. For each object query, STDAttn sparsely samples informative spatiotemporal locations across frames and aggregates the corresponding features. In a simplified matrix form, this operation can be expressed as:

$$\mathbf{Z}'' = \mathrm{STDAttn}(\mathbf{Z}', \mathbf{F}), \tag{2}$$

where $\mathbf{Z}'' \in \mathbb{R}^{M\times N\times D}$ represents the spatiotemporal object queries for all frames. Each attention head selects only

Table 1. Effect of the FDFA module across different feature scales on the ImageNet VID dataset.

| Model | Input Feature | mAP (%) |
|---|---|---|
| (a) | $C_3$ (Shallow-level) | 86.9 |
| (b) | $C_4$ (Middle-level) | 87.4 |
| (c) | $C_5$ (Deep-level) | 87.3 |
| (d) | $C_3$+$C_4$+$C_5$ (Multi-Scale) | **87.7** |

Table 2. Effect of each module in D2FANet on the EPIC-KITCHENS dataset.

| Method | Baseline | FDFA | SDFA | mAP (%) |
|---|---|---|---|---|
| A | ✓ | | | 37.4 |
| B | ✓ | ✓ | | 42.9 |
| C | ✓ | | ✓ | 43.4 |
| D | ✓ | ✓ | ✓ | **44.5** |

a small set of dynamically sampled positions, enabling efficient aggregation of object-centric motion and appearance cues while reducing computational cost. Finally, the spatiotemporal object queries $\mathbf{Z}''$ are passed through a feed-forward network (FFN) to produce the final spatiotemporal object queries, which are then used in subsequent concatenation operations.

## C. More Ablation Studies

**Effect of FDFA with Different Scales.** Table 1 summarizes the effect of the frequency-domain feature aggregation (FDFA) module with features from different scales, illustrating the impact of feature scales on accuracy.

Model (a) uses shallow-level features only, capturing fine-grained structural cues such as edges and contours, and it achieves 86.9% mAP.

Model (b) employs middle-level features only, which encode intermediate semantics, resulting in a slightly higher mAP of 87.4%.

Model (c) utilizes deep-level features that provide high-level contextual information, yielding 87.3% mAP.

Model (d) integrates multi-scale features, combining complementary information across different scales, and it achieves the best performance of 87.7% mAP. These results demonstrate that using multi-scale feature aggregation significantly enhances the robustness and effectiveness of FDFA by leveraging both low-level details and high-level semantic cues.

**Effect of Each Module in D2FANet on the EPIC-KITCHENS.** Table 2 summarizes the ablation results of D2FANet on the EPIC-KITCHENS dataset, demonstrating the effectiveness of each module.

Method A represents the baseline detector, Deformable DETR with the ResNet-101 backbone, which achieves 37.4% mAP.

Method B incorporates the frequency-domain feature aggregation (FDFA) module into method A, improving the accuracy to 42.9% mAP by capturing complementary multi-frequency features that enrich feature representations.

Method C integrates the spatiotemporal-domain feature aggregation (SDFA) module into method A, yielding 43.4% mAP by enhancing spatiotemporal dependency modeling with importance-guided aggregation.

Method D combines both FDFA and SDFA modules on top of method A, achieving the best accuracy of 44.5% mAP, as the two modules work collaboratively to provide more comprehensive and discriminative feature representations. The additional evaluations on the EPIC-KITCHENS dataset demonstrate that our D2FANet also generalizes well to complex video scenarios, delivering robust improvements in object localization and category discrimination.

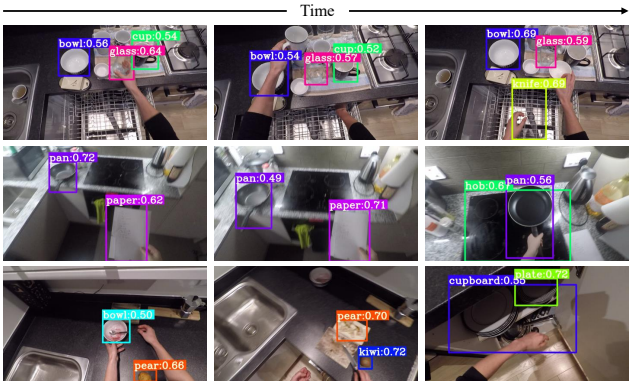


Figure 2. Visualization of the detection results of the proposed D2FANet on the EPIC-KITCHENS dataset across diverse kitchen environments.

## D. Visualizations on EPIC-KITCHENS

Figure 2 presents the visualized results of our D2FANet on the EPIC-KITCHENS dataset [14] across diverse kitchen environments. The results demonstrate the capability of our D2FANet to accurately detect and localize a wide variety of kitchen-related objects over time, with confidence scores provided for each bounding box. Despite challenges such as frequent hand-object interactions, partial occlusions, motion blur, and densely arranged objects with varying sizes and orientations, D2FANet still achieves stable object localization while maintaining high detection confidence. These observations further highlight the robustness and generalization capability of D2FANet, which confirms its reliability in handling complex and dynamic video scenarios, showing its potential for consistent performance across different types of cluttered kitchen environments.

## References


[1] Seungjun An, Seonghoon Park, Gyeongnyeon Kim, Jeongyeol Baek, Byeongwon Lee, and Seungryong Kim. Context enhanced transformer for single image object detection in video data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 682–690, 2024. 1

[2] Chenglizhao Chen, Guotao Wang, Chong Peng, Yuming Fang, Dingwen Zhang, and Hong Qin. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. *IEEE Transactions on Image Processing*, pages 3995–4007, 2021. 1

[3] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020. 1

[4] Chaorui Deng, Da Chen, and Qi Wu. Identity-consistent aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13434–13444, 2023. 1

[5] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *Proceedings of the European Conference on Computer Vision*, pages 18–34, 2020. 1

[6] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 3195–3215, 2021. 1

[7] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 817–825, 2016. 1

[8] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 817–825, 2016. 1

[9] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 2896–2907, 2017. 1

[10] Dongfang Liu, Yiming Cui, Zhiwen Cao, and Yingjie Chen. Indoor navigation for mobile agents: A multimodal vision fusion model. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2020. 1

[11] Zhifan Ni, Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Human–object interaction prediction in videos through gaze following. *Computer Vision and Image Understanding*, page 103741, 2023. 1

[12] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In *Proceedings of the European Conference on Computer Vision*, pages 732–747, 2022. 1

[13] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision*, pages 485–501, 2018. 1

[14] Liqi Yan, Qifan Wang, Yiming Cui, Fuli Feng, Xiaojun Quan, Xiangyu Zhang, and Dongfang Liu. Gl-rg: Global-local representation granularity for video captioning. *arXiv preprint arXiv:2205.10706*, 2022. 2

[15] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022. 1

[16] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 7853–7869, 2023. 1