

STF: Spatio-Temporal Fusion Module for Improving Video Object Detection

Noreen Anwar, Guillaume-Alexandre Bilodeau
*LITIV, Polytechnique Montréal,
Montréal, Canada*
{noreen. anwar, gabilodeau}@polymtl.ca

Wassim Bouachir
*Data Science Laboratory,
University of Quebec (TELUQ)
Montréal, Canada*
wassim.bouachir@teluq.ca

Abstract—Consecutive frames in a video contain redundancy, but they may also contain relevant complementary information for the detection task. The objective of our work is to leverage this complementary information to improve detection. Therefore, we propose a spatio-temporal fusion framework (STF). We first introduce multi-frame and single-frame attention modules that allow a neural network to share feature maps between nearby frames to obtain more robust object representations. Second, we introduce a dual-frame fusion module that merges feature maps in a learnable manner to improve them. Our evaluation is conducted on three different benchmarks including video sequences of moving road users. The performed experiments demonstrate that the proposed spatio-temporal fusion module leads to improved detection performance compared to baseline object detectors. Code is available at <https://github.com/noreenanwar/STF-module>

Keywords—Spatio-temporal object detection; feature fusion; spatio-temporal attention;

I. INTRODUCTION

Computer vision has made remarkable progress in object detection from a single frame for localizing and identifying objects ([1], [2]). However, relying solely on a single frame for detection is not always effective and sufficient as argued in several recent works ([3], [4], [5], [6], [7], [8]). Single-frame object detectors are subject to errors in the case of poor or improper visibility of objects that can be caused by occlusions, motion blur, or small object sizes. When objects are occluded or in the case of motion blur, their appearance features can be severely altered. The object detector should be robust to the fact that objects can exist on a spectrum of scales and sizes. Furthermore, small objects have less distinctive features making them harder to detect.

To address these problems, as in some previous work ([3], [4], [5], [6], [7], [8], [9], [10]), we propose to use multiple frames for better features representation. Given that we are interested in processing videos for road safety analysis, our work is in a context that fits well with object detection from multiple frames. Having several sequential frames to detect objects has the significant benefit of providing temporal complementary information about a given instance, generally observed over a short time. This kind of temporal information is utilized in some existing multiple-frame object detection methods by first applying a single-frame object detector and then integrating their bounding

boxes across frames using an off-the-shelf motion estimation method ([9], [10]). The performance improvement relies on heuristic post-processing and those methods do not capitalize on combining features from several frames to compensate for poor feature quality in some frames.

Another solution for multiple-frame detection is to fuse features from several frames together in a learnable manner for better feature alignment ([3], [6], [4], [5], [7]). Using multiple frames is not trivial, however as the features of consecutive frames are not always aligned or corresponding to the visibility state of an object that can change (e.g. not the same part of an object might be occluded). This means that there are no trivial ways to determine which features are more important for the detection. Therefore, feature fusion has to be done carefully.

Global contextual attention involves capturing long-range dependencies and relationships between different regions of an image to understand specific parts and the context. Therefore, we propose a global contextual attention model for feature selection and fusion from a pair of frames. We present an end-to-end framework that learns multiple frame information and fuses it without prior knowledge of motion or temporal relations. We aim to improve the detection accuracy by effectively utilizing temporal and spatial information from two frames, the current frame and the past frame. As mentioned above, it is important to consider that the features corresponding to the same object instances in two frames often lack spatial alignment across frames due to movements or occlusions. To take this into account, we introduced multi-frame (temporal) and single-frame (spatial) attention-based modules. Secondly, to handle small objects, we are considering the multiple layer resolution features from our backbone. Our attention modules operate on those multiple resolution layers.

Our proposed approach, STF (Spatio-temporal fusion), is based on per-frame feature learning through temporal and spatial fusion of features from the current and a past frame. To achieve this, we are proposing two new attention-based modules: the first applies multi-frame attention, while the second applies single-frame attention. Here, we hypothesize that global contextual information along with spatio-temporal information can address the detection problems better as compared to previous works, limited to single-

frames and multi-frame methods that fuse feature maps without attention ([3], [4], [11]). Then, our dual-frame fusion module helps to fuse the learned features from the past and current frames to improve detection accuracy under challenging conditions, like occlusion or motion blur. The effectiveness of our method is evaluated on three popular traffic-related datasets, including KITTI MOT [12], Cityscapes [13] and UAVDT [14] and we obtained competitive results compared to SOTA detectors.

Our main contribution is the introduction of an end-to-end learnable fusion module that combines the current and a past frame by utilizing their temporal, spatial and channel features information. Our specific contributions can be outlined as:

- A multi-frame attention (MFA) module with temporal convolutions used after the backbone feature extractor to efficiently use the feature maps of two frames, and enhance features for detecting occluded or blurred objects;
- A single-frame attention (SFA) module that weights the current frame feature maps in channel and spatial dimensions to reduce false positive detection;
- An efficient dual-frame fusion module to integrate single-frame and multi-frame feature maps at different scales.

II. RELATED WORK

Using multiple frames in object detection was studied in several previous works because it facilitates the association of detected objects, thereby improving the precision and resilience of the detection process. It consists of detecting objects in frames using their spatial and temporal features. Nevertheless, the study of video-based object detection is receiving comparatively less attention than single-frame detectors, yet their applications are numerous and impactful, which includes video surveillance for security, robot navigation, and autonomous driving. Sequential frames have significant complementary information about the same instances, generally observed in multiple frames during a short period. Existing multiple-frame object detection methods, such as those proposed by Kang et al. ([9]) and Lee et al. [10], readily capture this type of temporal information. These methods first apply single-frame object detectors and then integrate bounding boxes across frames using off-the-shelf motion estimation, which may compromise the quality of detection due to hand-crafted rules. The improvement in the performance depends on heuristic post-processing through box-level methods without end-to-end training.

Zhu et al. [3] introduced flow-guided feature aggregation, where optical flow warping is used to integrate feature maps from temporally adjacent frames in order to increase detection accuracy. There is another way, proposed in [8] that calculates the offsets between temporally adjacent frames. These offsets enable the sharing of features from

adjacent frames, improving the ability to perform the detection tasks. Similarly, there is another approach, known as FFAVOD (Feature Fusion Architecture for Video Object Detection) [6], which shares feature maps between nearby frames. FFAVOD proposes a feature fusion module that learns to merge feature maps to improve video-based object detection and classification. RN-VID[15] uses information from nearby frames and merges feature maps of similar dimensions using 1×1 convolution and re-ordering of channels to enhance detection. Zhou et al. [4] presented CenterTrack, a method that uses a point-based framework to perform simultaneous detection and tracking of objects. This method concatenates two frames and a prior heatmap as input and associates objects through time while performing the detection from the two frames. Previous research also explores using both motion and appearance cues of objects in a video sequence with models such as Recurrent Neural Networks (RNNs). Using an RNN, the method named Spatio-Temporal Memory module (STMM) [16] introduced a concatenated spatial-temporal memory across consecutive frames to improve detection. Additionally, Long Short-Term Memories (LSTMs) have been employed to interpolate feature maps, resulting in a notable improvement in inference speed [5]. The Recurrent Multi-frame Single Shot Detector (MF-SSD) method combines features extracted from multiple consecutive frames [7]. This is achieved through the integration of a recurrent convolutional module, enabling the integration of characteristics that extend across multiple frames.

The above mentioned works are mainly focusing on either concatenating or simply summing feature maps rather than using a more fully learnable way. Unlike these methods, our approach aims to train a learnable fusion-based module including temporal, spatial, and channel-based feature information, in a completely end-to-end manner, using the current frame and a past frame.

III. METHODOLOGY

A. Overview

The overview of our attention-based framework, STF, is shown in Figure 1. Given a pair of frames, a pre-trained HRNet [17], where we froze the first and third layers, is used to extract features. After that, the features go through two attention modules: 1) a multi-frame attention (MFA) module that uses the two extracted feature maps to perform temporal and spatial attention, assigning adaptive temporal weights to them, and 2) a single-frame attention (SFA) module that uses spatial and channel dimension attention for improving current frame feature maps. To use the temporally prior frame, the idea here is to combine in a learnable manner the extracted features of the past and current frames for object detection. To combine features from two frames after applying attention, our proposed network fuses temporal, channel, and spatial information by aggregating them at

the same time with our dual-frame fusion module. In the following, we introduce these modules in detail.

B. Multi-Frame Attention (MFA) module

Given an input video, the multi-scale feature maps of two frames (the current frame and a past frame) are extracted with the HRNet backbone. Then, our goal is to merge the features of these two frames. The Tada Convolution, introduced in the work by Huang et al. [18], efficiently addresses temporal modeling by introducing flexibility to the temporal invariance of 2D convolutions. This is achieved through the incorporation of adaptive temporal weights, which are superimposed onto the convolutional process. Similarly, Cao et al. proposed TCTrack [19], which exemplifies the application of Tada Convolution for improving object tracking. This approach employed Tada Convolutions to incorporate adaptive temporal weights, contributing to improved temporal modeling. Inspired by this previous research, to get adaptive temporal weights for each frame, we designed a Multi-Frame Attention (MFA) module (see Figure 2). The key idea is to adjust the model behavior in real-time as it processes each sequence of frames. This deals with size variations, movement, overlapping, or interaction of objects in frames.

Global information in object detection refers to semantic details that are consistent across frames, helping in identifying objects based on shared characteristics, while local temporal information involves using nearby frames to gather information, such as motion, helping to localize objects, especially in cases of uncertainty about their existence in a specific frame. This module improves the representation ability with multi-frame features by: 1) assigning adaptable weights to each frame to enhance the ability to detect and analyze changes over time, 2) combining both global and local information from multi-frames, and 3) better capturing both detail and broader spatial and temporal information using a multi-scale integrator.

Our MFA module works as follows. Let us assume that we have an input sequence of frames I_n and we get a sequence $X_n \in \mathbb{R}^{B \times C \times T \times H \times W}$ of features outputted by the HRNet backbone, where B is the batch size, C is the number of channels, T is the temporal dimension, and H and W are the spatial dimensions. For capturing the global spatial context, we start with global average pooling (GAP) across the spatial dimension of the past and current frame features. We then obtain frame descriptor $S_n = \text{GAP}_S(X_n)$, that encompasses global spatial context. To integrate local temporal context effectively, global average pooling across both spatial and temporal dimensions is applied to obtain spatio-temporal descriptor

$$T_n = \text{GAP}_{st}(X_n). \quad (1)$$

Global spatial context and local temporal information are then aggregated, and this combined information is passed

through a bottleneck block (BNB). The output of the bottleneck block results in obtaining local weights ω_t , as illustrated in Figure 2. These weights combine the spatial and temporal descriptors after the bottleneck block with

$$\omega_t = \text{BNB}(S_n + T_n). \quad (2)$$

Then, the total weights that we used in our model are the element-wise product of these weights ω_t and weights W_p that refer to the initial set of weights in the convolution kernel that is shared across all frames. Note that the local weights ω_t are set to 0 during initialization, which has the advantage of reducing the training time. An adaptive convolution is then applied to the current frame with

$$\hat{X}_o = (\omega_t \odot W_p) * X_n \quad (3)$$

where \odot denotes element-wise multiplication.

To effectively integrate spatio-temporal information and address the limitations in spatial features for a given frame, we finally apply a multi-scale integrator as shown in Figure 2. It is expressed as

$$X_o = \lambda(\hat{X}_o) + \gamma(\text{AvgPool}(\hat{X}_o)), \quad (4)$$

where \hat{X}_o is the output from the adaptive convolution. The operators λ and γ represent distinct normalization functions. The goal behind using an average pooling (AP) layer is to enlarge the receptive field to capture a wider range of spatial contexts.

C. Single-Frame Attention Module

Besides temporal attention, attention in the spatial and channel dimensions also provides a potential enhancement for feature maps derived from single-frame images. In the context of Convolutional Neural Networks (CNNs), the attention mechanism assigns an additional weight to individual pixels in a specific dimension, indicating the significance of particular information. These learned weights strengthen valuable features and weaken less useful ones, facilitating feature screening and enhancement. Furthermore, in videos with generally stable backgrounds, spatial and channel attention, as explained by the methodology proposed in Hou et al. [20], can efficiently suppress false positive detection in the background area.

Inspired by this work, we propose a Single-Frame Attention module (SFA) that uses channel and spatial attention mechanisms, as illustrated in Figure 3. The SFA module aims to refine feature representation within a single frame. In the SFA module, each frame denoted as I_n , is processed to enhance the channel and spatial information of its feature maps X_n . First, channel attention with average pooling (AP) and max pooling (MP) are applied to condense the spatial information. To help our model learn complex feature representation, we integrate 1×1 convolutional layer as

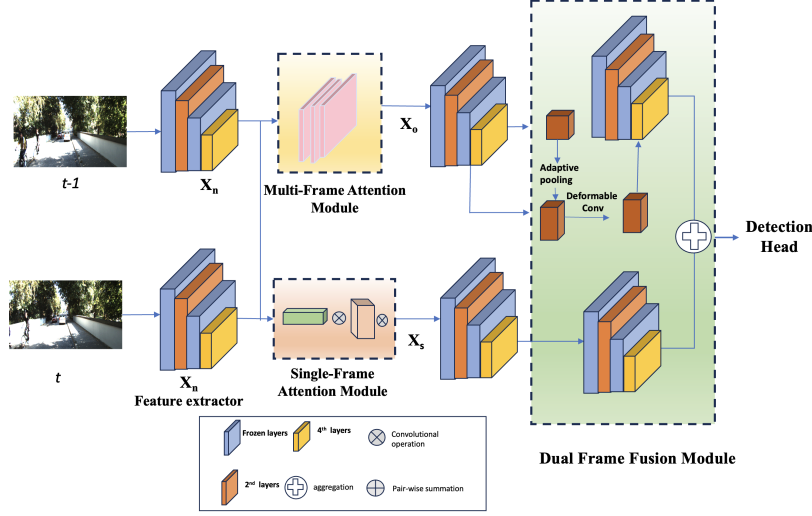


Figure 1. Overview of our spatio-temporal based fusion framework (STF), illustrating the key components: MFA, SFA, and dual-fusion module

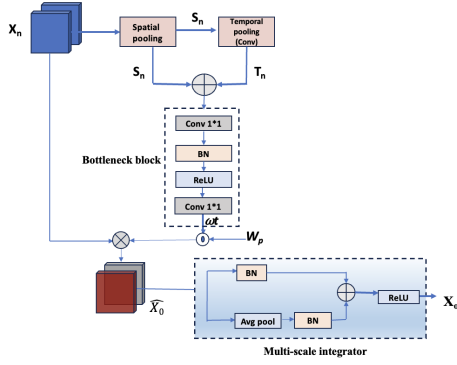


Figure 2. Multi-Frame Attention module with multi-scale integrator.

shown in Figure 3. This results in channel attention A_c formulated as

$$A_c = \text{Conv}_{1 \times 1}(\text{AP}(X_n)) + (\text{MP}(X_n)). \quad (5)$$

For spatial attention, a comparable approach is applied, but it operates within the spatial domain. Here, the features influenced by channel attention are subjected to average and max pooling operations (MP), focusing on spatial features. The resulting features are then concatenated and processed through a 5×5 convolutional layer to enhance the spatial aspects of the frame. This gives spatial attention A_s formulated as:

$$A_s = \text{Conv}_{5 \times 5} \left(\text{Conv}_{1 \times 1}(\text{AP}(A_c * X_n)) + (\text{MP}(A_c * X_n)) \right) \quad (6)$$

where $\text{Conv}_{5 \times 5}$ is the convolution operation using a 5×5 filter and $*$ symbolizes convolution.

Finally, the two attention tensors are concatenated with X_n to obtain the new features X_s . This fusion process

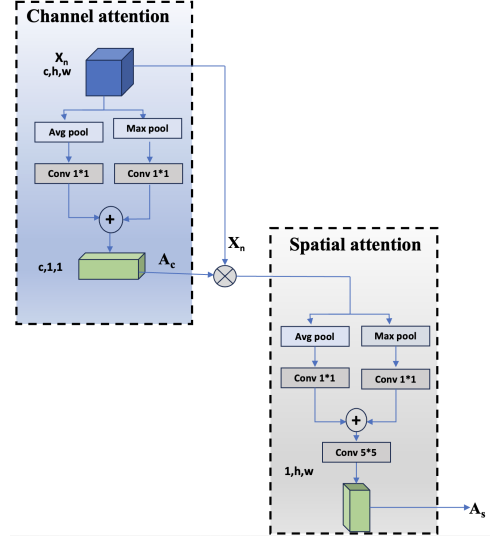


Figure 3. The channel and spatial attention modules of our proposed single-frame attention module

allows the model to focus on relevant information captured by the attention mechanisms, enhancing the representation of the feature maps. We observed that by using convolutional layers, the module can more effectively capture and enhance the intricate patterns in the features. This ensures that the model is capturing well the spatial features in each frame.

D. Dual-Frame Fusion Module

Figure 1 illustrates the feature maps X_o and X_s obtained after the SFA and MFA modules, which serve as inputs to our dual-frame fusion module. The proposed dual-frame fusion module combines semantic information of the high-level feature maps and spatial information of low-level

feature maps. Instead of traditional up-sampling, inspired by [21], we use Adaptive Feature Pooling for a more flexible approach. This offers an expanded receptive field, facilitating improved integration of both core and contextual semantics.

The high-level feature map is adaptively pooled to match the size of the low-level feature maps. These feature maps are then combined via pixel-wise summation and further processed through deformable convolutions. This offers better adaption to different object sizes, shapes, and other geometric deformations. The input has a total of four layers, and the aforementioned convolution and up-sampling process is iterated 2 times to have the final output. With the help of the above process, we obtained channel and spatial attention feature maps on single frames and temporal attention feature maps for multiple frames. These feature maps are aggregated to generate a fused feature map.

E. Detection Head

Our detection head is similar to CenterNet [22]. We performed the computation of the fused object probability heatmap on the merged feature maps. However, the size and offset of the bounding boxes are generated from single-frame features. The loss function comprises three components: a fusion heatmap loss calculated with Focal Loss, and two regression losses (offset and size) computed with L1 Loss. The formulation of each loss is as follows. L_Z is the unique fusion heatmap loss,

$$L_Z = -\frac{1}{M} \sum_{ij} \begin{cases} (1 - \hat{Q}_{ij})^\epsilon \log(\hat{Q}_{ij}) & \text{if } Q_{ij} = 1 \\ (1 - Q_{ij})^\zeta (\hat{Q}_{ij})^\epsilon \log(1 - \hat{Q}_{ij}) & \text{otherwise} \end{cases}, \quad (7)$$

where \hat{Q}_{ij} indicates the predicted heatmap value for each pixel, $Q_{ij} = 1$ signifies the pixel is the center of an object, and ϵ and ζ are the modified focal loss hyper-parameters. L_Y represents the loss for heatmap offset,

$$L_Y = \frac{1}{M} \sum_q |\hat{P}_q - T - \tilde{q}|, \quad (8)$$

where \hat{P}_q is the predicted offset, T is the position after down-sampling, and \tilde{q} is the actual center point. L_X calculates the loss for the size of the bounding box,

$$L_X = \frac{1}{J} \sum_{j=1}^J |\hat{R}_j - R_j|, \quad (9)$$

where \hat{R}_j is the predicted size and R_j is the ground truth size. The overall training objective is

$$L_{total} = L_Z + \lambda_{dim} L_X + \lambda_{pos} L_Y, \quad (10)$$

where λ_{dim} and λ_{pos} are the adjusted hyper-parameters for the size and offset loss components, respectively.

IV. EXPERIMENTS

In this section, we assess the performance of our proposed method compared to SOTA methods and perform an ablation study.

A. Datasets and Evaluation Metrics

Datasets: As our method relies on more than a frame, the evaluation requires the use of video datasets. Our selected evaluation domain focuses on traffic surveillance given its significant relevance to our research. We used datasets with videos, but some are not standard datasets for object detection. Nevertheless, they were used in previous work on video object detection. We chose: KITTI MOT (Multi-Object Tracking) [12] and Cityscapes [13], both not used for object detection usually but provide videos, and UAVDT [14] used for object detection in videos. Each of these datasets provides unique challenges and contains sequences at different viewpoints with different sizes of objects. As we are using non-standard datasets (KITTI, Cityscapes) for object detection, we needed to compute some results ourselves for competing SOTA methods for a fair comparison. However, this is not true for the UAVDT dataset, where we use the standard data training and test split.

Evaluation Metrics: We use Average Precision (AP) for multiple scales of objects and Mean Average Precision (mAP) across varying IoU thresholds and mAP50 and mAP75, respectively at 0.5 and 0.75 IoU thresholds, to evaluate detection accuracy. Intersection over Union (IoU) is used to evaluate bounding box precision on all datasets.

B. Implementation Details

For features extraction, we used HRNet [17], and pre-trained it on the COCO dataset [23], following the methodology described in [22]. Our global architecture follows CenterNet [22]. However, our training process is done in two steps. First, our backbone is fine-tuned on each dataset starting from the pre-trained weights on COCO. Then, the first and third layers of the backbone are frozen and the MFA, SFA, and dual-fusion modules as well as the network heads are trained. Training is conducted over 250 epochs utilizing the Adam optimizer, starting with a learning rate of 1×10^{-4} , which undergoes a decimation by a factor of 10 after the 130th and 140th epochs. To ensure training stability, we use gradient clipping. The same training protocol was used for the overall architecture as well as for all the base detectors to demonstrate the contribution of our approach.

C. Results and Discussion

Comparisons with SOTA methods on the Cityscapes dataset are reported in Table I.

They show that our attention-based fusion detector consistently outperforms the other SOTA detectors. There is a significant improvement in the detection results when using our STF model as compared to SOTA detectors. The

improvement in detection results is due to our two attention modules and our dual-fusion module, all contributing positively to detecting objects better (especially small or occluded ones). In Table I, we also compare our model with the vanilla HRNet as we use a feature extractor based on the HRNet architecture. This allows us to examine our results in comparison to vanilla HRNet to observe the impact of our STF module on a similar backbone. This comparison demonstrates a gain in accuracy for all sizes of objects. Furthermore, we changed the backbone of Centernet [22] to observe how HRNet affects its performance as it uses a detection similar as ours. It can be concluded from the results that using HRNet alone does not yield significant improvement. This is another demonstration that our method using a classification head similar to CenterNet performs better due to our SFA and MFA modules. By comparing our results with YOLOv5 and the recent YOLOX, our model shows improvement in terms of precision and accuracy, as well. Finally, we also perform better than PPNet which uses multiple frames.

Table II presents the results of the KITTI validation dataset. The conclusions are the same as for Cityscapes with similar improvements compared to baseline methods. By comparing it with other SOTA detectors, our proposed method outperforms them with improvements for all object size categories (Small, medium and large). Our method demonstrates an improvement in detection results when compared to SOTA single-frame and two-frame detectors.

Results on the UAVDT test dataset are reported in Table III. Our Spatio-Temporal Fusion (STF) module consistently outperforms the base detectors. As well, when compared to SOTA multi-frame detectors, such as FFAVOD and RN-VID that fuse features without attention, we can notice that although this helps compared to single-frame detectors, a more sophisticated fusion approach, like the one we propose is required to obtain even better results.

D. Ablation Study

An ablation study was performed to evaluate the contribution of the different parts of the proposed method: the multi-frame attention module, single-frame attention module as well as the single-frame and multi-frame attention with the dual-frame fusion module, and show the effect of each component in Table IV. We find that the method with the MFA module or the SFA module detects better than the baseline method (HRNet + CenterNet head). We also observe that the proposed method (STF) with both two modules and dual-frame fusion performs the best. According to the proposed STF module, our MFA module plays a crucial role in combining features from two frames. Similarly, the SFA module aims to improve the accuracy of detection within a single frame. This is achieved through the combination of single-frame channel and spatial attention, which effectively suppresses false positive detection in background regions.

In our observations, we noted that each module independently contributes to performance enhancement. Moreover, a synergistic effect is observed when both modules are combined, leading to a more significant improvement in results. Therefore, for better efficiency and accuracy, our proposed model demonstrates superior results as compared to other configurations.

To illustrate the specific contributions of our proposed dual-frame fusion method, we also conducted an ablation study on it. We aimed to understand the individual impact of different fusion strategies on the overall performance of our model. For that, we use different strategies of combining two frames, i.e. concatenation, median, mean, and max fusion. In all cases, that decreased the performance by a large margin as shown in Table ?? We attribute this to the misalignment of features across frames, necessitating a more intricate operation for aggregating these features. Admittedly, our model requires additional parameters to effectively learn the optimal combination of feature maps. However, as indicated in Table V, our findings strongly support the benefit of our dual-frame fusion method in integrating feature maps.

V. CONCLUSION

In this work, we designed a spatio-temporal fusion module as a new approach for multi-frame object detection. Specifically, we identified the ineffectiveness and inadequacy issues present in single-frame object detectors. Then, we proposed to solve these problems using multi-frame and single-frame attention modules, as well as a dual-frame fusion module to improve object representation. Our results show that by exploiting sequential frames, we can improve the efficiency and accuracy of detection under challenging conditions.

ACKNOWLEDGEMENT

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2020-04633].

REFERENCES

- [1] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.
- [2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [3] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 408–417.
- [4] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European conference on computer vision*. Springer, 2020, pp. 474–490.

Table I
COMPARISON OF OUR METHOD WITH SOTA METHODS ON THE CITYSCAPES VALIDATION DATASET. **BOLDFACE** INDICATES BEST RESULTS.

Method Type	Method	Backbone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA (Single Frame Detectors)	RetinaNet [24]	RetinaNet-50	92.8	94.1	93.2	43.1	58.2	94.0
	Vanilla HRNet*	HRNet	92.2	94.9	91.1	45.3	59.9	93.1
	CenterNet [24]	Hourglass-104	92.1	93.9	93.0	43.2	56.7	93.5
	CenterNet*	Resnet-18	92.7	92.5	92.7	44.8	57.6	93.2
	CenterNet*	HRNet	92.8	93.3	93.1	45.9	58.7	93.4
	YOLOv5 [24]	CSPDarknet53	93.6	93.4	91.8	43.7	59.5	95.1
	YOLOX[24]	CSPDarknet53	93.9	94.9	92.7	44.8	61.5	96.7
SOTA (Two Frame Detectors)	PPNet[24]	Resnet-50	94.8	96.2	92.5	43.9	57.4	95.8
	STF (Ours)	HRNet	95.7	97.2	95.3	49.3	65.3	97.3

*Trained by ourselves.

Table II
COMPARISON OF OUR METHOD WITH SOTA METHODS ON THE KITTI MOT VALIDATION DATASET. **BOLDFACE** INDICATES THE BEST RESULT.

Method Type	Method	Backbone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA methods (Single Frame Detectors)	RetinaNet [25]	RetinaNet-50	56.6	-	-	29.9	62.8	73.1
	Vanilla HRNet*	HRNet	79.1	81.6	70.0	48.3	65.2	74.3
	CenterNet [26]	Hourglass-104	-	85.3	-	-	-	-
	CenterNet*	Resnet-18	80.5	83.4	74.5	50.2	66.8	78.7
	CenterNet*	HRNet	81.7	83.3	74.1	50.0	66.8	77.4
	YOLOv5*	CSPDarknet53	84.3	86.8	76.3	52.9	70.4	83.5
	YOLOX*	CSPDarknet53	85.9	87.7	79.8	53.8	71.7	84.9
SOTA methods (Two Frame Detectors)	Mf-SSD [7]	SqueezeNet	83.0	-	-	-	-	-
	MFCN [27]	ResNet101	84.6	-	-	-	-	-
	PPNet [24]	ResNet50	86.2	-	-	-	-	-
	STF (Ours)	HRNet	88.7	90.0	82.9	57.1	74.6	88.1

*Trained by ourselves.

Table III
COMPARISON OF UAVDT TEST DATASET WITH DIFFERENT METHODS. **BOLDFACE** INDICATES THE BEST RESULT.

Method Type	Method	Backbone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA methods (Single Frame Detectors)	RetinaNet [28]	RetinaNet-50	16.2	34.0	13.7	8.8	30.1	23.8
	CenterNet [29]	Hourglass-104	16.4	29.7	16.6	12.2	25.1	11.3
	YOLOv5 [29]	CSPDarknet53	18.0	33.5	17.2	11.0	29.6	37.5
	YOLOX[30]	CSPDarknet53	26.0	43.3	21.4	-	-	-
	SpotNet [31]	U-Net	53.4	-	-	-	-	-
SOTA methods (Two Frame Detectors)	STDnet-ST [32]	RCN	13.3	36.4	-	-	-	-
	AdNet-MS [28]	Darknet53	13.3	43.5	18.3	12.1	37.9	27.9
	RN-VID [15]		39.4	-	-	-	-	-
	FFAVOD-SpotNet [6]		53.8	-	-	-	-	-
	FFAVOD-CenterNet [6]		52.1	-	-	-	-	-
	STF (Ours)	HRNet	58.4	79.5	46.3	35.8	59.4	61.9

Table IV
ABLATION STUDY ON THE MFA, SFA, AND DUAL-FUSION MODULES

Configuration	mAP (%)
Baseline (HRNet+CenterNet head)	92.10
Baseline + SFA	93.50
Baseline + MFA	94.91
Baseline (MFA+SFA)	95.73

Table V
ABLATION STUDY OF THE DIFFERENT FUSION STRATEGIES ON CITYSCAPES DATASET.

Fusion Methods	mAP
Concatenation	88.60
Median	91.50
Mean	91.70
Max	91.89
Dual-frame fusion (Ours)	95.73

- [5] M. Liu and M. Zhu, "Mobile video object detection with temporally-aware feature maps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5686–5695.
- [6] H. Perreault, G.-A. Bilodeau, N. Saunier, and M. H  ritier, "Ffavid: Feature fusion architecture for video object detection," *Pattern Recognition Letters*, vol. 151, pp. 294–301, 2021.
- [7] A. Broad, M. Jones, and T.-Y. Lee, "Recurrent multi-frame single shot detector for video object detection." in *BMVC*, 2018, p. 94.
- [8] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proceedings of the European Conference on Computer Vision*

- (ECCV), 2018, pp. 331–346.
- [9] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang *et al.*, “T-cnn: Tubelets with convolutional neural networks for object detection from videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.
 - [10] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, “Multi-class multi-object tracking using changing point detection,” in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 68–83.
 - [11] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, “An empirical study of spatial attention mechanisms in deep networks,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6688–6697.
 - [12] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
 - [13] G. Elsayed, P. Ramachandran, J. Shlens, and S. Kornblith, “Revisiting spatial invariance with low-rank local connectivity,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 2868–2879.
 - [14] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, “The unmanned aerial vehicle benchmark: Object detection and tracking,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.
 - [15] H. Perreault, M. H  ritier, P. Gravel, G.-A. Bilodeau, and N. Saunier, “Rn-vid: A feature fusion architecture for video object detection,” in *International Conference on Image Analysis and Recognition*. Springer, 2020, pp. 125–138.
 - [16] F. Xiao and Y. J. Lee, “Video object detection with an aligned spatial-temporal memory,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 485–501.
 - [17] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
 - [18] Z. Huang, S. Zhang, L. Pan, Z. Qing, M. Tang, Z. Liu, and M. H. Ang Jr, “Tada! temporally-adaptive convolutions for video understanding,” *arXiv preprint arXiv:2110.06178*, 2021.
 - [19] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, “Tc-track: Temporal contexts for aerial tracking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 798–14 808.
 - [20] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.
 - [21] Y. Zhang, Y. Bai, H. Wang, Y. Xu, and Y. Fu, “Look more but care less in video recognition,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 813–30 825, 2022.
 - [22] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.
 - [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll  r, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
 - [24] C.-J. Li, Z. Qu, and S.-Y. Wang, “Perspectivenet: An object detection method with adaptive perspective box network based on density-aware,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5419–5429, 2023.
 - [25]   . Lorente, I. Riera, and A. Rana, “Scene understanding for autonomous driving,” *arXiv preprint arXiv:2105.04905*, 2021.
 - [26] H. Wang, Y. Xu, Z. Wang, Y. Cai, L. Chen, and Y. Li, “Centernet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved centernet,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.
 - [27] D. Liu, Y. Cui, Y. Chen, J. Zhang, and B. Fan, “Video object detection for autonomous driving: Motion-aid feature calibration,” *Neurocomputing*, vol. 409, pp. 1–11, 2020.
 - [28] R. Zhang, S. Newsam, Z. Shao, X. Huang, J. Wang, and D. Li, “Multi-scale adversarial network for vehicle detection in uav imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 283–295, 2021.
 - [29] J. Liao, Y. Piao, J. Su, G. Cai, X. Huang, L. Chen, Z. Huang, and Y. Wu, “Unsupervised cluster guided object detection in aerial images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 204–11 216, 2021.
 - [30] X. Xu, Z. Feng, C. Cao, C. Yu, M. Li, Z. Wu, S. Ye, and Y. Shang, “Stn-track: Multiobject tracking of unmanned aerial vehicles by swin transformer neck and new data association method,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8734–8743, 2022.
 - [31] H. Perreault, G.-A. Bilodeau, N. Saunier, and M. H  ritier, “Spotnet: Self-attention multi-task network for object detection,” in *2020 17th Conference on Computer and Robot Vision (CRV)*. IEEE, 2020, pp. 230–237.
 - [32] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, “Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2021.