CompFeat: Comprehensive Feature Aggregation for Video Instance Segmentation

Yang Fu[1], **Linjie Yang**[2], **Ding Liu**[2], **Thomas S. Huang**[1], **Humphrey Shi**[1,3]
[1]University of Illinois at Urbana-Champaign
[2]ByteDance Inc
[3]University of Oregon

Abstract

Video instance segmentation is a complex task in which we need to detect, segment, and track each object for any given video. Previous approaches only utilize single-frame features for the detection, segmentation, and tracking of objects and they suffer in the video scenario due to several distinct challenges such as motion blur and drastic appearance change. To eliminate ambiguities introduced by only using single-frame features, we propose a novel comprehensive feature aggregation approach (**CompFeat**) to refine features at both frame-level and object-level with temporal and spatial context information. The aggregation process is carefully designed with a new attention mechanism which significantly increases the discriminative power of the learned features. We further improve the tracking capability of our model through a siamese design by incorporating both feature similarities and spatial similarities. Experiments conducted on the YouTube-VIS dataset validate the effectiveness of proposed CompFeat.

Introduction

Video instance segmentation (VIS) is a joint task of detection, segmentation and tracking of object instances in videos (Yang, Fan, and Xu 2019). Different from instance segmentation in image domain (Hariharan et al. 2014), video instance segmentation not only requires to segment object masks on individual frames, but also to track the identities of objects across different frames. Also, unlike semi-supervised video object segmentation (Voigtlaender et al. 2019a; Voigtlaender and Leibe 2017; Xu et al. 2018; Wug Oh et al. 2018; Oh et al. 2019; Xu et al. 2019), video instance segmentation does not require a ground truth mask in the first frame and all objects appear in the video should be processed. It has essential applications in many video-based tasks, including video editing, autonomous driving and augmented reality.

Video instance segmentation has several distinct challenges. For example, if an object is recognized as a wrong category in one frame of the video, tracking of this object will be extremely hard due to inconsistency of object categories. When there are multiple similar objects, finding the correspondences of them across the video is also challenging. VIS is an important but underexplored task. The pioneering work for VIS is MaskTrack-RCNN (Yang, Fan, and Xu 2019), which is built upon Mask-RCNN (He et al. 2017), a state-of-the-art method for image instance segmentation. A new tracking branch is tailored and added in order to track object instances. However, MaskTrack-RCNN relies on only single frame object features and neglects critical temporal information, *i.e.* temporal consistency of an object, motion pattern of different objects, etc, all of which can provide abundant information for category recognition, object detection and mask segmentation across video frames. And lots of recent video understanding work (Wang et al. 2018b; Wu et al. 2019) focused on how to utilize the temporal information. In addition, the proposed tracking head in MaskTrack-RCNN is preliminary and ignores the spatial layout of objects with simple object features, which has been proven crucial by modern object tracking algorithms (Zhang and Peng 2019; Zhu et al. 2018; Li et al. 2019) to improve the tracking of video instances.

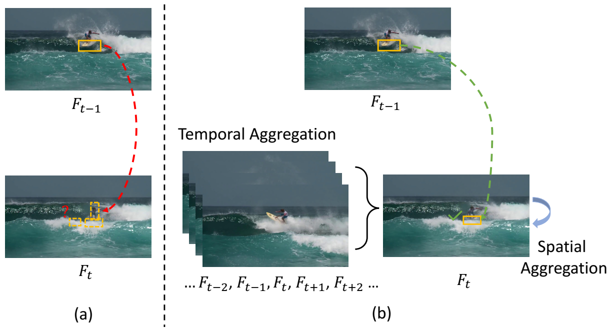


Figure 1: An illustration of the proposed comprehensive feature aggregation (CompFeat) method. F_t denotes the video frame at time t. (a) Previous video instance segmentation method without feature aggregation. (b) Our proposed comprehensive feature aggregation approach for video instance segmentation

In order to utilize the abundant information in videos and to harvest the benefits of modern object tracking approaches, we propose a comprehensive feature aggregation approach for video instance segmentation, termed **CompFeat**. The main idea of CompFeat is illustrated in Fig 1. As shown in Fig 1(a), the key object is not detected and fails to be tracked due to using only the unclear visual cues of a single frame, while other frames in the same video contain helpful information for locating and tracking the correct object. Hence, we propose a dual attention module with both temporal attention and spatial attention to aggregate contextual information from neighboring frames and other positions in the current frame as described in Fig 1(b). We also enhance the features of detected objects by extending the dual attention module to object level, which substantially improves the discriminative power of the object features, enabling more accuracy object detection and segmentation. In addition, we introduce a novel correlation-based tracking module to improve instance tracking across different frames. Instead of using a holistic similarity between a pair of detected object and reference object to determine object correspondence, our correlation-based module not only employs depth-wise correlation between an object pair to generate a matching score with spatial awareness, but also computes a correlation map between a reference object and the current frame to better localize the target object similar to Siamese object tracking.

To summarize, the main contributions of this work are threefold as follows:

- We propose a comprehensive feature aggregation approach for video instance segmentation, including temporal and spatial attention modules on both frame-level and object-level features.
- We introduce a correlation-based tracking module to track instances across frames, which predicts cross-correlation maps in both object-to-object and object-to-frame manners to produce multiple similarity cues for object tracking.
- We conduct extensive experiments and ablation study on YouTube-VIS (Yang, Fan, and Xu 2019) to demonstrate the effectiveness of our proposed framework and each of the individual components.

Related Work

In this section we review video instance segmentation and several closely-related tasks such as video object detection and video object tracking.

Video Object Detection. Video object detection aims to detect all objects in videos such as shown in the ImageNet VID challenge (Russakovsky et al. 2015; Han et al. 2016). Feature aggregation is widely used in video detection (Zhu et al. 2017; Feichtenhofer, Pinz, and Zisserman 2017; Chen et al. 2018; Liu et al. 2019). For instance, Zhu *et al.* proposed to aggregate features from nearby frames to enhance the feature quality of an input frame. However, its speed is pretty slow due to the dense detection and optical flow estimation. In (Chen et al. 2018), Chen *et al.* proposed to use a scale-time lattice to generate detection on sparse key frames and designed a temporal propagation approach for detection in an effective way. Inspired by these work, we propose to improve the feature quality for video instance segmentation via feature aggregation using attention mechanism.

Video Object Tracking. Video object tracking can be viewed as a sub-task of video instance segmentation, which has two scenarios: detection-based tracking and detection-free tracking. In detection-free tracking, given the ground truth location of the target object in the first frame, algorithm is required to track the target object through the whole video. Recently, the Siamese network based trackers have received significant attentions due to their well-balanced accuracy and efficiency (Li et al. 2019; Zhang and Peng 2019; Wang et al. 2018a; Valmadre et al. 2017). In particular, these trackers attempt to produce a similarity map from cross-correlation of the two feature branches, one for the target object and the other for the search region, where the similarity map embeds more semantic meanings. On the other hand, detection-based tracking (Sadeghian, Alahi, and Savarese 2017; Son et al. 2017; Shi 2018) simultaneously detect and track multiple video objects, which is more similar to the setting of video instance segmentation. In our proposed CompFeat, we borrow ideas from both detection-based tracking and detection-free tracking.

Video Instance Segmentation. MaskTrack-RCNN (Yang, Fan, and Xu 2019) is the first attempt to address the video instance segmentation problem. It proposes a large-scale video dateset named YouTube-VIS for benchmarking video instance segmentation algorithms. Several methods in the Large-Scale Video Object Segmentation Challenge achieve impressive results but they utilize large quantity of external data and complex algorithm pipelines (Wang et al. 2019; Dong et al. 2019; Luiten, Torr, and Leibe 2019). A closely related work is multi-object tracking and segmentation (Voigtlaender et al. 2019b) which is proposed to evaluate multi-object tracking along with instance segmentation. However, because of its limited data scale and few object categories, we do not compare with it in this paper. MaskTrack-RCNN only uses image features but not temporal information of video sequences. We extend this work with a more sophisticated comprehensive feature aggregation approach which greatly boosts the performance on video instance segmentation.

Proposed Method

In our framework, the video frames are sequentially processed. During the video processing, we randomly select one frame as the current frame and sample other several frames as support frames which are used for temporal feature aggregation. Meanwhile, a correlation-based tracking module is proposed to track object identities across frames with comprehensive cues. The overview of the proposed CompFeat framework is shown in Fig 2. The current frame and all support frames are first fed into ResNet50 (He et al. 2016) for feature extraction. Then, our proposed **temporal attention module** takes the features of the current frame and support frames as inputs for feature aggregation over different frames. Meanwhile, the features of the current frame are processed by a **spatial attention module** for global context feature aggregation on a single frame. The similar process is performed on object level with same model structures. In addition, we enhance the tracking branch of the network via a **correlation-based tracking module** for a more accurate object tracking. The correlation-based tracking module combines cross-correlation between a pair of reference object

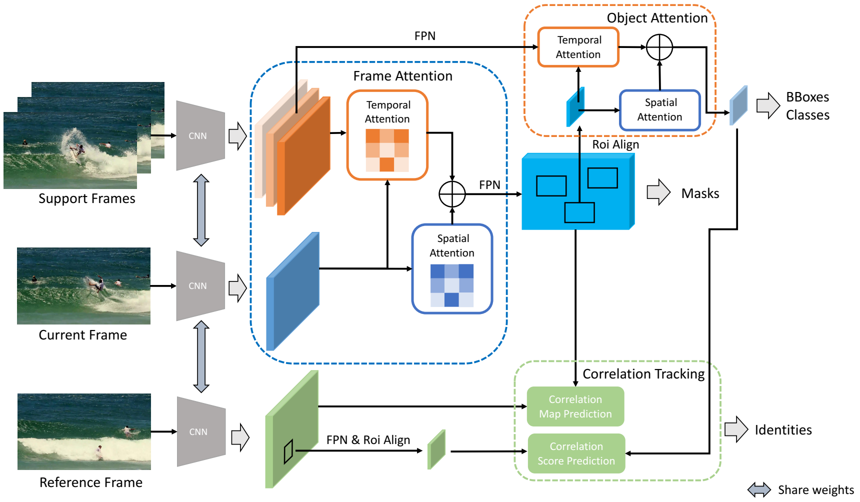


Figure 2: An overview of our proposed CompFeat approach for video instance segmentation. CompFeat consists of three major components: (a) A frame level attention module incorporating both temporal and spatial attention modules. (b) An object level attention module which has a similar structure as frame level attention. (c) A correlation-based tracking module to predict the correlation score and correlation map simultaneously.

and newly detected object and correlation between the reference object and the current frame (search region). Finally, we integrate the three proposed modules into a complete framework to perform three different tasks, object detection, mask segmentation and object tracking simultaneously. Next we describe each proposed component in details.

Temporal Attention Module

Inspired by the non-local networks (Wang et al. 2018b), we propose a novel temporal attention module to refine the features of the current frame via aggregating information from other support frames. Different from the original non-local block, which aims to model the long-range dependencies by attention mechanism, our proposed temporal attention module focuses more on embedding information from other frames and use them to refine features of the current frame by cross-attention mechanism. Specifically, the temporal attention module has three steps: embedding of current frame embedding, embedding of support frames and features aggregation as described in Fig 3(a).

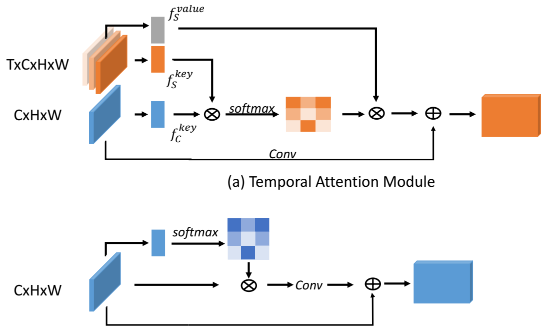
Embedding of current frame. Given the features of the current frame $f_C \in \mathbb{R}^{C \times H \times W}$, where H, W, C are the height, width, and the feature dimension of output feature map from the backbone network. We first feed it into a convolution layer to generate a feature map f_C^{key}, where $f_C^{key} \in \mathbb{R}^{\frac{C}{4} \times H \times W}$, and a non-linearity activation is applied. The f_C^{key} can store the key features of the current frame including which and where objects may exist. Therefore, the feature map f_C^{key} is learnt to encode the key information of visual semantics in the current frame, *i.e.* object categories, object locations and masks.

Embedding of support frames. Given a stack of feature maps obtained from support frames $\{f_{S_t} \in \mathbb{R}^{C \times H \times W}; t = 1 : T\}$ (T is the number of the support frames), each feature map f_{S_t} is first encoded into a pair of feature maps $f_{S_t}^{key}$ and $f_{S_t}^{value}$ by two parallel convolution layers, where $f_{S_t}^{key}, f_{S_t}^{value} \in \mathbb{R}^{\frac{C}{4} \times H \times W}$. Non-linear activation is applied as well. If there are more than one support frames ($T > 1$), we concatenate features of different frames along temporal dimension and obtain $f_S^{key}, f_S^{value} \in \mathbb{R}^{T \times \frac{C}{4} \times H \times W}$. f_S^{key} contains the information of key features of support frames. The similarities between f_C^{key} and f_S^{key} shows when-and-where the features of support frames are suitable to be aggregated for the current frame. In addition, the feature map f_S^{value} is learnt to represent the context information of all support frames.

Features aggregation. In the feature aggregation step, we first compute the attention weights by similarities between all pixel in feature map f_C^{key} and f_S^{key}. The similarities are computed as the correlation of every spatial location in the feature map f_C^{key} and every spatial-temporal location in feature map f_S^{key}. Then the attention weights are used to aggregate features from f_S^{value} to obtain context features from support frames. In addition, we perform a feature transformation by a 1×1 convolution layer. The whole process of feature aggregation for every position of the current frame can be summarized as the following equation,

$$F_{TA} \Leftrightarrow \begin{cases} X = f_S^{key} \odot f_C^{key} \\ f_A(:,j) = F(f_S^{value} \odot \dfrac{exp(X(:,j))}{\sum_{i=1}^{N_p} exp(X(:,i))}) \end{cases} \tag{1}$$

where X is the similarity matrix between the current frame and support frames with size of $HWT \times HW$, i, j are the indices of every position in the similarity matrix and the feature map, N_p is the total number of positions in the feature map, $\odot$ is dot product, F is a transformation function with



(a) Temporal Attention Module

(b) Spatial Attention Module

Figure 3: An illustration of (a) temporal attention module and (b) spatial attention module. Best viewed in color.

non-linear activation, f_A is the aggregated feature map after the transformation. Note that, all feature maps in above equation may be processed by some necessary reshaping or permutation operations. Finally, the refined feature map $\tilde{f_C}$ is obtained by summing up the aggregated feature map f_A and the feature map of the current frame f_C.

After aggregation, we can obtain a feature map $\tilde{f_C}$, which not only preserves some information key visual semantics of current frame, but also extracts useful contextual information existing in other frames within the same video.

Spatial Attention Module

Besides aggregating temporal contextual information from support frames, we also propose a spatial attention module to model the spatial context of the current frame. The proposed spatial attention module is based on the non-local block (Wang et al. 2018b) with several modifications. As shown in Fig 3(b), instead of computing the attention maps of all spatial positions for each channel, we simplify it by sharing a single attention weight for all spatial positions in each channel, which is known as the channel attention. As reported in (Cao et al. 2019), this simplified channel attention module can achieve very similar performance compared to original non-local block. Also, since it doesn't need to compute a specific attention weight for every spatial position, it can be more efficient. Then, we use two more convolution layers to transform the features. Finally, the channel attention is added back to the feature map of the current frame for global context feature aggregation.

Combining Two Attention Modules

Furthermore, we integrate the two attention modules into a dual attention module. The dual attention model takes the features of the current frame and support frames as inputs. The features of the current frame f_C is first processed by spatial attention model to obtain an attention map with context features of itself. And features of current frames f_C and features of support frames f_S are fed into the temporal attention module to generate another attention map with context information of support frames. Then, we aggregate the features of the current frame with two attention maps by adding them to the original feature map f_C as follows,

$$f_{agg} = F_{TA}(f_C, f_S) + F_{SA}(f_C) + f_C \qquad (2)$$

where F_{TA}, F_{SA} represent proposed temporal attention module and spatial attention module respectively.

Attention Module on Object Level

Attention modules described above are all processed on frame level features, which means they aggregate the context information for the whole feature map from the current frame and support frames. However, for instance segmentation with a two-stage framework, aggregating context information for each object proposal is also critical. Object proposals act as valuable candidates for the final object predictions and they encode more focused features for the individual objects. Proper feature aggregation onto these object proposals can elucidate confusing feature representations and improve recognition accuracy. We attempt to extend the two proposed attention modules to object level by applying similar operations to the object features produced by ROI Align (He et al. 2017). Since the proposed attention modules can be applied to feature maps with arbitrary size, we adopt the two proposed attention modules in a similar fashion. We denote the features of detected object proposals as f_{C_p}. $f_{C_p} \in \mathbb{R}^{P \times C \times h \times w}$, where P is the number of object proposals, C is the channel dimension, and h, w are height and width of the proposal. The features of each proposal are fed into a dual attention module along with feature maps of support frames. Since the size of the features of proposals is much smaller than the whole feature map, aggregating features on object level can be very efficient. In the next section, we show the performance gain of adding object-level attention modules is comparable or even better than adding attention module on frame level.

The proposed attention module is inspired by some early work of video object detection and video object segmentation (Wu et al. 2019; Oh et al. 2019), where the non-local blocks are used for extracting self-attention features. However, the motivation of proposed attention module is different as it is designed to aggregate the features from other frames (support frames) to current frame through feature matching, but not to itself by self-attention mechanisms. In addition, we extend the proposed attention module to object level, which largely reduces the computational complexity while improves the performance as frame-level attention module. Furthermore, frame-level and object-level module can be integrated into a single framework for a better performance. To our best knowledge, there are few work on enhancing object level feature by other frames. And the experiments in Sec. show the effectiveness of our proposed attention module.

By performing proposed dual attention module on both frame level and object level, we achieve feature aggregation on both spatial and temporal dimension, and in both local and global granularities, which is a comprehensive approach to aggregate and enhance intermediate features in a video instance segmentation framework.

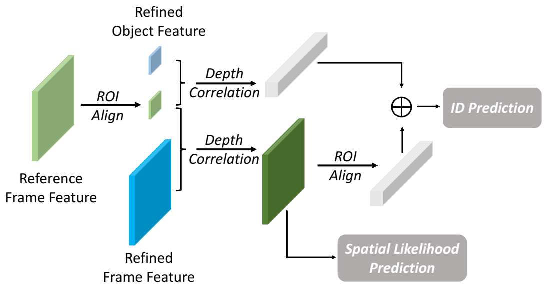


Figure 4: An illustration of the correlation-based tracking module. The convolution layers are omitted here.

Correlation-based Tracking Module

In order to track an object across frames in a more robust and consistent way, we propose a new correlation-based tracking module to generate both spatial likelihoods and object similarities, which is more powerful than the previous approach (Yang, Fan, and Xu 2019).

Recent works on object tracking shows the effectiveness of correlation-based tracking which has achieved state-of-the-art performance on several tracking benchmarks. Inspired by SiamRPN++ (Li et al. 2019), our proposed tracking module incorporates knowledge from both object similarities and cross-correlations of the target object and the search region. As shown in Fig 4, our tracking module can be abstracted into three procedures:(a) pair-wise similarity computation, which uses depth-wise correlation instead of matrix multiplication (Yang, Fan, and Xu 2019) to predict a similarity matrix between two groups of objects. (b) cross-correlation computation, which generates correlation maps with depth-wise correlation between a single object and the whole frame. To train the model to produce high quality correlation maps, we employ a pseudo likelihood map represented by a two dimensional gaussian distribution centered at the location of the object as the supervision signal. Then we compute the aggregated similarity vector for a detected object with ROI align within the corresponding bounding box on the correlation map. After that, we sum up the two similarity vectors to obtain the final depth-wise similarity vector. (c) The final similarity score is mapped from the similarity vector with two 1×1 convolution layers. The proposed correlation-based tracking module considers both feature similarity and correlation-based similarity scores on the raw image features to increase the robustness of tracking.

Experiments

In this section, we conduct extensive experiments on YouTube-VIS (Yang, Fan, and Xu 2019) to evaluate the effectiveness of each proposed component and compare our proposed method with previous approaches.

Data and Evaluation Metric

Data YouTube-VIS is the first and largest dataset for video instance segmentation, which is a subset of YouTube-VOS dataset (Xu et al. 2018). YouTube-VIS is comprised of 2,883 high resolution YouTube videos with 40 common object categories. In each video, several objects with bounding boxes and masks are labeled manually and the identities cross different frames are annotated as well. Since only the validation set is available for evaluation, all results reported in this paper are evaluated on the validation set.

Evaluation Metrics To evaluate the performance of the proposed method, we use the metrics mentioned in (Yang, Fan, and Xu 2019), which are average precision(AP) and average recall(AR) based on a spatial-temporal Intersection-over-Union (IoU). Following the COCO evaluation, AP is computed by averaging over multiple IoU thresholds, *e.g.* from from 50% to 95% at step 5% and AR is the maximum recall given some fixed number of segmented instances per video. Both metrics are first calculated for each category and then averaged over 40 categories.

Implementation Details

Training. Our proposed method is built on Mask-RCNN (He et al. 2017). The backbone network structure is ResNet50 with FPN (Lin et al. 2017), which is pre-trained on MSCOCO dataset (Lin et al. 2014). In the tracking branch, we use two convolution layers to refine the correlation features generated by depth-wise correlations, respectively. The first convolution layer has 256 channels which is the same dimension as the correlation features while the second one is used for correlation map prediction with only one output channel. Our model is implemented based on MMDetection (Chen et al. 2019) and the whole framework is trained end-to-end in 12 epochs with two NVIDIA 2080TI GPUs. We resize the original frame size to 640×360 for both training and testing. During training, the initial learning rate is set to 0.0125 and decays with a factor of 10 at epoch 8 and 11. For each input frame, we randomly select three frames from the same video, two used as support frames in the dual attention module and the other used as reference frame in the tracking module.

Testing. During evaluation, the testing video is processed by the proposed method frame by frame in an online fashion. For each input frame, four additional frames are sampled from the testing video as support frames. Note that the number of support frames used in training and testing can be different, since testing with more support frames can help improve performance. We conduct an ablation study on the number of support frames in the following section. For the tracking head, we ignore the correlation map and use the predicted correlation score as the tracking score. Then we follow the inference procedure described in (Yang, Fan, and Xu 2019) to predict the category, bounding box and mask for the object instance. In addition, we combine other cues, *i.e.* detection confidence, bounding box IoU, and category consistency, along with tracking scores to improve the tracking accuracy as a powerful post-processing.

Ablation Study

Baseline Model and Data Augmentation Since our proposed method is built on Mask-Track RCNN (Yang, Fan, and Xu 2019), we take Mask-Track RCNN as the baseline model to validate the effectiveness of each contribution. We

Methods	AP	$AP_{0.5}$	$AP_{0.75}$
Mask-Track RCNN (Yang, Fan, and Xu 2019)	21.1	37.7	23.6
Our Implementation	20.9	37.9	21.6
Our Implementation + MSCOCO	**24.1**	**42.6**	**24.9**

Table 1: Performance of the baseline model on YouTube-VIS validation set. “Our Implementation” means our reproduced results of Mask-Track RCNN.“Our Implementation + MSCOCO” is used as the baseline model in all ablation studies hereafter.

(a) **Attention on Frame Level**			
	AP	$AP_{0.5}$	$AP_{0.75}$
Baseline + Temporal Attention	25.2	43.9	25.6
Baseline + Spatial Attention	24.9	43.4	25.2
Baseline + Spatial-Temporal Attention	25.8	44.5	27.0
(b) **Attention on Object Level**			
	AP	$AP_{0.5}$	$AP_{0.75}$
Baseline + Temporal Attention	25.4	44.4	25.9
Baseline + Spatial Attention	24.8	43.3	25.1
Baseline + Spatial-Temporal Attention	26.1	45.6	26.7
(c) **Attention on Both Frame and Object Level**			
	AP	$AP_{0.5}$	$AP_{0.75}$
Baseline + Spatial-Temporal Attention	**27.5**	**46.1**	**28.9**

Table 2: Ablation Study of our proposed attention module on YouTube-VIS validation set. The best results are highlighted in bold.

Modules	AP	$AP_{0.5}$	$AP_{0.75}$
Baseline + CM	25.1	44.3	26.7
Baseline + CM + FDA	26.3	45.3	27.1
Baseline + CM + BDA	26.7	45.9	27.3
Baseline + CM + FDA + BDA (CompFeat)	**27.9**	**46.8**	**30.2**

Table 3: Ablation Study of the proposed track module on YouTube-VIS validation set. CM, FDA, BDA denote the proposed tracking module with correlation map, the frame level dual attention module and object level dual attention module, respectively. The best results are highlighted in bold.

Train/Test(Uniform)	2 frames	3 frames	4 frames	5 frames
Uniform 2 frames	27.1	27.4	27.6	27.3
Uniform 3 frames	26.2	26.6	26.6	26.8
Random 2 frames	27.3	27.7	27.9	27.4
Random 3 frames	26.1	26.8	26.8	26.7

Table 4: Performance of different sampling methods and different frames during training/testing on the validation set of Youtube-VIS. The performance is reported in AP.

first reproduce the Mask-Track RCNN with publicly available codes and the results are listed in Table 1. Our result is close to the result reported in original paper. Note that all results in Table 1 are without any post-processing.

In addition, we find that the number of object instances in YouTube-VIS dataset is limited. Since the proposed method does not depend on the temporal smoothness of a video, some image-based datasets can be adopted to increase the training samples. We choose MSCOCO (Lin et al. 2014) as external data which has a large overlap on the object categories with YouTube-VIS. In order to make use of the image data, we generate support frames and reference frame based on a single image by some affine transformations, *i.e. i.e.* rotation, translation and shearing. The identity annotations across different images can be generated automatically in this process. The performance after using external data is listed in Table 1 as well. We use this model as a baseline model for all the following ablation experiments.

Effectiveness of Attention Module We conduct ablation study to prove the effectiveness of the proposed attention module: temporal attention and spatial attention. The results are listed in Table 2. Note all results here are without post-processing. We first evaluate each attention module on frame level in Table 2(a). With the temporal attention module, we improve baseline model by 1.1% in AP and 1.3% in $AP_{0.5}$ respectively. Similarly, spatial attention module can also slightly improve the baseline performance by 1%. When combining both temporal and spatial attention together as the dual attention module, we further boost the baseline model by 1.7%, 1.9% and 2.1% in AP, $AP_{0.5}$ and $AP_{0.75}$. These experimental results prove that our proposed attention module can aggregate helpful context feature from both other frames and the input frame.

We then experiment with temporal and spatial attention on object level. Table 2(b) shows that proposed attention method on object level can always achieve comparable or even better results compared with the frame level one. For instance, by performing both temporal and spatial attention on object level, the performance gain becomes 2.0%, 3.0% and 1.8% in AP, $AP_{0.5}$ and $AP_{0.75}$ respectively.

Furthermore, Table 2(c) lists the performance of combining attention modules on both frame level and object level. Comparing with the performance only on frame or object level, the combination one is superior. Specifically, with both attention on two different levels, we achieve $AP/AP_{0.5} = 27.5\%/46.1\%$, which outperforms the performance with attention module on frame level by 1.7%/1.6%. This further improvement shows that by using attention module on frame level and object level, we can aggregate context information in a global-to-local manner, which can greatly improve the baseline model by 3.4%, 3.5% and 4.0% in AP, $AP_{0.5}$ and $AP_{0.75}$.

Effectiveness of Correlation-based Tracking Module The ablation study on the proposed correlation-based tracking module is shown in Table 3. Again, the results are without post-processing. Compared to the baseline model, the tracking module with correlation map outperforms the baseline model by more than 1% on AP, $AP_{0.5}$ and $AP_{0.75}$. This improvement indicates the cross-correlation map between objects and the whole frames contains more semantic information than the object features used by the baseline. In addition, when integrating it into the whole framework with dual attention on frame level or object level, we obtain consistently better performance. In particular, by using correlation maps and dual attention module on frame level, we improve the performance in AP from 25.8% to 26.3%. Finally,

Methods	AP	$AP_{0.5}$	$AP_{0.75}$	AR_1	AR_{10}
IoUTracker+ (Bochinski, Eiselein, and Sikora 2017)	23.6	39.2	25.5	26.2	30.9
OSMN (Yang et al. 2018)	27.5	45.1	29.1	28.6	33.1
DeepSORT (Wojke, Bewley, and Paulus 2017)	26.1	42.9	26.1	27.8	31.3
SeqTracker	27.5	45.7	28.7	29.7	32.5
MaskTrack R-CNN (Yang, Fan, and Xu 2019)	30.3	51.1	32.6	31.0	35.5
SipMask (Cao et al. 2020)	32.5	53.0	33.3	33.5	38.9
MaskTrack R-CNN + MSCOCO	32.2	52.1	34.6	31.8	37.2
CompFeat	**35.0**	**56.9**	**37.8**	**33.4**	**39.8**

Table 5: Comparison of the proposed approach with the state-of-the-arts on YouTube-VIS validation set. Note that all results in this Table including the post-processing. The best results are highlighted in bold.

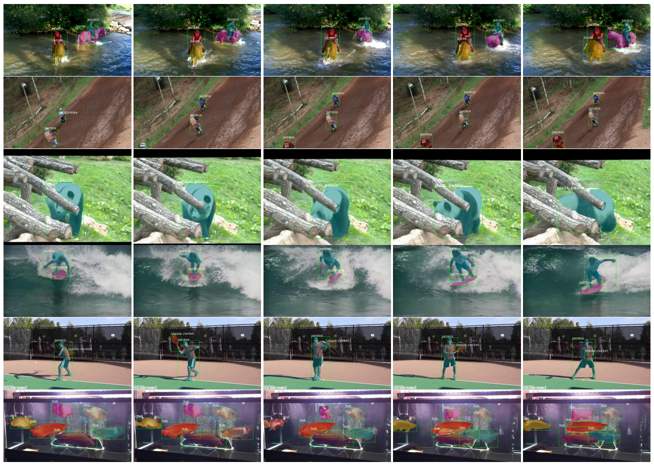


Figure 5: Visualization results of CompFeat. Each row has five sampled frames from a video sequence. Categories, bounding boxes and instance masks are shown for each object. Note objects with the same predicated identity across frames are marked with the same color. Zoom in to see details.

we evaluate the performance of video instance segmentation with all our proposed modules, *e.g.* frame level dual attention module, object level dual attention module and correlation maps. As shown in the last row in Table 3, we obtain the best performance. For instance, we achieve 27.9% and 46.8% on AP and $AP_{0.5}$ with all proposed modules, which surpasses the baseline model by 3.8% and 4.2%.

Sampling Method. The VIS performance under different sampling methods are shown in Table 4. Note that, for fairness, all experiment use the uniform sampling during testing. From Table 4, it can be observed that the random sampling is always powerful than the uniform one during training. The reason is that random sampling increases the variety of samples during training, which can lead our proposed ComFeat model discovers more temporal and spatial correspondences between current frames and support frames.

Comparison with the State-of-the-Arts

Mask-Track RCNN (Yang, Fan, and Xu 2019) is the first work on video instance segmentation. There are several work proposed in the Large-Scale Video Object Segmentation challenge (Luiten, Torr, and Leibe 2019; Wang et al. 2019; Dong et al. 2019), but it is hard to compare with them since they use different backbone networks and different external training data. We borrow experimental results of other approaches from (Yang, Fan, and Xu 2019). The comparison results are presented in Table 5. Mask-Track RCNN (Yang, Fan, and Xu 2019) is an online method which learns feature similarities for object matching. And SipMask (Cao et al. 2020) shares the similar structure while replace the instance segmentation branch with an one stage instance segmentation module. Compared with these methods, our proposed CompFeat achieves the best performance under all evaluation metrics. Compared with our baseline Mask-Track RCNN, the proposed CompFeat outperform it by a large margin. Note this performance gain is not from the additional training data since MaskTrack R-CNN with the same training data only achieves 32.2% and 52.1% on AP and $AP_{0.5}$.

Qualitative Results

Fig. 5 shows some qualitative results of our proposed CompFeat on YouTube-VIS validation set. Each row represents the predicted results on different frames in a video. The objects with the same identity are shown in the same color. As shown, CompFeat makes accurate predictions on object categories, bounding boxes, masks and identities under challenging conditions, *i.e.* multiple similar objects (row 1, 2), moderate occlusions (row 3), and drastic appearance changes (row 4). The last row shows a challenging case with six fish where our algorithm performs much better than MaskTrack-RCNN (Yang, Fan, and Xu 2019) although it misses a fish in the third image.

Conclusion

In this paper, we develop a comprehensive approach for feature aggregation for video instance segmentation, which is an underexplored direction in this area. Attention mechanisms are careful crafted for feature aggregations on both frame-level and object-level in both temporal and spatial manner. A new tracking module is designed to enhance local discriminative power of features with local and global correlation maps, in order to improve robustness of object tracking and re-identification. The effectiveness of the proposed modules is systematically evaluated with extensive experiments and ablation studies on the YouTube-VIS dataset.

References

Bochinski, E.; Eiselein, V.; and Sikora, T. 2017. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE.

Cao, J.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; Pang, Y.; and Shao, L. 2020. SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation. *Proc. European Conference on Computer Vision* .

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. GC-Net: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. *arXiv preprint arXiv:1904.11492* .

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* .

Chen, K.; Wang, J.; Yang, S.; Zhang, X.; Xiong, Y.; Change Loy, C.; and Lin, D. 2018. Optimizing video object detection via a scale-time lattice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7814–7823.

Dong, M.; Wang, J.; Huang, Y.; Yu, D.; Su, K.; Zhou, K.; Shao, J.; Wen, S.; and Wang, C. 2019. Temporal Feature Augmented Network for Video Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2017. Detect to track and track to detect. In *IEEE ICCV*.

Han, W.; Khorrami, P.; Paine, T. L.; Ramachandran, P.; Babaeizadeh, M.; Shi, H.; Li, J.; Yan, S.; and Huang, T. S. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* .

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2014. Simultaneous detection and segmentation. In *ECCV*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *IEEE ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE CVPR*, 770–778.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE CVPR*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *IEEE CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, M.; Zhu, M.; White, M.; Li, Y.; and Kalenichenko, D. 2019. Looking Fast and Slow: Memory-Guided Mobile Video Object Detection. *arXiv preprint arXiv:1903.10172* .

Luiten, J.; Torr, P.; and Leibe, B. 2019. Video Instance Segmentation 2019: A winning approach for combined Detection, Segmentation, Classification and Tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. *arXiv preprint arXiv:1904.00607* .

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* .

Sadeghian, A.; Alahi, A.; and Savarese, S. 2017. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *IEEE ICCV*.

Shi, H. 2018. Geometry-aware traffic flow analysis by detection and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 116–120.

Son, J.; Baek, M.; Cho, M.; and Han, B. 2017. Multi-object tracking with quadruplet convolutional neural networks. In *IEEE CVPR*.

Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; and Torr, P. H. 2017. End-to-end representation learning for correlation filter based tracking. In *IEEE CVPR*.

Voigtlaender, P.; Chai, Y.; Schroff, F.; Adam, H.; Leibe, B.; and Chen, L.-C. 2019a. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *IEEE CVPR*.

Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B. B. G.; Geiger, A.; and Leibe, B. 2019b. MOTS: Multi-object tracking and segmentation. In *IEEE CVPR*.

Voigtlaender, P.; and Leibe, B. 2017. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364* .

Wang, Q.; He, Y.; Yang, X.; Yang, Z.; and Torr, P. 2019. An Empirical Study of Detection-Based Video Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; and Maybank, S. 2018a. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *IEEE ICCV*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *IEEE CVPR*.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*.

Wu, C.-Y.; Feichtenhofer, C.; Fan, H.; He, K.; Krahenbuhl, P.; and Girshick, R. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 284–293.

Wug Oh, S.; Lee, J.-Y.; Sunkavalli, K.; and Joo Kim, S. 2018. Fast video object segmentation by reference-guided mask propagation. In *IEEE CVPR*.

Xu, K.; Wen, L.; Li, G.; Bo, L.; and Huang, Q. 2019. Spatiotemporal CNN for Video Object Segmentation. In *IEEE CVPR*.

Xu, N.; Yang, L.; Fan, Y.; Yue, D.; Liang, Y.; Yang, J.; and Huang, T. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327* .

Yang, L.; Fan, Y.; and Xu, N. 2019. Video Instance Segmentation. In *IEEE ICCV*.

Yang, L.; Wang, Y.; Xiong, X.; Yang, J.; and Katsaggelos, A. K. 2018. Efficient video object segmentation via network modulation. In *IEEE CVPR*.

Zhang, Z.; and Peng, H. 2019. Deeper and wider siamese networks for real-time visual tracking. In *IEEE CVPR*.

Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Flow-guided feature aggregation for video object detection. In *IEEE ICCV*.

Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; and Hu, W. 2018. Distractor-aware Siamese Networks for Visual Object Tracking. In *ECCV*.