# CONFIDENCE-GUIDED ADAPTIVE GATE AND DUAL DIFFERENTIAL ENHANCEMENT FOR VIDEO SALIENT OBJECT DETECTION

*Peijia Chen[1], Jianhuang Lai[1,2,3*], Guangcong Wang[1], Huajun Zhou[1]*

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
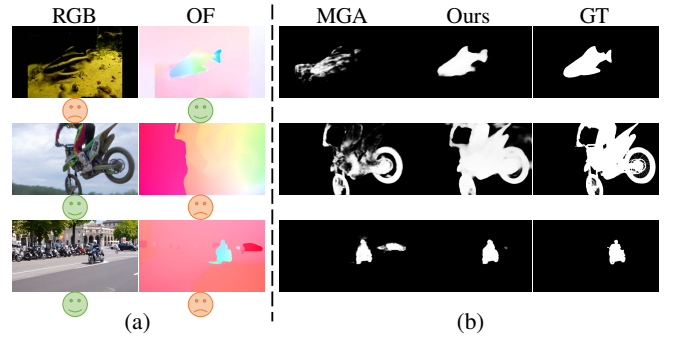chenpj8@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn

## ABSTRACT

Video salient object detection (VSOD) aims to locate and segment the most attractive object by exploiting both spatial cues and temporal cues hidden in video sequences. However, spatial and temporal cues are often unreliable in real-world scenarios, such as low-contrast foreground, fast motion, and multiple moving objects. To address these problems, we propose a new framework to adaptively capture available information from spatial and temporal cues, which contains Confidence-guided Adaptive Gate (CAG) modules and Dual Differential Enhancement (DDE) modules. For both RGB features and optical flow features, CAG estimates confidence scores supervised by the IoU between predictions and the ground truths to re-calibrate the information with a gate mechanism. DDE captures the differential feature representation to enrich the spatial and temporal information and generate the fused features. Experimental results on four widely used datasets demonstrate the effectiveness of the proposed method against thirteen state-of-the-art methods.

***Index Terms***— Confidence estimation, gate mechanism, differential feature, video salient object detection

## 1. INTRODUCTION

Video salient object detection (VSOD) aims to locate and segment the most attractive object in video sequences, which is widely used as an important pre-processing step to reduce computational burdens for many high-level computer vision tasks, such as video compression [1], video captioning [2] and person re-identification [3].

Different from salient object detection (SOD) that focuses on predicting saliency maps by exploiting spatial information extracted from one single image, VSOD further exploits temporal information hidden in video sequences. Although VSOD exploits more information to predict saliency maps, it
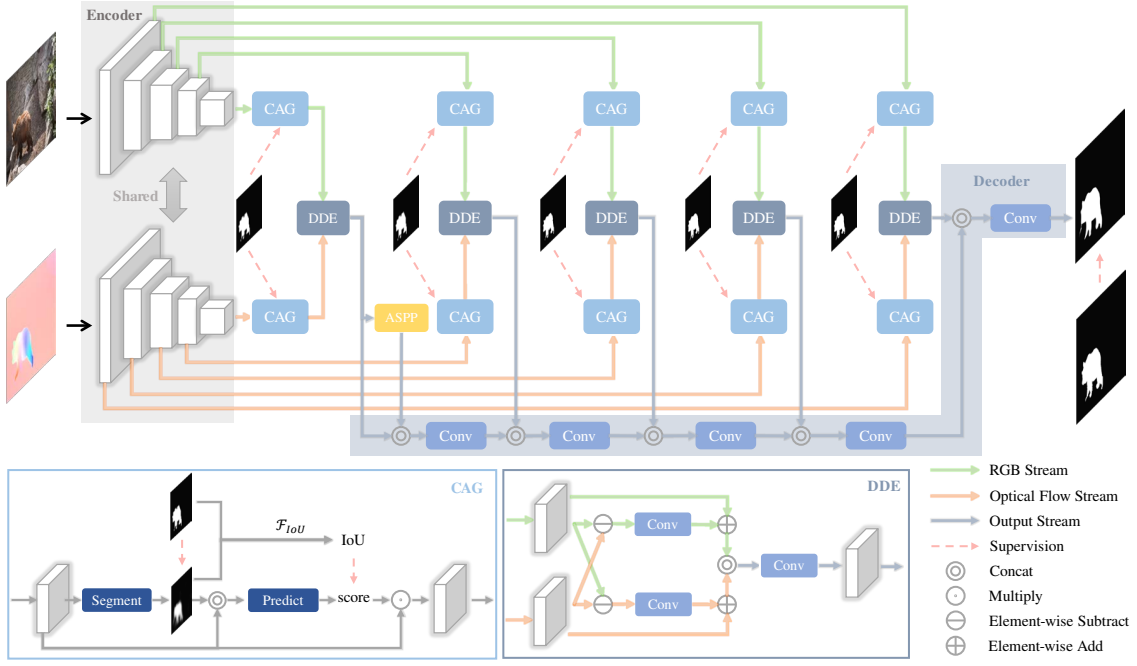
**Fig. 1**. The challenges for video salient object detection in scenes of low-contrast foreground (the first row), fast motion (the second row) and multiple moving objects (the third row). OF denotes the optical flow image and GT denotes the ground truth.

also brings extra challenges, as shown in Figure 1. **First**, spatial cues hidden in every single frame are often hard to be exploited when foreground and background share a similar feature representation. The first row shows that the RGB images of low contrast between the salient objects and backgrounds would bring in misleading information to predict background objects wrongly. **Second**, temporal cues hidden between different frames could be disturbed by fast motion, large displacements and illumination changes. The second row shows that the motocross is distinctive from the background in the RGB image but noisy in the optical flow image, which leads to the absence of the rider in the saliency map predicted by MGA. The third row shows that even the temporal information from the accurate optical flow images would confuse the spatial information in scenes of several moving objects. Driven by these two challenges, one would ask: how can we establish a model to automatically capture available spatial and temporal cues while suppress noisy ones?

The existing methods may offer a partial solution to this problem, which can be roughly classified into four groups, i.e., hand-crafted VSOD, long-term memory VSOD, attention

**Fig. 2**. The overall architecture of the proposed model.

VSOD, and parallel VSOD. **First**, hand-crafted VSOD methods try to combine the spatial information with motion cues based on the prior knowledge, such as the spatio-temporal background prior [4] and low-rank coherency [5], which yield poor performance limited by the hand-crafted low-level features. **Second**, long-term memory VSOD methods [6, 7, 8] extract the spatial information from the single image separately and model the temporal information through convolutional memory units such as ConvLSTM. **Third**, attention VSOD methods use a non-local mechanism to capture temporal information across several images [9]. The second and third groups are cascaded models that extract features for each frame and then model the dependence relationship between video sequences based on the extracted features. These "first spatial and then temporal" cascaded models cannot adaptively capture available cues from both spatial and temporal cues to cooperatively predict saliency maps. **Fourth**, instead of using cascaded ways, parallel VSOD methods [10] often adopt a two-stream framework where one stream is to extract features from every single frame and another stream is to process temporal information independently by using optical flow images. However, the existing parallel VSOD methods simply fuse spatial and temporal information without considering the confidence of the spatial and temporal information.

To address the above issues, we propose a new framework that adaptively captures available spatial and temporal information to predict completely saliency maps. Specifically, we introduce a Confidence-guided Adaptive Gate (CAG) module that can estimate the confidence score of the input features by

calculating IoU value. With the confidence score, CAG passes available information and suppresses the noises from spatial and temporal cues. We also propose a Dual Differential Enhancement (DDE) module that focuses on capturing differential information between spatial and temporal cues and generates the fused features.

Overall, our main contributions are summarized as follows:

- We propose a new framework to accurately predict the saliency maps for video salient object detection, which adaptively captures available information from spatial and temporal cues.

- We propose a Confidence-guided Adaptive Gate (CAG) module to suppress low-confidence information and a Dual Differential Enhancement (DDE) module to exploit discriminative features enhanced by the differential information from the spatial and temporal dual branch.

- Experiments on four widely used datasets demonstrate the superiority of our proposed method against state-of-the-art methods.

## 2. PROPOSED METHOD

In this section, we propose a Confidence-guided Adaptive Gate and Dual Differential Enhancement (CAG-DDE) method for video salient object detection, which passes the

available information while suppresses the unreliable information from spatial and temporal cues and enhances them with differential information to completely segment salient objects. As shown in Figure 2, the overall framework of the proposed method is built on the encoder-decoder architecture, which includes a shared encoder, an output decoder, a number of Confidence-guided Adaptive Gate (CAG) modules and Dual Differential Enhancement (DDE) modules. Given a pair of RGB and optical flow images (visualization of optical flow), the shared encoder outputs the RGB features and optical flow features at five different layers. Let $E_i^{RGB}$ and $E_i^{OF}$ denote RGB features and optical flow features, respectively. Here, $i$ indicates the $i$-th layer, $1 \leq i \leq 5$. Both $E_i^{RGB}$ and $E_i^{OF}$ are re-calibrated with their confidence scores estimated by the CAG modules, and then merged by the DDE modules to obtain the decoder features $D_i$, which can be formulated as

$$D_i = \mathcal{F}_{DDE}(\mathcal{F}_{CAG}(E_i^{RGB}), \mathcal{F}_{CAG}(E_i^{OF})), \quad (1)$$

where $\mathcal{F}_{CAG}$ and $\mathcal{F}_{DDE}$ denote a CAG operator and a DDE operator. Next, five decoder features $\{D_i\}_{i=1}^5$ at different layers are gradually combined by the output decoder. Following [10], we integrate an atrous spatial pyramid pooling (ASPP) module after the DDE module at the fifth layer. The output decoder is formulated as

$$D_i' = \begin{cases} \mathcal{F}_{conv}(D_i \odot \mathcal{F}_{upsample}(D_{i+1}')) & i = 1,2,3,4 \\ \mathcal{F}_{conv}(D_5 \odot \mathcal{F}_{ASPP}(D_5)) & i = 5 \end{cases} \quad (2)$$

where $\odot$ is a concatenation operator. $\mathcal{F}_{conv}$ and $\mathcal{F}_{upsample}$ are the operators of convolution and upsampling. Finally, we upsample the final decoder features as the predicted saliency map, which can be formulated as

$$P_f = \mathcal{F}_{upsample}(D_1'). \quad (3)$$

### 2.1. Confidence-guided Adaptive Gate Module

In real-world scenarios, both the spatial and temporal images contain unreliable information inevitably. How to measure the reliability of the informative features and noisy features is a key to the VSOD problem. To address the problem, we propose a Confidence-guided Adaptive Gate (CAG) module, which predicts the confidence score to represent the reliability of the features and re-calibrate the features, as shown in the bottom left of Figure 2.

Our CAG module is composed of two sub-networks, the segmentation sub-network and confidence score prediction sub-network. The segmentation sub-network consists of three convolution layers, which is used to predict the saliency map $P_i$ supervised by the ground truth $G_i$. The confidence score prediction sub-network consists of three convolution layers and one global average pooling layer, which aims to explicitly model the confidence score. We concatenate the predicted saliency map and the input features as the input of the confidence score prediction sub-network. Inspired by a segmentation quality metric Intersection over Union (IoU), we quantify

the confidence score as the IoU between the saliency map $P_i$ and the ground truth $G_i$. Given the input feature $E_i$, we can obtain the confidence score by

$$s_i = \sigma(\mathcal{F}_{pred}(\mathcal{F}_{seg}(E_i) \odot E_i)), \quad (4)$$

where $\sigma$ is a sigmoid function that scales the confidence score to $(0,1)$. $\mathcal{F}_{seg}$ and $\mathcal{F}_{pred}$ refer to the segmentation sub-network and confidence score prediction sub-network, respectively.

Under the guidance of the predicted confidence scores, our method adaptively re-calibrates the features based on a gate mechanism, which is given by

$$\mathcal{F}_{CAG}(E_i) = E_i \cdot s_i(E_i), \quad (5)$$

where $\cdot$ denotes the scalar multiplication. For each layer of the encoder, we employ a segmentation sub-network and a confidence score prediction sub-network to estimate the confidence scores of features at different levels.

### 2.2. Dual Differential Enhancement Module

Color saliency obtained by the RGB feature and motion saliency obtained by the optical flow feature are strongly complementary with each other. However, most complementary information is hidden in the difference between RGB and optical flow features. To make full use of their complementarity, we propose a Dual Differential Enhancement (DDE) module to discover differential information between the RGB and optical flow features, as shown at the bottom right of Figure 2. For each branch, we extract the differential information by subtracting the shared information and try to focus on learning branch-specific (spatial or motion) information. We then enhance the original information by adding the differential information. After the dual differential enhancement between the RGB and optical flow features, the merged feature is computed by

$$\mathcal{F}_{DDE} = \mathcal{F}_{conv}(\mathcal{F}_e(R_i^{RGB}, R_i^{OF}) \odot \mathcal{F}_e(R_i^{OF}, R_i^{RGB})), \quad (6)$$
$$\mathcal{F}_e(X,Y) = \mathcal{F}_{conv}(Y - X) + X \quad (7)$$

where $R_i^{RGB}$ and $R_i^{OF}$ denote the re-calibrated RGB and optical flow features, respectively. $\mathcal{F}_{DDE}$ refers to an operator of dual differential enhancement module and $\mathcal{F}_e$ refers to an operator of differential enhancement for a single branch.

### 2.3. Loss Function

Following [11], an effective loss function is adopted, which is given by

$$L_f = L_{BCE}(P_f, G) + L_{SSIM}(P_f, G) + L_{IoU}(P_f, G), \quad (8)$$

where $L_{BCE}$, $L_{SSIM}$, $L_{IoU}$ are binary cross entropy loss, structural similarity index measure loss and intersection over
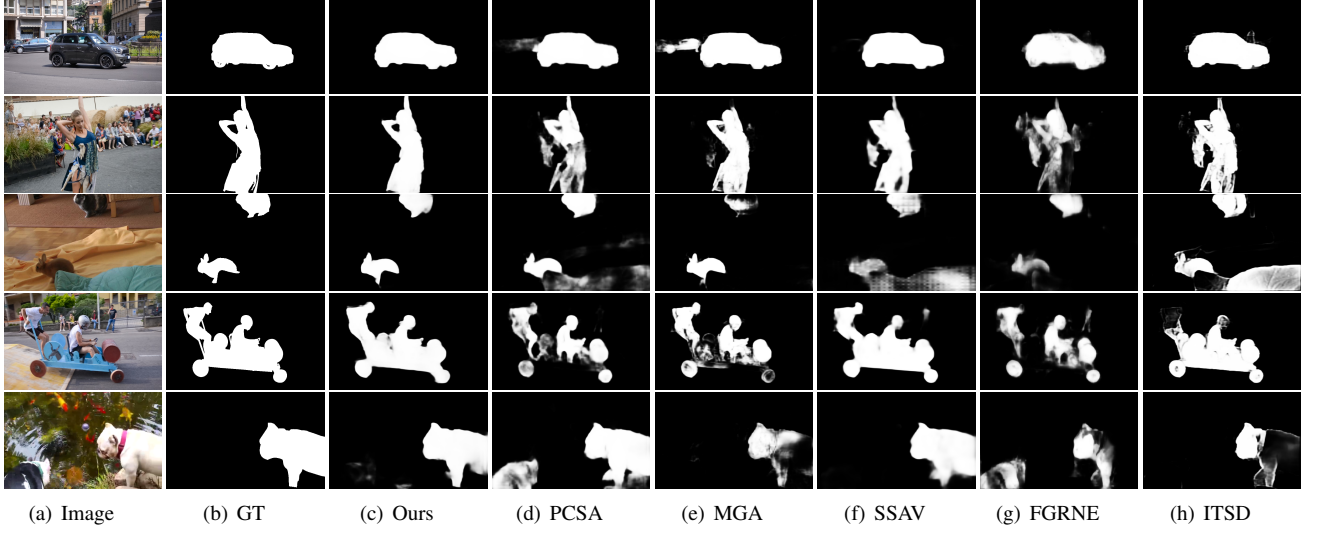
| (a) Image | (b) GT | (c) Ours | (d) PCSA | (e) MGA | (f) SSAV | (g) FGRNE | (h) ITSD |

**Fig. 3**. Qualitative comparison of our method and the state-of-the-art methods.

union loss, respectively. $P_f$ and $G$ denote the predicted saliency map and ground truth, respectively. These three losses are given by

$$L_{BCE} = -(GlogP + (1 - G)log(1 - P)), \qquad (9)$$

$$L_{SSIM} = 1 - \frac{(2\mu_P\mu_G + C_1)(2\sigma_{PG} + C_2)}{(\mu_P^2 + \mu_G^2 + C_1)(\sigma_P^2 + \sigma_G^2 + C_2)}, \quad (10)$$

$$L_{IoU} = 1 - \frac{TP}{TN + TP + FP}, \qquad (11)$$

where $P$ denotes the predicted saliency map. $\mu_P$ and $\sigma_P$ ($\mu_G$ and $\sigma_G$) are the mean and standard deviations of the predicted saliency map (ground truth), respectively. $\sigma_{PG}$ is the covariance of $P$ and $G$. TP, TN and FP represent true-positive, true-negative and false-positive, respectively.

For the CAG module at the i-th layer, we adopt the binary cross entropy loss to train the segmentation sub-network and $l_1$ loss to regress the confidence score prediction sub-network. Therefore, the loss function of CAG can be formulated as

$$L_i^t = L_{BCE}(P_i^t, G_i) + L_{l_1}(s_i^t, \mathcal{F}_{IoU}(P_i^t, G_i)), \qquad (12)$$

where $\mathcal{F}_{IoU}$ is a function to calculate the IoU. $t$ indicates the type of features, including $RGB$ for RGB features and $OF$ for optical flow features. $G_i$ is downsampled from the ground truth, which has the consistent size as the predicted saliency map $P_i$. Finally, the total loss function can be written as

$$L_{total} = L_f + \sum_t \sum_{i=1}^{N} L_i^t, \qquad (13)$$

where $N$ is 5.

## 3. EXPERIMENTS

### 3.1. Experimental Setup

**Datasets:** To evaluate the effectiveness of our method, we conduct experiments on four widely used public datasets, including SegV2 [17], FBMS [18], DAVIS [19] and DAVSOD [8] datasets.

**Evaluation Metrics:** To quantitatively evaluate the performance of VSOD, we adopt three metrics in our experiments, i.e., F-measure [20], S-measure [21] and Mean Absolute Error (MAE) [22]. F-measure is defined as

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \qquad (14)$$

where $\beta^2$ is set to 0.3 and we report the maximum F-measure for evaluation. S-measure takes both region-aware and object-aware structural similarity into consideration. MAE measures the pixel-level average absolute difference between the predicted map and ground truth.

**Implementation Details:** We implement our method on Pytorch. We use the pre-trained ResNet-101 [23] as our initial backbone. We employ RAFT [24] to render optical flow images. Following the previous methods [10, 9], we remove the CAG modules and the DDE modules, and pre-train our model with the training set of DUTS [25]. After pre-training, we use the training set of DAVIS and DAVSOD to train the whole network. We resize the input images to $448 \times 448$. We apply random horizontal flipping and scaling the input images with scales {0.75, 1, 1.25}. We use the Adam optimizer to train our model with an initial learning rate of 1e-5 with batch size of 4.

4

**Table 1**. Quantitative comparison with the state-of-the-art static and video salient object detection methods by three evaluation metrics. The best three results are highlighted in red, green, and blue respectively.

| Method | Year | DAVSOD | | | DAVIS | | | SegV2 | | | FBMS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | max $F_\beta$ ↑ | S ↑ | MAE ↓ | max $F_\beta$ ↑ | S ↑ | MAE ↓ | max $F_\beta$ ↑ | S ↑ | MAE ↓ | max $F_\beta$ ↑ | S ↑ | MAE ↓ |
| Static Salient Object Detection | | | | | | | | | | | | | |
| EGNet[12] | 2019 | 0.604 | 0.719 | 0.101 | 0.768 | 0.829 | 0.056 | 0.774 | 0.845 | 0.024 | 0.848 | 0.878 | 0.044 |
| CPD[13] | 2019 | 0.613 | 0.723 | 0.092 | 0.810 | 0.863 | 0.031 | 0.778 | 0.841 | 0.023 | 0.841 | 0.867 | 0.048 |
| BASNet[11] | 2019 | 0.597 | 0.708 | 0.109 | 0.812 | 0.858 | 0.031 | 0.774 | 0.838 | 0.023 | 0.822 | 0.858 | 0.047 |
| ITSD[14] | 2020 | 0.651 | 0.747 | 0.094 | 0.835 | 0.876 | 0.033 | 0.807 | 0.787 | 0.027 | 0.843 | 0.869 | 0.040 |
| Video Salient Object Detection | | | | | | | | | | | | | |
| STBP[4] | 2016 | 0.410 | 0.568 | 0.160 | 0.544 | 0.677 | 0.096 | 0.640 | 0.735 | 0.061 | 0.595 | 0.627 | 0.152 |
| SFLR[5] | 2017 | 0.478 | 0.624 | 0.132 | 0.727 | 0.790 | 0.056 | 0.745 | 0.804 | 0.037 | 0.660 | 0.699 | 0.117 |
| SCOM[15] | 2018 | 0.464 | 0.599 | 0.220 | 0.783 | 0.832 | 0.048 | 0.764 | 0.815 | 0.030 | 0.797 | 0.794 | 0.079 |
| SCNN[16] | 2018 | 0.532 | 0.674 | 0.128 | 0.714 | 0.783 | 0.064 | - | - | - | 0.762 | 0.794 | 0.095 |
| FGRNE[6] | 2018 | 0.573 | 0.693 | 0.098 | 0.783 | 0.838 | 0.043 | - | - | - | 0.767 | 0.809 | 0.088 |
| PDBM[7] | 2018 | 0.573 | 0.698 | 0.116 | 0.855 | 0.882 | 0.028 | 0.800 | 0.864 | 0.024 | 0.821 | 0.851 | 0.064 |
| SSAV[8] | 2019 | 0.603 | 0.724 | 0.092 | 0.861 | 0.893 | 0.028 | 0.801 | 0.851 | 0.023 | 0.865 | 0.879 | 0.040 |
| MGA[10] | 2019 | 0.655 | 0.751 | 0.081 | 0.892 | 0.912 | 0.022 | - | - | - | 0.906 | 0.910 | 0.026 |
| PCSA[9] | 2020 | 0.656 | 0.741 | 0.086 | 0.878 | 0.901 | 0.022 | 0.811 | 0.866 | 0.024 | 0.831 | 0.864 | 0.040 |
| Ours | - | 0.670 | 0.762 | 0.072 | 0.898 | 0.906 | 0.018 | 0.826 | 0.865 | 0.027 | 0.858 | 0.870 | 0.039 |

**Table 2**. Ablation studies of the proposed network architecture on DAVIS and DAVSOD datasets.

| Methods | DAVIS | | DAVSOD | |
|---|---|---|---|---|
| | max $F_\beta$ ↑ | MAE ↓ | max $F_\beta$ ↑ | MAE ↓ |
| Baseline (Concat) | 0.869 | 0.023 | 0.645 | 0.080 |
| Baseline + CAG | 0.887 | 0.022 | 0.662 | 0.072 |
| Baseline + DDE | 0.892 | 0.020 | 0.656 | 0.074 |
| Ours | 0.898 | 0.018 | 0.670 | 0.072 |

## 3.2. Comparison with State-of-the-art Methods

We compare our method against four state-of-the-art static salient object detection methods including EGNet [12], CPD [13], BASNet [11], ITSD [14]. We also compare our method against nine state-of-the-art video salient object detection methods including STBP [4], SFLR [5], SCOM [15], SCNN [16], FGRNE [6], PDBM [7], SSAV [8], MGA [10], PCSA [9]. In our experiments, we employ the evaluation code [8] to evaluate all the saliency maps for a fair comparison.

**Quantitative Evaluation.** We first conduct a quantitative evaluation, as shown in Table 1. It is observed that our method achieves state-of-the-art results on three datasets. Specifically, our method outperforms other state-of-the-art methods under the metric max $F_\beta$ on DAVIS, SegV2 and DAVSOD datasets, and achieves the third-best performance on FBMS dataset. Note that the leading performance of MGA on FBMS dataset is mainly induced by using the FBMS training set. Notably, our method achieves significant performance improvement on the most challenging dataset DAVSOD compared with the second-best method MGA (0.670 V.S. 0.655, 0.762 V.S. 0.751, 0.072 V.S. 0.081 in terms of max $F_\beta$, S and MAE respectively), which demonstrates the superior performance of our method in complex scenes.

**Qualitative Evaluation.** We conduct a qualitative evaluation in different scenes, as shown in Figure 3. The results show that our method can accurately locate and segment the salient objects in several complex scenes, such as multiple moving objects (the first row), cluttered foreground and background (the second and third rows), complex boundary (the fourth row) and saliency shifts (the last row).

## 3.3. Ablation Studies

The proposed network is composed of two main modules, the Confidence-guided Adaptive Gate (CAG) module and the Dual Differential Enhancement (DDE) module. To verify the effectiveness of each component, we conduct an ablation experiment on two large-scale datasets, i.e., DAVIS and DAVSOD. The experimental results are shown in Table 2. The first row is the baseline model that merges two features with the concatenation operation. The second and third rows show that both CAG and DDE can boost performance compared with the baseline. Moreover, the combination of the CAG and DDE modules can further improve the performance.

**Table 3**. Comparison with different fusion mechanisms on the DAVIS and DAVSOD datasets.

| Methods | DAVIS | | DAVSOD | |
|---------|-------|---|--------|---|
| | max $F_\beta$ ↑ | MAE ↓ | max $F_\beta$ ↑ | MAE ↓ |
| Cat | 0.869 | 0.023 | 0.645 | 0.080 |
| Add | 0.865 | 0.024 | 0.656 | 0.082 |
| Mul | 0.870 | 0.023 | 0.625 | 0.087 |
| Ours | 0.898 | 0.018 | 0.670 | 0.072 |

## 4. CONCLUSIONS

In this paper, we propose a new video salient object detection method to exploit available spatial and temporal information to predict robust saliency maps. We first propose a Confidence-guided Adaptive Gate (CAG) module to filters unreliable cues from spatial and temporal information. We then propose a Dual Differential Enhancement (DDE) module, which merges spatial and temporal information enhanced by differential features to capture complementary information. Experimental results on four widely used datasets demonstrate the effectiveness of our proposed method.

## 5. REFERENCES

[1] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *TIP*, vol. 13, no. 10, pp. 1304–1318, 2004.

[2] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *CVPR*, 2017, pp. 6504–6512.

[3] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[4] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *TIP*, vol. 26, no. 7, pp. 3425–3436, 2016.

[5] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *TIP*, vol. 26, no. 7, pp. 3156–3170, 2017.

[6] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *CVPR*, 2018, pp. 3243–3252.

[7] H. Song, W. Wang, S. Zhao, J. Shen, and K. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *ECCV*, 2018, pp. 715–731.

[8] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *CVPR*, 2019, pp. 8554–8564.

[9] Y. Gu, L. Wang, Z. Wang, Y. Liu, and S. Lu, "Pyramid constrained self-attention network for fast video salient object detection," *AAAI*, vol. 34, no. 7, pp. 10869–10876, 2020.

[10] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *ICCV*, 2019, pp. 7274–7283.

[11] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *CVPR*, 2019, pp. 7479–7489.

[12] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng, "Egnet: Edge guidance network for salient object detection," in *ICCV*, 2019, pp. 8779–8788.

[13] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *CVPR*, 2019, pp. 3907–3916.

[14] H. Zhou, X. Xie, J. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *CVPR*, 2020, pp. 9141–9150.

[15] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, "Scom: Spatiotemporal constrained optimization for salient object detection," *TIP*, vol. 27, no. 7, pp. 3345–3357, 2018.

[16] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li, "Weakly supervised salient object detection with spatiotemporal cascade neural networks," *TCSVT*, vol. 29, no. 7, pp. 1973–1984, 2018.

[17] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, 2013, pp. 2192–2199.

[18] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *TPAMI*, vol. 36, no. 6, pp. 1187–1200, 2013.

[19] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016, pp. 724–732.

[20] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*. IEEE, 2009, pp. 1597–1604.

[21] D. Fan, M. Cheng, Y Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.

[22] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*. IEEE, 2012, pp. 733–740.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*, 2020, pp. 402–419.

[25] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017, pp. 136–145.