Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection

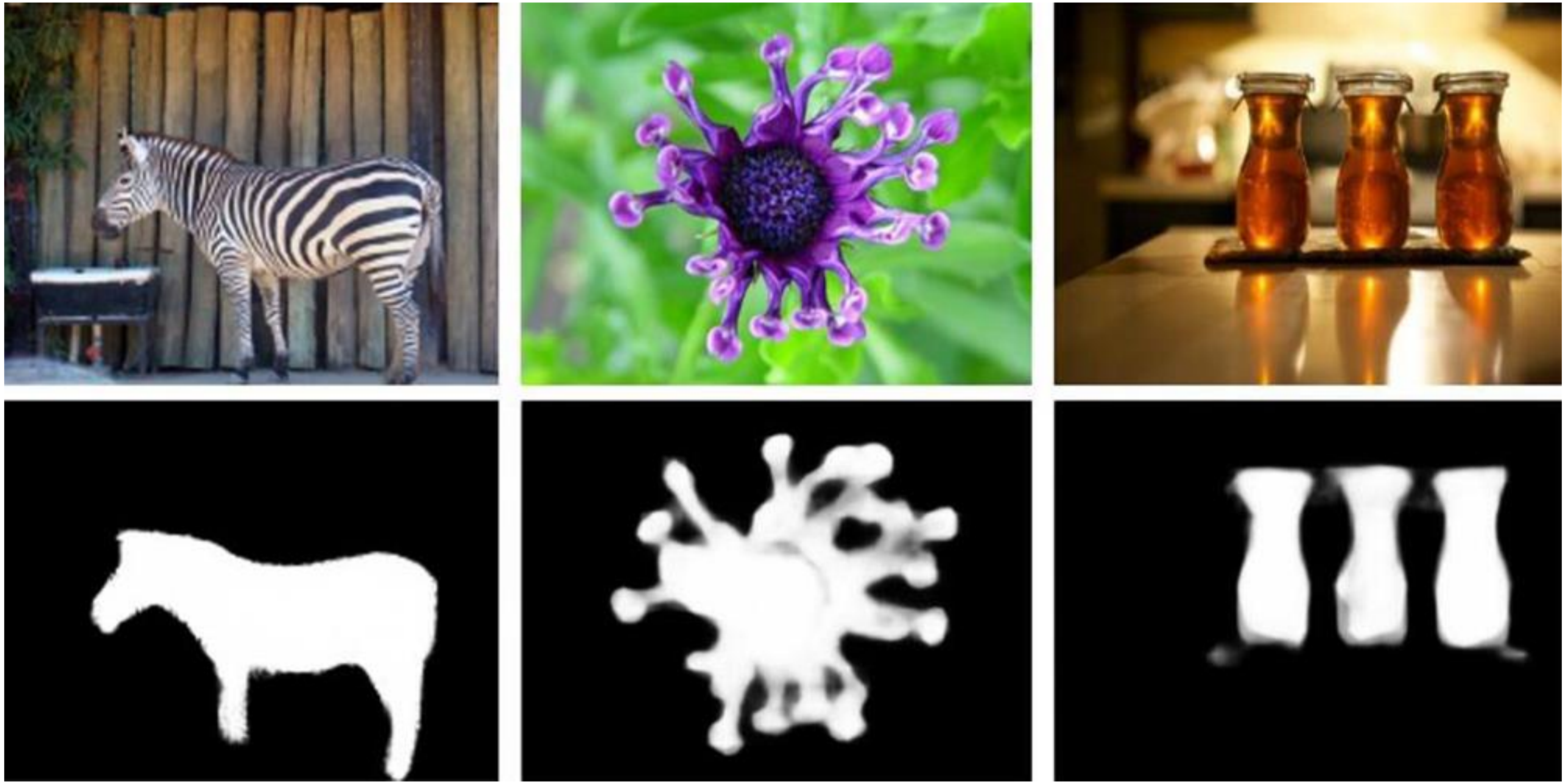
ECCV2018

2019.03.28

발표자 박성현

1 Introduction

Salient Object Detection



1 Introduction

ConvLSTM

기존의 LSTM에 Convolution을 적용한 모델

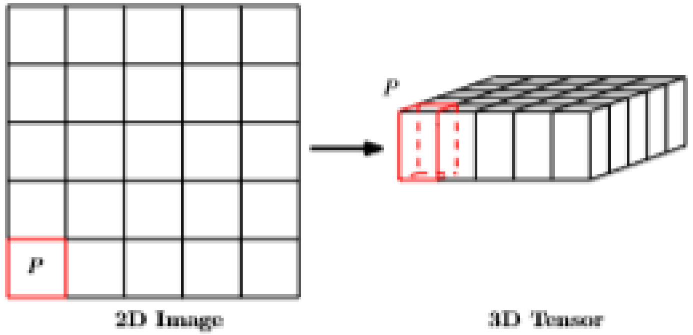


Figure 1: Transforming 2D image into 3D tensor

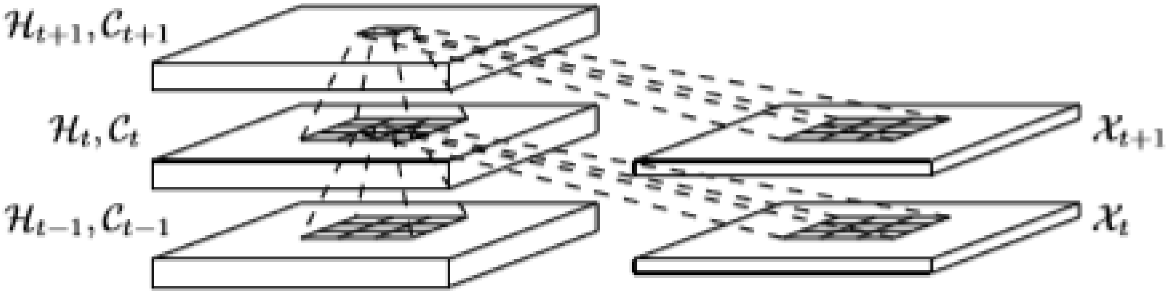
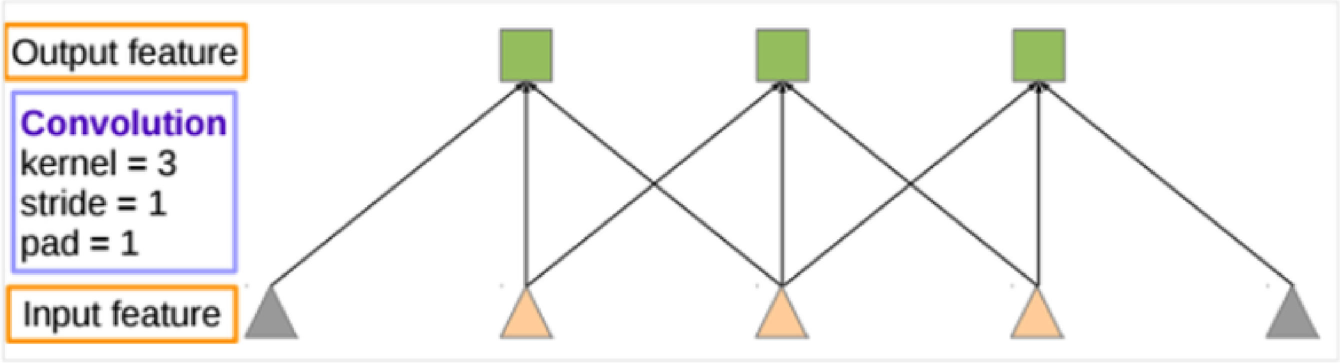

$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_t, \mathcal{C}_t$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_t$

Figure 2: Inner structure of ConvLSTM

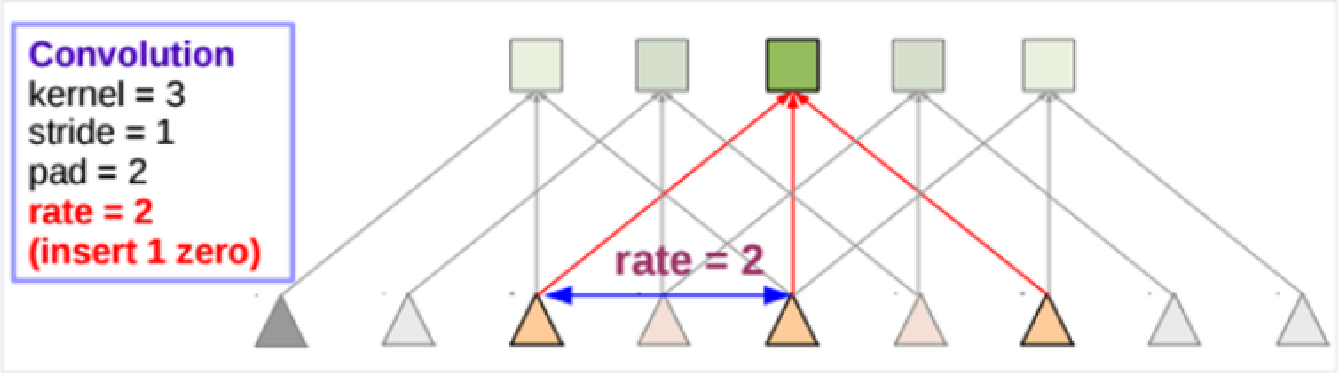
$$
\begin{aligned}
i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ \mathcal{C}_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ \mathcal{C}_{t-1} + b_f) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ \mathcal{C}_t + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$

2 Model

Atrous Convolution (=Dilated Convolution)



(a) Sparse feature extraction

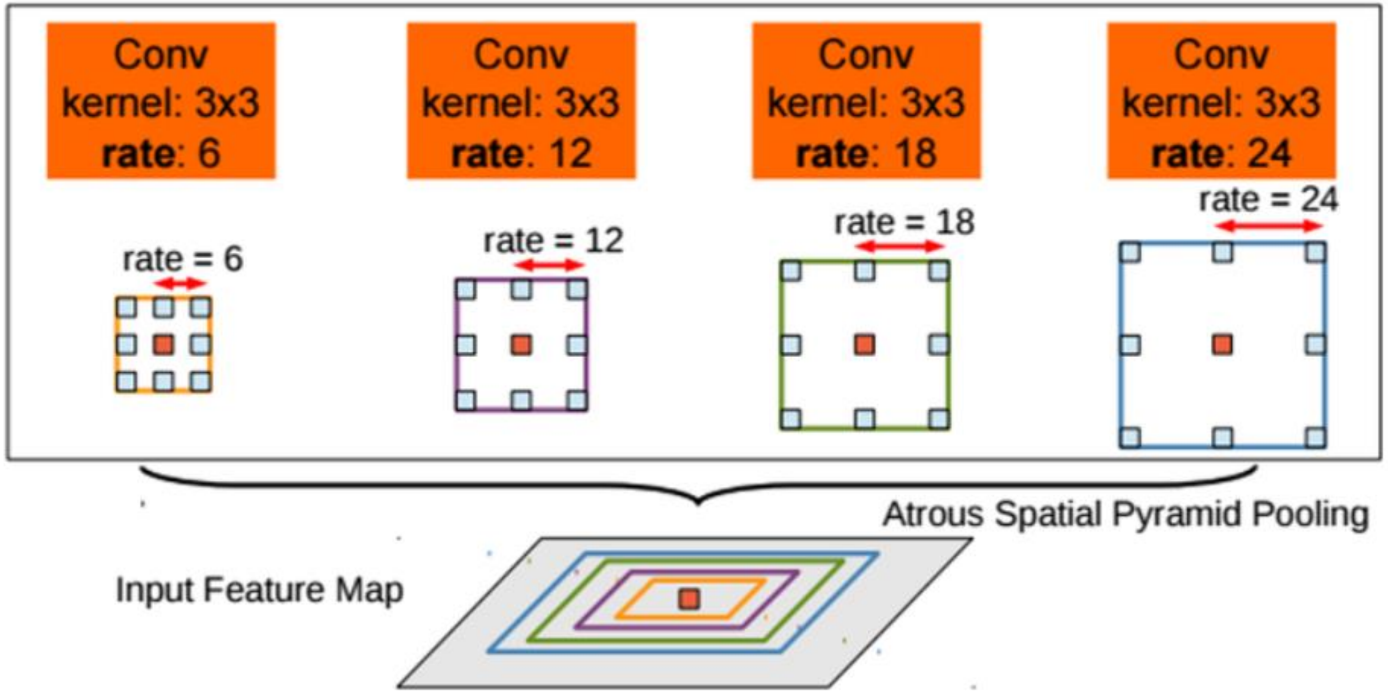


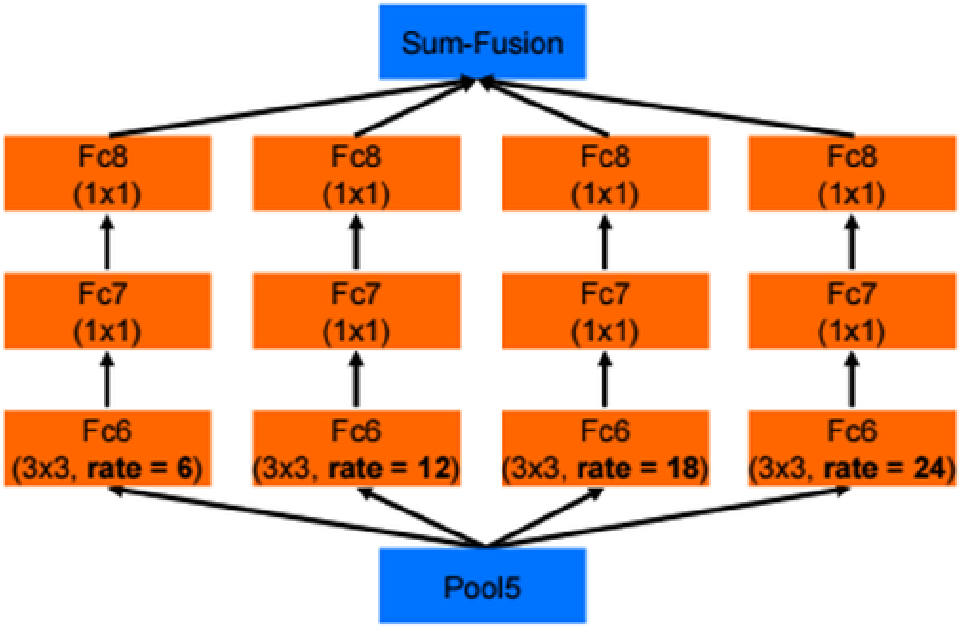
(b) Dense feature extraction

Atrous Convolution(Dilated Convolution)을 사용하면 kernel 크기는 동일하게 유지되어 연산량은 동일하지만, Receptive Field의 크기가 커지는 효과를 얻을 수 있음.
→ Segmentation or Detection 분야에서 자주 사용됨.

2 Model

Atrous Spatial Pyramid Pooling





(b) DeepLab-ASPP

2

Model

Pyramid Dilated Convolution (PDC) module

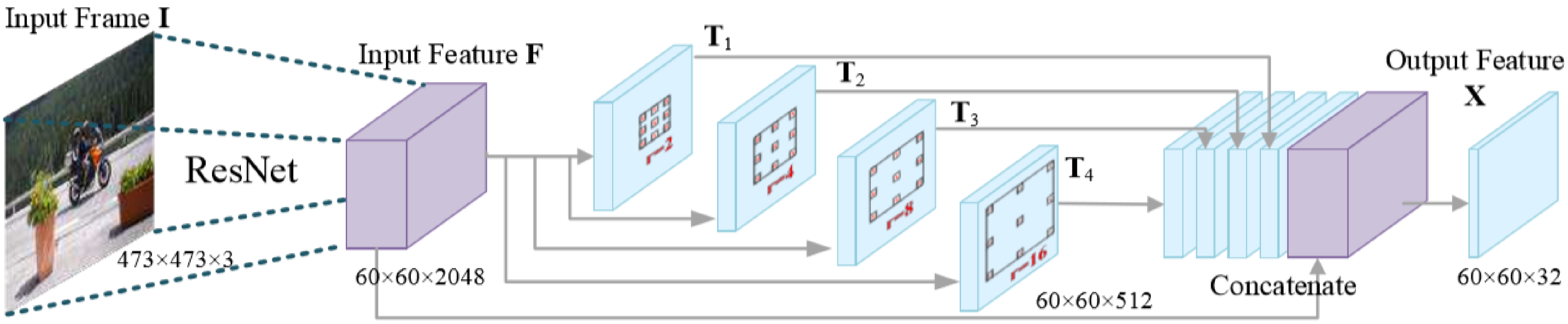


Fig. 2. Illustration of PDC module, where features from 4 parallel dilated convolution branches with different dilated rates are concatenated with the input features for emphasizing multi-scale spatial feature learning. See § 3.1 for details.

2 Model

Dilated Bidirectional(DB) ConvLSTM module

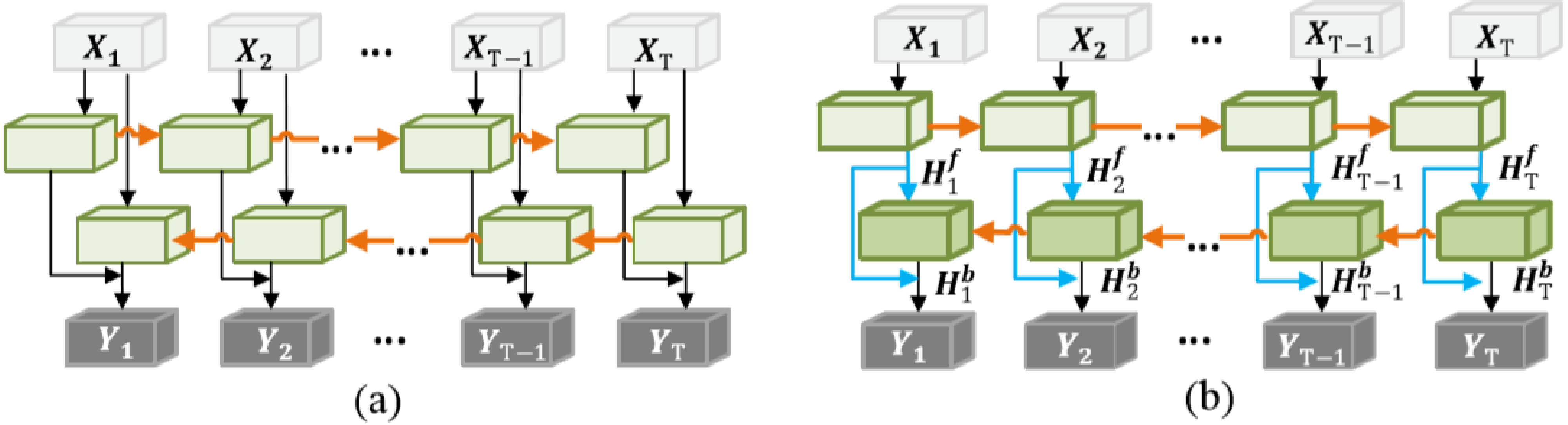
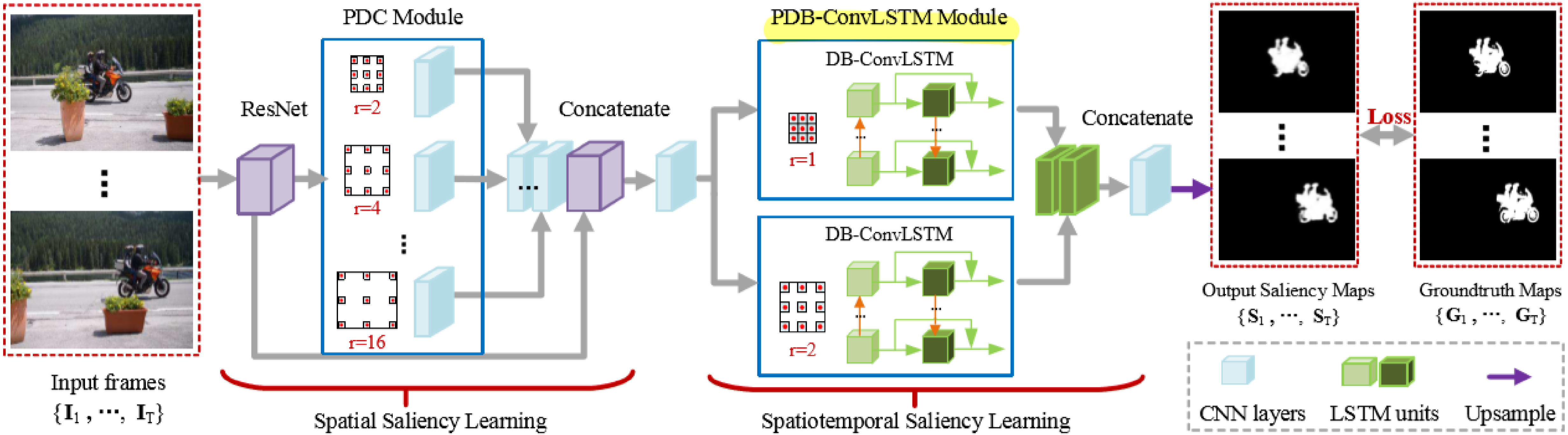


Fig. 3. Illustration of (a) Bidirectional ConvLSTM and (b) the proposed DB-ConvLSTM module. In PDB-ConvLSTM module, two DB-ConvLSTMs with different dilate rates are adopted for capturing multi-scale information and encouraging information flow between bi-directional LSTM units. See § 3.2 for details.

2

Model

Pyramid Dilated Bidirectional(PDB) ConvLSTM



2

Model

Pyramid Dilated Bidirectional(PDB) ConvLSTM

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_i^X * \mathbf{X}_t + \mathbf{W}_i^H * \mathbf{H}_{t-1}),\\
\mathbf{f}_t &= \sigma(\mathbf{W}_f^X * \mathbf{X}_t + \mathbf{W}_f^H * \mathbf{H}_{t-1}),\\
\mathbf{o}_t &= \sigma(\mathbf{W}_o^X * \mathbf{X}_t + \mathbf{W}_o^H * \mathbf{H}_{t-1}),\\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(\mathbf{W}_c^X * \mathbf{X}_t + \mathbf{W}_c^H * \mathbf{H}_{t-1}),\\
\mathbf{H}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t),
\end{aligned}
$$

[Vanilla ConvLSTM]

$$
\begin{aligned}
\mathbf{i}_t^b &= \sigma(\mathbf{W}_i^{H^f} * \mathbf{H}_t^f + \mathbf{W}_i^{H^b} * \mathbf{H}_{t+1}^b),\\
\mathbf{f}_t^b &= \sigma(\mathbf{W}_f^{H^f} * \mathbf{H}_t^f + \mathbf{W}_f^{H^b} * \mathbf{H}_{t+1}^b),\\
\mathbf{o}_t^b &= \sigma(\mathbf{W}_o^{H^f} * \mathbf{H}_t^f + \mathbf{W}_o^{H^b} * \mathbf{H}_{t+1}^b),\\
\mathbf{c}_t^b &= \mathbf{f}_t^b \circ \mathbf{c}_{t+1}^b + \mathbf{i}_t^b \circ \tanh(\mathbf{W}_c^{H^f} * \mathbf{H}_t^f + \mathbf{W}_c^{H^b} * \mathbf{H}_{t+1}^b),\\
\mathbf{H}_t^b &= \mathbf{o}_t^{\mathbf{b}} \circ \tanh(\mathbf{c}_t^{\mathbf{b}}).
\end{aligned}
$$

[Deeper Bidirectional ConvLSTM]

2

Model

Loss Function

$\mathbf{G} \in \{0,1\}^{473\times 473}$ and $\mathbf{S} \in [0,1]^{473\times 473}$ denote the groundtruth saliency map and predicted saliency

$$\mathcal{L}(\mathbf{S},\mathbf{G}) = \mathcal{L}_{cross\_entropy}(\mathbf{S},\mathbf{G}) + \mathcal{L}_{MAE}(\mathbf{S},\mathbf{G})$$

$$\mathcal{L}_{cross\_entropy}(\mathbf{S},\mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N}[g_i log(s_i) + (1-g_i)log(1-s_i)]$$

$$\mathcal{L}_{MAE}(\mathbf{S},\mathbf{G}) = \frac{1}{N}\sum_{i=1}^{N}|g_i - s_i|$$

3

Experiments

Video Salient Object Detection

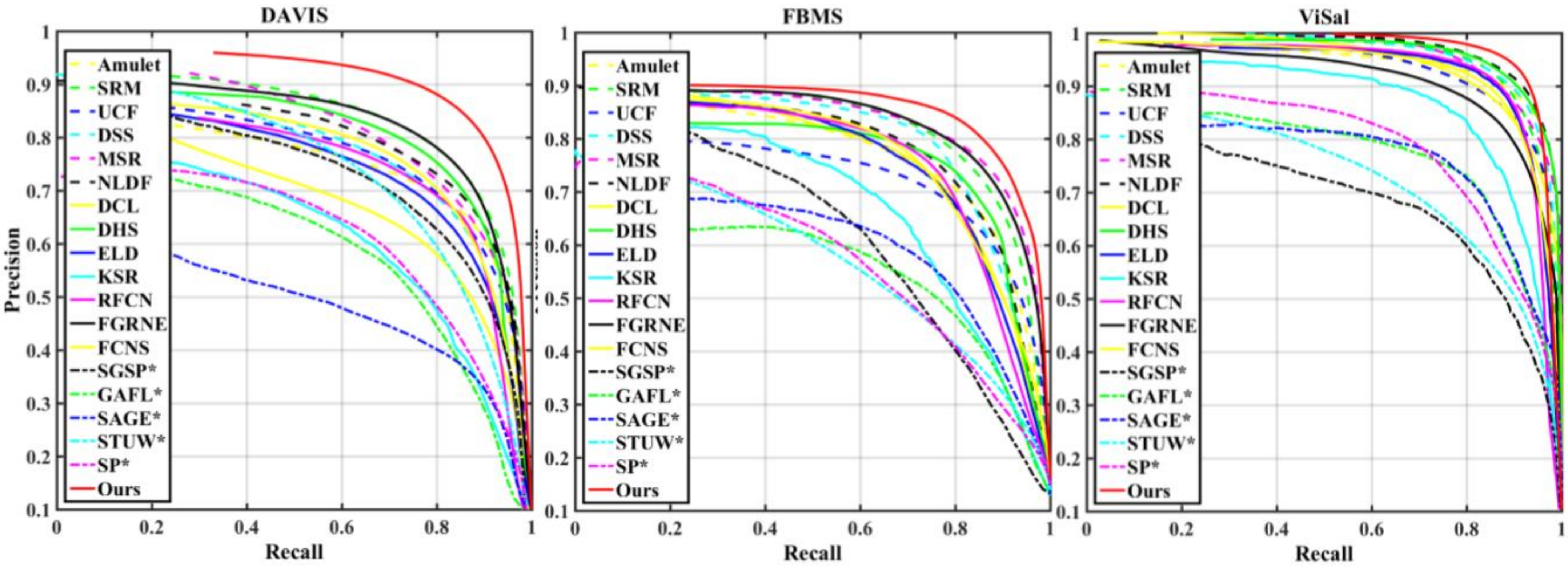


Fig. 4. Quantitative comparison against 18 saliency methods using PR curve on DAVIS [31], FBMS [2] and ViSal [43] datasets. Please see § 4.1 for more details.

3

Experiments

Video Salient Object Detection

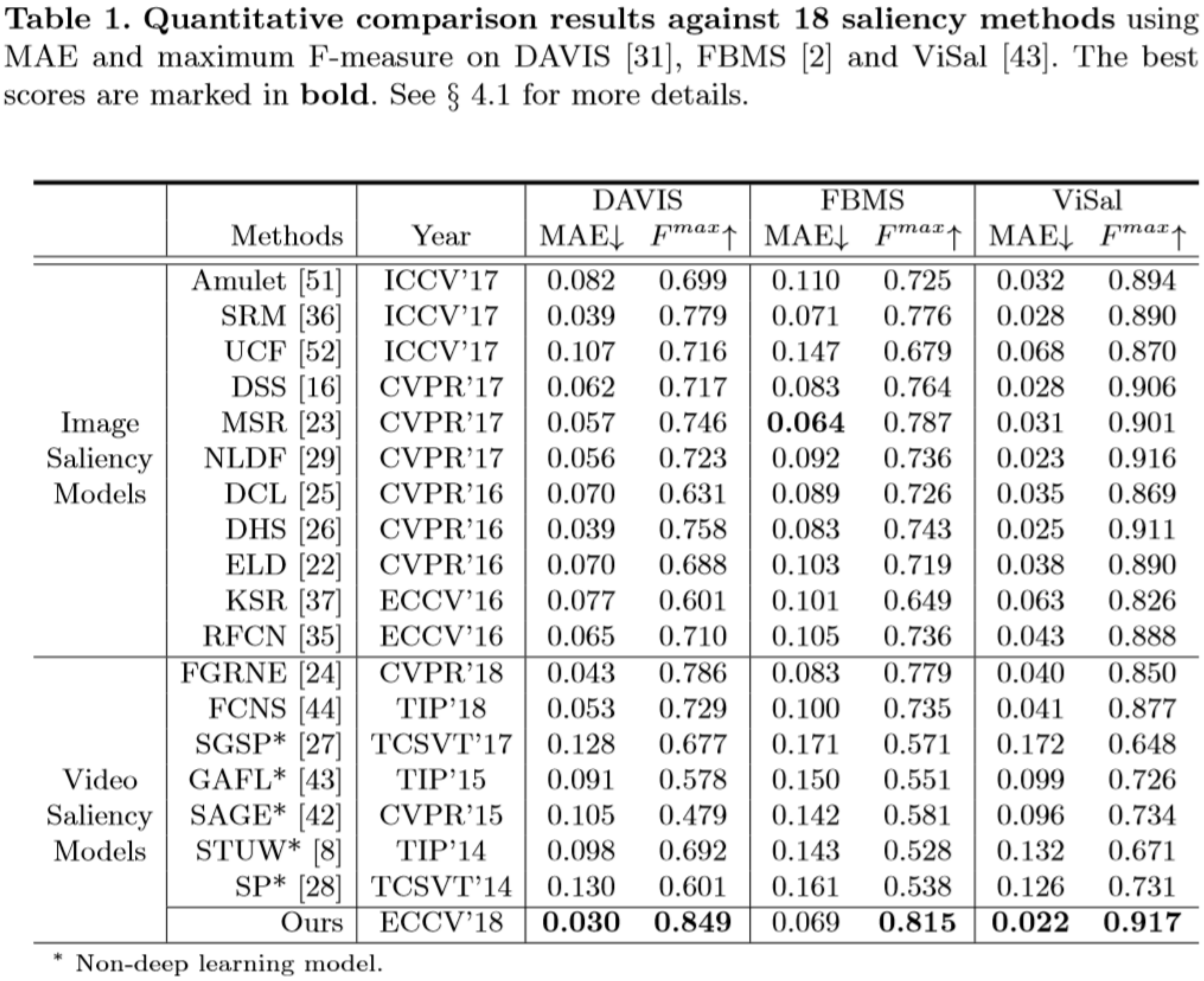


Table 1. Quantitative comparison results against 18 saliency methods using MAE and maximum F-measure on DAVIS [31], FBMS [2] and ViSal [43]. The best scores are marked in **bold**. See § 4.1 for more details.

			DAVIS		FBMS		ViSal	
	Methods	Year	MAE↓	F^{max}↑	MAE↓	F^{max}↑	MAE↓	F^{max}↑
Image Saliency Models	Amulet [51]	ICCV'17	0.082	0.699	0.110	0.725	0.032	0.894
	SRM [36]	ICCV'17	0.039	0.779	0.071	0.776	0.028	0.890
	UCF [52]	ICCV'17	0.107	0.716	0.147	0.679	0.068	0.870
	DSS [16]	CVPR'17	0.062	0.717	0.083	0.764	0.028	0.906
	MSR [23]	CVPR'17	0.057	0.746	**0.064**	0.787	0.031	0.901
	NLDF [29]	CVPR'17	0.056	0.723	0.092	0.736	0.023	0.916
	DCL [25]	CVPR'16	0.070	0.631	0.089	0.726	0.035	0.869
	DHS [26]	CVPR'16	0.039	0.758	0.083	0.743	0.025	0.911
	ELD [22]	CVPR'16	0.070	0.688	0.103	0.719	0.038	0.890
	KSR [37]	ECCV'16	0.077	0.601	0.101	0.649	0.063	0.826
	RFCN [35]	ECCV'16	0.065	0.710	0.105	0.736	0.043	0.888
Video Saliency Models	FGRNE [24]	CVPR'18	0.043	0.786	0.083	0.779	0.040	0.850
	FCNS [44]	TIP'18	0.053	0.729	0.100	0.735	0.041	0.877
	SGSP* [27]	TCSVT'17	0.128	0.677	0.171	0.571	0.172	0.648
	GAFL* [43]	TIP'15	0.091	0.578	0.150	0.551	0.099	0.726
	SAGE* [42]	CVPR'15	0.105	0.479	0.142	0.581	0.096	0.734
	STUW* [8]	TIP'14	0.098	0.692	0.143	0.528	0.132	0.671
	SP* [28]	TCSVT'14	0.130	0.601	0.161	0.538	0.126	0.731
	Ours	ECCV'18	**0.030**	**0.849**	0.069	**0.815**	**0.022**	**0.917**

* Non-deep learning model.

3

Experiments

Unsupervised Video Object Segmentation

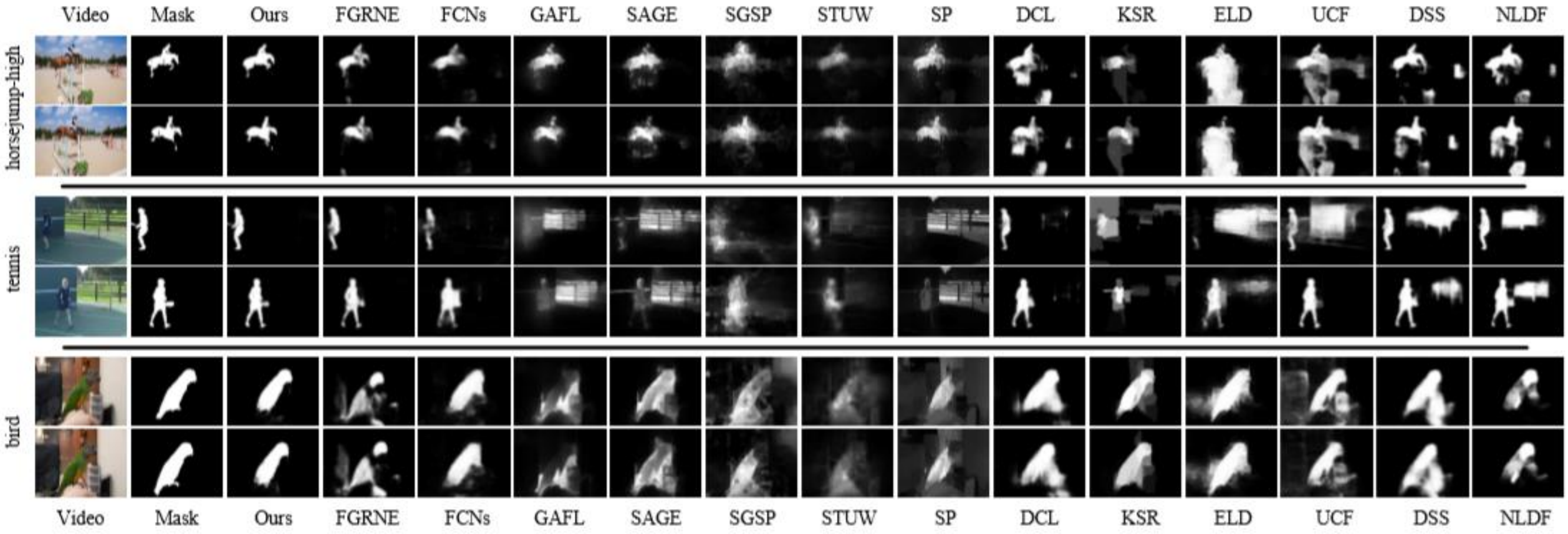


Fig. 5. Qualitative comparison against other top-performing saliency methods with groundtruths on three example video sequences. Zoom-in for details.

3 Experiments

Unsupervised Video Object Segmentation

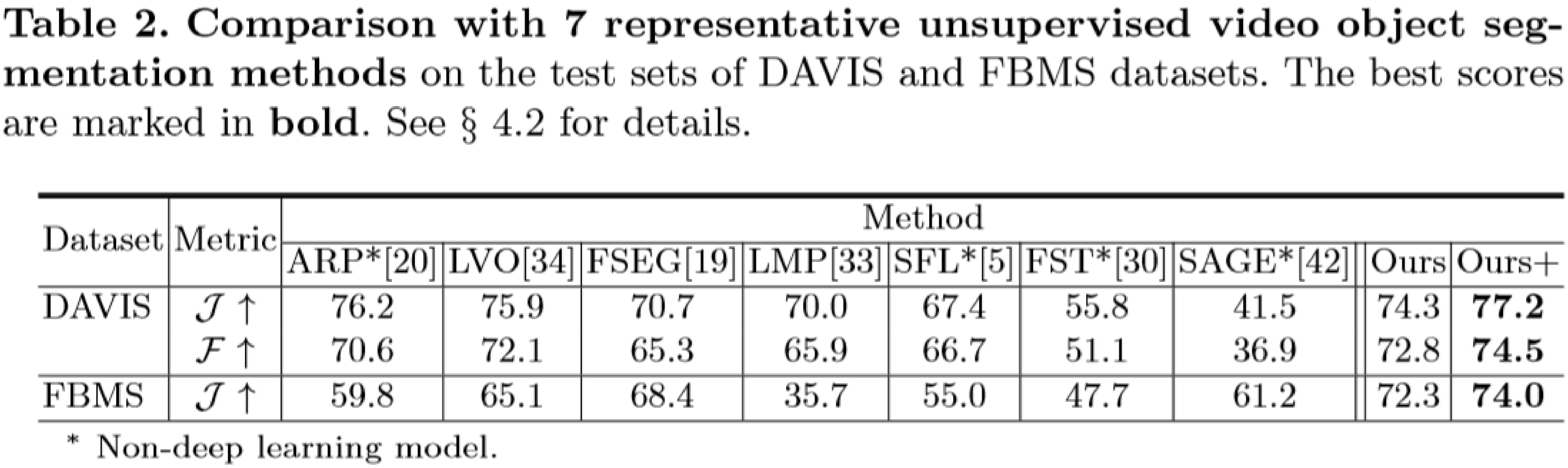


Table 2. Comparison with 7 representative unsupervised video object segmentation methods on the test sets of DAVIS and FBMS datasets. The best scores are marked in **bold**. See § 4.2 for details.

Dataset	Metric	Method								
		ARP*[20]	LVO[34]	FSEG[19]	LMP[33]	SFL*[5]	FST*[30]	SAGE*[42]	Ours	Ours+
DAVIS	$\mathcal{J} \uparrow$	76.2	75.9	70.7	70.0	67.4	55.8	41.5	74.3	**77.2**
	$\mathcal{F} \uparrow$	70.6	72.1	65.3	65.9	66.7	51.1	36.9	72.8	**74.5**
FBMS	$\mathcal{J} \uparrow$	59.8	65.1	68.4	35.7	55.0	47.7	61.2	72.3	**74.0**

* Non-deep learning model.

3 Experiments

Runtime Comparison

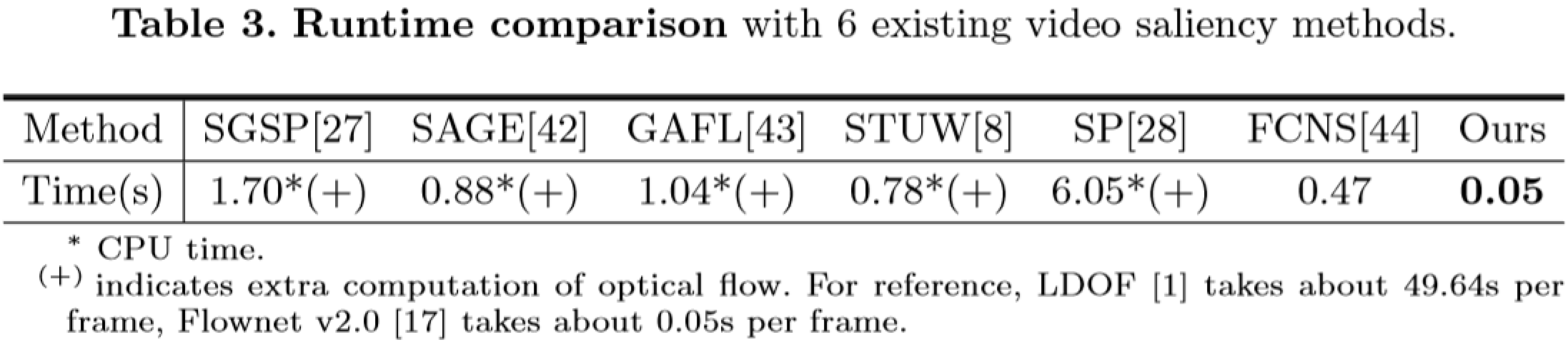


Table 3. Runtime comparison with 6 existing video saliency methods.

Method	SGSP[27]	SAGE[42]	GAFL[43]	STUW[8]	SP[28]	FCNS[44]	Ours
Time(s)	1.70*(+)	0.88*(+)	1.04*(+)	0.78*(+)	6.05*(+)	0.47	**0.05**

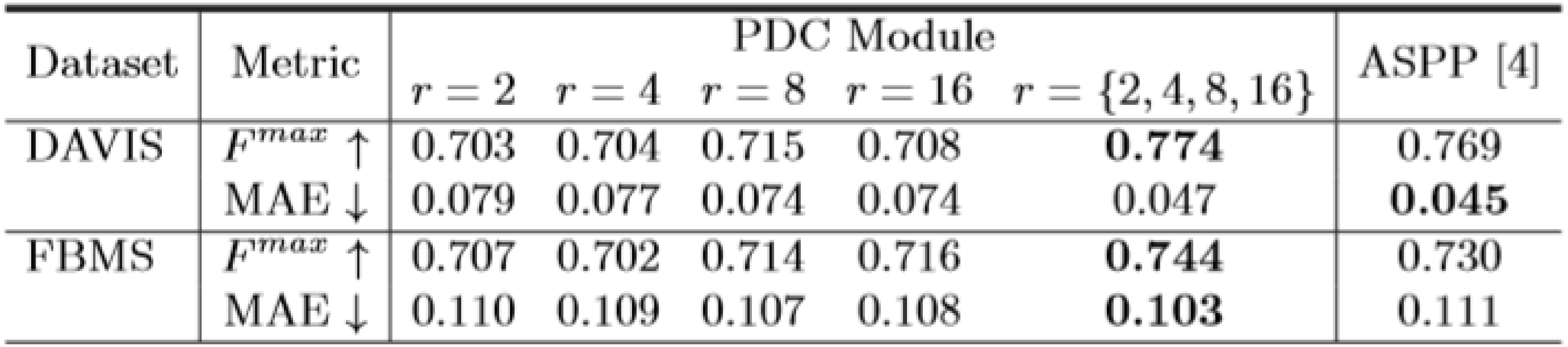
* CPU time.

(+) indicates extra computation of optical flow. For reference, LDOF [1] takes about 49.64s per frame, Flownet v2.0 [17] takes about 0.05s per frame.

3 Experiments

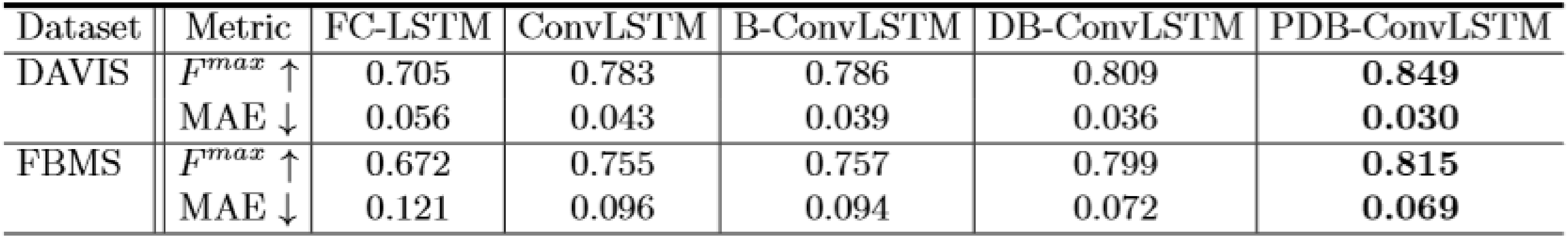
Ablation Study

Table 4. Ablation study for PDC module on DAVIS and FBMS datasets.



Dataset	Metric	PDC Module					ASPP [4]
		$r = 2$	$r = 4$	$r = 8$	$r = 16$	$r = \{2, 4, 8, 16\}$	
DAVIS	F^{max} ↑	0.703	0.704	0.715	0.708	**0.774**	0.769
	MAE ↓	0.079	0.077	0.074	0.074	0.047	**0.045**
FBMS	F^{max} ↑	0.707	0.702	0.714	0.716	**0.744**	0.730
	MAE ↓	0.110	0.109	0.107	0.108	**0.103**	0.111

Table 5. Ablation study for PDB-ConvLSTM on DAVIS and FBMS datasets.



Dataset	Metric	FC-LSTM	ConvLSTM	B-ConvLSTM	DB-ConvLSTM	PDB-ConvLSTM
DAVIS	F^{max} ↑	0.705	0.783	0.786	0.809	**0.849**
	MAE ↓	0.056	0.043	0.039	0.036	**0.030**
FBMS	F^{max} ↑	0.672	0.755	0.757	0.799	**0.815**
	MAE ↓	0.121	0.096	0.094	0.072	**0.069**