TSANET: TEMPORAL AND SCALE ALIGNMENT FOR UNSUPERVISED VIDEO OBJECT SEGMENTATION

Seunghoon Lee[1] Suhwan Cho[1] Dogyoon Lee[1] Minhyeok Lee[1] Sangyoun Lee[1,2]

[1]Yonsei University [2]Korea Institute of Science and Technology (KIST)

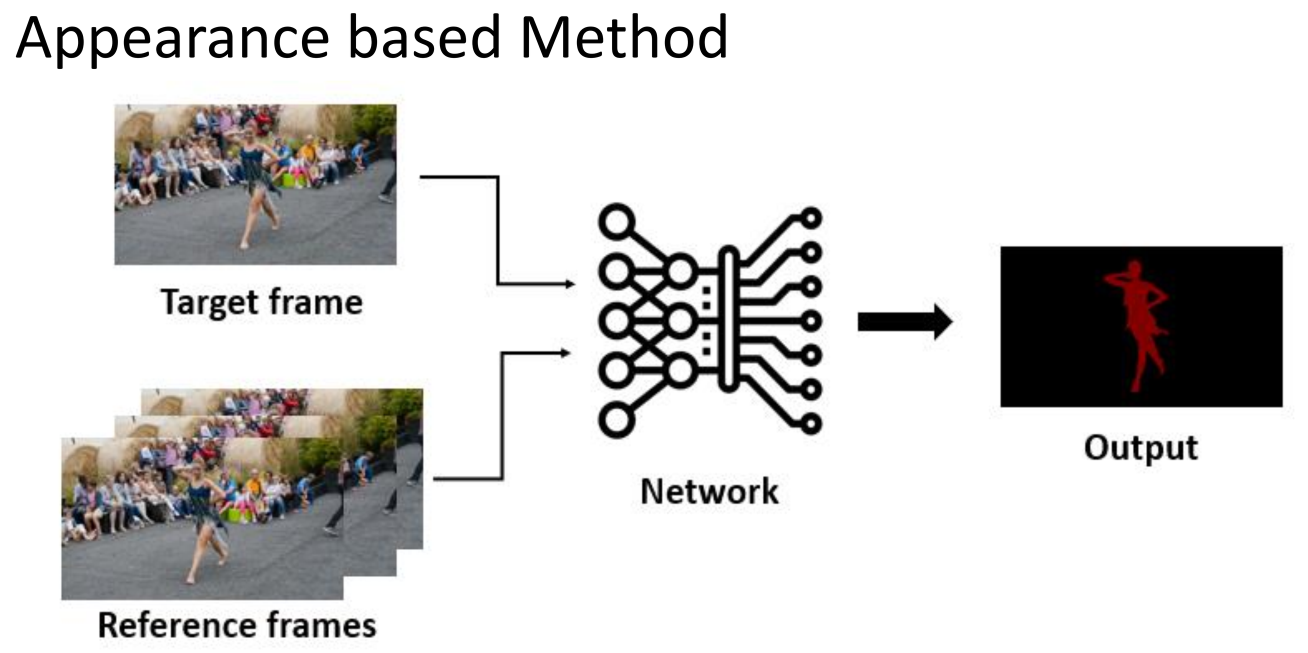
Introduction

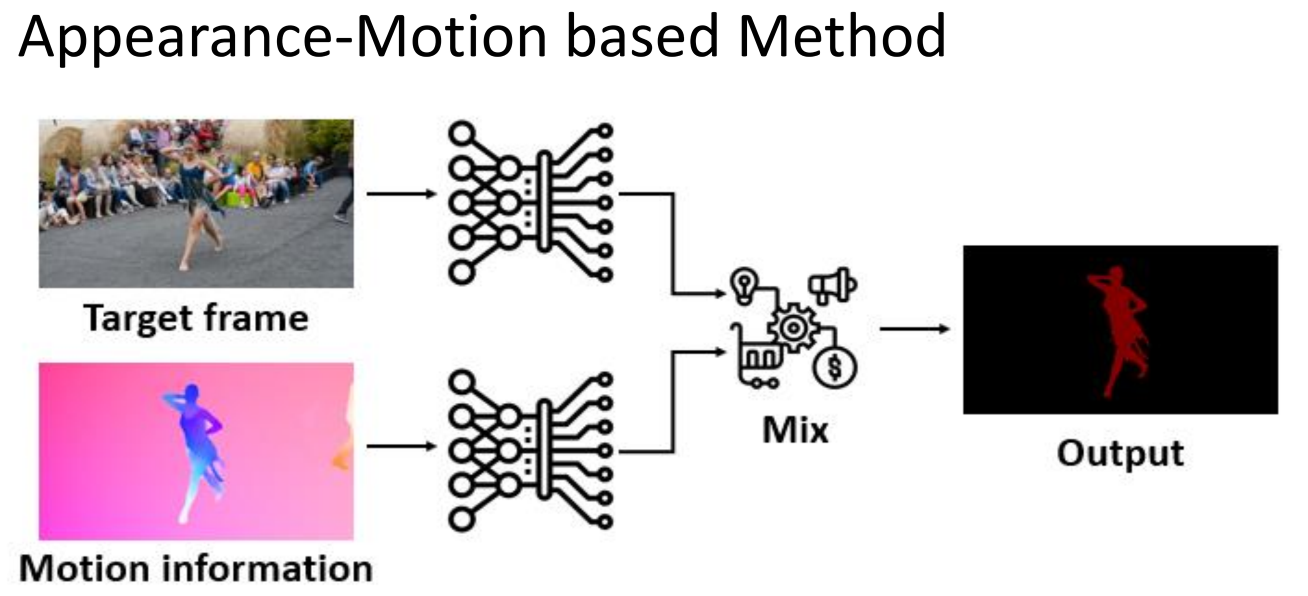
- **Task Definition**

Unsupervised Video Object Segmentation (UVOS) is a challenging task that **segments the prominent object in videos** without any manual guide.

- **Research Direction**

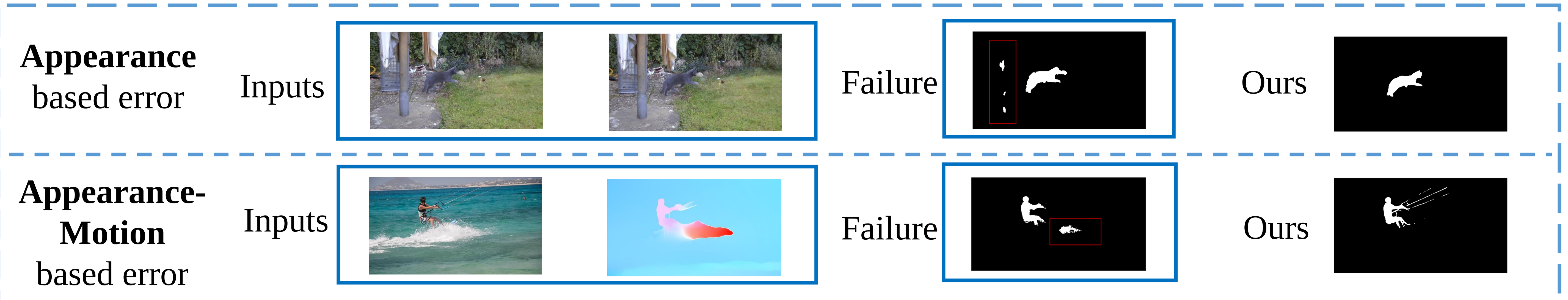
In recent works, there are two main approaches: **Appearance based**, **Appearance-Motion based methods**. Appearance-based methods cannot consider the motion of the target object although motion cues provides useful knowledge to detect the salient objects. Appearance-Motion based methods has a intrinsic drawback that the dependency on motion information.





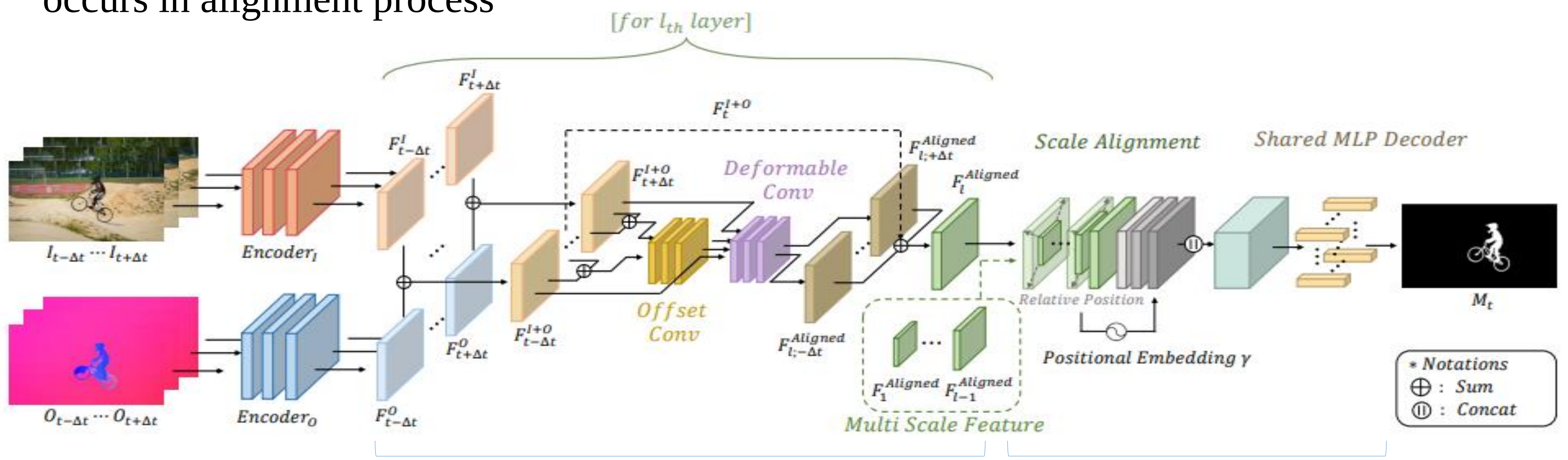
Motivation

- Appearance-based methods cannot consider the motion of the target object although motion cues provides useful knowledge to detect the salient objects
- Appearance-Motion based methods has a intrinsic drawback that the dependency on motion information. It causes the model to fail to capture detailed shape information of the object



Proposed Method

- We propose **Temporal Alignment Fusion (TAF)** that aligns features of adjacent frames to target features for leveraging the contextual information to the target frame
- To aggregate features which have different scales, we propose **Scale Alignment Decoder (SAD)**, introducing implicit neural representation using relative coordinate information that occurs in alignment process



Temporal Alignment Fusion Scale Alignment Decoder

Quantitative Results

- We evaluate our framework TSANet compared with the recent works. TSANet achieve **the-state-of-the-art performance** on DAVIS 2016
- TSANet also shows the promising performance on FBMS dataset (**second-best**)
- We demonstrate the effectiveness of our modules on ablation study

Model	Publication	PP	$\mathcal{J}$ Mean↑	$\mathcal{J}$ Recall↑	$\mathcal{J}$ Decay↓	$\mathcal{F}$ Mean↑	$\mathcal{F}$ Recall↑	$\mathcal{F}$ Decay↓	$\mathcal{J}\&\mathcal{F}$ Mean↑
AGS [12]	CVPR'19	✓	79.7	91.1	1.9	77.4	85.8	1.6	78.6
COSNet [1]	CVPR'19	✓	80.5	93.1	4.4	79.5	89.5	5	80.0
ADNet [13]	ICCV'19	✓	81.7	-	-	80.5	-	-	81.1
AGNN [14]	ICCV'19	✓	80.7	94.0	0.0	79.1	90.5	0.0	79.9
MATNet [4]	AAAI'20	✓	82.4	94.5	3.8	80.7	90.2	4.5	81.5
DFNet [15]	ECCV'20	✓	83.4	94.4	4.2	81.8	89.0	3.7	82.6
3DC-Seg [16]	BMVC'20	✓	84.2	95.8	7.4	84.3	92.4	5.5	84.2
F2Net [3]	AAAI'21		83.1	95.7	0.0	84.4	92.3	0.8	83.7
RTNet [5]	CVPR'21	✓	85.6	96.1	-	84.7	93.8	-	85.2
FSNet [17]	ICCV'21	✓	83.4	94.5	3.2	83.1	90.2	2.6	83.3
TransportNet [6]	ICCV'21		84.5	-	-	85.0	-	-	84.8
AMC-Net [7]	ICCV'21	✓	84.5	96.4	2.8	84.6	93.8	2.5	84.6
CFAM [2]	WACV'22		83.5	-	-	82.0	-	-	82.8
D^2Conv3D [18]	WACV'22		85.5	-	-	86.5	-	-	86.0
IMP [19]	AAAI'22		84.5	92.7	2.8	86.7	93.3	0.8	85.6
PMN [20]	WACV'23	✓	85.4	-	-	86.4	-	-	85.9
TMO [8]	WACV'23		85.6	-	-	86.6	-	-	86.1
TSANet (ours)			**86.6**	95.7	0.0	**88.3**	94.3	0.0	**87.4**

Table 1. Quantitative results on DAVIS 2016. PP indicates post-processing. Each color denotes **best** and second results.

Qualitative Results

horsejump-high parkour

*Predicted results overlaid on the video (RED)