JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Unified Mask Embedding and Correspondence Learning for Self-Supervised Video Segmentation

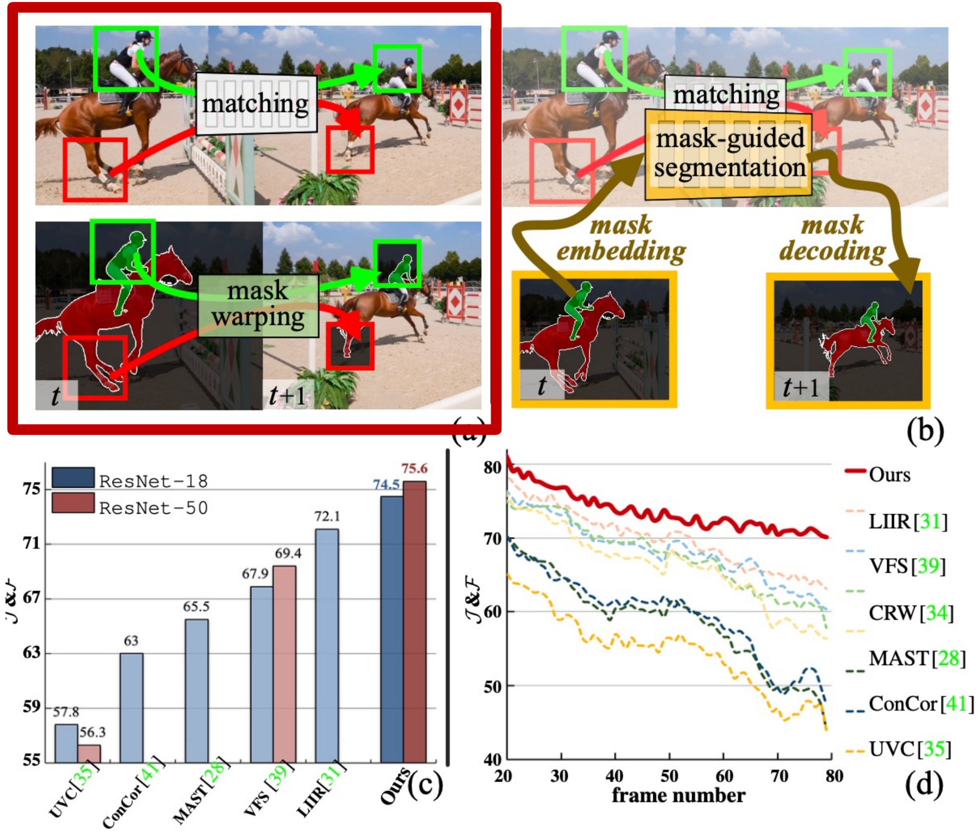
Liulei Li[1,4], Wenguan Wang[1†], Tianfei Zhou[2], Jianwu Li[3] , Yi Yang[1]

[1] ReLER, CCAI, Zhejiang University [2] ETH Zurich [3] Rbeijing Institute of Technology [4] Baidu VIS

FRI-AM-213

Overview

- **Correspondence Learning for Self-Supervised Video Segmentation**

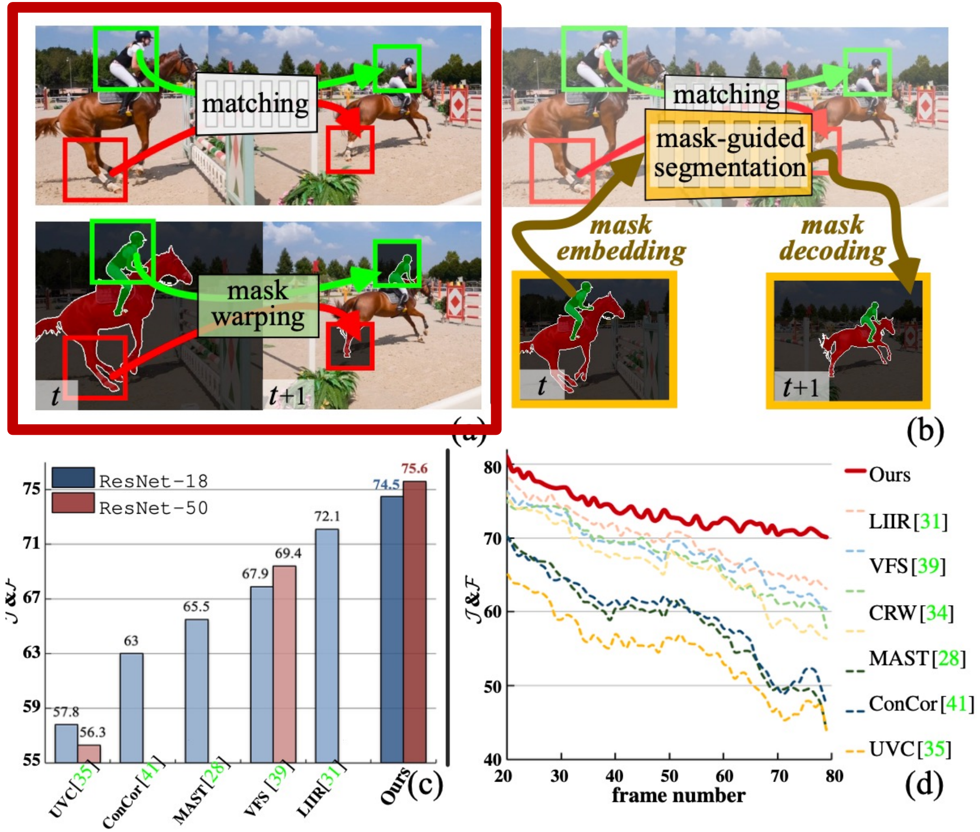


Existing solution:

unsupervised correspondence learning
+
non-learnable mask warping

Overview

- **Correspondence Learning for Self-Supervised Video Segmentation**



Existing solution:

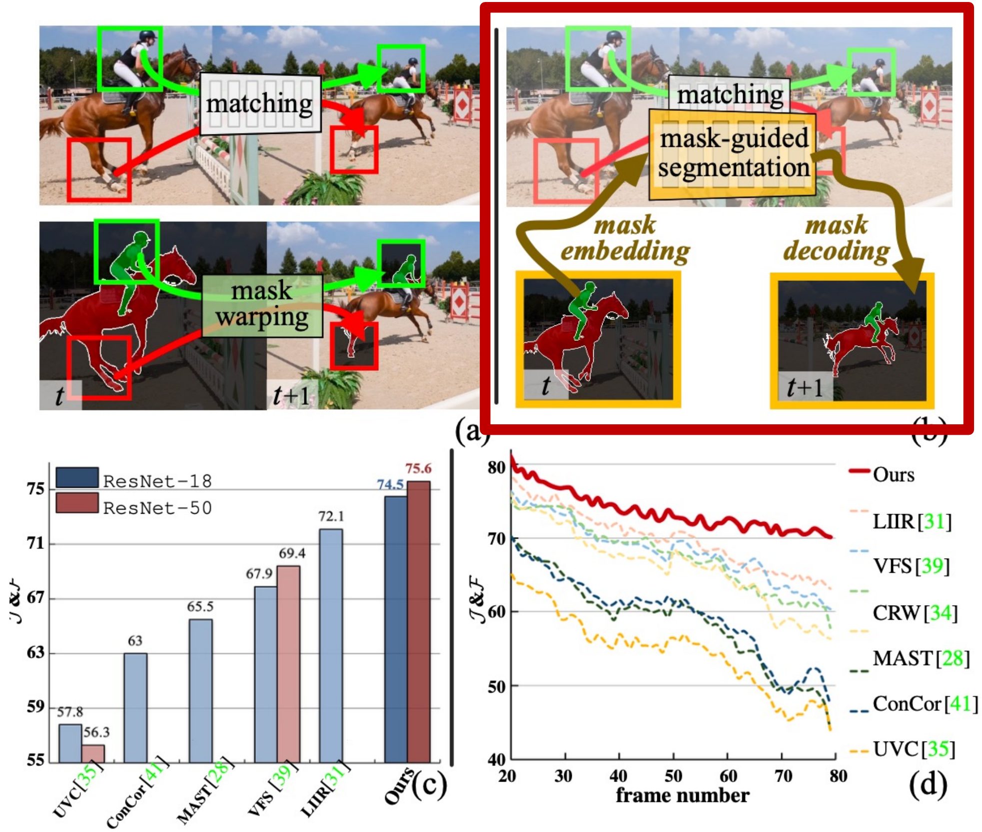
unsupervised correspondence learning
+
non-learnable mask warping

Limitations:

- leaving a significant gap between the training objective and task/inference setup.
- sensitive to outliers, resulting in error accum-ulation over time.

Overview

- **Unified Mask Embedding and Correspondence Learning**



Our solution:

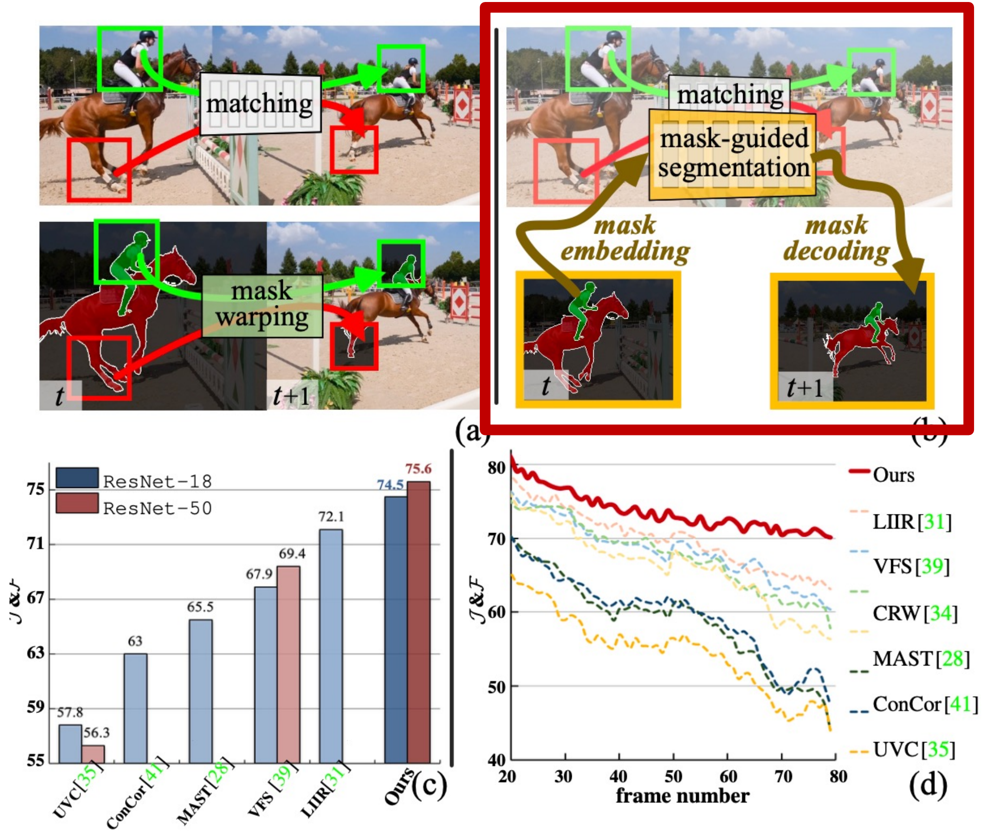
mask embedding learning

+

dense correspondence learning

Overview

- ## Unified Mask Embedding and Correspondence Learning



Our solution:

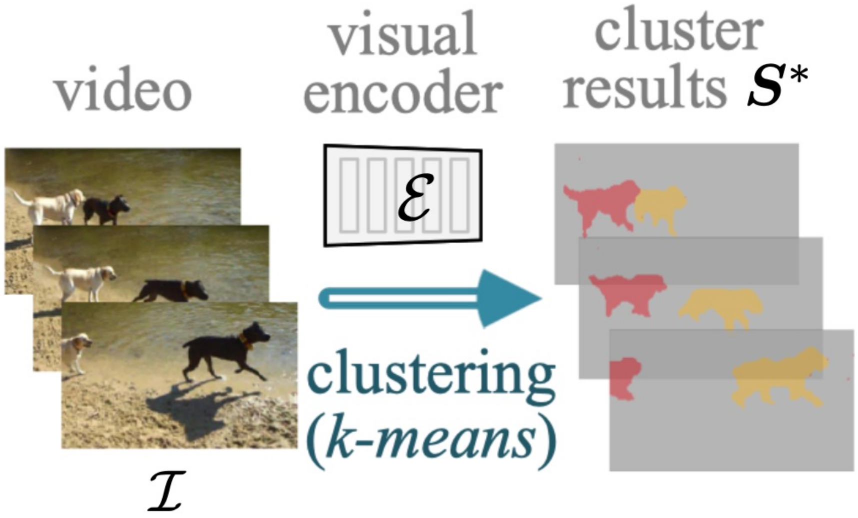
mask embedding learning
+
dense correspondence learning

Advantanges:

- **Aligned** training objective with the core nature of VOS.
- Target-oriented context can **reduce error accumulation** and perform more robust.
- Empowered by more **advanced VOS model designs** in the fully-supervised learning setting.

Our Method

- ## Space-time Clustering



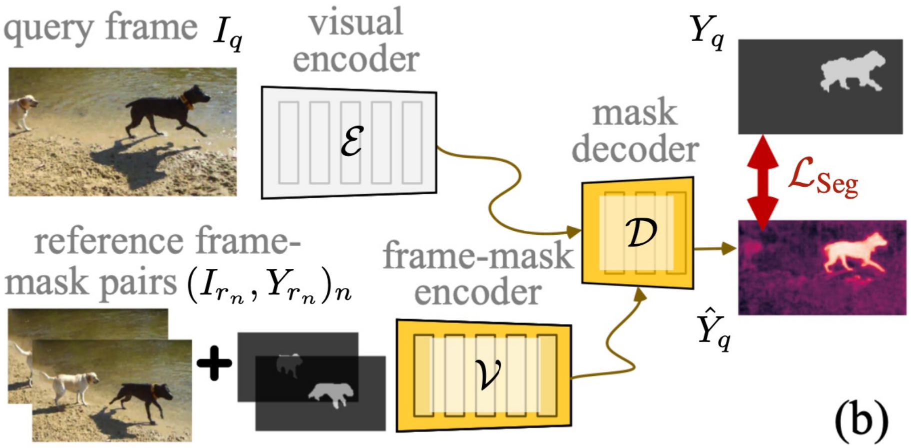
The goal of this step is to partition each training video into M space-time consistent segments, which can be achieved by solving the following optimization problem:

$$\min_{\boldsymbol{C},\boldsymbol{S}} \sum\nolimits_{i\in\mathcal{I}} \|\boldsymbol{i} - \boldsymbol{C}\boldsymbol{s}_i\|, \quad s.t.\ \ \boldsymbol{s}_i \in \{0,1\}^M, \quad \mathbf{1}^\top \boldsymbol{s}_i = 1.$$

Moreover, to pursue spatiotemporally compact clusters, for each pixel, we supply its embedding with a 3D sinusoidal position encoding vector.

Our Method

- **Mask-embedded Segmentation Learning**



We first apply our visual encoder and frame-mask encoder over each reference frame and each reference mask to obtain visual and target-specific embeddings:

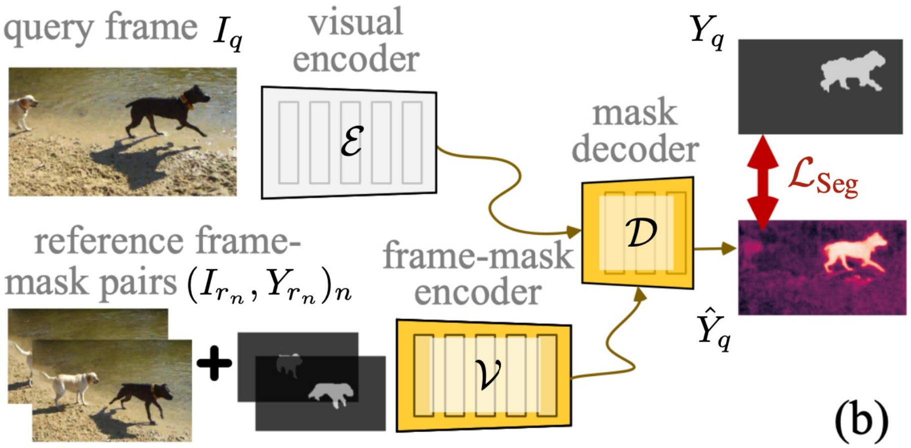
$$\boldsymbol{I}_{r_n} = \mathcal{E}(I_{r_n}) \in \mathbb{R}^{HW \times D},$$

$$\boldsymbol{V}_{r_n} = \mathcal{V}([I_{r_n}, Y_{r_n}]) \in \mathbb{R}^{HW \times D'}.$$

In this step, our model utilizes clustering results as pseudo ground-truths, to directly learn VOS as mask embedding and decoding.

Our Method

- **Mask-embedded Segmentation Learning**



We then estimate the affinity between the query and reference frames by:

$$A = \texttt{softmax}(\boldsymbol{I}_r \boldsymbol{I}_q^\top) \in \mathbb{R}^{NHW \times HW}.$$

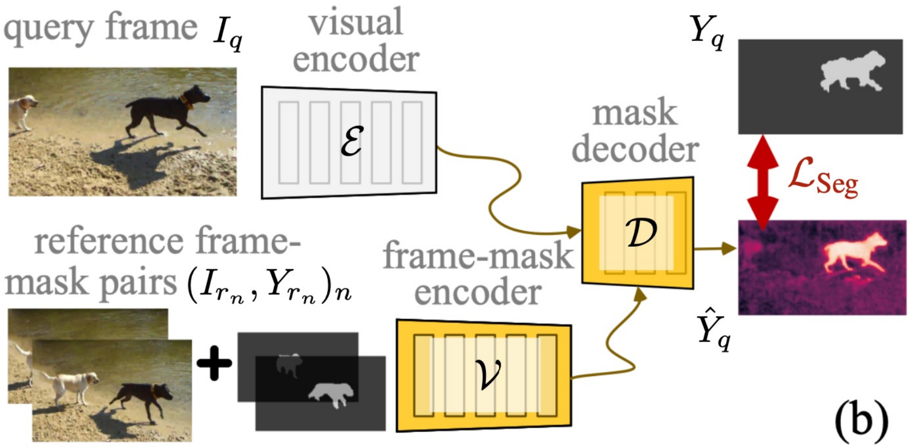
Target-specific, supportive features are accordingly assembled to yield:

$$\boldsymbol{V}_q = A^\top \boldsymbol{V}_r \in \mathbb{R}^{HW \times D'}.$$

In this step, our model utilizes clustering results as pseudo ground-truths, to directly learn VOS as mask embedding and decoding.

Our Method

▪ Mask-embedded Segmentation Learning



In this step, our model utilizes clustering results as pseudo ground-truths, to directly learn VOS as mask embedding and decoding.

We first construct a coarse mask for the query image by warping the reference masks w.r.t. the affinity:

$$\bar{Y}_q = A^\top [Y_{r_1}, Y_{r_2}, \cdots, Y_{r_n}] \in \mathbb{R}^{HW}.$$

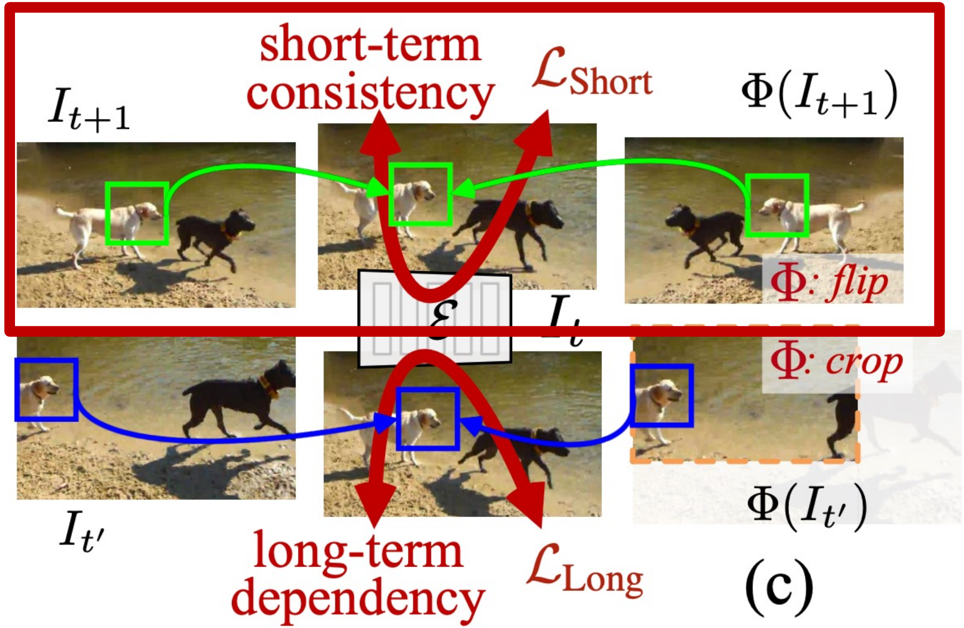
The segmentation prediction for the query image is made by:

$$\hat{Y}_q = \mathcal{D}([\boldsymbol{V}_q, \bar{\boldsymbol{V}}_q]),$$

$$\bar{\boldsymbol{V}}_q = \mathcal{V}([I_q, \bar{Y}_q]) \in \mathbb{R}^{HW \times D'}.$$

Our Method

- **Self-supervised Dense Correspondence Learning**



Short-term Appearance Consistency

❶ $\mathcal{E}(I_t) \approx \mathcal{E}(I_{t+1})$
short-term consistency

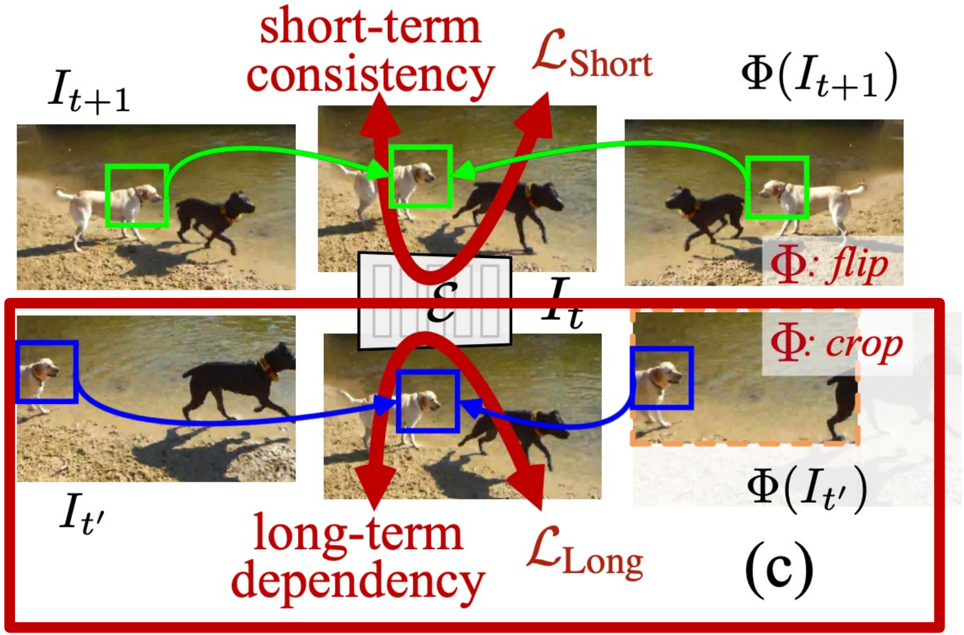
❷ $\mathcal{E}(\Phi(I_t)) = \Phi(\mathcal{E}(I_t))$
transformation-equivariance

$\Rightarrow \mathcal{E}(\Phi(I_t)) \approx \Phi(\mathcal{E}(I_{t+1}))$ ❸.

Given two successive frames $I_t, I_{t+1} \in \mathcal{I}$, their representations, delivered by the visual encoder, are constrained to be equivariant against geometric transformations (*i.e.*, scaling, flipping, and cropping).

Our Method

▪ Self-supervised Dense Correspondence Learning



Long-term Semantic Dependency

$$\left.\begin{array}{l} ❹\ \mathcal{E}(I_t) \approx A_{t'}^{t\top}\mathcal{E}(I_{t'}) \\ \text{long-term dependency} \\ ❷\ \mathcal{E}(\Phi(I_t)) = \Phi(\mathcal{E}(I_t)) \\ \text{transformation-equivariance} \end{array}\right\} \Rightarrow \mathcal{E}(I_t) \approx A_{\Phi(t')}^{t\top}\Phi(\mathcal{E}(I_{t'}))\ ❺.$$

Given two distant frames $I_t, I_{t'} \in \mathcal{I}$ (*s.t.* $|t-t'| \geq 5$), their representations, after being aligned *w.r.t.* the affinity $A_{t'}^{t}$, are constrained to be equivariant against geometric transformations (*i.e.*, scaling, flipping, and cropping).

Experiments

- Qulititative Results on DAVIS$_{17}$ val and YouTube-VOS val

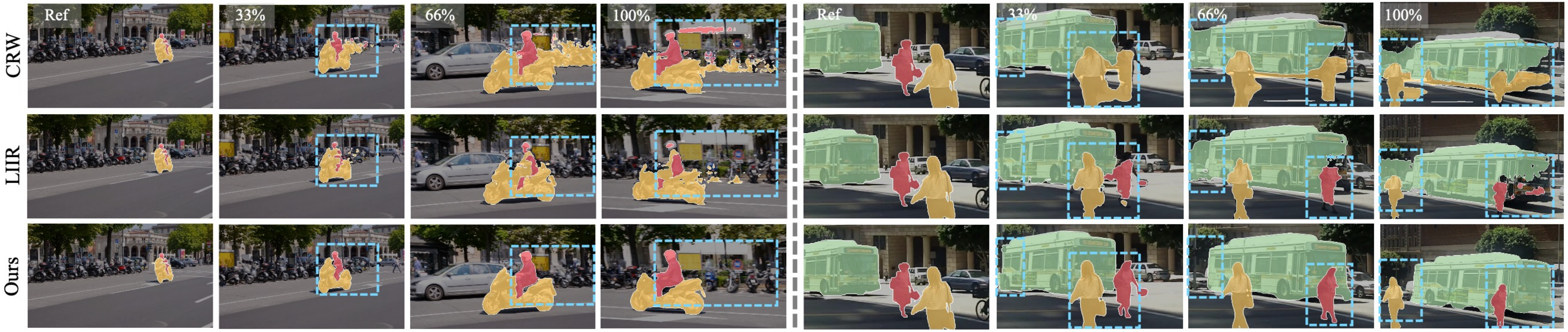


Figure 3. **Visual comparison results** (§4.1) on two videos from DAVIS$_{17}$ [42] `val` (left) and Youtube-VOS [52] `val` (right), respectively. CRW [34] and LIIR [31] suffer from error accumulation during mask tracking, due to the simple matching-based mask copy-paste strategy. However, our approach performs robust over time and yields more accurate segmentation results, by learning to embed target masks.

Experiments

- Quantitative Results on DAVIS$_{17}$ val

Method	Backbone	Dataset(size)	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{J}_r\uparrow$	$\mathcal{F}_m\uparrow$	$\mathcal{F}_r\uparrow$
Colorization [26] [ECCV18]	ResNet-18	Kinetics(- , 800 hours)	34.0	34.6	34.1	32.7	26.8
CorrFlow [27] [BMVC19]	ResNet-18	OxUvA(- , 14 hours)	50.3	48.4	53.2	52.2	56.0
TimeCycle [32] [CVPR19]	ResNet-50	VLOG(- , 344 hours)	48.7	46.4	50.0	50.0	48.0
UVC [35] [NeurIPS19]	ResNet-18	C+Kinetics(30K, 800 hours)	57.8	56.3	65.0	59.2	64.1
MuG [59] [CVPR20]	ResNet-18	OxUvA(- , 14 hours)	54.3	52.6	57.4	56.1	58.1
MAST [28] [CVPR20]	ResNet-18	Youtube-VOS(- , 5.58 hours)	65.5	63.3	73.2	67.6	77.7
CRW [34] [NeurIPS20]	ResNet-18	Kinetics(- , 800 hours)	68.3	65.5	78.6	71.0	82.9
ConCorr [41] [AAAI21]	ResNet-18	C+TrackingNet(30K, 300 hours)	63.0	60.5	70.6	65.5	73.0
CLTC [37] [CVPR21]	ResNet-18	Youtube-VOS(- , 5.58 hours)	70.3	67.9	78.2	72.6	83.7
JSTG [60] [ICCV21]	ResNet-18	Kinetics(- , 800 hours)	68.7	65.8	77.7	71.6	84.3
VFS [39] [ICCV21]	ResNet-18	Kinetics(- , 800 hours)	67.9	65.0	77.2	70.8	82.3
	ResNet-50		69.4	66.7	78.6	72.0	85.2
DINO [74] [ICCV21]	ResNet-50	I(1.28M, -)	56.2	54.5	58.1	57.9	60.3
	ViT-B/8		71.4	67.9	81.6	74.9	85.4
DUL [38] [NeurIPS21]	ResNet-18	Youtube-VOS(- , 5.58 hours)	69.3	67.1	81.2	71.6	84.9
SCR [40] [CVPR22]	ResNet-18	Kinetics(- , 800 hours)	70.5	67.4	78.8	73.6	84.6
LIIR [31] [CVPR22]	ResNet-18	Youtube-VOS(- , 5.58 hours)	72.1	69.7	81.4	74.5	85.9
OURS	ResNet-18	Youtube-VOS(- , 5.58 hours)	**74.5**	**71.6**	**82.9**	**77.4**	**86.9**
	ResNet-50		**75.6**	**73.3**	**83.6**	**77.8**	**87.3**
OSVOS [12] [CVPR17]	VGG-16	I+D(1.28M, 10k)	60.3	56.6	63.8	63.9	73.8
STM [10] [ICCV19]	ResNet-50	I+D+Youtube-VOS(1.28M, 164k)	81.8	79.2	88.7	84.3	91.8

- I: ImageNet [75]; C: COCO [76]; D: DAVIS$_{17}$ [42].

Table 1. **Quantitative segmentation results** (§4.1) on DAVIS$_{17}$ [42] `val`. For dataset size, we report (#*raw* images, length of *raw* videos) for self-supervised methods and (#image-level annotations, #pixel-level annotations) for supervised methods.

Experiments

- Quantitative Results on DAVIS$_{17}$ test-dev and YouTube-VOS val

Method	Backbone	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{J}_r\uparrow$	$\mathcal{F}_m\uparrow$	$\mathcal{F}_r\uparrow$
MAST [28] [CVPR20]	ResNet-18	54.3	50.7	58.9	57.8	64.5
CRW [34] [NeurIPS20]	ResNet-18	55.9	52.3	-	59.6	-
DUL [38] [NeurIPS21]	ResNet-18	57.0	53.5	60.4	60.5	67.6
SCR [40] [CVPR22]	ResNet-18	59.9	55.9	-	64.0	-
LIIR [31] [CVPR22]	ResNet-18	57.5	55.2	63.1	59.8	68.6
OURS	ResNet-18	**61.3**	**59.4**	**66.5**	**63.1**	**73.7**
	ResNet-50	**62.4**	**60.6**	**66.9**	**64.2**	**74.3**
RGMP [17] [CVPR18]	ResNet-50	52.9	51.3	-	54.4	-
STM [10] [ICCV19]	ResNet-50	72.2	69.3	-	75.2	-

Table 2. **Quantitative results** (§4.1) on DAVIS$_{17}$ [42] `test-dev`.

Method	Backbone	$\mathcal{J}\&\mathcal{F}_m\uparrow$	Seen		Unseen	
			$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$
Colorization [26] [ECCV18]	ResNet-18	38.9	43.1	38.6	36.6	37.4
CorrFlow [27] [BMVC19]	ResNet-18	46.6	50.6	46.6	43.8	45.6
MAST [28] [CVPR20]	ResNet-18	64.2	63.9	64.9	60.3	67.7
CRW [34] [NeurIPS20]	ResNet-18	68.7	67.4	69.1	65.1	73.2
CLTC [37] [CVPR21]	ResNet-18	67.3	66.2	67.9	63.2	71.7
DUL [38] [NeurIPS21]	ResNet-18	69.9	69.6	71.3	65.0	73.5
LIIR [31] [CVPR22]	ResNet-18	69.3	67.9	69.7	65.7	73.8
OURS	ResNet-18	**71.6**	**71.0**	**74.2**	**66.0**	**75.3**
	ResNet-50	**72.4**	**71.7**	**74.6**	**67.0**	**76.2**
OSVOS [12] [CVPR17]	VGG-16	58.8	59.8	60.5	54.2	60.7
STM [10] [ICCV19]	ResNet-50	79.4	79.7	84.2	73.5	80.9

Table 3. **Quantitative results** (§4.1) on YouTube-VOS [52] `val`.

Experiments

- Ablative studies on $DAVIS_{17}$

Loss	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$
$\mathcal{L}_{\text{Short}}$	57.4	55.8	58.9
$\mathcal{L}_{\text{Long}}$	67.2	64.9	69.5
$\mathcal{L}_{\text{Short}} + \mathcal{L}_{\text{Long}}$	68.8	66.7	70.9
$\mathcal{L}_{\text{Seg}}$	62.3	60.5	64.0
$\mathcal{L}_{\text{Seg}}+\mathcal{L}_{\text{Short}}+\mathcal{L}_{\text{Long}}$	**74.5**	**71.6**	**77.4**

(a) loss design

#Ref. Frame	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$
First	68.8	65.7	71.9
First + Last 1:15	73.2	70.4	76.0
First + Last 1:20	**74.5**	**71.6**	**77.4**
First + Last 1:25	73.5	70.9	76.1
First + Last 1:30	72.8	70.2	75.3

(b) number of reference frames

#Centroid	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$
$M = 2$	67.5	65.2	69.8
$M = 3$	71.6	69.0	74.2
$M = 5$	**74.5**	**71.6**	**77.4**
$M = 8$	72.5	69.6	75.4
$M = 10$	70.1	67.3	72.9

(c) number of cluster centers

Mask update	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$
No update	71.1	68.3	73.9
Per 20 epoch	72.8	69.9	75.7
Per 15 epoch	73.9	70.8	77.0
Per 10 epoch	**74.5**	**71.6**	**77.4**
Per 5 epoch	72.5	69.5	75.5
Every epoch	69.7	66.7	72.6

(d) pseudo mask update

Round	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$	FPS
0	69.7	67.3	72.1	1.86
1	72.6	69.8	75.4	1.84 (-1.1%)
2	73.9	71.1	76.7	1.80 (-3.2%)
3	**74.5**	**71.6**	**77.4**	1.77 (-4.8%)
4	74.3	71.2	**77.3**	1.73 (-7.0%)
5	74.0	71.0	77.0	1.69 (-9.2%)

(e) recurrent refinement

Strategy	Loss	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m\uparrow$	FPS
photometric	MAST [28]	65.5	63.3	67.6	1.13
reconstruction	MAST [28]+$\mathcal{L}_{\text{Seg}}$	**69.0** (+3.5)	**66.4**	**71.6**	1.01
cycle-consistency	CRW [34]	67.6	64.6	70.6	1.86
tracking	CRW [34]+$\mathcal{L}_{\text{Seg}}$	**71.8** (+4.2)	**68.3**	**75.3**	1.77
contrastive	$\mathcal{L}_{\text{Corr}}$ (ours)	68.8	66.7	70.9	1.86
matching	$\mathcal{L}_{\text{Corr}}+\mathcal{L}_{\text{Seg}}$	**74.5** (+5.7)	**71.6**	**77.4**	1.77

(f) correspondence learning schema

Table 4. A set of ablative studies on $DAVIS_{17}$ [42] `val` (§4.2). The adopted settings are marked in red.

Thank you!