SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Fast Target-aware Learning for Few-shot Video Object Segmentation

In this supplementary material for reviewer, we present a derivation of the steepest descent iterations used in Section 3.3 of the research paper. Next, we present more details about the object-conditional RDL (removing discriminative learning) alteration compared in Section 4.1. More quantitative results are also given in this appendix.

Appendix A Derivation of the Steepest Descent Iterations

In this section, we derive the employed steepest decent iterations defined as Equation 8, 9 in the research paper. To simplify the derivation, we first convert the loss defined in Equation 6 into a matrix formulation. Concretely, $M_t * \tau$ can be written in matrix form as $\vec{M}_t\vec{\tau}$, where $\vec{M}_t \in \mathbb{R}^{1\times HWK}$ and $\vec{\tau} \in \mathbb{R}^{HWK\times HWC}$. Further, Y_t and W_t are vectorized as $\vec{Y}_t \in \mathbb{R}^{1\times HWC}$ and $\vec{W}_t \in \mathbb{R}^{HWC\times HWC}$, respectively. Note that, $\vec{W}_t$ is a diagonal matrix where the element-wise wights are arrayed on the diagonal. Accordingly, Equation 6 can be written as

$$\mathcal{L}_{tar} = \frac{1}{2N} \parallel \vec{W}_t(\vec{M}_t\vec{\tau} - \vec{Y}_t) \parallel^2 + \frac{\lambda}{2} \parallel \vec{\tau} \parallel^2 . \tag{A1}$$

For each iteration, the optimization can be expressed as $\vec{\tau} = \vec{\tau} - \alpha\vec{g}$ in the gradient direction $\vec{g}$ with step length α. Using the chain rule,

$$\begin{aligned} \vec{g} = \nabla_{\vec{\tau}}\mathcal{L}_{tar}(\vec{\tau}) &= \frac{1}{N}\vec{W}_t^{\mathrm{T}}\vec{M}_t^{\mathrm{T}}\vec{W}_t(\vec{M}_t\vec{\tau} - \vec{Y}_t) + \lambda\vec{\tau} \\ &= \mathrm{vec}(\frac{1}{N} M_t *^{\mathrm{T}} (W_t^{\cdot 2} \cdot (M_t * \tau - Y_t)) + \lambda\tau), \end{aligned} \tag{A2}$$

where vec denotes a vectorization operation, and $*^{\mathrm{T}}$ is the transposed convolution operation. Thus,

$$g = \frac{1}{N} M_t *^{\mathrm{T}} (W_t^{\cdot 2} \cdot (M_t * \tau - Y_t)) + \lambda\tau). \tag{A3}$$

Next, the optimal step length α can be obtained by minimizing $\mathcal{L}_{tar}$ in the computed gradient direction, written as

$$\alpha = \arg\min_{\alpha} \mathcal{L}_{tar}(\vec{\tau} - \alpha\vec{g}). \tag{A4}$$

Since the L2 loss $\mathcal{L}_{tar}$ is convex, it has a globally optimal solution when $\nabla_\alpha\mathcal{L}_{tar}(\vec{\tau} - \alpha\vec{g}) = 0$. Again using the chain rule,

$$\begin{aligned} &\nabla_\alpha\mathcal{L}_{tar}(\vec{\tau} - \alpha\vec{g}) = (\frac{\mathrm{d}(\vec{\tau} - \alpha\vec{g})}{\mathrm{d}\alpha})^{\mathrm{T}}\nabla_{(\vec{\tau}-\alpha\vec{g})}\mathcal{L}_{tar}(\vec{\tau} - \alpha\vec{g}) \\ &= (-\vec{g})^{\mathrm{T}}(\frac{1}{N}\vec{M}_t^{\mathrm{T}}\vec{W}_t^2(\vec{M}_t(\vec{\tau} - \alpha\vec{g}) - \vec{Y}_t) + \lambda(\vec{\tau} - \alpha\vec{g})) \\ &= (\vec{g})^{\mathrm{T}}(\alpha(\frac{1}{N}\vec{M}_t^{\mathrm{T}}\vec{W}_t^2\vec{M}_t\vec{g} + \lambda\vec{g}) - \vec{g}) \\ &= \alpha(\frac{1}{N} \parallel \vec{W}_t\vec{M}_t\vec{g} \parallel^2 + \lambda \parallel \vec{g} \parallel^2) - \parallel \vec{g} \parallel^2 = 0, \end{aligned} \tag{A5}$$

and, α can be computed as

$$\begin{aligned} \alpha &= \frac{\parallel \vec{g} \parallel^2}{\frac{1}{N} \parallel \vec{W}_t\vec{M}_t\vec{g} \parallel^2 + \lambda \parallel \vec{g} \parallel^2} \\ &= \frac{\parallel g \parallel^2}{\frac{1}{N} \parallel W_t \cdot (M_t * g) \parallel^2 + \lambda \parallel g \parallel^2}. \end{aligned} \tag{A6}$$

Email:

Appendix B Object-conditional RDL Alteration

In this section, we give the details about the compared RDL alteration without the employed discriminative learning. Similar to the work in [1, 2], we construct object-conditional dictionaries according to the ground truth. That is, we use K-means algorithm to obtain one individual dictionary $\mu_c = \{\mu^{c1}, ..., \mu^{cK}\}$ for each object with the c_{th} class label. Enabling the model to account for intra-class variations, we consider the most similar visual word encouraging each pixel to resemble only one relevant word. Thus, the similarity score between pixel x and the c_{th} object is computed as

$$s(\hat{y} = c|x) = \max_{k \in \{1,\dots,K\}} \exp(\cos(\mathcal{F}(x|\omega), \mu^{ck})), \tag{B1}$$

and the probability of assigning the pixel to that object is normalized as

$$p(\hat{y} = c|x) = \frac{s(\hat{y} = c|x)}{\sum_{c^\star=1}^{C} s(\hat{y} = c^\star|x)}, \tag{B2}$$

where C denotes the count of object classes.

Finally, we employ the same loss function defined in [1] for meta training, written as

$$\begin{aligned} \mathcal{L}_{meta} = &-\frac{1}{|X|} \sum_{x \in X} \log[p(\hat{y} = y|x)] \\ &- \frac{1}{|X|(C-1)} \sum_{x \in X} \sum_{c=1, c \neq y}^{C} \log[1 - p(\hat{y} = c|x)], \end{aligned} \tag{B3}$$

where the training set is given in the form of pair (x, y). Here, the first term is the standard pixel-wise cross-entropy loss and the second term further reduces the probability of incorrect classes. Note that, even though RDL is trained by the bilevel-optimization strategy, there is no needed $\mathcal{L}_{tar}$ definition for this alteration. Since the target model has been obtained, i.e., the constructed object-conditional dictionaries.

Appendix C More Quantitative Results

Table C1 Quantitative results on DAVIS-2016 validation set.

Method	ATD	$\mathcal{J}\&\mathcal{F}$(%)	$\mathcal{J}$(%)	$\mathcal{F}$(%)	Speed(s)
MSK [3]	IM	77.6	79.7	75.4	12.0
OnAVOS [4]	IM	85.5	86.1	84.9	13.0
PReMVOS [5]	IM	**86.8**	84.9	**88.6**	32.8
OSMN [6]	IM	73.5	74.0	72.9	0.14
RGMP [7]	IM	81.8	81.5	82.0	0.14
GC [8]	IM	86.6	**87.6**	85.7	0.04
AGAME [2]	IM	81.9	81.5	82.2	0.07
STM [9]	IM	86.5	84.8	88.1	0.16
CINM [10]	-	84.2	83.4	85.0	30.0
OSVOS [11]	-	80.2	79.8	80.6	10.0
OSVOS-S [12]	-	**86.6**	**85.6**	**87.5**	4.5
MLDVW [1]	-	82.1	81.5	82.7	0.25
FEELVOS [13]	-	81.7	80.3	83.1	0.45
FRTM [14]	-	81.7	-	-	0.05
Ours(fast)	-	78.3	77.5	79.1	0.04
Ours	-	82.4	81.3	83.5	0.12

ATD: Additional training datasets used in the respective method. Specifically, 'IM' indicates static image datasets. Ours(fast): Our approach without online adaptation.

Table C2 Per-sequence results on DAVIS-2016 validation set.

	blackswan	bmx-trees	breakdance	camel	car-roundabout	car-shadow	cows
$\mathcal{J}$(%)	94.2	56.8	68.4	82.7	96.1	94.3	94.1
$\mathcal{F}$(%)	98.7	73.2	72.1	84.3	92.3	98.4	97.7

	dance-twirl	dog	drift-chicane	drift-straight	goat	horsejump-high	kite-surf
$\mathcal{J}$(%)	82.5	93.9	69.8	80.6	88.6	84.1	64.3
$\mathcal{F}$(%)	83.9	94.2	75.4	78.1	91.8	92.4	71.0

	libby	motocross-jump	paragliding-launch	parkour	scooter-black	soapbox	Overall
$\mathcal{J}$(%)	78.7	77.0	63.1	92.1	77.5	87.9	81.3
$\mathcal{F}$(%)	85.4	63.9	56.2	94.8	78.1	88.2	83.5

For a comprehensive comparison, Table C1 also presents the results on DAVIS-2016 that is a subset of DAVIS-2017 [15], containing 30 training videos and 20 validation videos labeled with only a single object. Though our method doesn't outperform all the compared approaches, the proposed approach perform comparable to these state-of-the-art algorithms in terms of speed/accuracy tradeoffs. Especially, it obtains the third best rank among the methods do not using additional training data, where OSVOS-S [12] and CINM [10] benefited considerably from the time-consuming finetunning step. In conclusion, the proposed method learns a new object well in a fast way just from few-shot information. The per-sequence results on DAVIS-2016 and DAVIS-2017 are also detailed in Table C2, C3 for reference. Note that, the proposed method is evaluated on DAVIS-2016 by only training the model on the 30 DAVIS-2016 training sequences.

References

1 Behl H S, Najafi M, Arnab A, et al. Meta Learning Deep Visual Words for Fast Video Object Segmentation. In: Proceedings of the 2019 Conference on Neural Information Processing Systems Machine Learning for Autonomous Driving Workshop, 2019.

2 Johnander J, Danelljan M, Brissman E, et al. A Generative Appearance Model for End-To-End Video Object Segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8945-8954

3 Perazzi F, Khoreva A, Benenson R, et al. Learning Video Object Segmentation from Static Images. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.

Table C3 Per-sequence results on DAVIS-2017 validation set.

	bike-packing_1	bike-packing_2	blackswan	bmx-trees_1	bmx-trees_2	breakdance	camel	car-roundabout	car-shadow	cows	dance-twirl	dog	dogs-jump_1	dogs-jump_2	dogs-jump_3	drift-chicane
$\mathcal{J}(\%)$	68.9	74.7	95.1	40.9	67.8	74.2	83.6	97.2	96.0	94.7	84.1	94.1	21.1	54.8	87.8	76.3
$\mathcal{F}(\%)$	83.0	72.0	98.8	81.4	82.2	82.6	88.1	95.4	98.8	98.4	85.6	94.7	18.3	71.3	94.6	76.8

	drift-straight	goat	gold-fish_1	gold-fish_2	gold-fish_3	gold-fish_4	gold-fish_5	horsejump-high_1	horsejump-high_2	india_1	india_2	india_3	judo_1	judo_2	kite-surf_1	kite-surf_2
$\mathcal{J}(\%)$	82.3	88.1	65.8	68.9	68.4	68.0	75.6	80.7	69.4	59.7	54.2	66.2	84.6	80.9	38.0	29.9
$\mathcal{F}(\%)$	79.6	91.2	60.8	71.9	57.4	65.1	61.6	91.6	85.4	57.6	52.4	58.3	87.2	81.8	77.6	39.9

	kite-surf_3	lab-coat_1	lab-coat_2	lab-coat_3	lab-coat_4	lab-coat_5	libby	loading_1	loading_2	loading_3	mbike-trick_1	mbike-trick_2	motocross-jump_1	motocross-jump_2	paragliding-launch_1	paragliding-launch_2
$\mathcal{J}(\%)$	75.9	18.9	20.8	68.8	41.7	64.0	84.1	95.1	82.1	47.0	64.2	78.5	50.7	69.1	80.6	66.7
$\mathcal{F}(\%)$	93.0	58.4	55.8	63.5	39.3	60.9	87.1	92.4	85.0	56.9	75.2	80.6	62.2	60.6	91.0	89.7

	paragliding-launch_3	parkour	pigs_1	pigs_2	pigs_3	scooter-black_1	scooter-black_2	shooting_1	shooting_2	shooting_3	soapbox_1	soapbox_2	soapbox_3	Overall
$\mathcal{J}(\%)$	30.1	92.0	64.5	55.1	94.5	63.7	78.0	38.9	72.6	56.0	81.1	78.8	77.4	68.5
$\mathcal{F}(\%)$	67.4	94.9	65.9	72.6	82.7	78.9	77.5	33.5	69.9	71.9	81.9	84.0	88.2	74.8

4 Voigtlaender P, Leibe B. Online Adaptation of Convolutional Neural Networks for Video Object Segmentation. In: Proceedings of the 2017 British Machine Vision Conference, 2017.

5 Luiten J, Voigtlaender P, Leibe B. PReMVOS:Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. In: Proceedings of the 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops, 2018.

6 Yang L, Wang Y, Xiong X, et al. Efficient Video Object Segmentation via Network Modulation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6499-6507

7 Oh S W, Lee J, Sunkavalli K, et al. Fast Video Object Segmentation by Reference-Guided Mask Propagation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7376-7385

8 Li Y, Shen Z R, Shan Y. Fast Video Object Segmentation using the Global Context Module. In: Proceedings of the 2020 European Conference on Computer Vision, 2020.

9 Oh S W, Lee J, Xu N, et al. Video Object Segmentation Using Space-Time Memory Networks. In: Proceedings of the 2019 IEEE International Conference on Computer Vision, 2019. 9225-9234

10 Bao L C, Wu B Y, Liu W. CNN in MRF: Video Object Segmentation via Inference in A CNN-Based Higher-Order Spatio-Temporal MRF. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.

11 Caelles S, Maninis K K, Pont-Tuset J, et al. One-Shot Video Object Segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5320-5329

12 Maninis K, Caelles S, Chen Y, et al. Video Object Segmentation without Temporal Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 41(6):1515-1530

13 Voigtlaender P, Chai Y, Schroff F, et al. FEELVOS: Fast End-To-End Embedding Learning for Video Object Segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019. 9473-9482

14 Robinson A, Lawin A J, Danelljan M, et al. Learning Fast and Robust Target Models for Video Object Segmentation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020. 7404-7413

15 Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 DAVIS Challenge on Video Object Segmentation. arXiv, 2017. 1704.00675