Implicit Motion Handling for Video Camouflaged Object Detection – Supplementary Material –

[*]Xuelian Cheng[1], [*]Huan Xiong[3], [†]Deng-Ping Fan[4], Yiran Zhong[6,7],
Mehrtash Harandi[1,8], Tom Drummond[1], Zongyuan Ge[1,2,5]
[1]Faculty of Engineering, Monash University, [2]eResearch Centre, Monash University
[3]Mohamed bin Zayed University of Artificial Intelligence, [4]CVL, ETH Zurich,
[5]Airdoc Research Australia, [6]SenseTime Research, [7]Shanghai AI Laboratory, [8]Data61, CSIRO

Abstract

In this supplementary material, we provide our short-term correlation pyramid details, semi-supervised training procedure, training details and dataset curation.

1. Short-term Correlation Pyramid Details

To enable the network to learn detailed information, a correlation pyramid $\mathbf{C}^i, i \in \{2,3,4\}$ is construct by incorporating multi-scale features. Thus for a sequence of frame features $\{\mathcal{F}_\theta(\mathbf{I}_t), \mathcal{F}_\theta(\mathbf{I}_{t+1})\} \in \mathbb{R}^{C\times H/2^{i+1}\times W/2^{i+1}}$, our short-term correlation pyramid can be denoted as $\mathbf{C}^i(\mathbf{I}_t, \mathbf{I}_{t+1}) \in \mathbb{R}^{H/2^{i+1}\times W/2^{i+1}\times H/2^{i+1}\times W/2^{i+1}}$. It outputs an aggregated feature map $f'^{(i)}_{t\leftarrow t+1}(\mathbf{I}_{t\leftarrow t+1})$ at the pyramid scale $i, i \in \{2,3,4\}$, which has the same dimension as the reference frame feature $\mathcal{F}_\theta(\mathbf{I}_t)$. For downsampled neighboring frames, we set the $k = \{2,4,8\}$ with max-pooling kernels of growing size. We also repeat the correlative aggregation once on every other neighboring frame. In this way, we obtain aggregated feature maps $f'^{(i)}_{t\leftarrow t+1}(\mathbf{I}_{t\leftarrow t+2})$.

2. Semi-supervised Training Procedure

As the annotations are provided in the form of dense segmentation masks for every five frames, we adopt a bi-directional consistency check strategy to generate pseudo masks for unlabelled frames. Given five consecutive frames $\{\mathbf{I}_t, \mathbf{I}_{t+1}, \mathbf{I}_{t+2}, \mathbf{I}_{t+3}, \mathbf{I}_{t+4}\}$ and labelled ground-truth $\mathbf{gt}_t$, we first estimate forward and backward optical flow fields between frame $\mathbf{I}_t$ and $\mathbf{I}_{t+n}, n \in [1,4]$. Then we can produce the warped ground-truth $\hat{\mathbf{gt}}_{t+n}$ with the inverse warping from ground-truth $\mathbf{gt}_t$.

1.Flow Estimation. We take the ground-truth mask of the reference frame $\mathbf{I}_t$ as an example, to generate pseudo ground-truth of its immediate following frame $\mathbf{I}_{t+1}$. The optical flow estimation module[1] $\mathcal{O}$ takes $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ and predicts the optical flow field:

$$\mathbf{u}^x_{t,t+1},\ \mathbf{u}^y_{t,t+1} = \mathcal{O}(\mathbf{I}_t, \mathbf{I}_{t+1}), \tag{1}$$

where $\mathbf{u}^x_{t,t+1}$ and $\mathbf{u}^y_{t,t+1}$ denote the x, y components of the estimated flow field, respectively. The flow field maps each pixel $(x,\ y)$ in $\mathbf{I}_{t+1}$ to its corresponding coordinates $(x',\ y') = (x + \mathbf{u}^x_{t,t+1}(x),\ y + \mathbf{u}^y_{t,t+1}(y))$ in $\mathbf{I}_t$.

(a) Forward (b) Backward (c) Bi-directional

Figure 1. Illustration of forward-backward consistency check. After bi-directional check, undesirable ghosting artifacts, *i.e.,* the nose (red box) of the elephant in forward direction and the tail (blue box) in backward direction, and occlusions can be effectively removed.

2.Forward/Backward Pseudo Labeling. Given the forward optical flow sequences $(\mathbf{flow}_t, \mathbf{flow}_{t+n}), n \in 1,2,3,4$, we can obtain the aligned neighboring frame $\hat{\mathbf{gt}}_{t+n}$ by a warping interpolation on $\mathbf{gt}_t$ using the mapped coordinates. After repeating the explicit alignment step for the preceding frame, we acquire the sequence of warped input frames $\{\mathbf{gt}_t, \hat{\mathbf{gt}}_{t+1}, \hat{\mathbf{gt}}_{t+2}, \hat{\mathbf{gt}}_{t+3}, \hat{\mathbf{gt}}_{t+4}\}$. The backward pseudo ground-truth sequences are obtained by performing warping ground-truth masks with backward optical flows in the reverse order.

3.Bidirectional Consistency Check. To identify valid masks, we adopt forward-backward consistency check

[1]In practice, we make use of RAFT [4] to obtain the optical flow.

to eliminate inconsistent regions. Under the forward-backward consistency assumption [3], traversing flow vector forward and then backward should arrive at the same position. We mark pixels as invalid whenever this constraint is violated. As shown in Figure 1, the invalid regions emphasized by the orange boxes are marked as background.

3. Training Details

We implement both long-term and short-term architecture in PyTorch. The input images are resized to 352×352. We train the short-term architecture with a batch size of 8 on an NVIDIA V100 GPU and use Adam optimizer with initial learning rate of 1e-4, decreasing every 50k iterations. For the long-term optimization, our model takes 10 frames as the input at one time with the frame sampling rate 1. For our pseudo ground-truth generation, we exploit RAFT [4] as the optical flow estimation module and pre-trained weights on Sintel dataset [1].

4. Data Curation

- **Remove Invalid Scenes.** We first select and exclude scenarios in that animals are obvious and easy to identify from the background at our first glance. After cleaning the dataset, our new subset includes 87 video sequences, 22,939 frames in total.
- **Segmentation Masks.** For annotations, we further provide accurate human-labeled segmentation masks for every five frames. Thus our GT consists of two formats, that is 4,691 bounding box annotations as well as 4,691 pixel-level masks.
- **Pseudo Masks.** We use a bidirectional optical flow-based strategy to generate the pseudo GT masks, refer to the *SM*. Note that these pseudo masks still contain motion estimation errors, requiring algorithms to have the capability to handle noise labels when using them.
- **Dataset Split.** The whole dataset is split into 71 sequences, 19,313 frames for training, and 16 sequences, 3,626 frames selected for testing. The summary of each sub-sequence distribution could be found in Fig. 3.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

[2] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020.

[3] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[5] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *ICCV*, 2019.

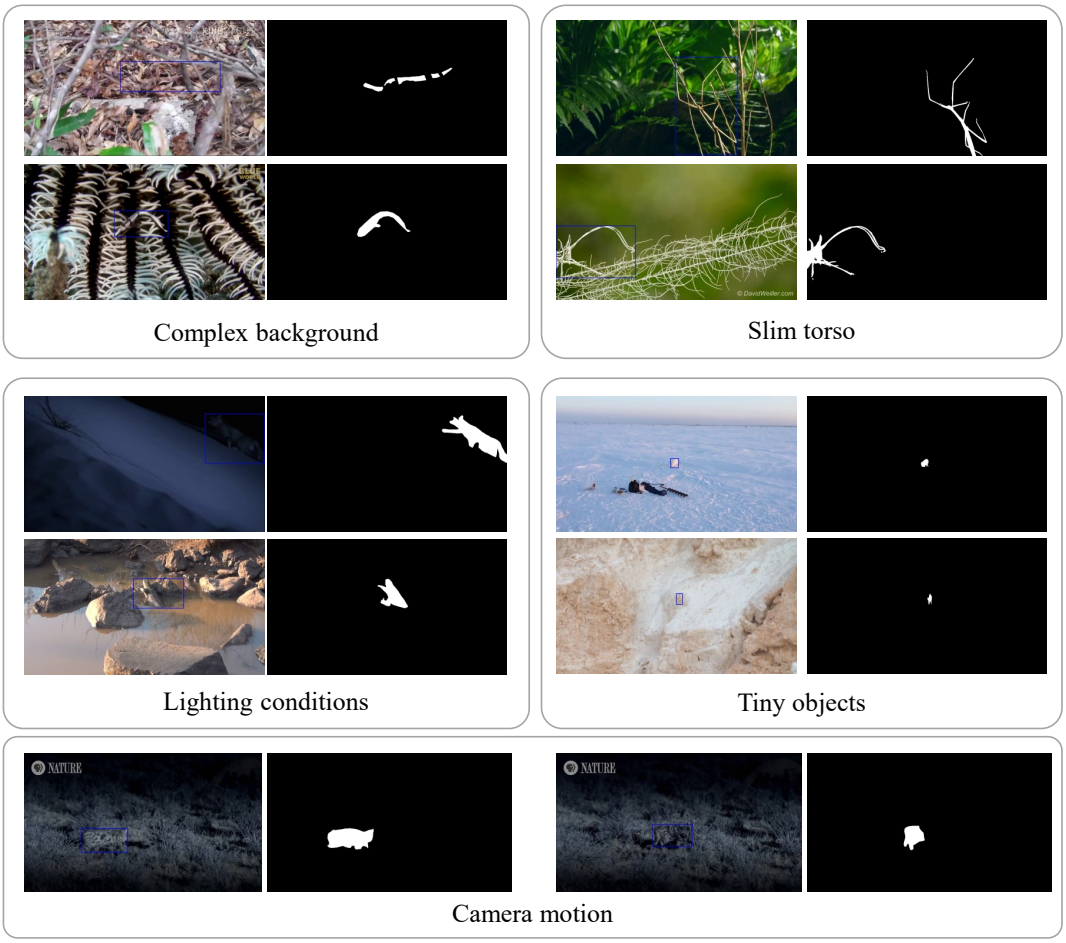

Figure 2. Representative samples from MoCA-Mask. The dataset is quite challenging including diverse scenes, suash as various lighting conditions, *i.e.,* dark and sunny, complex background, camera motions, small ratio of animals and tiny body structures, such as slim torso /limbs.

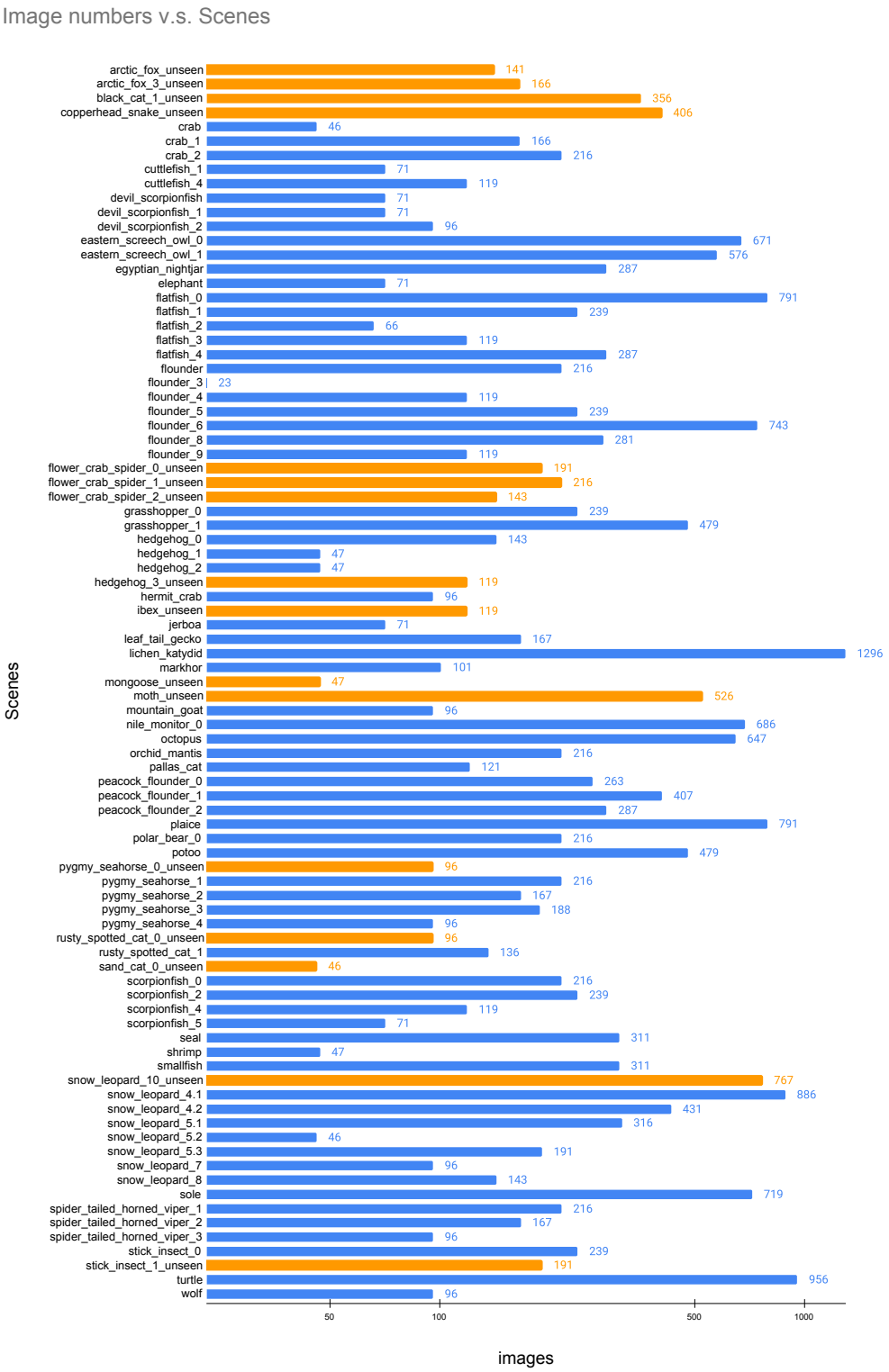

Figure 3. Summary for training and test set distribution. Our MoCA-Mask dataset includes 87 video sequences in total, in which 16 sequences were tagged as "unknow" (colored in orange). This split is used to validate the sensitivity of different models on novel samples. Zoom-in for details.

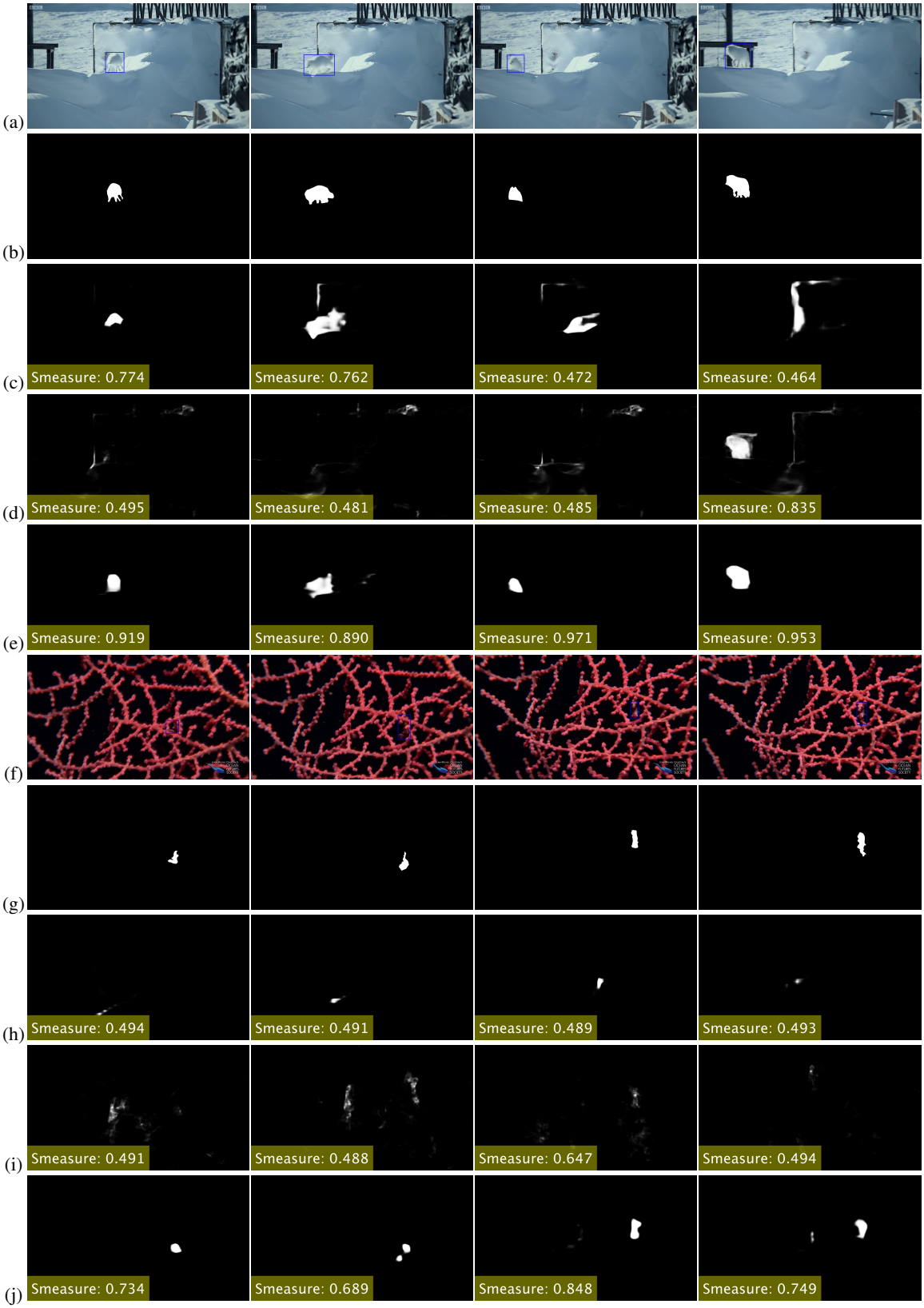

Figure 4. Comparison of our proposed network with two top-performing baselines on MoCA-Mask test dataset. Example squences of each row means: (a) (f) Frames, (b) (g) GT, (c) (h) SINet [2], (d) (i) RCRNet [5], (e) (j) SLT-Net (Ours).