The Making and Braking of Camouflage
Supplementary Material

Hala Lamdouar[1] Weidi Xie[1,2] Andrew Zisserman[1]
[1]Visual Geometry Group, University of Oxford [2]CMIC, Shanghai Jiao Tong University
{lamdouar,weidi,az}@robots.ox.ac.uk

We first provide examples from all the camouflage datasets included in our study in Section A. Next, we present a class analysis in Section B. In Section C, we present more results of ranking camouflage datasets and compare the rankings obtained by the different scores in terms of kendall-τ metric [6]. Then in Section D, we present examples of the camouflage image and mask pairs generated via our proposed Generative Adversarial Networks setting with and without training with the $\mathcal{L}_{\mathcal{F}}$ loss term as well as examples of the generated synthetic camouflage sequences. Finally, in Section E, we provide a detailed description of our model's architecture and show qualitative results of our motion segmentation model trained on the synthetic dataset.

Contents

A. Camouflage Datasets

We provide random samples of the camouflage benchmarks included in our study in Fig. 1. Further examples of our synthetically generated camouflage datasets are shown in Fig. 8 and Fig. 9.

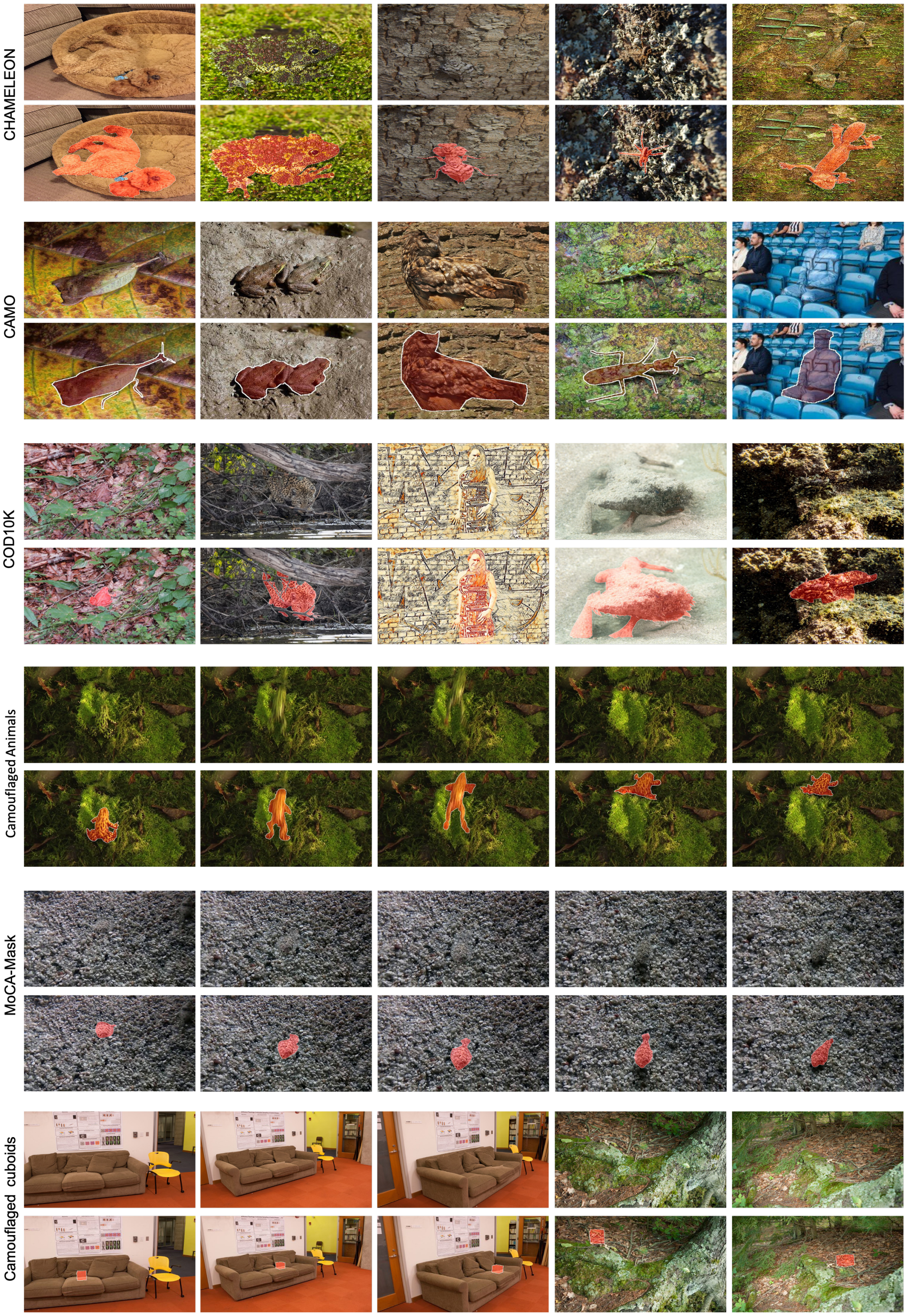

Figure 1: **Randomly sampled examples from the camouflage datasets included in our work**. For the video datasets, Camouflaged Animals and MoCA-mask, we show an example sequence. For the multiple-view dataset Camouflaged cuboids (bottom), we show example views from two scenes of the 4-view texture synthesis method from [5].

B. Class Analysis on COD10K

We compute our scores for each of the classes from COD10K [4] and report the results in Tab. 1. Overall, the highest scored examples, according to our S_α, were from the Aquatic class. We found the Stingaree subclass camouflage to be the most effective in terms of background matching, S_{R_f}=0.885, and the worst in terms of boundary visibility. The Best subclass for the boundary score is the Lion S_b=0.502 followed by the caterpillar S_b=0.492.

Classes	$S_{R_f} \uparrow$	$S_b \uparrow$	$S_\alpha\uparrow$
Aquatic	**0.711**	0.427	**0.612**
Terrestrial	0.599	**0.443**	0.545
Flying	0.668	0.426	0.584
Amphibian	0.678	0.432	0.592
Other	0.606	0.439	0.548
Best subclass	Stingaree(0.885)	Lion (0.502)	Flounder (0.725)
Worst subclass	Ant (0.294)	Stingaree (0.391)	Ant (0.347)

Table 1: Results per class from COD10K [4]. S_{R_f} stands for the reconstruction fidelity score, S_b stands for the boundary score and $S_\alpha$stands for the combined perceptual score.

C. Ranking Camouflage Images

We present, in this section, examples of highest and lowest scored images and sequences from the camouflage datasets with respect to the different proposed camouflage assessment scores and we further report their pairwise correlations in terms of the kendall-τ metric [6].

C.1. Correlation of the proposed score rankings

In this section, we present the results of comparison of the ranking obtained using our scores in terms of kendall-τ metric, on CAMO [9] in Tab. 2, COD10K [4] in Tab. 3, MoCA-Mask [8, 3] in Tab. 4, CHAMELEON [10] in Tab. 6, Camouflaged Animals [1] in Tab. 5. Note that we have inverted the rankings obtained with the $d_\mathcal{F}$ distance for consistency with the other score rankings. With our choice for $\alpha = 0.35$, therefore privileging reconstruction fidelity over boundary score, $S_\alpha$is correlated with S_{R_f}. Our proxi score $d^2_\mathcal{F}$ shows correlation with $S_\alpha$on COD10k and MoCA-Mask with the highest correlation on Camouflaged Animals $\tau = 0.245$.

Rankings	S_{R_f}	S_b	S_α	$d^2_\mathcal{F}$
S_{R_f}	1	0.012	0.035	-0.014
S_b	0.012	1	0.012	0.002
S_α	0.035	0.012	1	-0.040
$d^2_\mathcal{F}$	-0.014	0.002	-0.040	1

Table 2: Kendall-τ of rankings of CAMO [9]

Rankings	S_{R_f}	S_b	S_α	$d^2_\mathcal{F}$
S_{R_f}	1	0.0	0.004	0.003
S_b	0.0	1	0.007	-0.010
S_α	0.004	0.007	1	0.013
$d^2_\mathcal{F}$	0.003	-0.010	0.013	1

Table 3: Kendall-τ of rankings of COD10K [4]

Rankings	S_{R_f}	S_b	S_α	$d^2_{\mathcal{F}}$
S_{R_f}	1	0.083	0.102	0.097
S_b	0.083	1	0.081	0.057
S_α	0.102	0.081	1	0.091
$d^2_{\mathcal{F}}$	0.097	0.057	0.091	1

Table 4: Kendall-τ of rankings of MoCA-Mask [8, 3]

Rankings	S_{R_f}	S_b	S_α	$d^2_{\mathcal{F}}$
S_{R_f}	1	0.291	0.349	0.229
S_b	0.291	1	0.287	0.343
S_α	0.349	0.287	1	0.245
$d^2_{\mathcal{F}}$	0.229	0.343	0.245	1

Table 5: Kendall-τ of rankings of Camouflaged Animals [1]

Rankings	S_{R_f}	S_b	S_α	$d^2_{\mathcal{F}}$
S_{R_f}	1	-0.032	0.067	-0.084
S_b	-0.032	1	-0.057	-0.076
S_α	0.067	-0.057	1	-0.015
$d^2_{\mathcal{F}}$	-0.084	-0.076	-0.015	1

Table 6: Kendall-τ of rankings of CHAMELEON [10]

C.2. Example of ranked images

We present the rankings of COD10K in Fig. 2 and Fig. 3, for MoCA-Mask in Fig. 4 and Fig. 5 and for CAMO in Fig. 6 and Fig. 7.

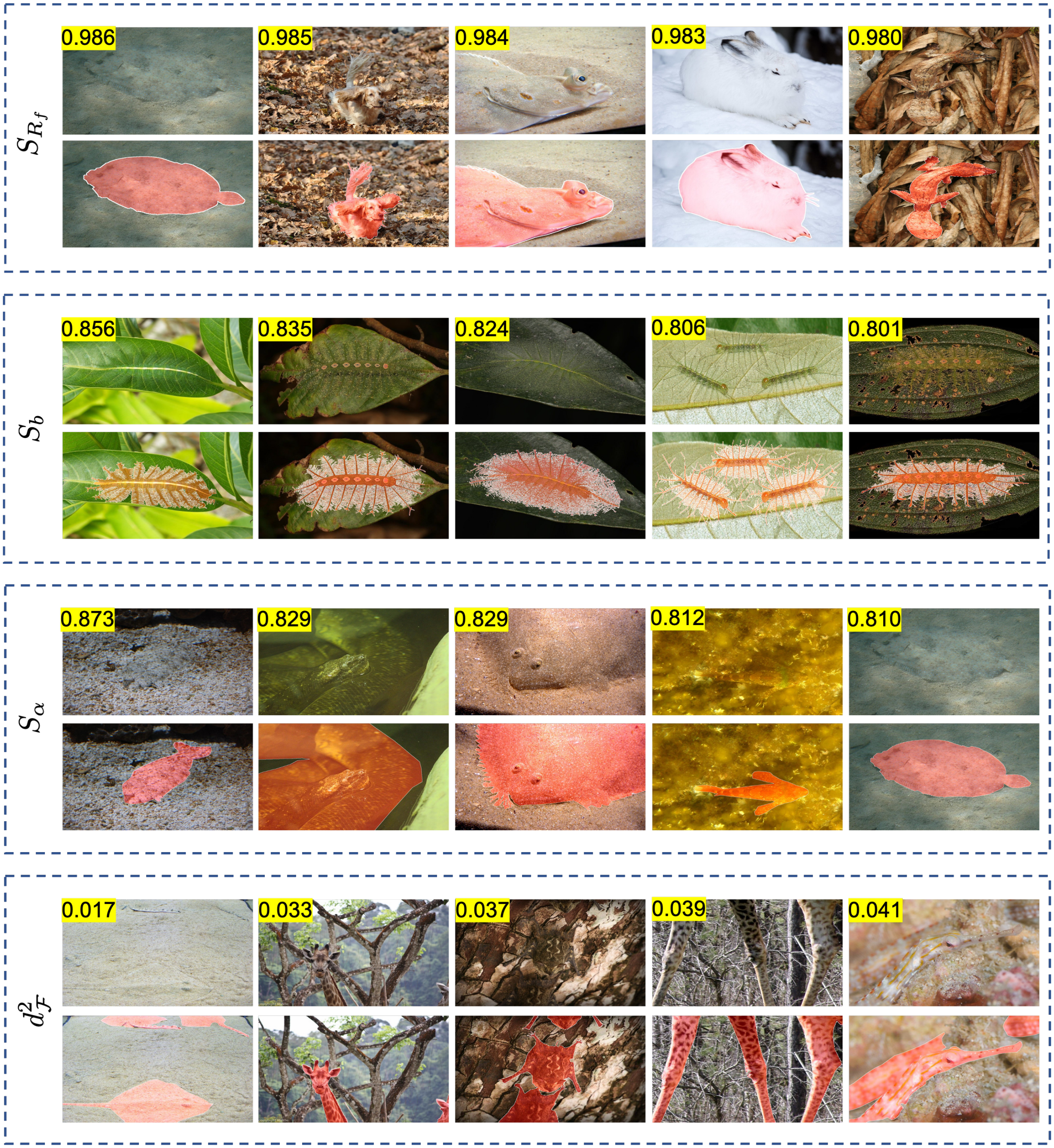

Figure 2: **Top scored examples from COD10K.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each image, we show the corresponding groundtruth mask and the computed score.

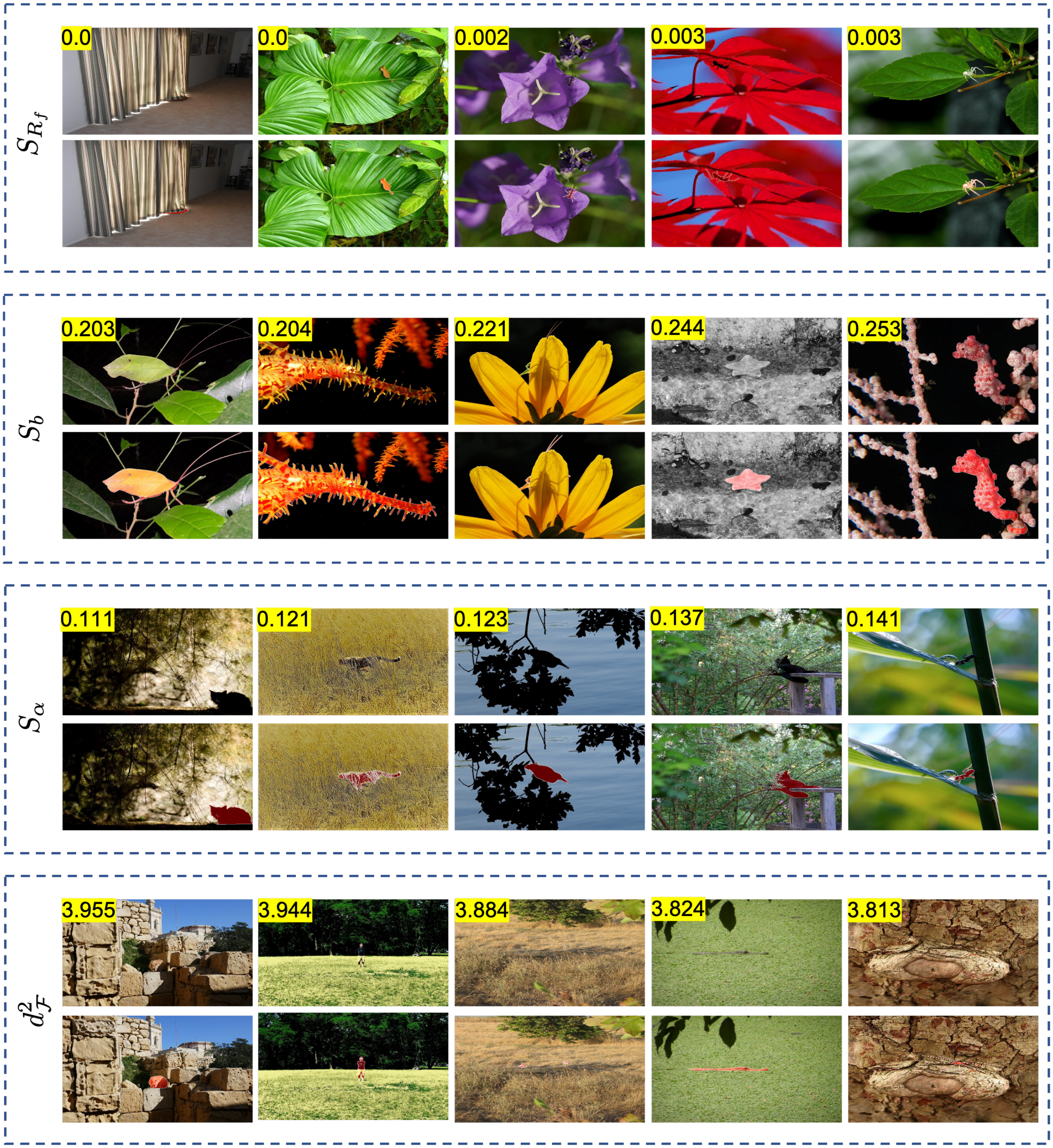

Figure 3: **Lowest scored examples from COD10K.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each image, we show the corresponding groundtruth mask and the computed score.

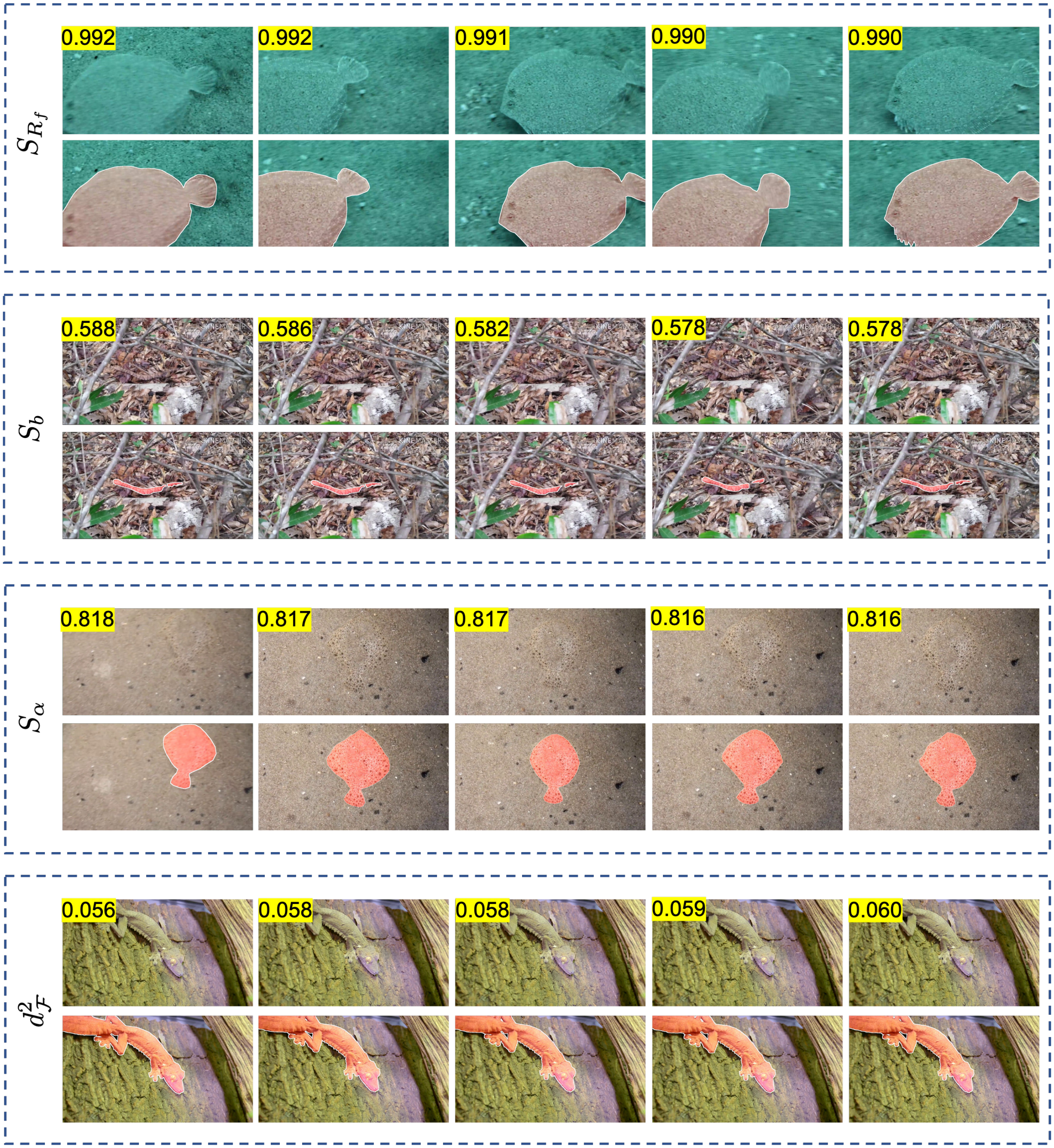

Figure 4: **Top scored examples from MoCA-Mask.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each frame, we show the corresponding groundtruth mask and the computed score.

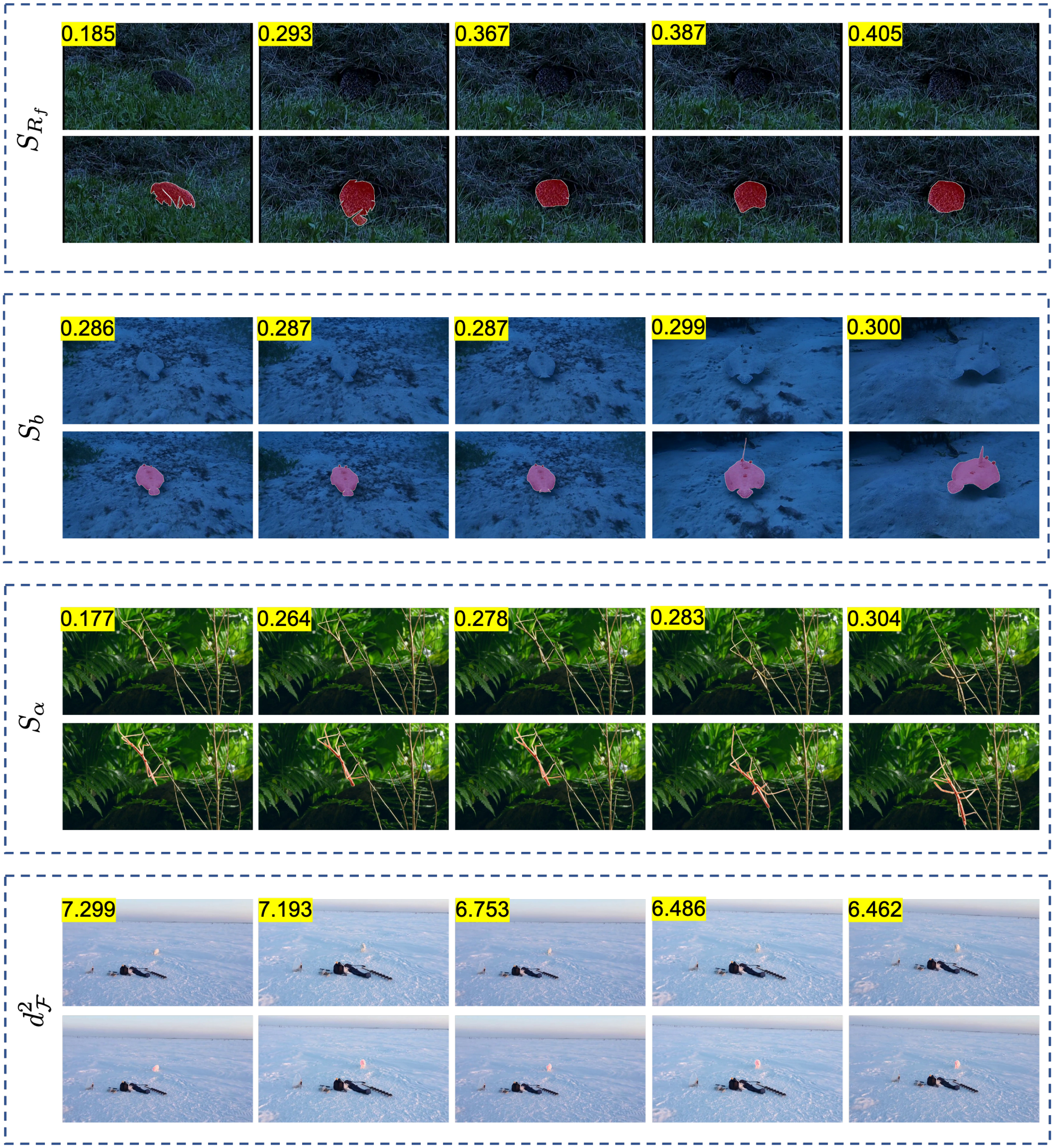

Figure 5: **Lowest scored examples from MoCA-Mask.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each frame, we show the corresponding groundtruth mask and the computed score.

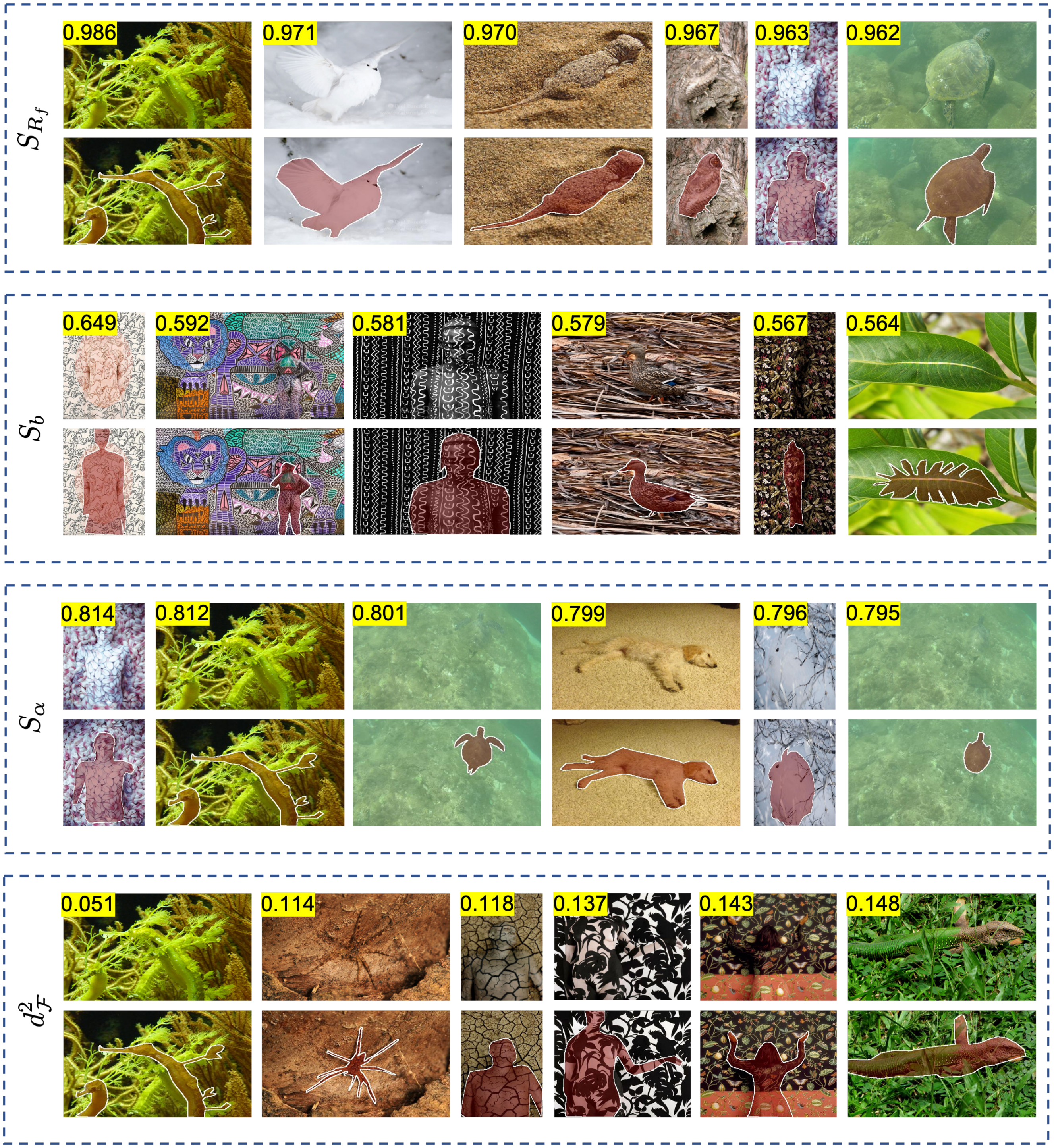

Figure 6: **Top scored examples from CAMO.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each image, we show the corresponding groundtruth mask and the computed score.

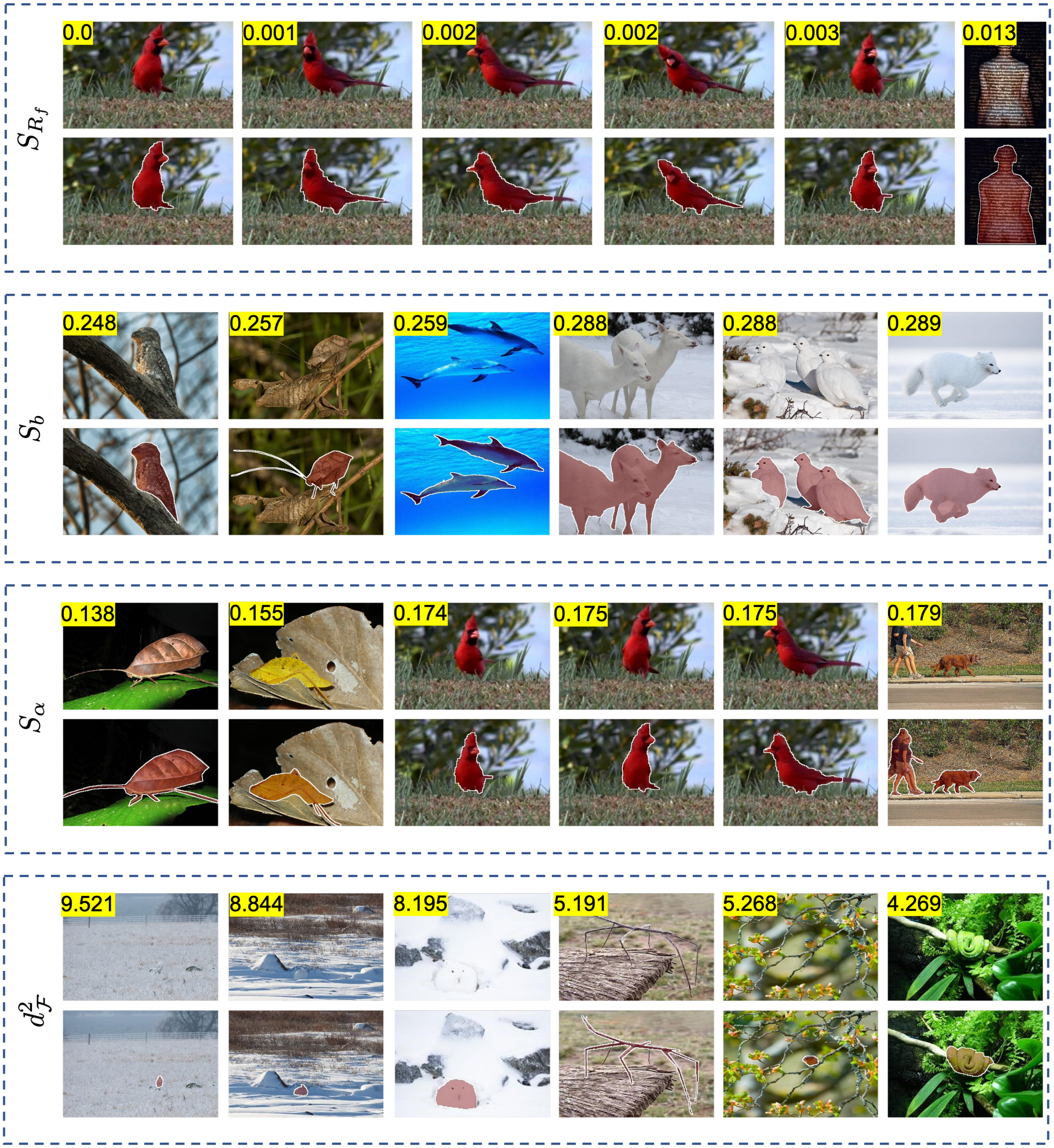

Figure 7: **Lowest scored examples from CAMO.** From top to bottom: S_{Rf}, S_b, S_α and $d^2_{\mathcal{F}}$ rankings. For each frame, we show the corresponding groundtruth mask and the computed score. While CAMO is a single image dataset, we found multiple images of a red bird with low S_{R_f} and $S_\alpha$scores. Note that these are not part of a sequence.

D. Synthetic Camouflage Dataset

In this section, we present samples of the synthetically generated camouflage images using our trained generative model and we provide examples of the created video sequences.

D.1. Example of generated image samples

We present in Fig 8 an example of image and mask samples created using our trained generator. When comparing generated samples, we notice a significant improvement of the visual blending of the animal with its background as a result of training the generator with additional intra-image Fréchet loss term.

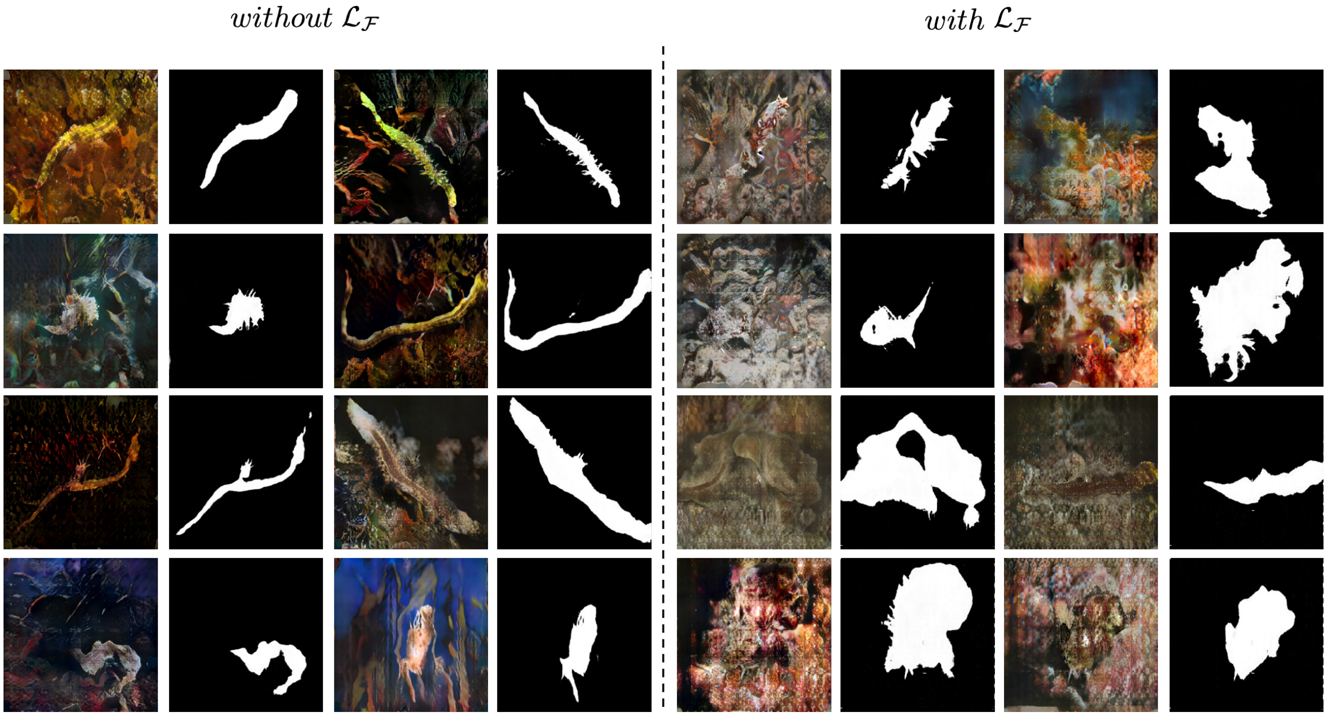
$without\ \mathcal{L}_{\mathcal{F}}$
$with\ \mathcal{L}_{\mathcal{F}}$

Figure 8: **Generated camouflage image and mask samples**: For the images on the right, the generator was trained using additional $\mathcal{L}_{\mathcal{F}}$ auxiliary loss.

D.2. Example of generated sequences

Fig. 9 presents examples of generated camouflaged animals sequences. We use translational motion and incorporate static subsequences.

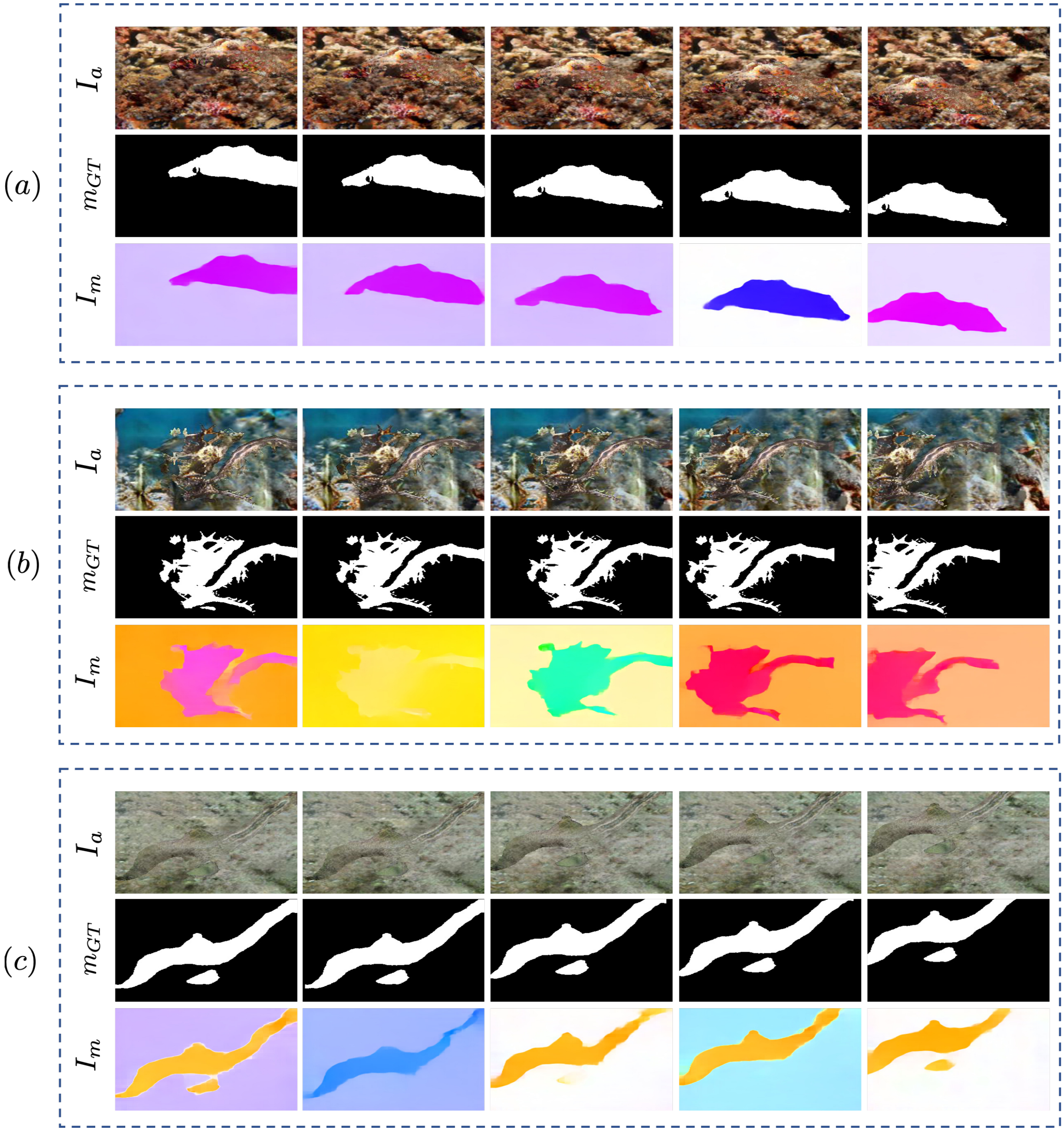

Figure 9: **Example of generated camouflage sequences.** I_a denotes a sequence of RGB images and I_m the corresponding optical flow sequence.

E. Moving Camouflaged Animal Segmentation

E.1. Model architecture

We detail our motion network architecture in Tab. 7. We build on the motion segmentation architecture [7] that takes a sequence $T = 5$ of optical flow frames and uses spatio-temporal self-attention to learn motion features. The output of the transformer encoder is fed to a pixel decoder with skip connections and further aggregated with motion features via a transformer decoder described in Tab. 8. The adopted architecture for the appearance encoder, *i.e.* SINetV2 [4] is described in Tab. 8. Based on a ResNet50 backbone, this architecture incorporates three levels of resolution: Low-level features ($d = 512$), Mid-level-features ($d = 1024$), and High level features ($d = 2048$). Each feature range is processed by a residual convolutional block. Note that we have omitted the kernel sizes for simplicity. In each residual block, kernel sizes are gradually increased $k \in \{1, 3, 5, 7\}$. To aggregate the appearance and motion features, we use a Transformer Decoder with learnable query features and masked attention [2]. The final predicted mask is obtained by multiplying the outputs of linear_2 and conv_2.

	stage	operation	output sizes
	input	–	$T \times 3 \times 256 \times 256$
Motion Encoder	conv1	$[3 \times 3, 64] \times 2$	$T \times 64 \times 256 \times 256$
	mp1	maxpool, stride = 2	$T \times 64 \times 128 \times 128$
	conv2	$[3 \times 3, 128] \times 2$	$T \times 128 \times 128 \times 128$
	mp2	maxpool, stride = 2	$T \times 128 \times 64 \times 64$
	conv3	$[3 \times 3, 256] \times 2$	$T \times 256 \times 64 \times 64$
	mp3	maxpool, stride = 2	$T \times 256 \times 32 \times 32$
	conv4	$[3 \times 3, 512] \times 2$	$T \times 512 \times 32 \times 32$
	mp4	maxpool, stride = 2	$T \times 512 \times 16 \times 16$
	conv5	$[3 \times 3, 512] \times 2$	$T \times 512 \times 16 \times 16$
Transformer Encoder	t_{pos}	Embedding $[T \times 512]$	$512 \times T$
	x_{pos}	Embedding $[16 \times 512]$	512×16
	y_{pos}	Embedding $[16 \times 512]$	512×16
	transEnc	input $= 512$, n_l=3 n_h=8,fwd=1024	$(T \times 16 \times 16) \times 512$
Decoder	conv1	$[3 \times 3, 512] \times 2$	$T \times 512 \times 16 \times 16$
	$\text{conv}^T 1$	$3 \times 3, 256, \text{stride} = 2$	$T \times 256 \times 32 \times 32$
	conv2	$[3 \times 3, 256] \times 2$	$T \times 256 \times 32 \times 32$
	$\text{conv}^T 2$	$3 \times 3, 128, \text{stride} = 2$	$T \times 128 \times 64 \times 64$
	conv3	$[3 \times 3, 128] \times 2$	$T \times 128 \times 64 \times 64$
	$\text{conv}^T 3$	$3 \times 3, 64, \text{stride} = 2$	$T \times 64 \times 128 \times 128$
	conv4	$[3 \times 3, 64] \times 2$	$T \times 64 \times 128 \times 128$
	$\text{conv}^T 4$	$3 \times 3, 64, \text{stride} = 2$	$T \times 64 \times 256 \times 256$

Table 7: **Architecture of the Motion Network.** In this work, we adopt a ConvNet as motion encoder, and use Transformer Encoder for aggregating the motion temporal information along $T = 5$ frames, followed by ConvNet-based decoder to recover the resolution, similarly to [7]. Unless specified otherwise, all convolution layers have stride $= 1$ and padding $= 1$.

	stage	operation	output sizes
	input	–	$T \times 3 \times 256 \times 256$
Appearance Encoder	Backbone	ResNet50	$T \times 512 \times 32 \times 32$ $T \times 1024 \times 16 \times 16$ $T \times 2048 \times 8 \times 8$
	res_block_1	$\begin{bmatrix}\text{conv}, 32\end{bmatrix} \times 15$	$T \times 32 \times 32 \times 32$
	res_block_2	$\begin{bmatrix}\text{conv}, 32\end{bmatrix} \times 15$	$T \times 32 \times 16 \times 16$
	res_block_3	$\begin{bmatrix}\text{conv}, 32\end{bmatrix} \times 15$	$T \times 32 \times 8 \times 8$
	conv_block_1	$\begin{bmatrix} \begin{bmatrix}\text{Upsample} \\ \text{conv}, 32\end{bmatrix} \times 4 \\ \begin{bmatrix}\text{conv}, 64\end{bmatrix} \times 2 \\ \begin{bmatrix}\text{conv}, 96\end{bmatrix} \times 2 \\ \text{conv}, 1 \end{bmatrix} \times 1$	$T \times 1 \times 32 \times 32$
	conv_block_2	$\begin{bmatrix} \begin{bmatrix}\text{Upsample} \\ \text{conv}, 32 \\ \text{conv}, 1\end{bmatrix} \times 3 \end{bmatrix} \times 3$	$T \times 4 \times 256 \times 256$
Transformer Decoder	conv_1_1	$[1 \times 1, 512] \times 3$	$T \times 512 \times 256 \times 256$
	transDec	input $= 512$, n$_l$=3 n$_h$=8,fwd=2048	512×1
	query$_{feat}$	Embedding $[1 \times 512]$	512×1
	query$_{pos}$	Embedding $[1 \times 512]$	512×1
	linear	$[512] \times 2$	512×1
	linear_2	[256]	256×1
	conv_2	$[3 \times 3, 256]$	$T \times 256 \times 256 \times 256$

Table 8: **Architecture of the appearance Network and aggregation.** We adopt a SINetV2 [4] architecture for encoding appearance. For aggregating the appearance and motion features, we use a Transformer Decoder with learnable query features and masked attention [2]. The final predicted mask is obtained by multiplying the outputs of linear_2 and conv_2.

E.2. Qualitative results

We show in Fig. 10 qualitative results of our motion segmentation network. Our model learns motion and appearance cues from the synthetically generated dataset in order to segment the moving camouflaged animal. Note that our model is able to provide accurate segmentations even in the presence of degraded optical flow, *e.g*. the first sequence, or static animal with respect to the background, *e.g*. the momentarily static stick insect depicted in the second sequence and the partially static flower crab spider in the third sequence.

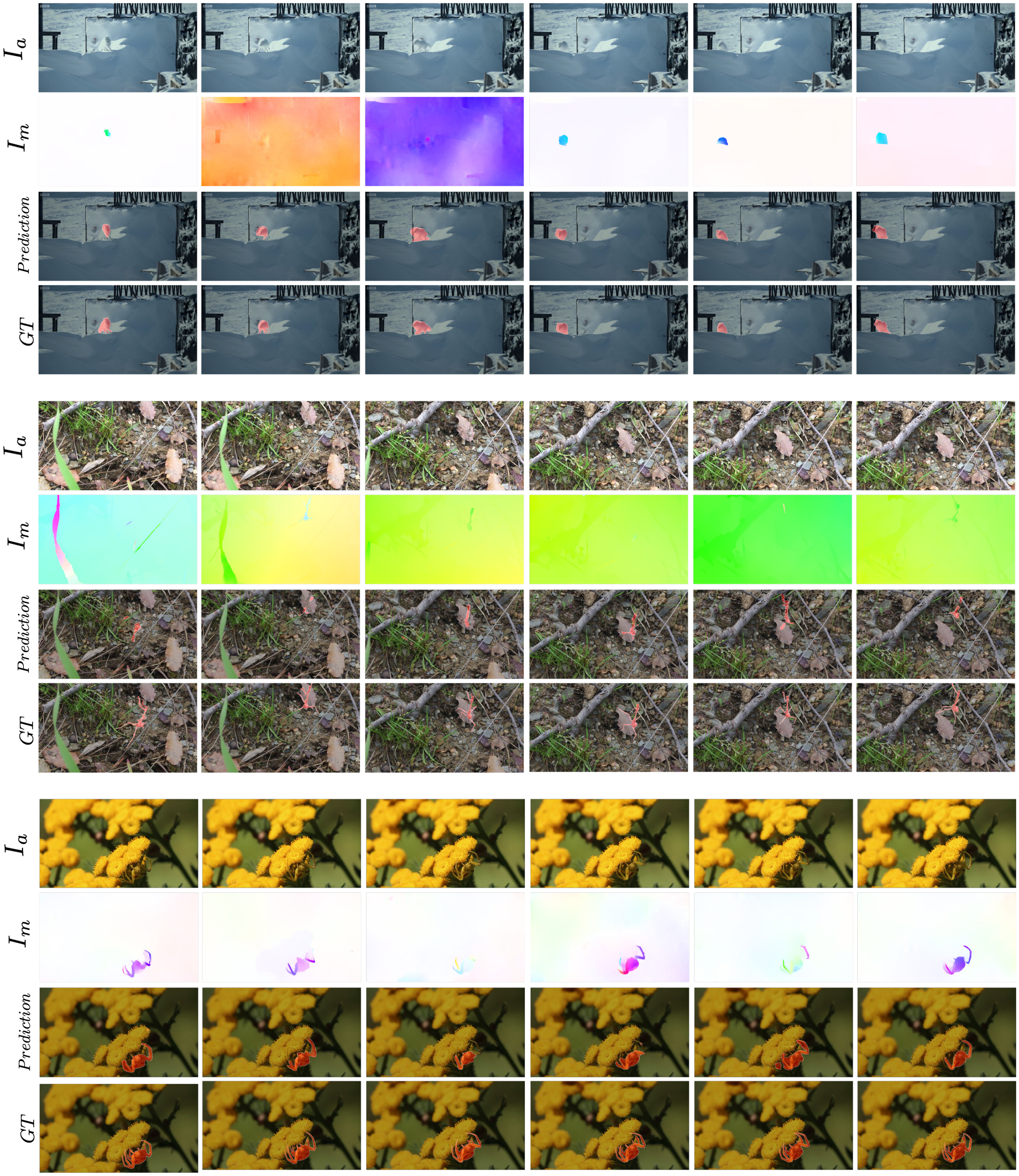

Figure 10: **Results of our motion segmentation model on sequences from the test subset of MoCA-Mask**. Our model takes as input a sequence of RGB images denoted I_a and a sequence of optical flow I_m and predicts segmentation masks of the moving camouflaged animal.

References

[1] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, pages 433–449, 2016. 3, 4

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 13, 14

[3] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *CVPR*, pages 13864–13873, 2022. 3, 4

[4] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 13, 14

[5] Rui Guo, Jasmine Collins, Oscar de Lima, and Andrew Owens. Ganmouflage: 3d object nondetection with texture fields. In *arXiv preprint arXiv:2201.07202*, 2022. 2

[6] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 1, 3

[7] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. Segmenting invisible moving objects. In *Proc. BMVC*, 2021. 13

[8] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proc. ACCV*, 2020. 3, 4

[9] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, pages 45–56, 2019. 3

[10] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2018. URL: https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/. 3, 4