Supplementary Materials: "Multi-modal Salient Object Detection via a Unified Diffusion Model"

In this supplementary material, we provide additional implementation details and experimental results. Our code is publicly available at: https://github.com/QuantumScriptHub/diffSOD.

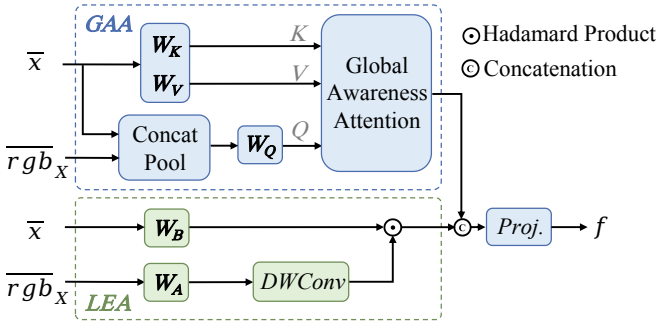


Figure 1: Diagrammatic details on how to conduct interactions between RGB and depth features (or thermal features). It contains three parts, which are responsible for global, local details, and extract surface information respectively. The base module is omitted for simplification.

1 IMPLEMENTATION DETAILS

Multi-resolution noise. Prior research has investigated variations from the classical DDPM approaches, introducing elements like atypical noise distributions or shortcuts that diverge from the Markovian sequence. The framework we introduce, along with the detailed fine-tuning procedure, accommodates alterations to the noise sequence during the fine-tuning phase. We discovered that a synergistic blend of noise applied at multiple resolutions, coupled with a gradually decreasing schedule, accelerates convergence and significantly enhances outcomes compared to the conventional DDPM approach. This layered noise approach involves layering a series of Gaussian noise patterns at varying scales, each scaled up to match the resolution required for the input of U-Net. Our suggested graduated schedule smoothly transitions from this layered, multi-resolution noise at $t = T$ to the typical Gaussian noise at $t = 0$. Furthermore, the gradual reduction of multi-resolution noise contributes to an additional improvement. We have also observed that training with multi-resolution noise results in more consistent predictions, even when different initial noise is used during inference. The annealing process further enhances this consistency.

Test-time ensembling. The probabilistic nature of the inference pipeline results in diverse predictions, contingent upon the initialization noise present in z_T^y. Leveraging this, we introduce a test-time ensemble approach capable of amalgamating numerous inference iterations on the same input For every input sample, there is the option to perform inference up to N times. This approach provides a versatile balance between computational efficiency and predictive accuracy, determined by the chosen value of N. We have observed a consistent improvement as the ensemble size increases. However, this improvement begins to diminish after reaching 10 predictions per sample.

Denoising steps. We assess the impact of re-spaced inference denoising steps guided by the DDIM scheduler. Despite being trained with 1000 DDPM steps, opting for 50 steps proves adequate for accurate inference results. As anticipated, employing more denoising steps leads to superior outcomes. We note that the point of diminishing returns with additional denoising steps varies across datasets but consistently remains below 10 steps. This suggests that denoising steps can be further reduced to 10 or fewer for increased efficiency while maintaining comparable performance. Interestingly, this threshold is lower than the typical requirement of 50 steps for diffusion-based image generators.

2 ATTENTION FEATURE INTERACTION MODULE

Building Block. Our building attention feature interaction module(AFIM) is mainly composed of the global awareness attention (GAA) module and the local enhancement attention (LEA) module and builds interaction between the RGB and depth modalities(or thermal modalities). GAA incorporates depth information(or thermal information) and aims to enhance the capability of object localization from a global perspective, while LEA adopts a large-kernel convolution to capture the local clues from the depth features(or thermal features), which can refine the details of the RGB representations. The AFIM is an attention mechanism, leveraging depth information(or thermal information) to enhance the 3D awareness of RGB features, for better global context understanding, while the latter uses depth features (or thermal features) to augment the geometry details of RGB ones by a Hadamard product. The details of the interaction modules are shown in Fig. 1. Besides, we also preserve a base operation that only process RGB features to extract the pure color cues. Different from the heavy cross-attention mechanisms that present quadratic relationship to the number of input pixels or tokens, our block can efficiently use the depth or thermal cues for scene understanding and shape awareness. This makes our encoder able to interact RGB and depth or thermal cues densely. Hadamard product to aggregate global context information from the feature maps. where the first one leverages features extracted from depth(or thermal) and rgb as attention to help semantic discrimination and the other utilize the geometry shape information in depth feature (or thermal feature) to refine the rgb feature towards better awareness of object shapes. Instead of using heavy attention operations, where the computation cost present quadratic growth as the pixels or tokens increase, we implement the attention mechanism under a small fixed size and the enhancement for 3D geometry via Hadamard product, which are efficient yet effective. Thus, we can apply the interaction between RGB and depth(or thermal) in each blocks without heavy burdens.

Our GAA fuses depth (or thermal) and RGB features to build relationships across the whole scene, enhancing 3D awareness

and further helping capture semantic objects. Different from the self-attention mechanism that introduces quadratic computation growth as the pixels or tokens increase, the Query (Q) in GAA is down-sampled to a fixed size and hence the computational complexity can be reduced. It illustrates that fusing depth features(or thermal features) with Q is adequate and there is no need to combine them with K or V, which brings computation increment but no performance improvement. And we implement this interaction in a fixed small size. So, Q comes from the concatenation of the RGB features and depth features (or thermal features), while key (K) and value (V) are extracted from RGB features. Given the RGB features $\bar{x}_i$ and depth features $\overline{rgb}_{X_i}$, the above process can be formulated as:

$$Q = \text{Linear}(\text{Pool}_{k\times k}([\bar{x}_i, \overline{rgb}_{X_i}])), K = \text{Linear}(\bar{x}_i), V = \text{Linear}(\bar{x}_i), \tag{1}$$

where $[\cdot,\cdot]$ denotes the concatenation operation along the channel dimension, $\text{Pool}_{k\times k}(\cdot)$ performs adaptively average pooling operation across the spatial dimensions to $k \times k$ size, and Linear is linear transformation. Based on the generated $Q \in \mathbb{R}^{k\times k\times C^d}$, $K \in \mathbb{R}^{h\times w\times C^d}$, and $V \in \mathbb{R}^{h\times w\times C^d}$, where h and w are the height and width of features in the current stage, we formulate the GAA as follows:

$$X_{GAA} = \text{UP}(V \cdot \text{Softmax}(\frac{Q^\top K}{\sqrt{C^d}})), \tag{2}$$

where $\text{UP}(\cdot)$ is a bilinear upsampling operation that converts the spatial size from $k \times k$ to $h \times w$. In practical use, Eq.2 can also be extended to a multi-head version, as done in the original self-attention, to augment the feature representations.

We also design the LEA module to capture more local details, which can be regarded as a supplement to the GAA module. Unlike most previous works that use addition and concatenation to fuse the RGB features and depth features(or thermal features). We conduct a depth-wise convolution with a large kernel on the depth features(or thermal features) and use the resulting features as attention weights to reweigh the RGB features via a simple Hadamard product . This is reasonable in that adjacent pixels with similar depth values (or thermal information) often belong to the same object and the 3D geometry information thereby can be easily embedded into the RGB features. To be specific, the calculation process of LEA can be defined as follows:

$$X_{LEA} = \text{DConv}_{k\times k}(\text{Linear}(\overline{rgb}_{X_i})) \odot \text{Linear}(\bar{x}_i), \tag{3}$$

where $\text{DConv}_{k\times k}$ is a depth-wise convolution with kernel size $k\times k$ and $\odot$ is the Hadamard product.

To preserve the diverse appearance information, we also build a base module to transform the RGB features $\bar{x}_i$ to x_{Base}, which has the same spatial size as X_{GAA} and X_{LEA}. The calculation process of x_{Base} can be defined as follows:

$$x_{Base} = \text{DConv}_{k\times k}(\text{Linear}(\bar{x}_i)) \odot \text{Linear}(\bar{x}_i). \tag{4}$$

Finally, the features, X_{GAA}, X_{LEA}, x_{Base}, are fused together by concatenation and linear projection to update the interaction features f.

3 EXPERIMENTAL RESULTS

From the results of methods based on non-transformer and transformers approaches presented in the paper, we can intuitively discern that the information from both global and local contexts is exceedingly beneficial for the integrity of salient object detection (SOD) predictions. Consequently, the state-of-the-art (SOTA) methods to date still predominantly rely on transformer-based approaches. However, existing methods continue to struggle with generating complete predictions, particularly in terms of edge refinement and fine details. This is primarily due to two factors: First, CNN-based encoders are not effective in capturing long-range dependencies for global context, leading to incomplete predictions. Second, downsampling the ground truth to fit the size of the prediction introduces inaccuracies, as details of the ground truth are lost during the interpolation or pooling process. We introduce Res2Nets as an auxiliary branch of our diffSOD encoder to better model long-term dependencies.The ablation experiments discussed in the main text regarding the addition of Res2Nets auxiliary branch also demonstrate that although our original diffusion model structure is already capable of yielding satisfactory results, the detection outcomes for salient targets can be further elevated by incorporating the Res2Nets auxiliary branch. As shown in Fig. 4, our network achieved the best results in precision and recall calculated across the benchmark datasets. It demonstrates outstanding performances of the proposed diffSOD. Due to space limitations, we presented the three different experiments together in the main text. In this supplementary material, we present them separately and provide additional experimental results with more comparative methods, offering a clearer demonstration of the superior performance of our approach. You can also use the provided code and pre-trained weights to perform inference on any RGB-D or RGB-T data pairs you have. If you encounter any issues, feel free to contact us at any time. We also provide more comparative experiments on RGB, RGB-D, and RGB-T. From a visual perspective, our method demonstrates significant advantages.

To provide more intuitive and equitable visual experimental results, we conducted inference on datasets unseen by the model. From these datasets, we randomly selected 150 images for display. For ease of layout, each display comprises three rows: the original image in the first row, the ground truth in the second row, and our model's prediction in the third row. We uniformly scaled each image, resulting in some appearing slightly distorted; however, upon enlargement, remarkable effects become evident. Evidently, our fine-tuned diffusion model leverages rich visual prior knowledge to impeccably capture features and edge information of salient targets. Remarkably, it even detects finer edge details than the ground truth, showcasing formidable zero-shot capabilities. Our diffSOD fully leverages the prior knowledge of salient objects in stable diffusion, and this fine-tuning strategy offers valuable insights for adapting diffusion models to other tasks. We have fully open-sourced both our training and inference code, allowing easy adaptation to other tasks with minimal input modifications. We hope to contribute even more to the open-source community through this effort.

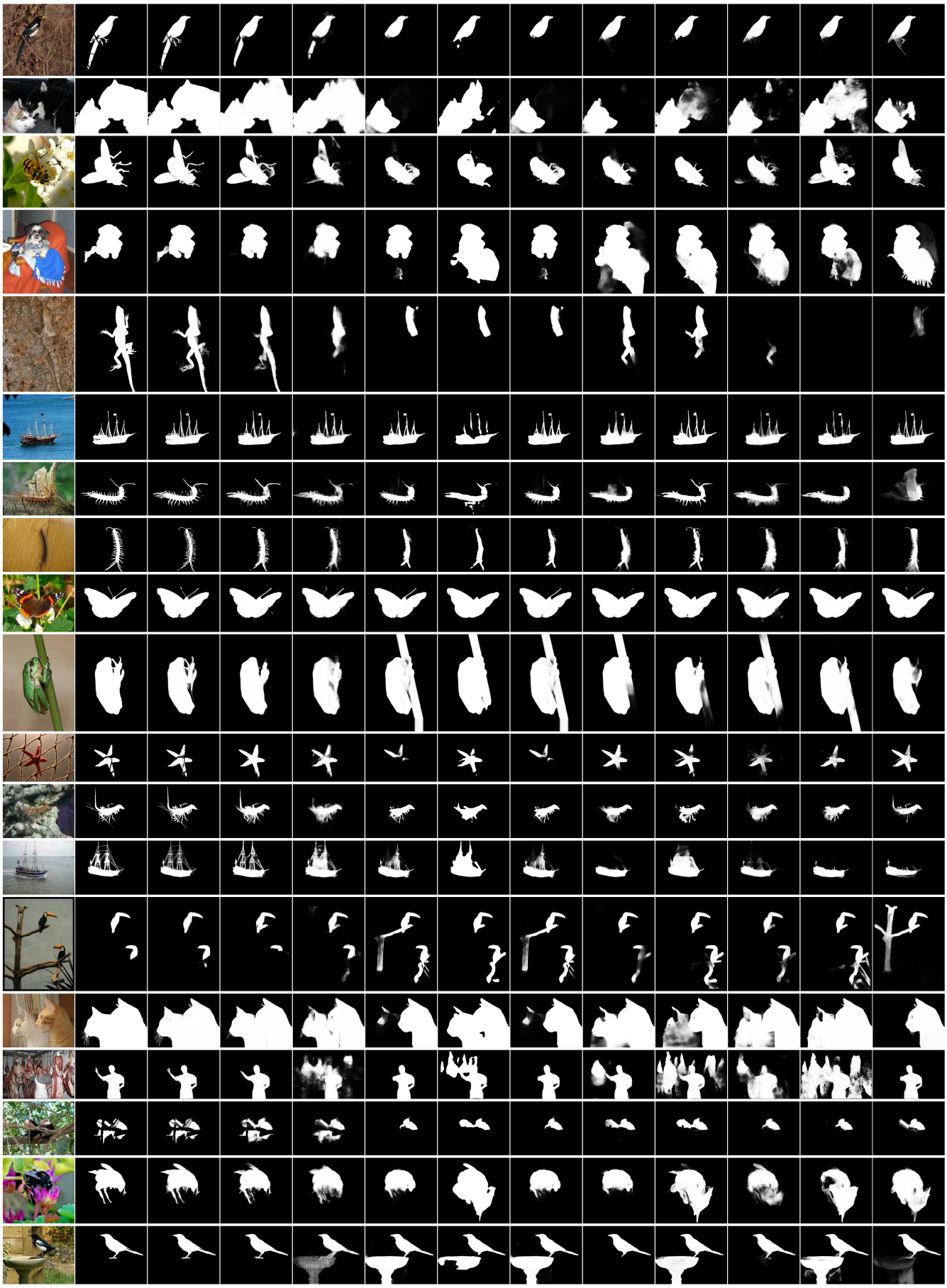


Image
GT
Ours*
M3Net*
VST*
SHNet
MSF
PFS
EDN
F3Net
USOD
MINet
U2Net

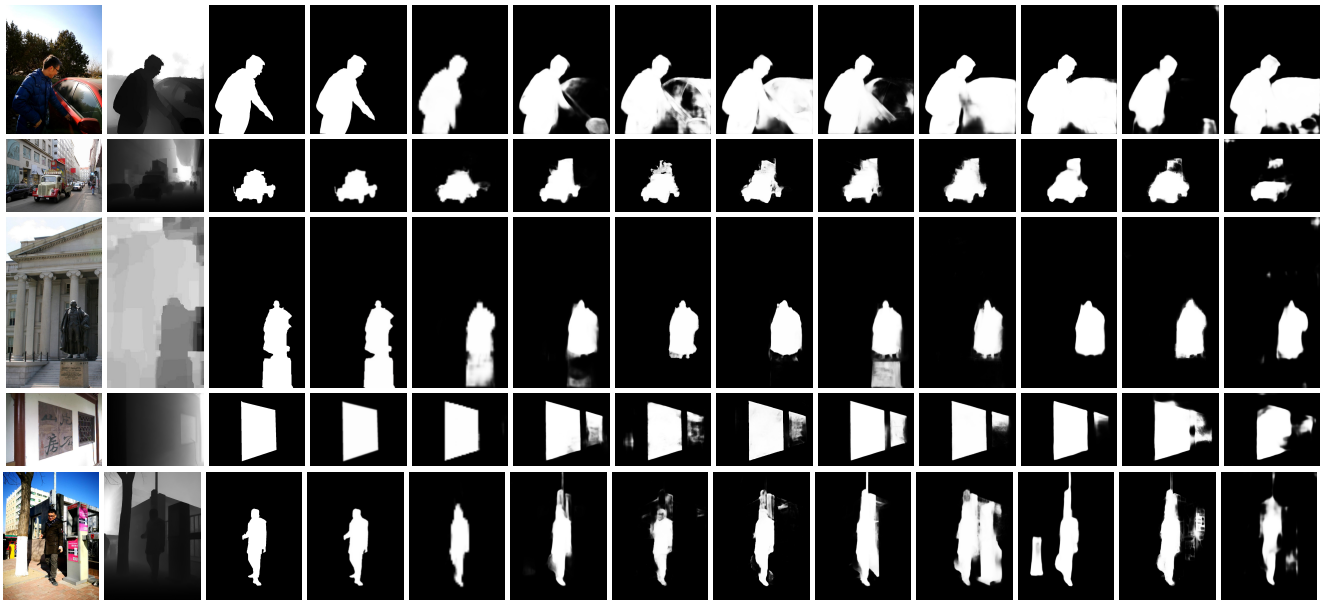


Figure 2: Visual comparison of saliency map results generated by various methods for RGB-D SOD.

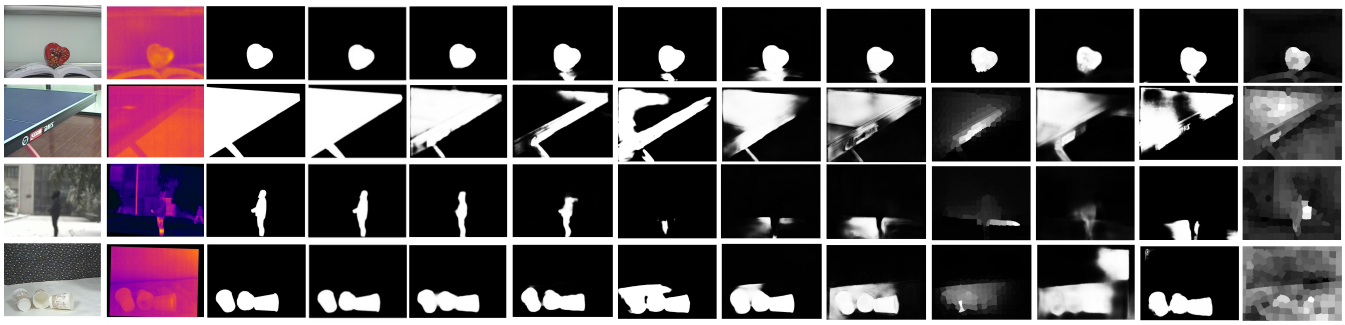
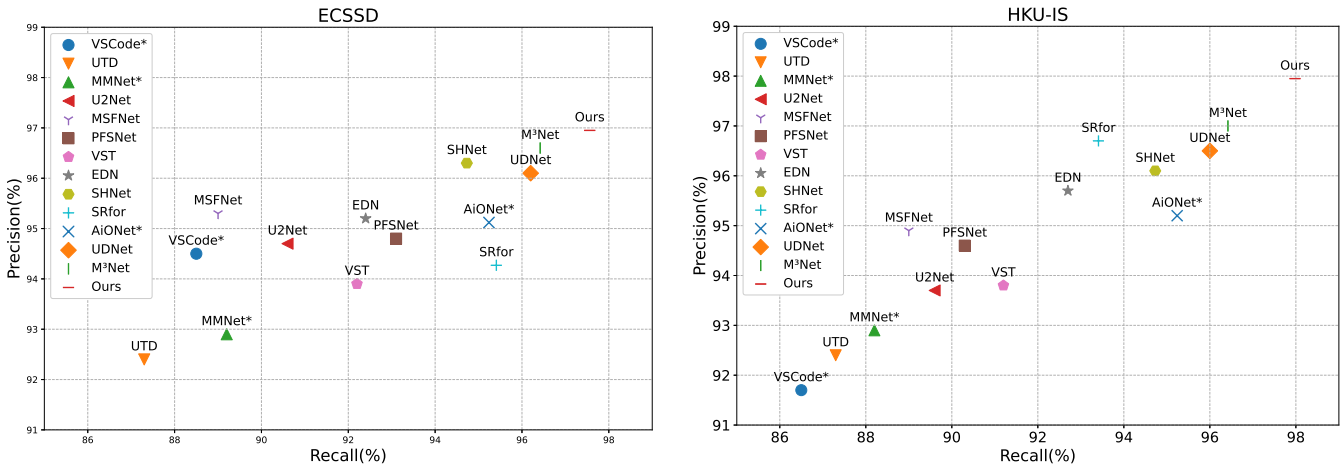


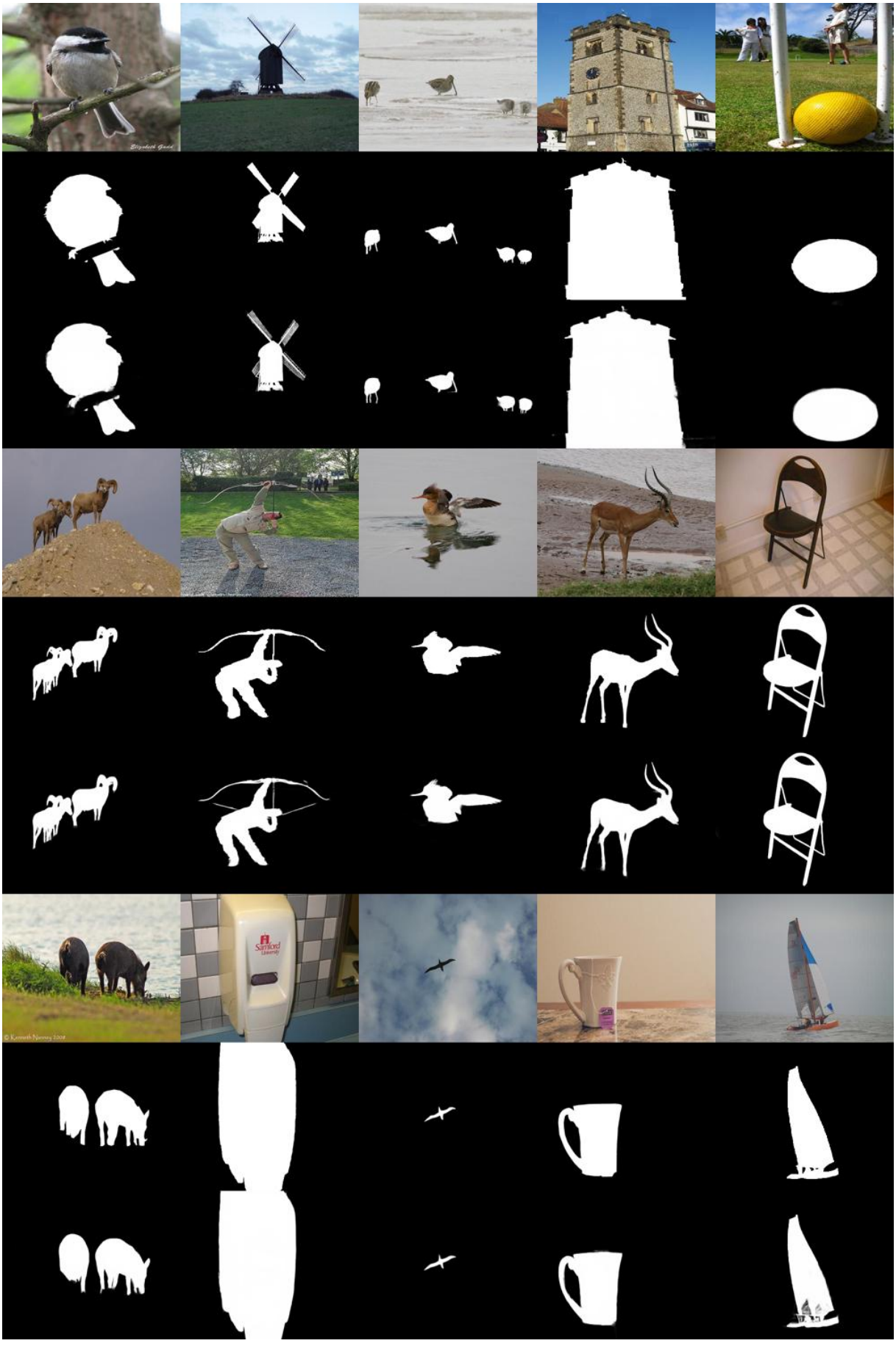
Figure 3: Visual comparison of saliency map results generated by various methods for RGB-T SOD.



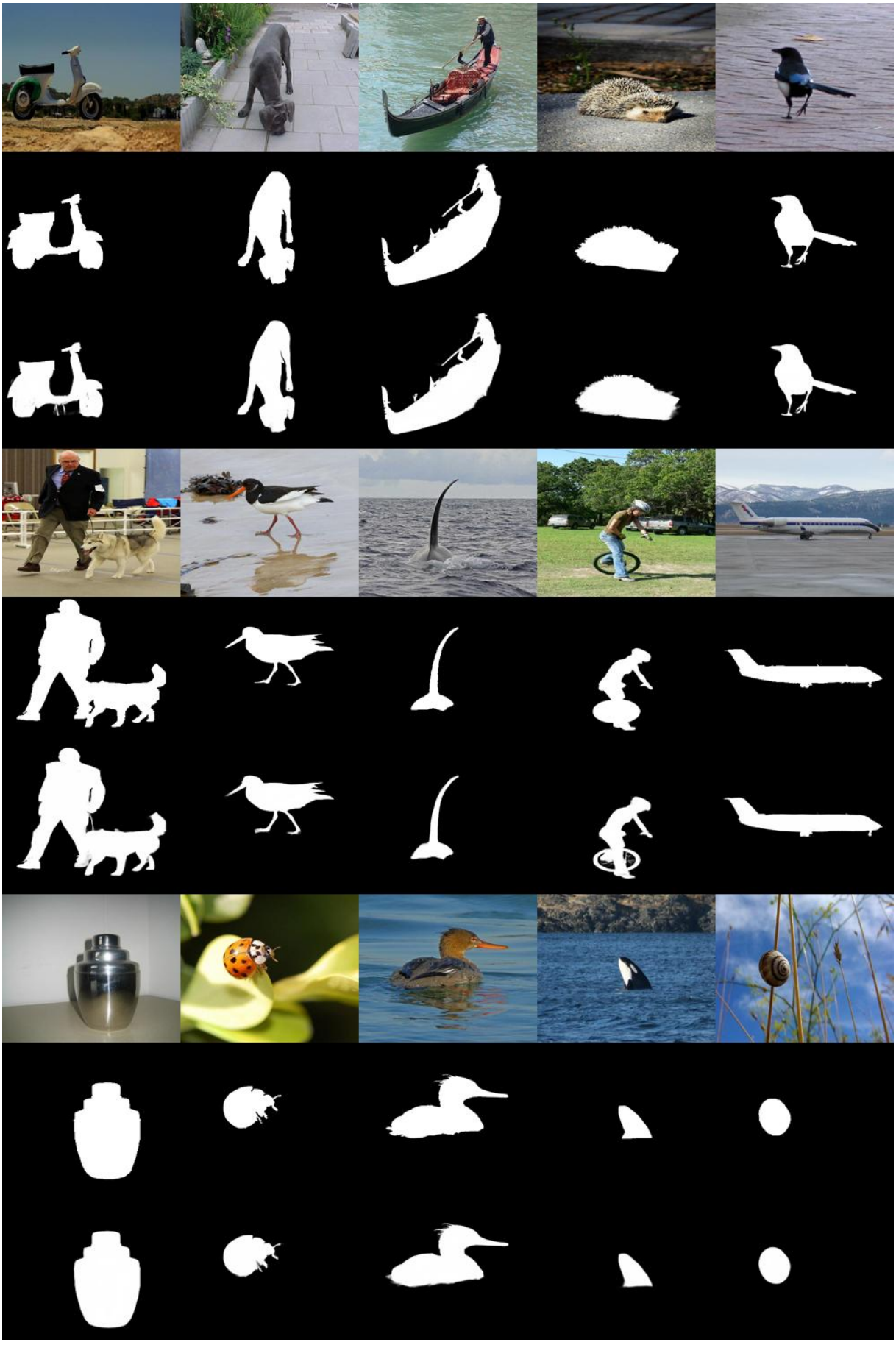
(a) The precision and recall on ECSSD.

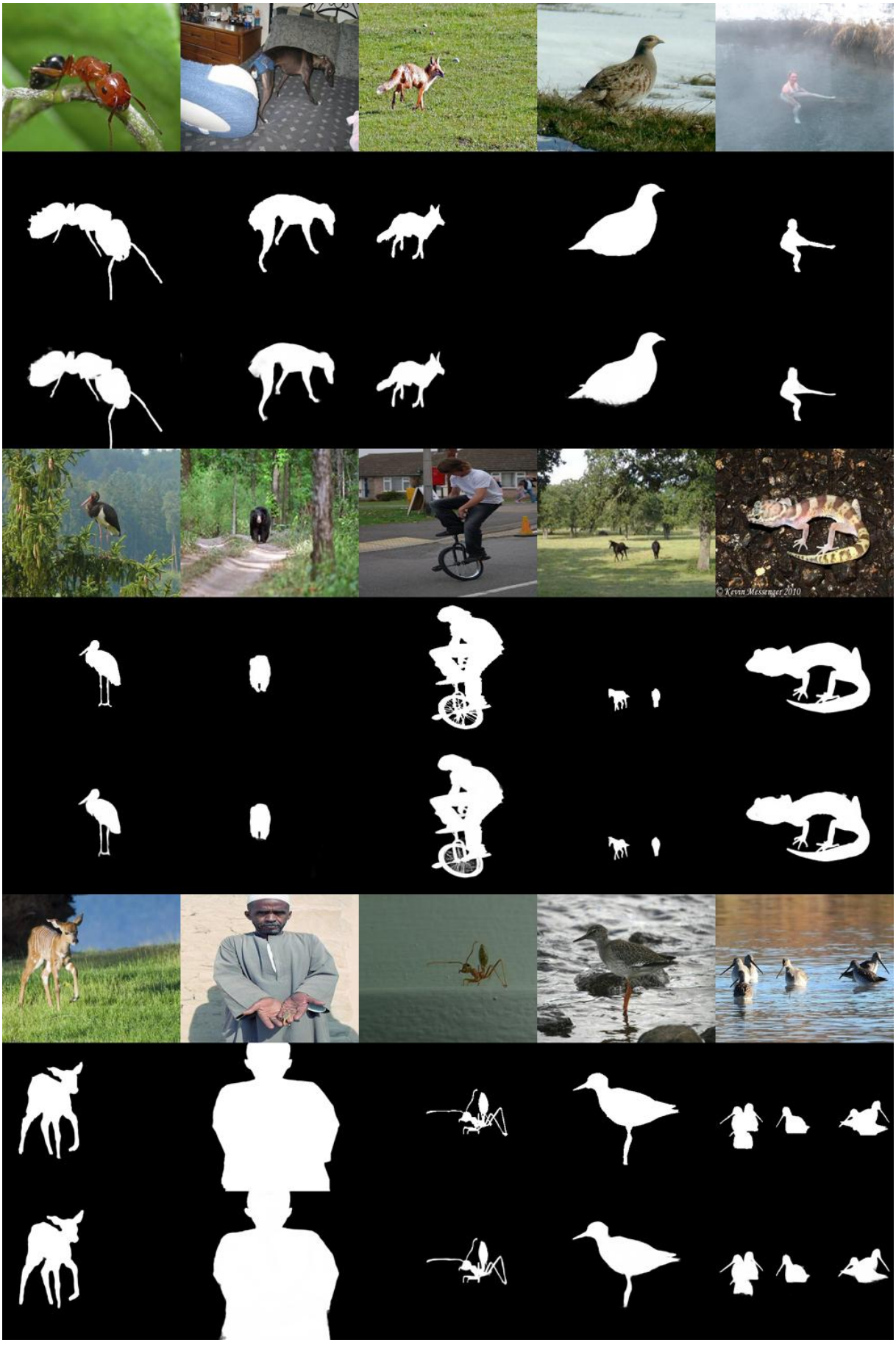
(b) The precision and recall on HKU-IS.

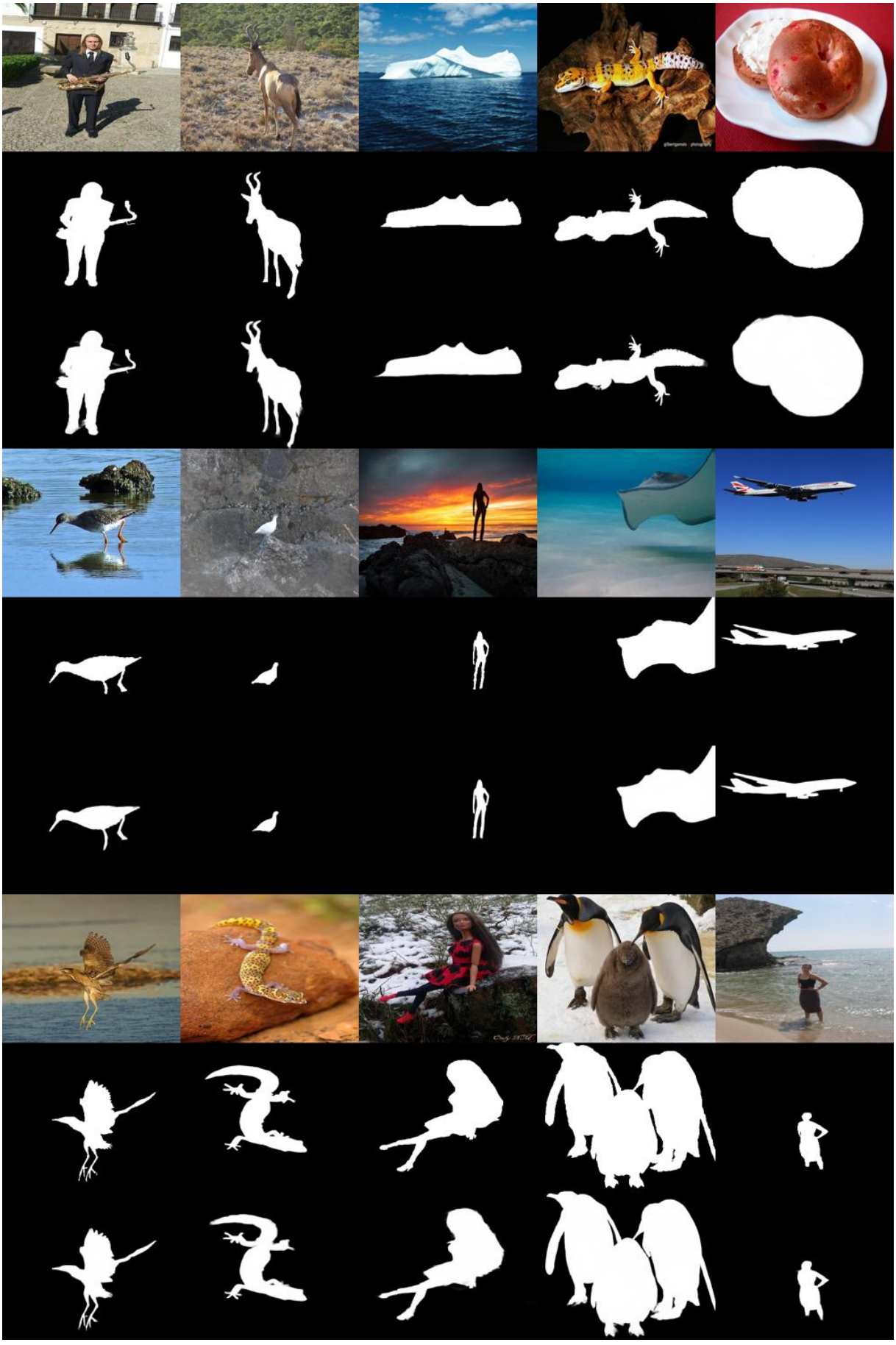
Figure 4: The precision and recall of all the methods.

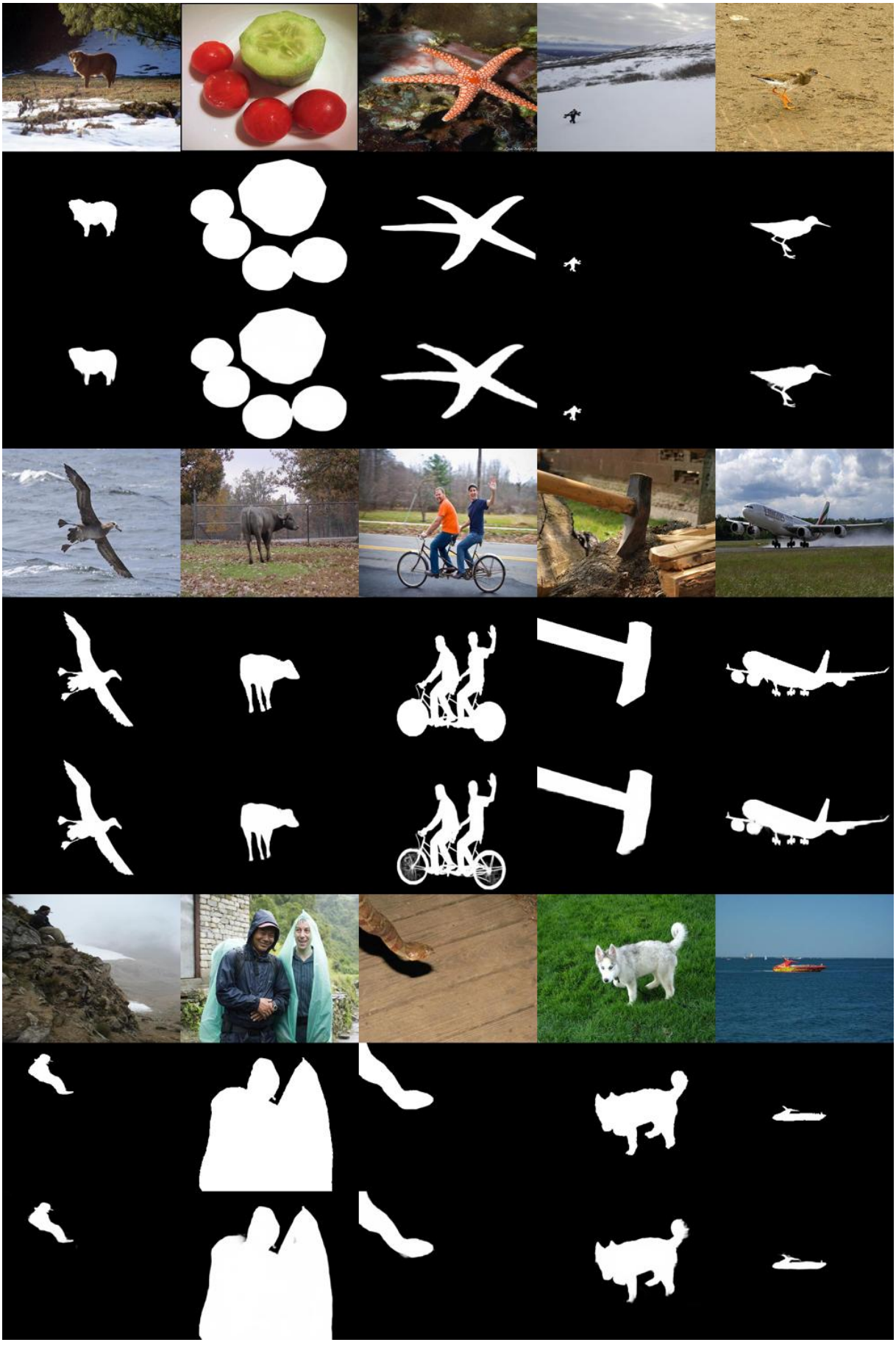


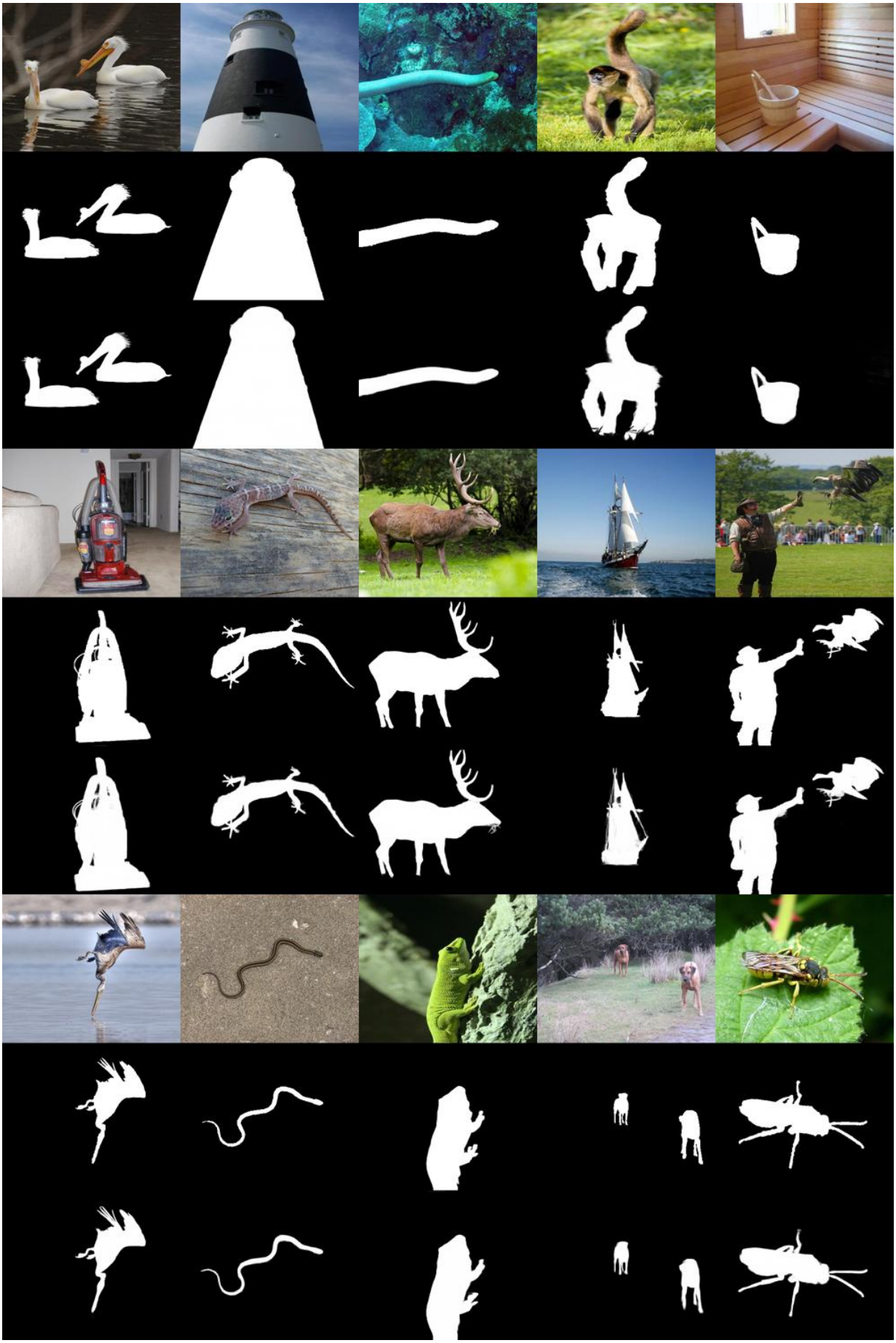

ROCHESTER
IS BIG & TALL IS

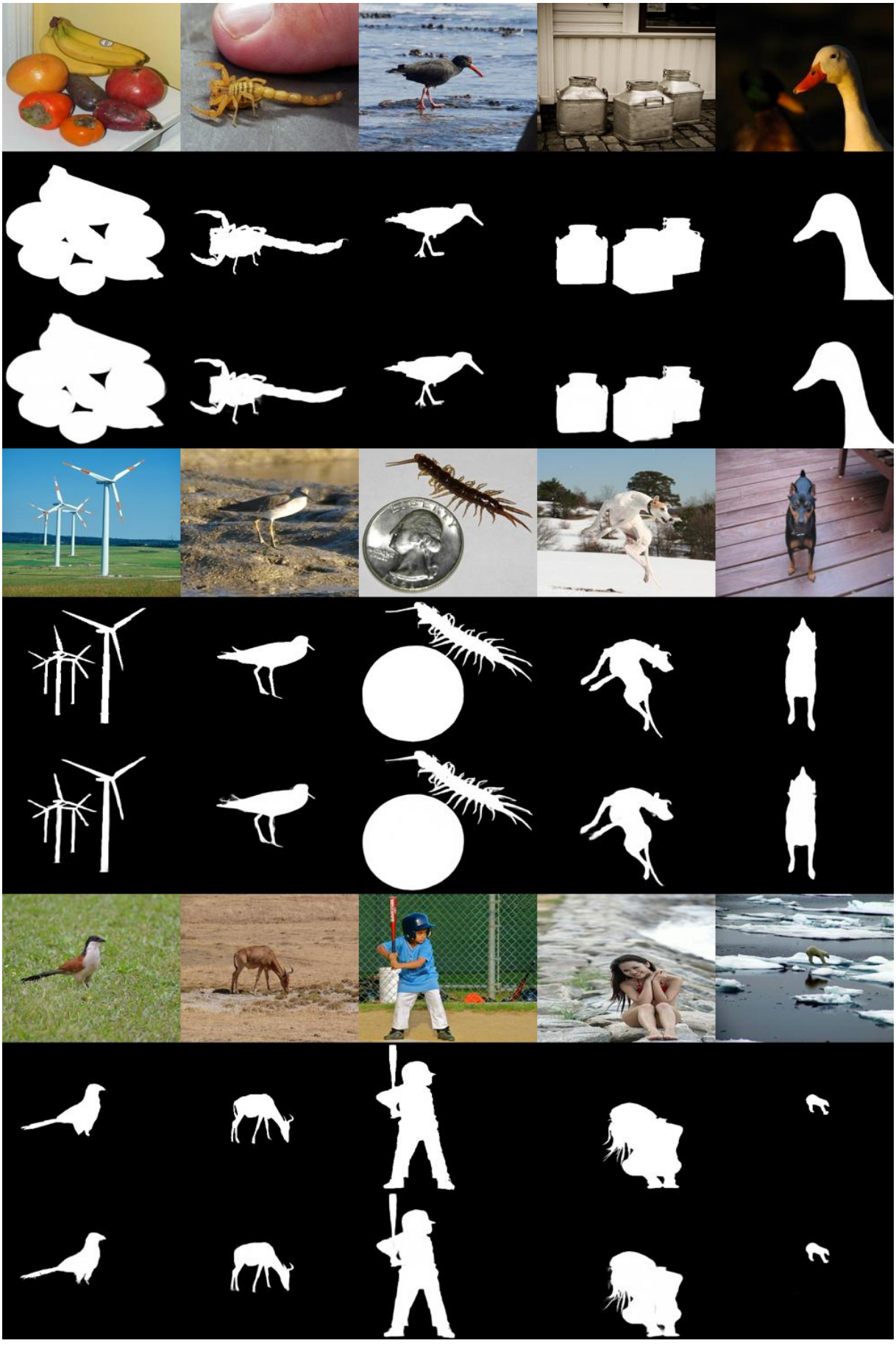


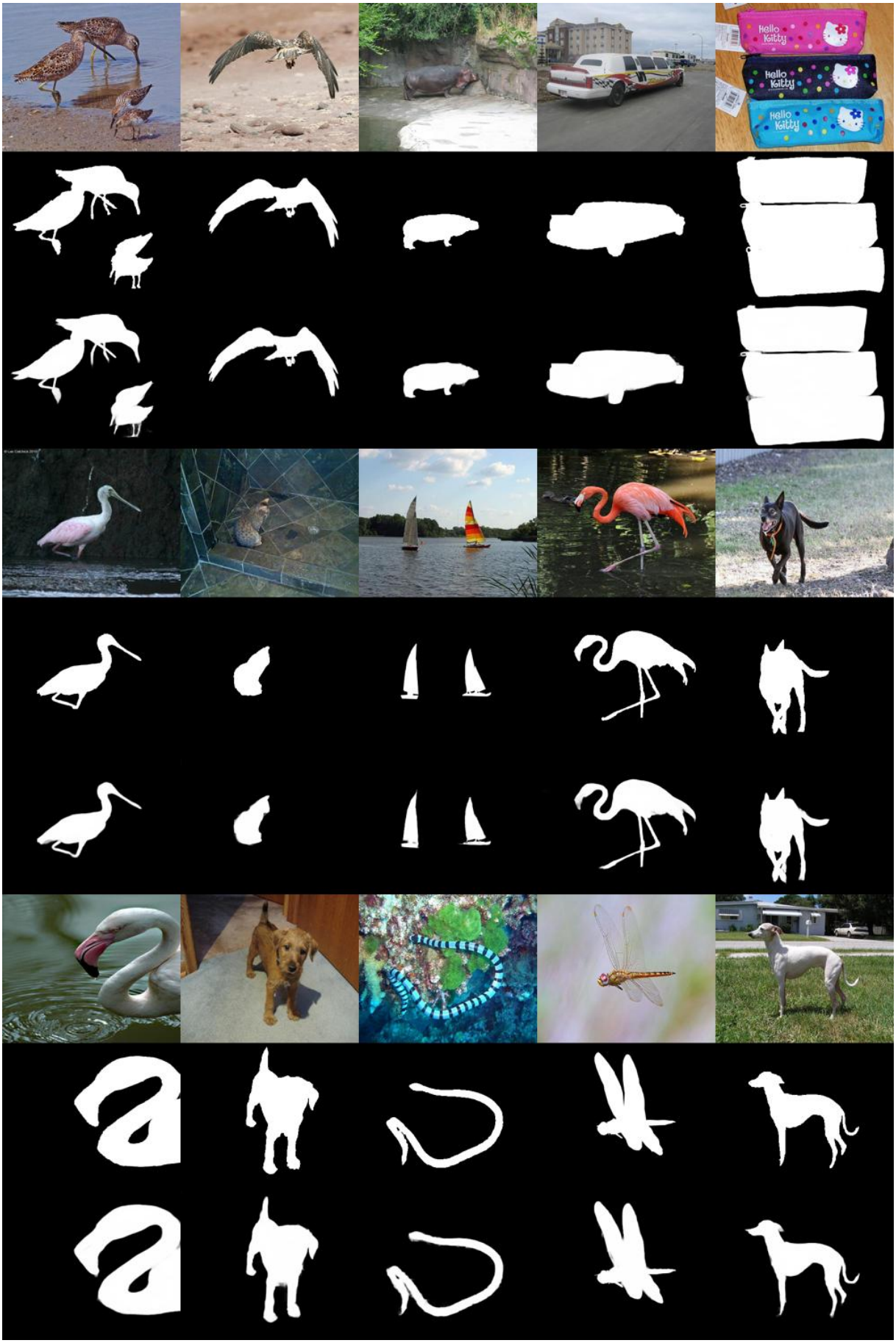

© Kevin Messenger 2010










Hello Kitty
Hello Kitty
Hello Kitty