A Multi-step Fusion Network Based on Environmental Knowledge Graph for Camouflaged Object Detection

Zheng Wang, Wenjun Huang, Ruoxun Su, Xinyu Yan, Meijun Sun

College of Intelligence and Computing, Tianjin University, Tianjin, China

ABSTRACT

Due to the high similarity in color and texture between camouflaged objects and noise backgrounds, existing single-step detection methods often fail especially when the camouflage level of objects is high. However, with prior knowledge of the environment, humans can effectively distinguish camouflaged objects, for example, when humans see snowy ground, they spontaneously associate that white rabbits might be concealed there. In this paper, we propose an Environmental Knowledge-guided Multi-step Network (EKNet) to simulate this mechanism. To extract prior knowledge of the background, we construct a knowledge graph with information extracted from the image and generate a relevance score matrix (RS) for prior knowledge and the camouflaged object with GCN as the correlation scoring matrix generation module (CSM). After that, we fuse the RS with Canny edge-enhanced features, which guides the model to detect camouflaged objects more accurately by observing the background information with edge semantics as the knowledge integration module (KIM). To our knowledge, this work is the first to introduce environmental knowledge to guiding camouflaged object detection(COD). Extensive experiments on three benchmark datasets show that our EKNet outperforms 15 existing state-of-the-art methods under four widely-used evaluation metrics.

INTRODUCTION

Camouflage is a unique method of concealment. A camouflaged object may disguise itself by mimicking the color or texture of another object, such as imitating the appearance of the surrounding environment or using disruptive coloring.

Overall, three major difficulty problems exist in camouflaged object detection: the wide variety of camouflaged objects, the obscured boundaries, and the obstruction in front of objects. It can be seen in Figure 1 that the above-mentioned challenges cannot be solved completely by these single-step recognition methods, thus, we mine semantic knowledge of the background and fuse it into a multi-step network.

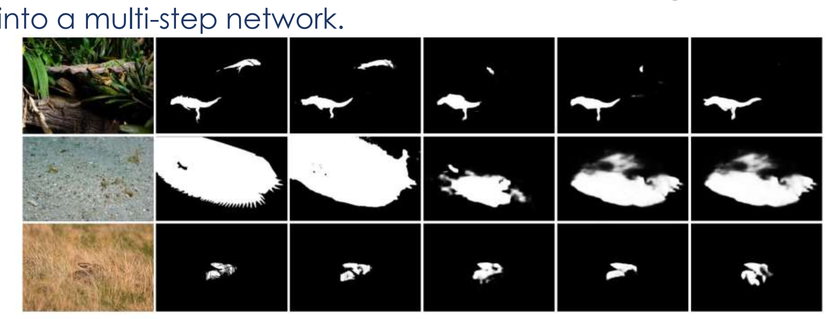

Figure 1: From top to bottom, three challenging camouflage scenarios with multiple objects, indefinable boundaries, and occluded objects are listed. Our model outperforms SINet[3] ,PFNet[23] and CF2Net[30] under these challenging scenarios.

OUR METHOD

The overall model architecture of the proposed EKNet is shown in Figure 3. Firstly, in the correlation score matrix(CSM) module, with small-scale manual annotations, we extract the environmental knowledge by fine-tuning the object detection network. The information is leveraged to construct a knowledge graph. Then, we generate the relevance score(RS)matrix using GCN[15]. Secondly, in the knowledge integration module(KIM), the RS matrix is fused with Canny edge features to generate a more complete and detailed segmentation result. Each module will be described in later sections in more detail.

Correlation scoring matrix generation module(CSM)

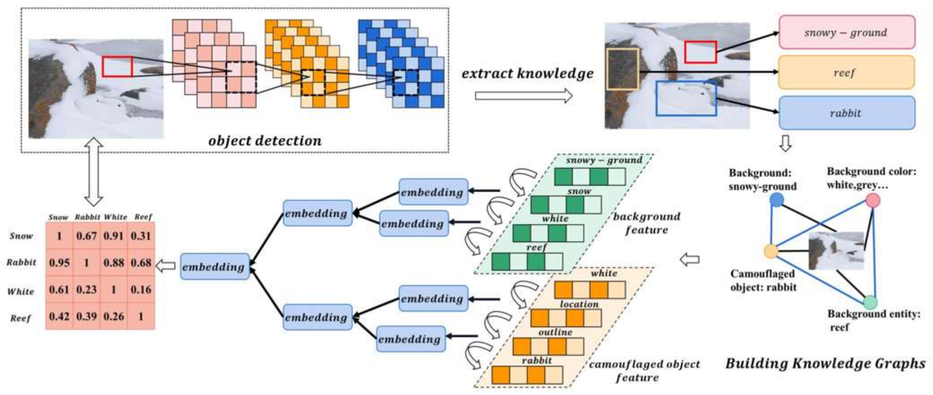

Figure 2: Correlation scoring matrix generation module (CSM). Firstly, extract the knowledge in the graph by the object detection algorithm. Secondly, construct the environment-based visual knowledge graph (EVKG) and generate the relevance score matrix (RS) by the graph convolution network embedding

Firstly, a visual object detection method is used to detect semantic entities in the images. In this paper, YOLO[1] is applied to extract the semantic information of the objects and backgrounds in the dataset. To ensure that these object detection methods can effectively extract the features needed, we fine-tuned object detection model.

Then, the fine-tuned object detection model is used to detect semantic entities in the images, these entities and their relationships are used to construct the EVKG.

Knowledge Integration Module (KIM)

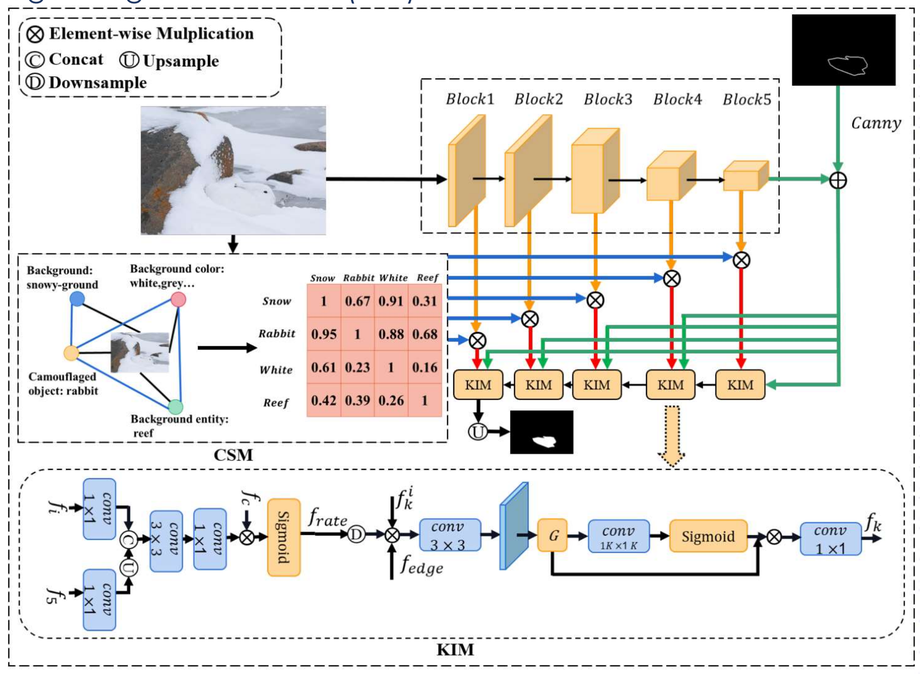

Figure 3: Complete model architecture(EKNet). The complete network design and the detailed design of the KIM module are shown here.

The design of KIM is shown in Figure 3. This module aims to integrate edge cue information and RS into representation learning to enhance the implicit knowledge reasoning ability and object-structured semantic feature representation.

$$f_m = F_{conv3\times3}((f_i \otimes f_{rate}) \otimes (D(f_{edge}) \oplus f_5))$$

We utilize channel-wise global average pooling (GAP) to aggregate the convolutional feature fm, and then obtain corresponding channel attention weights.

$$f_k = F_{conv1\times1}(Sigmoid(f^a_{1K}(G(f_m))) \otimes f_m$$

Loss Function

$$L_{sum} = \sum_{i=1}^{5}(L_G + L^{\omega}_{BCE}(R_i, S_m) + L^{\omega}_{IOU}(R_i, S_m)) + \varepsilon L_{dice}(R_e, S_e)$$

EXPERIMENTS

We compare EKNet with 15 state-of-the-art methods. Table 1 shows the comparison results. For a fair comparison, all the predictions of these methods are either provided by the authors or produced by models retrained with open-source codes.

Method	Pub./Year	CAMO-Test				COD10K-Test				NC4K			
		$S_\alpha\uparrow$	$E_\phi\uparrow$	$F^\omega_\beta\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F^\omega_\beta\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F^\omega_\beta\uparrow$	$\mathcal{M}\downarrow$
EGNet[38]	ICCV2019	0.732	0.796	0.601	0.107	0.736	0.802	0.515	0.059	0.777	0.842	0.639	0.078
PraNet[4]	MICCAI2020	0.769	0.824	0.676	0.094	0.784	0.863	0.642	0.056	0.797	0.889	0.685	0.073
F3Net[33]	AAAI2020	0.711	0.741	0.564	0.109	0.739	0.795	0.544	0.051	0.780	0.824	0.656	0.070
SINet[3]	CVPR2020	0.745	0.804	0.704	0.092	0.776	0.864	0.645	0.043	0.809	0.872	0.753	0.058
PFNet[23]	CVPR2021	0.782	0.841	0.695	0.085	0.800	0.868	0.660	0.040	0.829	0.887	0.745	0.053
R-MGL[37]	CVPR2021	0.775	0.812	0.673	0.088	0.814	0.851	0.666	0.035	0.833	0.867	0.739	0.053
TANet[40]	AAAI2021	0.781	0.847	0.678	0.087	0.793	0.848	0.635	0.043	-	-	-	-
C2FNet[30]	IJCAI2021	0.796	0.857	0.730	0.078	0.813	0.889	0.691	0.036	0.840	0.896	0.771	0.048
UGTR[36]	ICCV2021	0.785	0.822	0.685	0.086	0.818	0.852	0.667	0.035	0.839	0.876	0.746	0.052
JCSOD[18]	CVPR2021	0.800	0.859	0.728	0.073	0.809	0.884	0.684	0.035	0.841	0.898	0.771	0.047
OCENet[36]	WACV2022	0.807	0.866	0.744	0.075	0.829	0.890	0.721	0.034	0.848	0.899	0.785	0.046
SegMaR[12]	CVPR2022	0.811	0.868	0.749	0.073	0.831	0.899	0.722	0.033	0.841	0.896	0.781	0.046
CubeNet[41]	PR2022	0.788	0.838	0.682	0.085	0.795	0.865	0.643	0.041	-	-	-	-
ERRNet[11]	PR2022	0.779	0.842	0.679	0.085	0.786	0.867	0.630	0.043	0.827	0.887	0.737	0.054
SAM[16]	arXiv2023	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050	0.767	0.776	0.696	0.078
EKNet(Ours)		0.821	0.879	0.749	0.073	0.833	0.900	0.727	0.032	0.850	0.904	0.785	0.044

Table 1: Quantitative comparison with state-of-the-art methods for COD on three benchmarks using four widely used evaluation metrics (*i.e.*, S_α, E_ϕ, F^ω_β, $\mathcal{M}$). "↑" / "↓" indicates that larger/smaller is better. The top three results are highlighted in red, green, and blue.

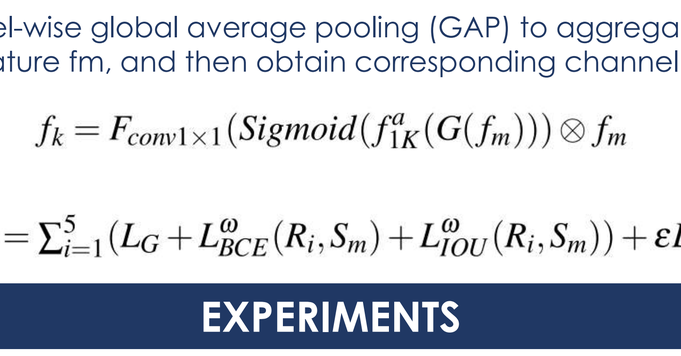

Figure 4: Visual comparison of the proposed model with five state-of-the-art COD methods. Obviously, our method is capable of accurately segmenting various camouflaged objects with more clear boundaries.

Figure 4 shows the qualitative comparisons of different COD methods on several typical samples from the COD10K[2] dataset. It is obvious that our method provides accurate camouflaged object predictions with finer and more complete object structure and boundary details.

REFERENCES

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[2] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2777–2787, 2020.