A novel deep neural network for hidden target detection in images

Rabeb Hendaoui *, Vasif Nabiyev
Dept. of Computer Engineering, Karadeniz Technical University, Turkey
**Corresponding author: rabeb@ktu.edu.tr*

Abstract

The significant similarity between the hidden target and the background makes it difficult to find camouflaged people, such as warriors in warfare, or even camouflaged objects in natural environments. Hence, it is hard to ascertain these concealed targets. To address this issue, a novel deep neural network is proposed in this paper that produces an estimated mask within the hidden target for an input image. Our approach consists of two phases: hidden target segmentation and hidden target identification. For the first phase, we propose the Multilevel Attention Network (MA-Net), which generates the camouflaged target mask based on a Multi-Attention Module (MAM) that helps distinguish the hidden people from the background. Later on, the concealed target will be highlighted in the second phase. Experimental results on the camouflaged people dataset demonstrate that our proposed method can achieve state-of-the-art performance for hidden target detection.

Keywords: Concealed people; hidden target; neural network; target identification; target segmentation

1. Introduction

Computer vision applications have been well explored in the literature. In particular, many notable object detection methods (He *et al.*, 2017; Redmon *et al.*, 2016) have already been studied by various researchers.

At times, objects conceal their signatures and generate disguises in their surrounding environment. The presence of camouflage makes the identification of objects more difficult. Camouflage is the capacity of the prey to hide from predators by adjusting their pattern, texture, and coloration according to the background. This phenomenon was adopted by human beings and broadly utilized on the battlefield. Human vision systems cannot sufficiently recognize a hidden target. Certain animals have distinct biological capabilities that conceal them in their surroundings. The visual characteristics of a disguised object (like color/texture) resemble the background, making detecting procedures complicated. Hence, a camouflaged target cannot be identified by state-of-the-art methods for object detection. Consequently, the study of hidden target detection in the sector is required.

Owing to the complexity of the issue, less work has been suggested to detect camouflaged people. Existing studies (Pan *et al.*, 2011; Song & Geng, 2010; Bhajantri & Nagabhushan, 2006; Galun *et al.*, 2003; Tankus & Yeshurun, 2001) investigated the matter with low-level features. Generally, these methods use texture, brightness, color, and edge features to distinguish objects from backgrounds. Disguised target detection is not fully explored; most research is concerned with detecting the foreground region despite some parts of its texture being similar to that of the background. These techniques rely on hand-crafted characteristics and are effective in a limited number of situations where images have an essential and non-uniform backdrop. Additionally, their effectiveness in detecting and segmenting camouflage is poor when the foreground and background have a high degree of resemblance. Recently, a new camouflaged people detection has been proposed via a dense deconvolution network (Zheng *et al.*, 2018). The authors introduced a dense deconvolution network to fuse the extracted features in deep CNN. Then super-pixel segmentation was applied in detection optimization. Within this context, using high-level features, we propose a new hidden target detection and demonstrate through experiments that it outperforms the state-of-the-art methods.

In summary, we present the following contributions: 1) We adopt an inception module to enhance the ability to excavate the interior of visual features and focus on feature representation, 2) We design a multi attention module that can compensate for the loss of perceptual details, emphasize hidden targets, and better identify small-scale disguised objects, 3) We evaluate our model and compare it with state-of-the-art methods. Results show that our approach performs favorably over all the others.

This paper proceeds as follows: Section 2 introduces the proposed framework. Section 3 outlines the experiments. In Section 4, we discuss the study's limitations. Lastly, we reach conclusions.

2. Proposed method

Hidden target detection is a fundamentally difficult task because the camouflage strategy works by misleading the observer's visual perceptual system. A substantial amount of visual perception information is necessary to remove the uncertainties produced by the substantial inherent similarities between the target and the background. As seen in the natural environment, predators and prey animals use camouflage to conceal their location and prevent bringing attention to themselves, making it harder to find them. Based on this observation, we were inspired to detect the disguised target through an attention mechanism to bring attentiveness and pay attention to the target. In the first phase, we aim at segmenting the hidden target where we implement an inception module to enhance the feature representation. Then we introduce a Multi-Attention Module to improve the detection of the ambiguous hidden target further from the background and generate finer details. In the second phase, we identify the concealed target based on the predicted mask.

In this section, we explain our proposed approach for hidden target detection. As shown in Figure 1, our method involves two steps: hidden target segmentation and hidden target extraction.

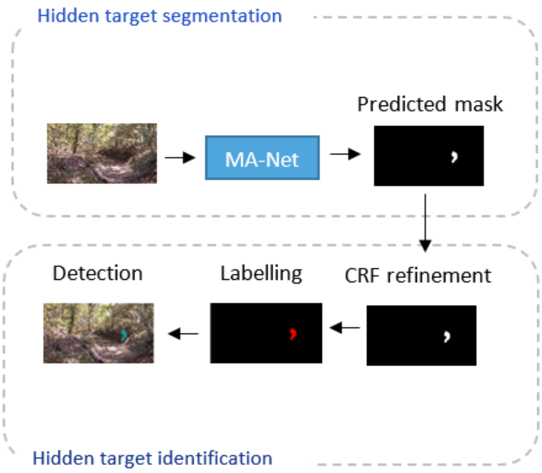

Fig. 1. Our overall framework for hidden target detection.

2.1 Hidden target segmentation

As shown in Figure 2, we propose the MA-Net model for hidden target segmentation. First, we extract features in a bottom-up way from the input images based on the ResNet (He *et al.*, 2016) backbone network. Then features are enhanced based on an inception module with multi-scale receptive fields. Furthermore, we fuse generated feature maps with multilevel semantic information in a top-down way. Finally, we employ the MAM for every level, and we combine the findings of the predictions from all layers for an outcome.

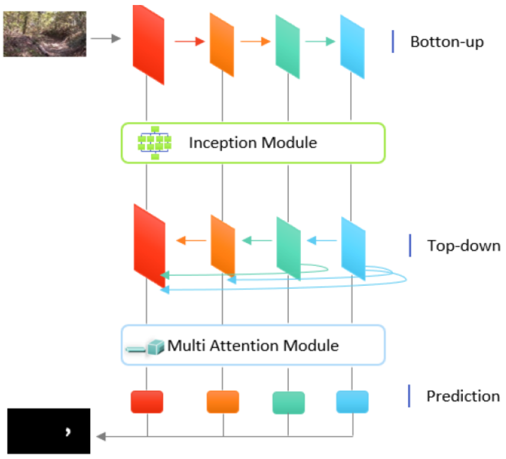

Fig. 2. The architecture of our proposed MA-Net Network.

2.1.1 CNN feature extraction and enhancement

To extract features from multiple levels, we employ ResNet (He *et al.*, 2016) as the backbone because of its fast convergence compared to VGG (Simonyan & Zisserman, 2015). ResNet-101 consists of 101 convolutional layers with five convolutional blocks, an average pooling layer, and one fully connected layer. We made some changes to it in order to adapt it to our camouflage target prediction task. First, the fully connected layers, which are built explicitly for classification tasks, are removed. The number of parameters is also considerably reduced as a result of this. Second, because the final feature map size of the original ResNet is 32 times less than the input, if we directly upsample on it, the results will be too coarse. To overcome this, in levels 4 and 5, we utilize dilated convolution (Chen *et al.*, 2016) which allows us to keep the same receptive field without lowering the size of the feature map or adding any additional parameters. As a result, the feature map size at these two levels is only eight times less than the input. The extracted features from different levels are fed into an inception module to enhance feature representation by extracting multi-scale receptive field features. As shown in Figure 3, the designed module consists of 4 branches with a 1 × 1 convolution at the beginning to reduce the number of channels. The outputted features from three branches go through 1 × y dilated convolution and then through y × 1 dilated convolution with a rate equal to 3. Here, y= 3, 5, and 7. In order to localize small and irregular objects, a deformable convolution (Dai *et al.*, 2017) layer is added at the end of each branch. The features from all branches will be concatenated, go through 1 × 1 convolution, and have a residual connection with input features for faster optimization. To encode more contextual information, we densely connect the features in a top-down way. High-level feature maps are reused multiple times to add more contextual

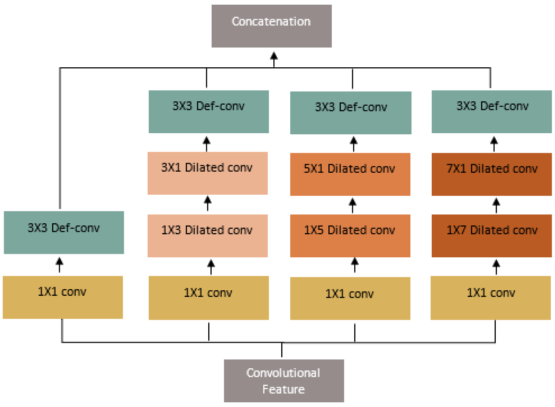

Fig. 3. Our inception module.

information to low levels.

2.1.2 Feature refinement and prediction

As in the natural environment, animals utilize camouflage to not draw attention to themselves; thereby, it is difficult to find them. Based on this fact, we constructed a MAM to better detect the disguised target from the background and create more delicate features to draw attention to this hidden target. Our suggested MAM highlights essential features in the image by disregarding less critical information. As illustrated in Figure 4, initially, a Pyramid Pooling Module (Zhao *et al.*, 2017) is applied to the input feature, and the output feature is used as an input feature for the MAM. This latter consists of three attention blocks: the Transformation attention block, the Channel Attention Block, and the Spatial Attention Block. The first block attempts to represent feature transformations by using deformable convolution (Dai *et al.*, 2017). It can improve the network's attention to foreground areas. The feature map is processed by a 3 × 3 deformable convolutional layer, followed by a normalization layer, ReLu, and a second 3 × 3 deformed convolutional layer. The channel attention block highlights the camouflaged target and reduces the inaccuracies caused by duplicated channel features. The feature map is re-allocated using two 1 × 1 convolutions and a global pooling operation. This global attention map explicitly makes the positions of the hidden objects known on feature maps. The spatial attention block attempts to investigate where to concentrate on a feature map more. The spatial attention module is used as a complement to the channel attention module to produce efficient features. Finally, the refined features from all the blocks are concatenated into a final attention map.

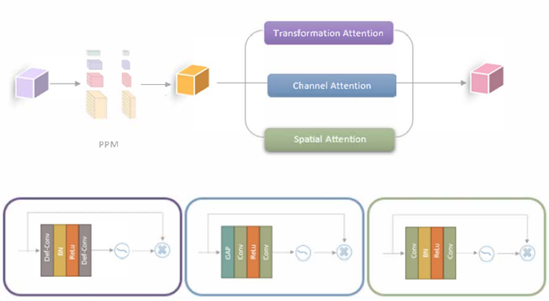

Fig. 4. Multi Attention Module.

2.2 Hidden target identification

In order to make the extracted camouflaged objects more accurate, we use the Dense CRF (Krähenbühl & Koltun, 2011) method to refine the camouflaged target contours. A conditional random field (CRF) is a probabilistic graph modeled by a Gibbs distribution as follows:

$$P\left(X|\,Q\right) = \frac{1}{Z\left(Q\right)}\exp\left(-E\left(X|\,Q\right)\right) \tag{1}$$

where Q is the global observation (image), $Z\left(Q\right)$ is the normalization factor, and E(X) denotes the Gibbs energy. In Dense CRF, the energy function is defined as:

$$E\left(X\right) = i\sum\psi_u\left(x_i\right) + i < j\sum\psi_p(x_i, x_j) \tag{2}$$

where x_i and x_j denote the vertices of CRF, ψ_u is the unary potential, and ψ_p is the pairwise potential. The unary potential is calculated based on the predicted segmentation map while the pairwise potential $\psi_p\left(x_i\right)$ is given by:

$$\psi_p\left(x_i, x_j\right) = \mu\left(x_i, x_j\right) m\sum w^{(m)}k^{(m)}(f_i, f_j) \tag{3}$$

where $\mu\left(x_i, x_j\right) = 1$ if $x_i \neq x_j$ and equal to 0 otherwise. f_i and f_j are feature vectors. Specifically, the kernel k is defined as:

$$w_1 \exp\left(\frac{-|p_i - p_j|^2}{2\sigma_\alpha^2} - \frac{|q_i - q_j|^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\sigma_\gamma^2}\right) \tag{4}$$

where p_i, indicate the pixel's location, q_i, q_j the pixel's spectral features. σ_α, σ_β, and σ_γ are three key hyper-parameters controlling the degree of connectivity and similarity.
After the camouflaged map is refined, the connected components are decided to identify each object in the image. The corresponding bounding box for each connected component is then computed. Figure 5 demonstrates the final target detection results of our proposed approach.

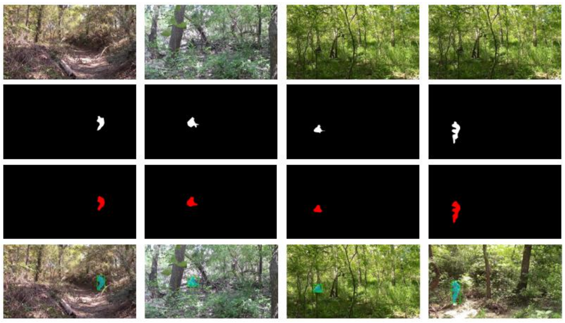

Fig. 5. Some examples of target detection results of our proposed approach. Image (row 1), GT (row 2), labeled image from our segmentation mask (row 3), and final detection (row 4).

3. Experiments

3.1 Dataset

To evaluate our method, we used the camouflaged people dataset (Zheng *et al.*, 2018). It contains 1000 images of size 480 × 854, including camouflaged people with ten different kinds of camouflage patterns like Arid Fleck, Desert, and different scenes like woodlands and snowfields. From the dataset, 80% of images are randomly selected for training, and the remaining 20% are used for testing. We apply data augmentation to the selected training images. Mirror reflection and rotation techniques were used.

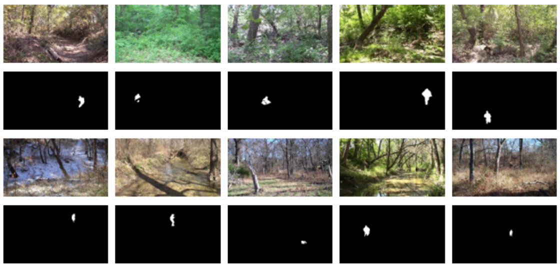

Fig. 6. A few examples from the Camouflaged People Dataset with corresponding ground truth labels.

3.2 Implementation Details

Our model is implemented based on the Caffe framework (Jia *et al.*, 2014), using GPU Nvidia GTX 1080. A stochastic gradient descent optimization (SGD) algorithm was used for training with a momentum value of 0.9 and a weight decay of 0.0005. We set the base learning rate to 1e-10 with a mini-batch size of one. After 20K iterations, we stopped the training.

3.3 Evaluation Metrics

In our experiments, we use the following evaluation metrics: Mean Absolute Error (MAE) (Perazzi *et al.*, 2012) , F-Measure (Achanta *et al.*, 2009), E-measure (Fan *et al.*, 2018) and Structure Measure (S-Measure) (Fan *et al.*, 2017), which are explained below.

MAE: is a metric to directly calculate the average absolute error between the prediction maps S and the corresponding ground truth maps G. The formula is as follows:

$$MAE = \frac{1}{H \times W} Wx = 1 \sum Hy = 1 \sum |S\,(x, y) - G(x, y)| \tag{5}$$

where W and H are the width and height of the input image. In general, a lower MAE indicates a better result.

F-measure: is defined as the weighted harmonic mean of recall and precision metrics, with an anon negative weight of β. The F-measure is defined as:

$$F_\beta = \frac{\left(1+\beta^2\right)\ Precision \times Recall}{\beta^2 Precision + Recall} \tag{6}$$

where we set β^2 to a fixed value of 0.3 as suggested in (Achanta *et al.*, 2009) to emphasize precision over recall. Note that, unlike MAE, a higher F_β indicates a better performance.

E-measure: is a perceptual-inspired criterion and is defined as:

$$E = \frac{1}{H \times W} Wx = 1 \sum Hy = 1 \sum \phi_{FM}\left(x,y\right) \tag{7}$$

in which ϕ_{FM} is an enhanced alignment matrix. The greater the E Score, the better the performance.

S-measure: is to measure the structural similarity between the predicted map and the ground-truth map.

$$S_\alpha = \left(1-\alpha\right) S_o + \alpha S_r \tag{8}$$

in which S_r indicates the region-aware structural similarity and S_o denotes the object-aware structural similarity. As suggested in (Fan *et al.*, 2017) , we set $\alpha = 0.5$. Note that the higher the S-measure score, the better the model performs. The significantly larger the S-score, the better the model is.

3.4 Ablation Study

To investigate the impact of the different modules in our method, we conducted an ablation study. The experiment selects ResNet-101 as the baseline (B). Then we add the Inception Module (IM), Transformation Attention Module (TAM), Chanel Attention Module (CAM), and Spatial Attention Module (SAM) into the network in turn. As shown in Table 1, with the addition of modules, the test performance gradually improves. All these modules boost the model performance. When these modules are combined, we can get the best results. It demonstrates that all components are necessary for the proposed framework.

Table 1. Component analysis. Note that a lower MAE and higher F, S, and E correspond to better results.

Settings	MAE	F	E	S
B	0.01	0.847	0.954	0.922
B + IM	0.008	0.853	0.967	0.930
B + IM + TAM	0.006	0.854	0.969	0.933
B + IM + TAM + CAM	0.005	0.856	0.972	0.934
B+ IM + TAM + CAM + SAM	0.004	0.859	0.974	0.937

3.5 Baseline Models

We select deep learning baseline models according to different categories such as edge, FCN, and high-resolution-based techniques. The chosen models are as follows:

- HDFN (Zhang *et al.*, 2019) utilizes a densely hierarchical feature fusion network that predicts the most critical area and segments the associated objects.
- AFNet (Feng *et al.*, 2019) predicts salient objects with entire structures and exquisite boundaries.
- HRSOD (Zeng *et al.*, 2019) leverages global semantic information and local high-resolution details to detect salient objects accurately in high-resolution images.
- SFCN (Zhang *et al.*, 2018) uses asymmetrical FCN to learn complementary visual features under the guidance of lossless feature reflection.

- Amulet (Zhang *et al.*, 2017) aggregates multi-level features into multiple resolutions.
- UCF (Zhang *et al.*, 2017) uses an encoder-decoder architecture to produce finer-resolution predictions. It learns uncertainty through a reformulated dropout in the decoder and avoids artifacts using a hybrid up-sampling scheme.

3.6 Comparison

We compared our model to DDCN (Zheng *et al.*, 2018), a technique for detecting camouflaged people, as well as to other state-of-the-art deep learning detection approaches, including Amulet (Zhang *et al.*, 2017), UCF (Zhang *et al.*, 2017), SFCN (Zhang *et al.*, 2018), HDFN (Zhang *et al.*, 2019), HRSOD (Zeng *et al.*, 2019) and AFNet (Feng *et al.*, 2019). Table 2 shows the comparison results of all methods on the four evaluation metrics. Obviously, our method outperforms competing approaches with a large margin across all the evaluation metrics, which demonstrates the superiority of the proposed model. Compared with the state-of-the-art method DDCN, our method improves F-measure and E-measure by 3.9% and 3.2%, respectively. Also, our model significantly lowers the MAE scores. This indicates that our model is more convinced of the predicted target regions and provides more accurate mask maps. S-measure, the most recent evaluation measure, has been used to emphasize the deficiencies of traditional evaluation metrics. When mask maps are evaluated, conventional metrics use pixels, which provide inadequate overall structural information. In the present study, our model still maintains the outstanding S-measure performance with an improvement of 6.5%. All quantitative results show that our model yields improved performance.

Table 2. Quantitative results on camouflaged people dataset. The best two scores are shown in red and blue colors, respectively.

Model	F	MAE	E	S
Amulet	0.349	0.081	0.587	0.630
UCF	0.254	0.168	0.497	0.589
SFCN	0.303	0.130	0.528	0.617
HDFN	0.313	0.140	0.534	0.618
HRSOD	0.489	0.018	0.741	0.690
AFNet	0.451	0.020	0.715	0.728
DDCN	0.820	0.007	0.942	0.872
Ours	0.859	0.004	0.974	0.937

In order to more intuitively illustrate the advantages of the proposed method, we visualize the prediction results of our network with DDCN in different scenarios. As shown in Figure 7, we observe that the proposed method highlights the hidden target completer and is more precise compared to DDCN. It excels in dealing with various challenging scenarios, different scales, and postures of people (rows 1 and 3), small objects (row3), occlusion (rows 2 and 4), and also accurately locating hidden targets (rows 5 and 6). From this comparison, the segmentation maps produced by our method are sharper and more accurate. Our model consistently outperforms DDCN. This can also illustrate the effectiveness of the proposed approach.

4. Discussion

It is noteworthy that disguised target detection is more complicated than salient object detection. The goal of salient object predictors is to identify and segment prominent features in images. To segment salient objects, we just need to focus on detecting such remarkable and discriminative areas. On the other hand, hidden targets fuse too much with their surroundings. It becomes even more difficult to discriminate between them from the background. The boundaries of the concealed target are therefore challenging to detect.

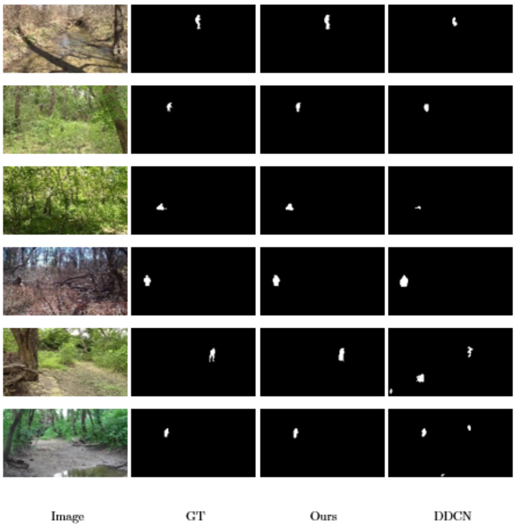

Fig. 7. Qualitative visual comparison of segmentation masks of our proposed model and DDCN.

There are two significant limitations in this study that could be addressed in future research. First, the study focused on evaluating the model on a single dataset. This is because the problem of hidden target detection is not well explored in the literature. The lack of diverse datasets remains the main issue. Thus, to promote advancements in hidden target detection and its evaluation, we aim to build a new, more challenging dataset for future work. Second, through the experimental evaluation and visual observation of the segmentation results, we found that not all the mask results of the images are satisfactory, and the segmentation results of some scenes are still deficient. Therefore, in future work, we plan to enhance the network for better detection.

5. Conclusion

In this paper, we have proposed a novel hidden target detection network that segments and effectively identifies the concealed target for accurate detection. Our model first extracts and enhances the features via enlarging receptive fields with different kernels. Then the multi-attention module is used to differentiate the disguised target from the background even more. Finally, based on the predicted mask result, the hidden target is identified. Experiments on the camouflaged people dataset demonstrate that our model is an effective detection model and outperforms state-of-the-art methods both qualitatively and quantitatively.

References

Achanta, R., Hemami, S., Estrada, F. & **Susstrunk, S. (2009).** Frequency-tuned salient region detection. In: Conference on Computer Vision and Pattern Recognition,1597–1604.

Bhajantri, N.U. & **Nagabhushan, P. (2006).** Camouflage Defect Identification: A Novel Approach. In Proc. International Conference on Information Technology, 145-148.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & **Yuille, A.L. (2016).** Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Dai, J., Qi, H., Xiong Y. ,Li Y. , Zhang, G., Hu, H. & **Wei, Y. (2017).** Deformable convolutional networks. International Conference on Computer Vision.

Fan, D.-P.,Cheng, M.-M., Liu, Y., Li, T. & **Borji, A. (2017).** Structure-measure: a new way to evaluate foreground maps. IEEE International Conference on Computer Vision, 4558–4567.

Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M. & Borji, A. (2018). Enhanced-alignment measure for binary foreground map evaluation. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 698–704.

Feng M., Lu H. & Ding, E. (2019). Attentive feedback network for boundary-aware salient object detection. IEEE Conference on Computer Vision and Pattern Recognition, 1623–1632.

Galun, M., Sharon, E., Basri, R. & **Brandt A. (2003).** Texture segmentation by multiscale aggregation of filter responses and shape elements. International Conference on Computer Vision, 716–723.

He, k., Gkioxari, G., Dollár, P. & **Girshick, R. B. (2017).** Mask R-CNN. IEEE International Conference on Computer Vision (ICCV).

He, K., Zhang, X., Ren, S. & **Sun, J. (2016).** Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama,S. & **Darrell, T.(2014).** Caffe: Convolutional architecture for fast feature embedding. ACM international conference on Multimedia, 675–678.

Krähenbühl, P. & **Koltun,V. (2011).** Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems, 109–117.

Pan,Y., Chen,Y., Fu, Q., Zhang, P. & **Xu, X. (2011).** Study on the camouflaged target detection method based on 3d convexity. Mod. Appl. Sci. 5 (152).

Perazzi, F., Krähenbühl, P., Pritch, Y. & **Hornung, A. (2012).** Saliency filters: Contrast based filtering for salient region detection. IEEE conference on computer vision and pattern recognition, 733–740.

Redmon, J., Divvala, S., Girshick R. & **Farhadi, A. (2016).** You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Simonyan, K. & **Zisserman, A. (2015).** Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

Song, L. & **Geng, W. (2010).** A New Camouflage Texture Evaluation Method Based on WSSIM and Nature Image Features. In Proc. International Conference on Multimedia Technology, 1-4.

Tankus, A. & **Yeshurun, Y. (2001).** Convexity-Based Visual Camouflage Breaking. Computer Vision and Image Understanding, 82(3),208-237.

Zeng, Y., Zhang, P., Zhang, J., Lin, Z. & **Lu, H. (2019).** Towards High-Resolution Salient Object Detection. IEEE International Conference Computer Vision, 7234–7243.

Zhang, P., Liu, W.,Lei, Y. & **Lu, H. (2019).** Hyperfusion-Net: Hyper-densely reflective feature fusion for salient object detection. Pattern Recognition, 521-533.

Zhang, P., Liu, W., Lu, H. & **Shen, C. (2018).** Salient object detection by lossless feature reflection. International Joint Conference on Artificial Intelligence, 1–8.

Zhang, P., Wang, D., Lu, H., Wang, H. & **Ruan, X. (2017).** Amulet: Aggregating Multi-Level Convolutional Features for Salient Object Detection. IEEE International Conference on Computer Vision.

Zhang, P., Wang, D., Lu, H., Wang, H. & **Yin, B. (2017).** Learning Uncertain Convolutional Features for Accurate Saliency Detection. IEEE International Conference on Computer Vision.

Zhao, H., Shi, J., Qi, X., Wang, X. & **Jia, J.(2017).** Pyramid scene parsing network. IEEE Conference on Computer Vision and Pattern Recognition, 2881–2890.

Zheng, Y., Zhang, X., Wang, F., Cao, T. & **Sun, M. (2018).** Detection of People with Camouflage Pattern Via Dense Deconvolution Network. IEEE Signal Processing Letters, 14(8), 29-33.

Submitted: 19/07/2021
Revised: 03/09/2021
Accepted: 12/10/2021
DOI: 10.48129/kjs.15249