Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection

Youwei Pang†, Xiaopi Zhao†, Tian-Zhu Xiang, Lihe Zhang* and Huchuan Lu
1Dalian University of Technology, China 2Peng Cheng Laboratory, China
3Inception Institute of Artificial Intelligence, UAE
{larptang, zxq}@mail.dlut.edu.cn, tianzhu.xiang19@gmail.com, {zhanglihe, lhchuan}@dlut.edu.cn

Abstract

The recently proposed camouflaged object detection (COD) attempts to segment objects that are visually blended into their surroundings, which is extremely complex and difficult in real-world scenarios. Apart from high intrinsic similarity between the camouflaged objects and their background, the objects are usually diverse in scale, fuzzy in appearance, and even severely occluded. To deal with these problems, we propose a mixed-scale triplet network, ZoomNet, which mimics the behavior of humans when observing vague images, i.e., zooming in and out. Specifically, our ZoomNet employs the zoom strategy to learn the discriminative mixed-scale semantics by the designed scale integration unit and hierarchical mixed-scale unit, which fully explores imperceptible clues between the candidate objects and background surroundings. Moreover, considering the uncertainty and ambiguity derived from indistinguishable textures, we construct a simple yet effective regularization constraint, uncertainty-aware loss, to promote the model to accurately produce predictions with higher confidence in candidate regions. Without bells and whistles, our proposed highly task-friendly model consistently surpasses the existing 23 state-of-the-art methods on four public datasets. Besides, the superior performance over the recent cutting-edge models on the SOD task also verifies the effectiveness and generality of our model. The code will be available at https://github.com/larptang/ZoomNet.

1. Introduction

Camouflaged objects are often “seamlessly” integrated into the environment by changing their appearance, coloration or pattern to avoid detection, such as chameleons, cuttlefishes and flatfishes. This is mainly due to their self-protection mechanism in the harsh living environment.

†These authors contributed equally to this work.
*Corresponding author.

Broadly speaking, camouflaged objects also refer to the objects that are extremely small in size, highly similar to the background, or heavily obscured. They subtly hide themselves in the surroundings, making them difficult to be found, e.g., soldiers wearing camouflaged uniforms and lions hiding in the grass. Camouflaged object detection (COD) is far more complex and challenging than traditional salient object detection or other object segmentation. Recently, it has attracted ever-growing research interest from the computer vision community and facilitates many valuable real-life applications, such as search and rescue [6], species discovery [36], and medical image analysis [8, 9, 12, 18, 55].

Recently, numerous deep learning-based methods have been proposed and achieved significant progress. Nevertheless, they are still struggled to accurately and reliably detect camouflaged objects, due to visual insignificance of camouflaged objects, and high diversity in scale, appearance.
and occlusion. By observing our experiments, it is found that the current COD detectors are susceptible to distractors from background surroundings. Thus it is difficult to excavate discriminative and subtle semantic cues for camouflaged objects, resulting in the inability to clearly segment the camouflaged objects from the chaotic background and the predictions of some uncertain (low-confidence) regions. Taking these into mind, in this paper, we summarize the COD issue into two aspects: 1) How to accurately locate camouflaged objects under conditions of inconspicuous appearance and various scales? 2) How to suppress the obvious interference from the background and infer camouflaged objects more reliably? Intuitively, to accurately find the vague or camouflaged objects in the scene, humans may try to refer to and compare the changes in the shape or appearance at different scales by zooming in and out (re-scaling) the image. This specific behavior pattern of human beings motivates us to identify camouflaged objects by mimicking the zooming in and out strategy.

With this inspiration, in this paper, we propose a mixed-scale triplet network, ZoomNet, which significantly improves the existing camouflaged object detection performance. **Firstly**, for accurate object location, we employ scale space theory [23, 24, 43] to imitate zooming in and out strategy. Specifically, we design two key modules, i.e., the scale integration unit (SIU) and the hierarchical mixed-scale unit (HMU). As shown in Fig. 1, our model extracts differentiated camouflaged object features at different “zoom” scales using the triplet architecture, then adopts SIUs to screen and aggregate scale-specific features, and utilizes HMUs to further reorganize and enhance mixed-scale features. Thus, our model is able to mine the accurate and subtle semantic clues between objects and background under the mixed scales, and produce accurate predictions. Besides, we use the shared weight strategy, which achieves a good balance of efficiency and effectiveness. **Secondly**, it is related to reliable prediction in complex scenarios. Although the object is accurately located, the indistinguishable texture and background will easily bring negative effects to the model learning, e.g. predicting uncertain/ambiguity regions, which greatly reduces the detection performance and cannot be ignored. This can be seen in Fig. 6 (Row 3 and 4) and Fig. 1 in the supp. To this end, we design an uncertainty-aware loss (UAL) to guide the model training, which is only based on the prior knowledge that a good COD prediction should have a clear polarization trend. Its GT-independent characteristic makes it suitable for enhancing the GT-based BCE loss. This targeted enhancement strategy can force the network to optimize the prediction of the uncertain regions during the training process, enabling our ZoomNet to distinguish the uncertain regions and segment the camouflaged objects reliably.

Our contributions can be summarized as follows: 1) For the COD task, we propose a mixed-scale triplet network, ZoomNet, which can credibly capture the objects in complex scenes by characterizing and unifying the scale-specific appearance features at different “zoom” scales and the purposeful optimization strategy. 2) To obtain the discriminative feature representation of camouflaged objects, we design SIUs and HMUs to distill, aggregate and strengthen the scale-specific and subtle semantic representation for accurate COD. 3) We propose a simple yet effective optimization enhancement strategy, UAL, which can significantly suppress the uncertainty and interference from the background without increasing extra parameters. 4) Our model greatly surpasses recent 23 state-of-the-art methods under seven metrics on four COD datasets. Furthermore, it shows good generalization in the SOD task and the superior performance compared with the existing SOD methods.

### 2. Related Work

**Camouflaged Object.** The study of camouflage has a long history in biology. This behavior of creatures in nature can be regarded as the result of natural selection and adaptation. In fact, in human life and other parts of society, it also has a profound impact, e.g., arts, popular culture, and design. More details can be found in [40]. In the field of computer vision, research on camouflaged objects is often associated with salient object detection (SOD), which mainly deals with those salient and easily observed objects in the scene. In general, saliency models are designed for the general observation paradigm (i.e., finding visually prominent objects). They are not suitable for the specific observation (i.e., finding concealed objects). Therefore, it is necessary to establish models based on the essential requirements and specific data of the task to learn the special knowledge.

**Camouflaged Object Detection (COD).** Different from the traditional SOD task, the COD pays more attention to the undetectable objects (mainly because of too small size, occlusion, concealment or self-disguise). Due to the differences in the attributes of the objects of interest, the goals of the two tasks are different. The difficulty and complexity of the COD far exceed the SOD due to the high similarity between the object and the environment. Some valuable attempts have been made in recent years. Recent works [19, 29, 48] construct the multi-task learning framework in the prediction process of camouflaged objects and introduce some auxiliary tasks like classification and edge detection. Some uncertainty-aware methods [20, 47] are proposed to model and cope with the uncertainty in data annotation or COD data itself. In the other two methods [31, 41], contextual feature learning also plays an important role. There are also a number of bio-inspired methods, such as [7, 46]. They capture camouflaged objects by imitating the behavior process of hunters or changing the viewpoint of the scene. Although our method can also be at-
tributed to the last category, ours is different from the above methods. Our method simulates the behavior of humans to understand complex images by zooming in and out strategy. The proposed method explores the scale-specific and imperceptible semantic features under the mixed scales for accurate predictions, with the supervision of BCE and our proposed uncertainty-aware loss. Accordingly, our method achieves a more comprehensive understanding of the scene, and accurately and robustly segments the camouflaged objects from the complex background, which even can be transferred to the SOD task effectively and smoothly.

Scale Space Integration. The scale-space theory aims to promote an optimal understanding of image structure, which is an extremely effective and theoretically sound framework for addressing naturally occurring scale variations. Its ideas have been widely used in computer vision, including the image pyramid [2] and the feature pyramid [22]. Due to the structural and semantic differences at different scales, the corresponding features play different roles. However, the commonly-used inverted pyramid-like feature extraction structures [13, 52] often cause the feature representation to lose too much texture and appearance details, which are unfavorable for dense prediction tasks [27, 38] that emphasize the integrity of regions and edges. Thus, some recent CNN-based COD methods [3, 7, 31, 41, 48] and SOD methods [14–17, 21, 25, 34, 35, 53, 54, 56] explore the combination strategy of inter-layer features to enhance the feature representation. These bring some positive gains for accurate localization and segmentation of objects. However, for the COD task, the existing approaches overlook the performance bottleneck caused by the ambiguity of the structural information of the data itself that makes it difficult to be fully perceived at a single scale. Different from them, we mimic the zoom strategy to synchronously consider differentiated relationships between object and background at multiple scales, thereby fully perceiving the camouflaged objects and confusing scenes. Besides, we also further explore the fine-grained feature scale space between channels.

3. Proposed Method

In this section, we first elaborate on the overall architecture of the proposed ZoomNet, and then present the details of each module and the uncertainty-aware loss.

3.1. Overall Architecture

The overall architecture of the proposed ZoomNet is illustrated in Fig. 2. Inspired by the zoom strategy from human beings when observing confusing scenes, we argue that different zoom scales often contain their specific information. Aggregating the differentiated information on different scales will benefit exploring the inconspicuous yet valuable clues from confusing scenarios, thus facilitating COD. To implement it, intuitively, we resort to the image pyramid. Specifically, we customize an image pyramid based on the single scale input to identify the camouflaged objects. The scales are divided into a main scale (i.e., the input scale) and two auxiliary scales. The latter is obtained by re-scaling the former to imitate the operation of zooming in and out. We utilize the shared triplet feature encoder to extract features on different scales and feed them to the scale merging layer. To integrate these features that contain rich scale-specific information, we design a series of scale integration units (SIUs) based on the attention-aware filtering mechanism. Thus, these auxiliary scales are integrated into the main scale, i.e., information aggregation of ”zoom in and out” operation. This will largely enhance the model to distill crit-
ical and informative semantic cues for capturing difficult-to-detect camouflaged objects. After that, we construct hierarchical mixed-scale units (HMUs) to gradually integrate multi-level features in a top-down manner to enhance the mixed-scale feature representation. It further increases the receptive field range and diversifies feature representation within the module. The captured fine-grained and mixed-scale clues promote the model to accurately segment the camouflaged objects in the chaotic scenes. Besides, to overcome the uncertainty in the prediction caused by the inherent complexity of the data, we design an uncertainty-aware loss (UAL) to assist the BCE loss, enabling the model to distinguish these uncertain regions and produce an accurate and reliable prediction.

3.2. Triplet Feature Encoder

We start by extracting deep features through a shared triplet feature encoder for the group-wise inputs, which consists of the feature extraction and the channel compression networks, i.e., E-Net and C-Net. For the trade-off between efficiency and effectiveness, the main scale and the two auxiliary scales are empirically set to $1.0 \times$, $1.5 \times$ and $0.5 \times$. E-Net is constituted by the commonly-used ResNet-50 [11] that is removed the structure after “layer4”. C-Net is cascaded to further optimize computation and obtain a more compact feature. For more details about it, please see the supp. Thus, three sets of 64-channel feature maps corresponding to three input scales are produced, i.e., $\{f^k_i\}_{k=1}^5$, $k \in \{0.5, 1.0, 1.5\}$. Next, these features are fed successively to the scale merging layer and the hierarchical mixed-scale decoder for subsequent processing.

3.3. Scale Merging Layer

We design an attention-based SIU to screen (weight) and combine scale-specific information, as shown in Fig. 3. Several such units make up the scale merging layer. Through filtering and aggregation, the expression of different scales is self-adaptively highlighted. Before scale integration, the features $f^{1.5}_i$ and $f^{0.5}_i$ are first resized to be consistent resolution with the main scale feature $f^{1.0}_i$. Specifically, for $f^{1.5}_i$, we use a hybrid structure of “max-pooling + average-pooling” to down-sample it, which helps to preserve the effective and diverse responses for camouflaged objects in high-resolution features. For $f^{0.5}_i$, we directly up-sample it by the bi-linear interpolation. Then, these features are fed into the “attention generator”, and a three-channel feature map is calculated through a series of convolutional layers. After a softmax activation layer, the attention map $A^K (k \in \{0.5, 1.0, 1.5\})$ corresponding to each scale can be obtained and used as respective weights for the final integration. The process is formulated as:

$$A_i = \text{softmax}(\Psi(\mathcal{U}(f^{0.5}_i, f^{1.0}_i, \mathcal{D}(f^{1.5}_i)), \phi)), \tag{1}$$

where $\Psi(\star, \phi)$ indicates the stacked “Conv-BN-ReLU” layers in the attention generator, and $\phi$ means the parameters of these layers. $\mathcal{\star}$ represents the concatenation operation. $\mathcal{D}$ and $\mathcal{U}$ refer to the hybrid pooling and bi-linear interpolation operations mentioned above, respectively. Note that some operations before and after the sampling operation are not shown in Equ. 1 for simplicity but can be seen in Fig. 3. These designs aim to selectively aggregate the scale-specific information to explore subtle but critical semantic cues at different scales, boosting the feature representation.

3.4. Hierarchical Mixed-scale Decoder

After SIUs, the auxiliary-scale information is integrated into the main-scale branch. Similar to the multi-scale case, different channels also contain differentiated semantics. Thus, it is necessary to excavate valuable clues contained in different channels. To this end, we design HMUs to conduct information interaction and feature refinement between channels, which strengthen features from coarse-grained group-wise iteration to fine-grained channel-wise modulation in the decoder, as depicted in Fig. 4. The input $f_i$ of the HMU$_i$ contains the multi-scale fused feature
Table 1. Comparisons of different methods on COD datasets. The best three results are highlighted in red, green and blue. ‘—’—Not available; ‘*’—Using more datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>CAMO</th>
<th>CHAMELEON</th>
<th>COD10K</th>
<th>NC4K</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANet-SRM [19]</td>
<td>0.682</td>
<td>0.484</td>
<td>0.126</td>
<td>0.541</td>
</tr>
<tr>
<td>SiNe [6]</td>
<td>0.745</td>
<td>0.644</td>
<td>0.092</td>
<td>0.702</td>
</tr>
<tr>
<td>SLSR [29]</td>
<td>0.787</td>
<td>0.696</td>
<td>0.080</td>
<td>0.744</td>
</tr>
<tr>
<td>MGLR [48]</td>
<td>0.775</td>
<td>0.673</td>
<td>0.088</td>
<td>0.726</td>
</tr>
<tr>
<td>PSNet [31]</td>
<td>0.782</td>
<td>0.695</td>
<td>0.083</td>
<td>0.746</td>
</tr>
<tr>
<td>USGC [20]</td>
<td>0.800</td>
<td>0.728</td>
<td>0.073</td>
<td>0.723</td>
</tr>
<tr>
<td>MIRNet [46]</td>
<td>0.785</td>
<td>0.719</td>
<td>0.077</td>
<td>0.754</td>
</tr>
<tr>
<td>C-UNet [11]</td>
<td>0.789</td>
<td>0.719</td>
<td>0.080</td>
<td>0.762</td>
</tr>
<tr>
<td>UGTR [47]</td>
<td>0.784</td>
<td>0.684</td>
<td>0.076</td>
<td>0.735</td>
</tr>
<tr>
<td>Ours</td>
<td>0.820</td>
<td>0.752</td>
<td>0.066</td>
<td>0.794</td>
</tr>
</tbody>
</table>

Camouflaged Object Detection

\[ f_i \text{ from the } \text{SIU}_i \text{ and the feature } f_{i+1} \text{ from the HMU}_{i+1}: \]
\[ f_i = f_i + U(f_{i+1}). \]  

Group-wise Iteration. We adopt a \( 1 \times 1 \) convolution to extend the channel number of feature map \( f_i \). The features are then divided into \( G \) groups \( \{g_{ij}\}_{j=1} \) along the channel dimension. Feature interaction between groups is carried out in an iterative manner. Specifically, the first group \( \{g_{i1}\} \) is split into three feature sets \( \{g_{i1}^{1k}\}_{k=1} \) after a convolution block. Among them, the \( g_{i1}^{1k} \) is adopted for information exchange with the next group, and the other two are used for channel-wise modulation. In the \( g_{ij}^{th} \) \((1 < j < G)\) group, the feature \( g_{ij} \) is concatenated with the feature \( g_{i-1,j} \) from the previous group along the channel, followed by a convolution block and a split operation, which simultaneously divides this feature group into three feature sets. It is noted that the output of the group \( G \) with the similar input form to the previous groups only contains \( g_{i1}^{G} \) and \( g_{i1}^{G} \). Such an iterative mixing strategy learns to identify the critical clues from different channels and obtain a powerful feature representation. From another perspective, the iterative structure in HMU can be equivalent to a kernel pyramid structure.

Channel-wise Modulation. The features \( \{[g_{i1}^{G}]_{j=1}^{G}\} \) are concatenated and converted into the feature modulation vector \( \alpha \) by a small convolutional network, which is employed to weight another concatenated feature \( \{[g_{i1}^{3G}]_{j=1}^{G}\} \). The weighted feature is then processed by a convolutional layer, which is defined as:
\[ \tilde{f}_i = A(f_i + N(T(\alpha \cdot ([g_{i1}^{3G}]_{j=1}^{G})))). \]  

where \( A, N \) and \( T \) represent the activation layer, the normalization layer and the convolutional layer, respectively.

Based on five cascaded HMUs and several stacked convolutional layers, a single-channel logit map is obtained. The final confidence map \( P \) that highlights the camouflaged objects is then generated by a sigmoid function.

3.5. Loss Functions

The binary cross entropy loss (BCEL) is widely used in various image segmentation tasks and its mathematical form is:
\[ L_{BCEL} = -g_{ij} \log p_{ij} - (1 - g_{ij}) \log (1 - p_{ij}), \]  

where \( g_{ij} \in \{0, 1\} \) and \( p_{ij} \in [0, 1] \) denote the ground truth and the predicted value at position \((i, j)\), respectively. As shown in Fig. 6, due to the complexity of the COD data, if trained only under the BCEL, the model produces serious ambiguity and uncertainty in the prediction and fails to accurately capture objects, of which both will reduce the reliability of COD. To force the model to enhance “confidence” in decision-making and increase the penalty for fuzzy prediction, we design a strong constraint as the auxiliary of the BCEL, i.e., the uncertainty-aware loss (UAL).

In the final probability map of the camouflaged object, the pixel value range is \([0, 1]\), where 0 means the pixel belongs to the background, and 1 means it belongs to the camouflaged object. Therefore, the closer the predicted value is to 0.5, the more uncertain the determination about the property of the pixel is. To optimize it, a direct way is to use the ambiguity as the supplementary loss for these difficult samples. To this end, we first need to define the ambiguity measure of the pixel \( x \), which maximizes at \( x = 0.5 \) and
minimizes at \( x = 0 \) or \( x = 1 \). And as a loss, the function should be smooth and continuous with only a finite number of non-differentiable points. For brevity, we empirically consider two forms, \( \Phi^\text{pow}_{\alpha}(x) = 1 - |2x - 1|^\alpha \) based on the power function and \( \Phi^\text{exp}_{\alpha}(x) = e^{-(\alpha(x-0.5))^2} \) based on the exponential function. Besides, inspired by the form of the weighted BCE loss, we also try to use \( \omega = 1 + \Phi^\text{pow}_2(x) \) as the weight of BCE loss to increase the loss of hard pixels. After massive experiments (Sec. 4.3), the proposed UAL is formulated as \( l_{U,AL} = 1 - |2p_{i,j} - 1|^2 \). Finally, the total loss function can be written as:

\[
L = L_{BCEL} + \lambda \times L_{U,AL},
\]

where \( \lambda \) is the balance coefficient and we design three adjustment strategies of \( \lambda \), i.e., a fixed constant value, an increasing linear strategy, and an increasing cosine strategy in Sec. 4.3. The different forms and corresponding results are listed in the supp. From the results, we find that the increasing strategies, especially “cosine”, do achieve better performance. So, the cosine strategy is used by default.

4. Experiments
4.1. Experiment Setup
Datasets. We use four COD datasets, CAMO [19], CHAMELEON [39], COD10K [7] and NC4K [29]. CAMO consists of 1,250 camouflaged and 1,250 non-camouflaged images. CHAMELEON contains 76 hand-annotated images. COD10K includes 5,066 camouflaged, 3,000 background and 1,934 non-camouflaged images. NC4K is another large-scale COD testing dataset including 4,121 images from the Internet. Following the data partition of [7,29,31,48], we use all images with camouflaged objects in the experiments, in which 3,040 images from COD10K and 1,000 images from CAMO are used for training, and the rest ones for testing. Besides, we also show the performance on five SOD datasets in the supp.

Implementation Details. The proposed ZoomNet is implemented with PyTorch. As the settings in recent methods [7,29,31,48], the encoder is initialized with the parameters of ResNet-50 pretrained on ImageNet, and the remaining parts are randomly initialized. SGD with momentum 0.9 and weight decay 0.0005 is chosen as the optimizer. The learning rate is initialized to 0.05 and follows a linear warm-up and linear decay strategy. The entire model is trained for 40 epochs with a batch size of 8 in an end-to-end manner on an NVIDIA 2080Ti GPU. During training and inference, the main scale is \( 384 \times 384 \). Random flipping and rotating are employed to augment the training data.

4.2. Comparisons with State-of-the-arts
COD is an emerging field, so we introduce some methods for salient object detection and medical image segmentation for comparison. The results of all these methods come from existing public data or are generated by models.
that are retrained based on the code released by the authors.

**Quantitative Evaluation.** Tab. 1 shows the detailed comparison results. It can be seen that the proposed model consistently and significantly surpasses recent methods on all datasets without relying on any post-processing tricks. Compared with the recent best COD method UJSC, although it introduces extra SOD data for training and has suppressed other existing methods, our method still shows the obvious performance improvement on these datasets. Especially, our approach has more advantages on the metrics $F^p$, $MAE$, and $F_\beta$. On four datasets, the proposed method arewidely outperforms the second-best method C2F-Net by 19.3% in terms of $MAE$ and the average gains in terms of $F^p$ and $F_\beta$ are 4%. Besides, PR and $F_\beta$ curves shown in the *supp.* also demonstrate the effectiveness of the proposed method. The flatness of the $F_\beta$ curve reflects the consistency and uniformity of the prediction. Our curves are almost horizontal, which can be attributed to the effect of the proposed UAL. It drives the predictions to be more polarized and reduces the ambiguity.

**Qualitative Evaluation.** Visual comparisons of different methods on several typical samples are shown in Fig. 5. They present the complexity in different aspects, such as big objects (Col. 1), middle objects (Col. 2-8), small objects (Col. 9-13), occlusions (Col. 2 and 10), background interference (Col. 10-13), and indefinite boundaries (Col. 1, 2, 6-13). These results intuitively show the superior performance of the proposed method. In addition, it can be noticed that our predictions have clearer and more complete object regions and sharper contours.

### 4.3. Ablation Studies

In this section, we perform comprehensive ablation analyses on different components. Because COD10K is the most widely-used large-scale COD dataset, and contains various objects and scenes, all subsequent ablation experiments are carried out on it.

#### Effectiveness of SIUs and HMUs.
In the proposed model, both the SIU and the HMU are very important structures. We install them one by one on the baseline model to evaluate their performance. The results are shown in Tab. 2. Our baseline ① and other models ② and ③ only use the inputs of the main scale. As can be seen, our baseline shows a good performance, probably due to the proper training setup and the more reasonable network architecture detailed in the *supp.* From ①-③, it can be seen that the two proposed modules make a significant contribution to the performance when compared to the baseline. Besides, the results in Fig. 6 show that the two modules can benefit each other and reduce their errors (e.g., Col. 1, 2, 5 and 7) to locate and distinguish objects more accurately. These components effectively help the model to excavate and distill the critical and valuable semantics and improve the capability of distinguishing hard objects. Under the cooperation between the proposed modules and loss functions, ZoomNet can completely capture the camouflaged objects of different scales and generate the predictions with higher contrast and consistency. In addition, in Tab. 2, the model ④ is a simple extension of the baseline model ① using some standard convolutional blocks, to make the similar number of parameters and FLOPs with ④. The model ④ still achieves better performance, which reflects the effectiveness of the proposed modules and the rationality of the design.

#### Number of Groups in HMUs.
In Tab. 2, we also show the effects of different group numbers in the proposed HMU. It can be seen from the results that the best performance appears when the number of groups is equal to 6. Also, it achieves a good balance between performance and efficiency. So, in other experiments, we set the number of groups to 6.
Table 4. Different forms of the proposed UAL. Form 0 is our model without UAL. “—”: Unable to converge. Form 1.5 is used by default because of its balanced performance. Their curves are shown in the supp.

<table>
<thead>
<tr>
<th>No.</th>
<th>Form</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\omega$</th>
<th>$\beta$</th>
<th>MAE↑</th>
<th>$F_{\beta}$↑</th>
<th>$E_{\max}$↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>×</td>
<td>×</td>
<td>0.843</td>
<td>0.694</td>
<td>0.032</td>
<td>0.743</td>
<td>0.909</td>
<td></td>
</tr>
<tr>
<td>1.1, 1.2, 1.3</td>
<td>1/8, 1/4, 1/2</td>
<td>0.844</td>
<td>0.716</td>
<td>0.029</td>
<td>0.757</td>
<td>0.903</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>$\Phi_{\text{sm}}(x) = 1 - (2x - 1)^2$</td>
<td>1</td>
<td>0.838</td>
<td>0.729</td>
<td>0.029</td>
<td>0.766</td>
<td>0.911</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>2</td>
<td>0.834</td>
<td>0.727</td>
<td>0.029</td>
<td>0.763</td>
</tr>
<tr>
<td>1.6</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>4</td>
<td>0.833</td>
<td>0.725</td>
<td>0.029</td>
<td>0.760</td>
</tr>
<tr>
<td>1.7</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>8</td>
<td>0.833</td>
<td>0.720</td>
<td>0.029</td>
<td>0.757</td>
</tr>
<tr>
<td>2.1</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>1/8</td>
<td>0.844</td>
<td>0.698</td>
<td>0.032</td>
<td>0.744</td>
</tr>
<tr>
<td>2.2</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>1/4</td>
<td>0.845</td>
<td>0.700</td>
<td>0.032</td>
<td>0.746</td>
</tr>
<tr>
<td>2.3</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>1/2</td>
<td>0.843</td>
<td>0.701</td>
<td>0.031</td>
<td>0.746</td>
</tr>
<tr>
<td>2.4</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>1</td>
<td>0.842</td>
<td>0.713</td>
<td>0.030</td>
<td>0.754</td>
</tr>
<tr>
<td>2.5</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>2</td>
<td>0.839</td>
<td>0.720</td>
<td>0.030</td>
<td>0.761</td>
</tr>
<tr>
<td>2.6</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>4</td>
<td>0.839</td>
<td>0.706</td>
<td>0.032</td>
<td>0.752</td>
</tr>
<tr>
<td>2.7</td>
<td>$\Phi_{\text{sm}}(x) = e^{-</td>
<td>x-x_0</td>
<td>^2}$</td>
<td>8</td>
<td>0.841</td>
<td>0.698</td>
<td>0.032</td>
<td>0.745</td>
</tr>
<tr>
<td>3</td>
<td>BCE w/ $\omega = 1 + \Phi_{\text{sm}}(x)$</td>
<td>2</td>
<td>0.844</td>
<td>0.697</td>
<td>0.032</td>
<td>0.744</td>
<td>0.913</td>
<td></td>
</tr>
</tbody>
</table>

Mixed-scale Input Scheme. Our model is designed to mimic the behavior of "Zoom In&Out". The feature expression is enriched by combining the scale-specific information from different scales. In Fig. 1, the intermediate features and attention maps show that our mixed-scale scheme plays a positive and important role in locating the camouflaged object. Considering that the objects in COD10K are mainly small objects [7], which may limit the role of $0.5 \times$ input to some extent, we list average results on COD10K and CAMO in Tab. 3. The proposed scheme performs better than the single-scale one and simply mixed one. This verifies the rationality of such a design for the COD task.

Options of Setting $\lambda$. We compare three strategies and the results are listed in the supp., in which the increasing cosine strategy achieves the best performance. This may be due to the advantage of its smooth change process. This smooth intensity warm-up strategy of UAL motivates the model to improve the learning process and to mitigate the possible negative interference of UAL on BCEL due to the lower accuracy of the model during the early stage of training.

Forms of UAL. Different forms of UAL are listed in Tab. 4 and the corresponding curves are illustrated in the supp. As can be seen, Form 1.5 has a more balanced performance. Also, it is worth noting that, when approaching 0 or 1, the form which can maintain a larger gradient will obtain better performance in terms of $F_{\beta}$, MAE and $F_{\beta}$. This may provide some reference for designing a better loss.

Effectiveness of UAL. The results of Fig. 6 intuitively show that the UAL greatly reduces the ambiguity caused by the interference from the background. Besides, we visualize the histogram maps of all results on CHAMELEON and the intermediate features from different stages in the decoder in the supp. In the stacked histogram map “w/o UAL”, a large number of pixels appear in the middle area, which corresponds to more visually blurred/uncertain predictions. Besides, in the corresponding feature visualization, especially in the region inside the red box, there is clear background interference due to the complex scenarios and blurred edges, which are extremely prone to yield false positive predictions. However, when UAL is introduced, it can be seen that the middle interval of “w UAL” is flatter than the one of “w/o UAL”, that is, most pixel values approach two extremes. And the feature maps become more discriminative and present a more compact and complete response in the regions of camouflaged objects.

4.4 Discussion on SOD and COD

It can be seen from the experiments in the supp. that our method not only performs well on COD, but also shows outstanding performance on SOD. Considering the difference between these two tasks, we may wonder why our method consistently performs well on such two seemingly different tasks. We attribute this to the generality and rationality of the designed structure. In fact, SOD and COD have a clear commonality, i.e., the accurate segmentation has to depend on multi-scale and category-free discriminative features. By integrating rich scale-specific features, our model can extract critical and informative cues from scenes and objects, which helps precise localization and smooth segmentation of objects. In addition, the proposed UAL can mitigate the ambiguity of predictions caused by the inherent complexity of scenes. Although the objects in COD are salient, it can also benefit from UAL due to the vagueness introduced by the CNN model itself in the detailed information recovery process. All of these components are built on the common demand of the two tasks, which provides a solid foundation for the performance.

5. Conclusion

In this paper, we propose the ZoomNet by imitating the behavior of human beings to zoom in and out on images. This process actually considers the differentiated expressions about the scene from different scales, which helps to improve the understanding and judgment of camouflaged objects. We first filter and aggregate scale-specific features through the scale merging layer to enhance feature representation. Next, in the hierarchical mixed-scale decoder, the strategies of grouping, mixing and fusion further mine the mixed-scale semantics. Lastly, we introduce the uncertainty-aware loss to penalize the ambiguity of the prediction. Extensive experiments verify the effectiveness of the proposed method in both the COD and SOD tasks with superior performance to existing state-of-the-art methods.

Acknowledgements This work was supported by the National Natural Science Foundation of China #61876202 and #61829102, the Liaoning Natural Science Foundation #2021-KF-12-10, and the Fundamental Research Funds for the Central Universities #DUT20ZD212.
References


[51] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance


[53] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In AAAI Conference on Artificial Intelligence, 2022. 3


