https://doi.org/10.26599/AIR.2024.9150038 CAAI Artificial Intelligence Research

Bilateral Reference for High-Resolution Dichotomous Image Segmentation

Peng Zheng[1, 2], Dehong Gao[3], Deng-Ping Fan[1] ✉, Li Liu[4], Jorma Laaksonen[5], Wanli Ouyang[2], and Nicu Sebe[6]

ABSTRACT

We introduce a novel bilateral reference framework (BiRefNet) for high-resolution dichotomous image segmentation (DIS). It comprises two essential components: the localization module (LM) and the reconstruction module (RM) with our proposed bilateral reference (BiRef). LM aids in object localization using global semantic information. Within the RM, we utilize BiRef for the reconstruction process, where hierarchical patches of images provide the source reference, and gradient maps serve as the target reference. These components collaborate to generate the final predicted maps. We also introduce auxiliary gradient supervision to enhance the focus on regions with finer details. In addition, we outline practical training strategies tailored for DIS to improve map quality and the training process. To validate the general applicability of our approach, we conduct extensive experiments on four tasks to evince that BiRefNet exhibits remarkable performance, outperforming task-specific cutting-edge methods across all benchmarks. Our codes are publicly available at https://github.com/ZhengPeng7/BiRefNet.

KEYWORDS

dichotomous image segmentation; camouflaged object detection; salient object detection; bilateral reference; high-resolution segmentation

With the advancement in high-resolution (HR) image acquisition, image segmentation technology has evolved from traditional coarse localization to achieving high-precision object segmentation. This task, whether it involves salient[1] or concealed object detection[2, 3], is referred to as HR dichotomous image segmentation (DIS)[4] and has attracted widespread attention and use in the industry, e.g., by Samsung, Adobe, and Disney. Corresponding image quality assessment methods for HR segmentation have also been proposed for better and fairer evaluation[5–8].

For the new DIS task, recent works have considered strategies such as intermediate supervision[4], frequency prior[9], and unite-divide-unite[10], and have achieved favorable results. Essentially, they either split the supervision[4, 10] at the feature-level or introduce an additional prior[9] to enhance feature extraction. These strategies are, however, still insufficient to capture very fine features (see Fig. 1). Based on our observations, we found that fine and non-salient features in image objects can be well reflected by obtaining gradient features through derivative operations on the original image. In addition, when certain positions exhibit high similarity in color and texture to the background, the gradient features are probably too weak. For such cases, we further introduce ground-truth (GT) features for side supervision, allowing the framework to learn the characteristics of these positions. We name the incorporation of the image reference and the introduction of both the gradient and GT references as bilateral reference.

We propose a novel progressive bilateral reference network BiRefNet to handle the HR DIS task with separate localization and reconstruction modules. For the localization module, we extract hierarchical features from vision transformer backbone, which are combined and squeezed to obtain corase predictions in low resolution in deep layers. For the reconstruction module, we further design the inward and outward references as bilateral references (BiRef), in which the source image and the gradient map are fed into the decoder at different stages. Instead of resizing the original images to lower-resolution versions to ensure consistency with decoding features at each stage[11, 12], we keep the original resolution for intact detail features in inward reference and adaptively crop them into patches for compatibility with decoding features. In addition, we investigate and summarize practical strategies for the training on HR data, including long training and region-level loss for better segmentation in parts of fine details, and multi-stage supervision to accelerate learning of them.

Our main contributions are summarized as follows:

(1) We present a bilateral reference network (BiRefNet), which is a simple yet strong baseline to perform high-quality dichotomous image segmentation.

(2) We propose a bilateral reference module, which consists of an inward reference with source image guidance and an outward reference with gradient supervision. It shows great efficacy in the reconstruction of the predicted HR results.

(3) We explore and summarize various practical strategies

1 College of Computer Science, Nankai University, Tianjin 300350, China
2 Shanghai AI Laboratory, Shanghai 200232, China
3 School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China
4 College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China
5 Department of Computer Science, Aalto University, Espoo FI-02150, Finland
6 Department of Information Engineering and Computer Science, University of Trento, Trento I-38122, Italy
Address correspondence to Deng-Ping Fan, dengpfan@gmail.com

Fig. 1 Visual comparison between the results of our proposed BiRefNet and the latest state-of-the-art methods (e.g., IS-Net[4] and UDUN[10]) for high-resolution dichotomous image segmentation (DIS). Details of segmentation are zoomed in for better display.

tailored for DIS to easily improve performance, prediction quality, and convergence acceleration.

(4) The proposed BiRefNet shows its excellent performance and strong generalization capabilities to achieve state-of-the-art performance on not only the DIS5K task but also on HRSOD and COD with 6.8%, 2.0%, and 5.6% average S_m[13] improvements, respectively.

1 Related Work

1.1 High-resolution class-agnostic segmentation

High-resolution class-agnostic segmentation has been a typical computer vision objective for decades, and many related tasks have been proposed and attracted much attention, such as dichotomous image segmentation (DIS)[4], high-resolution salient object detection (HRSOD)[14], and concealed object detection (COD)[3]. To provide standard HRSOD benchmarks, several typical HRSOD datasets (e.g., HRSOD[14], UHRSD[15], HRS10K[16]) and numerous approaches[11, 14, 15, 17] have been proposed. Zeng et al.[14] employed a global-local fusion of the multi-scale input in their network. Xie et al.[15] used cross-model grafting modules to process images at different scales from multiple backbones (lightweight[18] and heavy[19]). Pyramid blending was also used in Ref. [11] for a lower computational cost. Concealed objects are difficult to locate due to similar-looking surrounding distractors[20]. Therefore, image priors, such as frequency[21], boundary[22], gradient[23], etc., are used as auxiliary guidance to train COD models. Furthermore, a higher resolution has been found beneficial for detecting targets[23–25]. To produce more precise and fine-detail segmentation results, Yin et al.[26] employed progressive refinement with masked separable attention. Li et al.[12] incorporated the original images at different scales to aid in the refining process.

High-resolution DIS is a newly proposed task that focuses more on the complex slender structure of target objects in high-resolution images, making it even more challenging. Qin et al.[4] proposed the DIS5K dataset and IS-Net with intermediate supervision to alleviate the loss of fine areas. In addition, Zhou et al.[9] embedded a frequency prior to their DIS network to capture more details. Pei et al.[10] applied a label decoupling strategy[27] to the DIS task and achieved competitive segmentation performance in the boundary areas of objects. Yu et al.[28] used patches of HR images to accelerate the training in a more memory-efficient way. Unlike previous models that used compressed/resized images to enhance HR segmentation, we utilized intact HR images as supplementary information for better predictions in high resolution.

1.2 Progressive refinement in segmentation

In the image matting task, trimaps have been used as a pre-positioning technique for more precise segmentation results[29, 30]. In Fig. 2, we illustrate different approaches of relevant networks and compare the differences[4, 9, 10, 31, 34]. Many approaches have been proposed based on the progressive refinement strategy. Yu et al.[35] used the predicted LR alpha matte as a guide for refined HR maps. In BASNet[36], the initial results are revised with an additional refiner network. The CRM[37] continuously aligns the feature map with the refinement target to aggregate detailed features. In ICNet[32], the original images are also downscaled and added to the decoder output at different stages for refinement. In ICEG[38], the generator and the detector iteratively evolve through interaction to obtain better COD segmentation results. In addition to images and GT, auxiliary information is also used in existing methods. For example, Tang et al.[39] cropped patches on the boundary to further refine them. In the LapSRN network[40, 41] for image super-resolution, Laplacian pyramids are also generated to help with image reconstruction at higher resolution. Although these methods successfully employed refinement to achieve better results, models are not guided to focus on certain areas, which is a problem in DIS. Therefore, we introduce gradient supervision in our outward reference to guide features sensitive to areas with richer fine details.

2 Methodology

2.1 Overview

As shown in Fig. 2d, our proposed BiRefNet is different from the previous DIS methods. On the one hand, our BiRefNet explicitly decomposes the DIS task on HR data into two modules, i.e., a localization module (LM) and a reconstruction module (RM). On

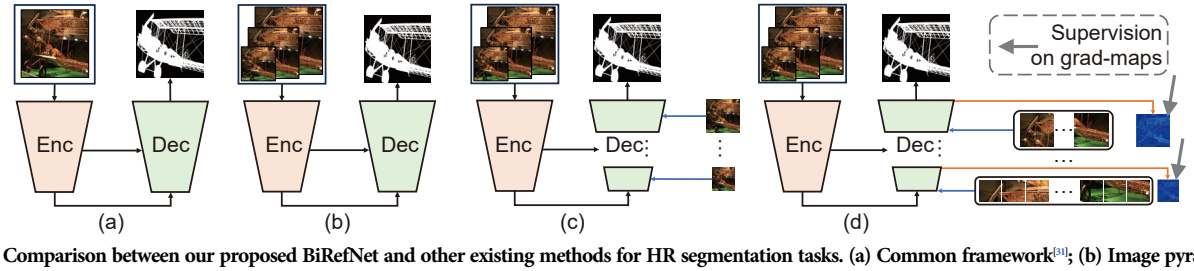

Fig. 2 Comparison between our proposed BiRefNet and other existing methods for HR segmentation tasks. (a) Common framework[31]; (b) Image pyramid as input[32, 33]; (c) Scaled images as inward reference[11, 12]; (d) BiRefNet: patches of images at original scales as inward reference and gradient priors as outward reference. Enc = encoder, Dec = decoder.

the other hand, instead of directly adding the source images[33] or the priors[9] to the input, BiRefNet employs our proposed bilateral reference in the RM, making full use of the source images at the original scales and the gradient priors. The complete framework of our BiRefNet is illustrated in Fig. 3.

2.2 Localization module

For a batch of HR images $\mathcal{I} \in \mathbb{R}^{N\times 3\times H\times W}$ as input, the transformer encoder[19] extracts features at different stages, i.e., $\mathcal{F}_1^e, \mathcal{F}_2^e, \mathcal{F}_3^e, \mathcal{F}^e$ with resolutions as $\left\{\left[\frac{H}{k}, \frac{W}{k}\right], k = 4, 8, 16, 32\right\}$. The features of the first four stages $\{\mathcal{F}_i^e\}_{i=1}^3$ are transferred to the corresponding decoder stages with lateral connections (1×1 convolution layers). Meanwhile, they are stacked and concatenated in the last encoder block to generate $\mathcal{F}^e$.

The encoder output feature $\mathcal{F}^e$ is then fed into a classification module, where $\mathcal{F}^e$ is led into a global average pooling layer and a fully connected layer for classification with the category C to obtain a better semantic representation for localization. HR features are squeezed in the bottleneck. To enlarge the receptive fields to cover features of large objects and focus on local features for high precision simultaneously[42], which is important for HR tasks involved, we employ ASPP modules[43] here for multi-context fusion. $\mathcal{F}^e$ is squeezed to $\mathcal{F}^d$ for transfer to the reconstruction module.

2.3 Reconstruction module

The setting of the receptive field (RF) has been a challenge of HR segmentation. Small RFs lead to inadequate context information to locate the right target on a large background, whereas large RFs often result in insufficient feature extraction in detailed areas. To achieve balance, we propose the reconstruction block (RB) in each BiRef block as a replacement for the vanilla residual blocks. In RB, we employ deformable convolutions[44] with hierarchical receptive fields (i.e., 1×1, 3×3, 7×7) and an adaptive average pooling layer to extract features with RFs of various scales. These features extracted by different RFs are then concatenated as $\mathcal{F}_i^\theta$, followed by a 1×1 convolution layer and a batch normalization layer to generate the output feature of RM $\mathcal{F}_i^{d'}$. In the reconstruction module, the squeezed feature $\mathcal{F}^d$ is fed into the BiRef block for the feature $\mathcal{F}_3^d$. With $\mathcal{F}_3^l$, the first BiRef block predicts coarse maps, which are then reconstructed into higher-resolution versions through the following BiRef blocks. Following Ref. [45], the output feature of each BiRef block $\mathcal{F}_i^d$ is added with its lateral feature $\mathcal{F}_i^l$ of the LM at each stage, i.e., $\{\mathcal{F}_i^{d+} = \text{Upsample}\uparrow(\mathcal{F}_i^d + \mathcal{F}_i^l), i =, 3, 2, 1\}$. Meanwhile, all BiRef blocks generate intermediate predictions $\{\mathcal{M}_i\}_{i=3}^1$ by multi-stage supervision, with resolutions in ascending order. Finally, the last decoding feature $\mathcal{F}_1^{d+}$ is passed through an 1×1 convolution layer to obtain the final predicted maps $\mathcal{M} \in \mathbb{R}^{N\times 1\times H\times W}$.

2.4 Bilateral reference

In DIS, HR training images are very important for deep models to learn details and perform highly accurate segmentation. However, most segmentation models follow previous works[31, 45] to design the network architecture in an encoder-decoder structure with down-sampling and up-sampling, respectively. Besides, due to the large size of the input, concentrating on the target objects becomes more challenging. To deal with these two main problems, we propose bilateral reference, consisting of an inward reference (InRef) and an outward reference (OutRef), which is illustrated in Fig. 4. Inward reference and outward reference play the roles of supplementing HR information and drawing attention to areas with dense details, respectively.

In InRef, images $\mathcal{I}$ with original high resolution are cropped to patches $\{\mathcal{P}_{k=1}^N\}$ of consistent size with the output features of the corresponding decoder stage. These patches are stacked with the original feature $\mathcal{F}_i^{d+}$ to be fed into the RM. Existing methods with similar techniques either add $\mathcal{I}$ only at the last decoding stage[10] or resize $\mathcal{I}$ to make it applicable with original features in low resolution. Our inward reference avoids these two problems through adaptive cropping and supplies the necessary HR information at every stage.

In OutRef, we use gradient labels to draw more attention to areas of richer gradient information, which is essential for the segmentation of fine structures. First, we extract the gradient maps of the input images as G_i^{gt}. Meanwhile, $\mathcal{F}_i^\theta$ is used to generate the feature $\mathcal{F}_i^G$ to produce the predicted gradient maps $\hat{G}_i$. With this gradient supervision, $\mathcal{F}_i^G$ is sensitive to the gradient. It passes through a conv and a sigmoid layer and is used to generate the

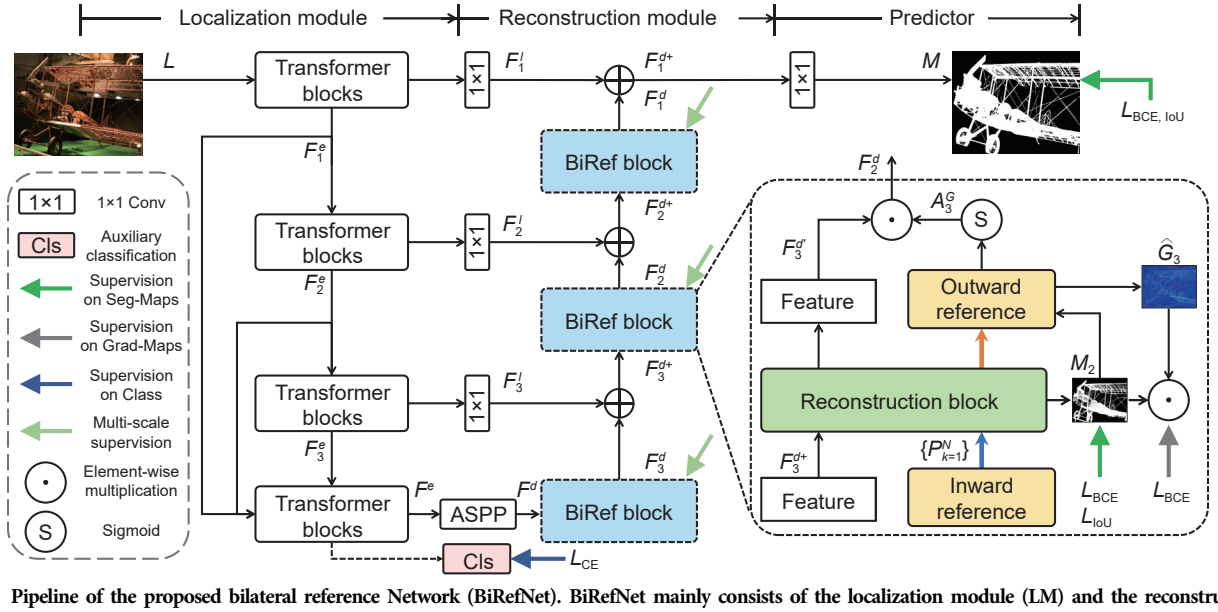

Fig. 3 Pipeline of the proposed bilateral reference Network (BiRefNet). BiRefNet mainly consists of the localization module (LM) and the reconstruction module (RM) with bilateral reference (BiRef) blocks. Please refer to Section 3.1 for details.

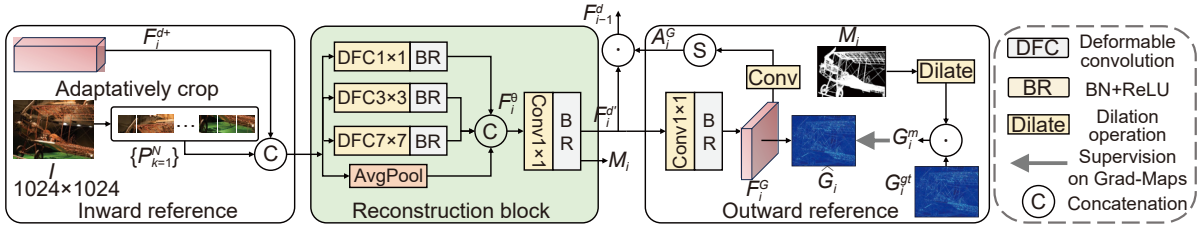

Fig. 4 Pipeline of the proposed bilateral reference blocks. The source images at the original scale are combined with decoder features as the inward reference and fed into the reconstruction block, where deformable convolutions with hierarchical receptive fields are employed. The aggregated features are then used to predict the gradient maps in the outward reference. Gradient-aware features are then turned into the attention map to act on the original features.

gradient referring attention $\mathcal{A}_i^G$, which is then multiplied by $\mathcal{F}_i^{d'}$ to generate output of the BiRef block as $\mathcal{F}_{i-1}^d$.

Considering that the background may have non-target noise with a lot of gradient information, we apply a masking strategy to alleviate the influence of non-target areas. We perform morphological operations on intermediate predictions $\mathcal{M}_i$ and use dilated $\mathcal{M}_i$ as a mask. The mask is used to multiply the gradient map G_i^{gt} to generate G_i^m, where the gradients outside the mask area are removed.

2.5 Objective function

In HR segmentation tasks, using only pixel-level supervision (BCE loss) usually results in the deterioration of detailed structural information in HR data. Inspired by the great results in Ref. [36] which uses a hybrid loss, we use BCE, IoU, SSIM, and CE losses together to collaborate for the supervision on the levels of pixel, region, boundary, and semantic, respectively. The final objective function is a weighted combination of the above losses and can be formulated as:

$$\begin{aligned} L &= L_{\text{pixel}} + L_{\text{region}} + L_{\text{boundary}} + L_{\text{semantic}} = \\ &\lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{IoU}} + \lambda_3 L_{\text{SSIM}} + \lambda_4 L_{\text{CE}} \end{aligned} \tag{1}$$

where λ_1, λ_2, λ_3, and λ_4 are respectively set to 30, 0.5, 10, and 5 to keep all the losses on the same quantitative level at the beginning of the training. The final objective function consists of binary cross-entropy (BCE) loss, intersection over union (IoU) loss, structural similarity index measure (SSIM) loss, and cross-entropy (CE) loss. The complete definition of losses can be found below.

● BCE loss: pixel-aware supervision, which is used for pixel-level supervision for the generation of binary maps:

$$L_{\text{BCE}} = -\sum_{(i,j)} [G(i,j)\log(M(i,j)) + (1-G(i,j))\log(1-M(i,j))] \tag{2}$$

where $G(i,j)$ and $\mathcal{M}(i,j)$ denote the value of the GT and binarized predicted maps, respectively, at pixel (i,j).

● IoU loss: region-aware supervision for the enhancement of binary map predictions:

$$L_{\text{IoU}} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} M(i,j)G(i,j)}{\sum_{r=1}^{H}\sum_{c=1}^{W} [M(i,j) + G(i,j) - M(i,j)G(i,j)]} \tag{3}$$

● SSIM loss: boundary-aware supervision to improve the accuracy in boundary parts. Given GT maps G and predicted maps $\mathcal{M}$, $y = \{y_j : j = 1,2,\ldots,N^2\}$ and $x = \{x_j : j = 1,2,\ldots,N^2\}$ represent the pixel values of two corresponding $N \times N$ patches derived from G and $\mathcal{M}$, respectively. $\text{SSIM}(x,y)$ is defined as

$$L_{\text{SSIM}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

where μ_x, μ_y and σ_x, σ_y are the means and standard deviations of x and y, respectively, and σ_{xy} is their covariance. C is used to avoid division by zero.

● CE loss: semantic-aware supervision, which is used to learn better semantic representation:

$$L_{\text{CE}} = -\sum_{c=1}^{N} y_{o,c}\log(p_{o,c}) \tag{5}$$

where N is the number of classes, $y_{o,c}$ states whether class label c is the correction classification for observation o, and $p_{o,c}$ denotes the predicted probability that o is of class c.

2.6 Training strategies tailored for DIS

Due to the high cost of training models on HR data, we have explored training tricks for HR segmentation tasks to improve performance and reduce training costs.

First, we found that our model converges relatively quickly in the localization of targets and the segmentation of rough structures (measured by F-measure[46], S-measure[13]) on DIS5K (e.g., 200 epochs). However, the performance in segmenting fine parts is still increasing after very long training (e.g., 400 epochs), which is reflected in metrics such as F_β^w and HCE_γ. Second, though long training can easily achieve great results in terms of both structure and edges, it consumes too much computation; we found that multi-stage supervision can dramatically accelerate the learning on segmenting fine details and make the model achieve similar performance as before but with only 30% training epochs. Third, we also found that fine-tuning with only region-level losses can easily improve the binarization of predicted results and those metric scores (e.g., F_β^w, E_φ^m, HCE) that are closer to practical use. Finally, we used context feature fusion and image pyramid inputs on the backbone, which are commonly used tricks to process HR images with deep models. In experiments, these two modifications to the backbone achieved a general improvement in DIS and similar HR segmentation tasks.

As shown in Table 1, we show the effectiveness of training epochs and the multi-stage supervision. As the results show, our BiRefNet can achieve relatively good results after 200 epochs of training. Continuous training in 400 epochs can increase a small portion of metrics measuring structural information (e.g., F_β^x, S_m), while bringing a larger improvement in metrics measuring fine details (e.g., HCE_γ).

Although simple long training can achieve better results, the improvement is relatively small, concerning its high computational cost on HR data. We investigated the multi-stage supervision (MSS), which is a widely used training strategy used in binary segmentation works[4, 47]. Different from MSS in these works for higher precision, it plays a role in accelerating the training convergence. As the results in Table 1 show, our BiRefNet trained for 200 epochs with MSS can achieve similar performance with it trained for 400 epochs. MSS successfully cut training time in half

Table 1 Quantitative ablation studies of the proposed multi-stage supervision for acceleration and training epochs. "↑" ("↓") means that the higher (lower) is better.

Setting		$F_\beta^x \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\varphi^m \uparrow$	$\text{HCE}_\gamma \downarrow$
MSS	Epoch						
	200	0.875	0.848	0.041	0.886	0.914	1207
	400	0.897	0.863	0.036	0.905	0.937	1039
√	200	0.892	0.858	0.037	0.901	0.932	1043

and can be used for further HR segmentation tasks for more efficient training.

3 Experiment

3.1 Datasets

Training sets. For DIS, we follow Refs. [4, 9, 10] to use DIS5K-TR as our training set in experiments. For HRSOD, we follow Ref. [15] to set different combinations of HRSOD, UHRSD, and DUTS as the training set. For COD, we follow Refs. [3, 24] to use the concealed samples in CAMO-TR and COD10K-TR as the training set.

Test sets. To obtain a complete evaluation of our BiRefNet, we tested it on all test sets in DIS5K (DIS-TE1, DIS-TE2, DIS-TE3, and DIS-TE4). We also conducted an evaluation of BiRefNet on the HRSOD test sets (DAVIS-S[14], HRSOD-TE[14], and UHRSD-TE[15]) and the COD test sets (CAMO-TE[48], COD10K-TE[20], and NC4K[49]). Low-resolution SOD test sets (DUTS-TE[50] and DUT-OMRON[51]) are additionally used for supplementary experiments.

3.2 Evaluation protocol

For a comprehensive evaluation, we employ the widely used metrics, i.e., S-measure[13] (S_m), max/mean/weighted F-measure[46] ($F_\beta^x/F_\beta^m/F_\beta^\omega$), max/mean E-measure[52] (E_ξ^x/E_ξ^m), mean absolute error (MAE), and relax HCE[4] (HCE_γ) to evaluate performance. Detailed descriptions of these metrics can be found as follows.

- **S-measure**[13] (structure measure, S_α) is a structural similarity measurement between a saliency map and its corresponding GT map. Evaluation with S_α can be obtained at high speed without binarization. The S_α-measure is computed as

$$S_\alpha = \alpha \cdot S_o + (1-\alpha) \cdot S_r \tag{6}$$

where S_o and S_r denote object-aware and region-aware structural similarity, and α is set to 0.5 by default, as suggested by Fan et al. in Ref. [13].

- **F-measure**[46] (F_β) is designed to evaluate the weighted harmonic mean value of precision and recall. The output of the saliency map is binarized with different thresholds to obtain a set of binary saliency predictions. The predicted saliency maps and GT maps are compared to obtain precision and recall values. F-measure can be computed as

$$F_\beta = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{7}$$

where β^2 is set to 0.3 to emphasize precision over recall, following Ref. [53]. The maximum F-measure score obtained with the best threshold for the entire dataset is used and denoted as F_β^x.

- **E-measure**[52] (enhanced-alignment measure, E_ξ) is designed as a perceptual metric to evaluate the similarity between the predicted maps and the GT maps both locally and globally. E-measure is defined as

$$E_\xi = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\varphi_\xi(x,y) \tag{8}$$

where φ_ξ indicates the enhanced alignment matrix. Similarly to the F-measure, we also adopt the maximum E-measure (E_ξ^x) and also the mean E-measure (E_ξ^m) as our evaluation metrics.

- **MAE** (mean absolute error, ϵ) is a simple pixel-level evaluation metric that measures the absolute difference between non-binarized predicted results $\mathcal{M}$ and GT maps G. It is defined as

$$\epsilon = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}|\hat{Y}(x,y) - G(x,y)| \tag{9}$$

- HCE_γ[4] (human correction efforts, HCE_γ) is a newly proposed metric aimed at evaluating the human efforts required to correct faulty predictions to satisfy specific accuracy requirements in real-world applications. Specifically, HCE is quantified by the approximate number of mouse clicks. In practical applications, minor prediction errors can be tolerated. Therefore, relax HCE (HCE_γ) is introduced, where γ denotes tolerance. In experiments, we use HCE_5 (relax HCE with $\gamma = 5$) to stay consistent with that of the original paper[1], where detailed descriptions of HCE_γ are provided.

3.3 Implementation details

All images are resized to 1024×1024 for training and testing. The generated segmentation maps are resized (i.e., bilinear interpolation) to the original size of the corresponding GT maps for evaluation. Horizontal flip is the only data augmentation used in the training process. The number of categories C is set to 219, as given in DIS-TR. {The proposed BiRefNet is trained with Adam optimizer[54] for DIS/HRSOD/COD tasks for 600/150/150 epochs, respectively. The model is fine-tuned with the IoU loss for the last 20 epochs. The initial learning rate is set to 10^{-4} and 10^{-5} for DIS and others, respectively. Models are trained with PyTorch[55] on eight NVIDIA A100 GPUs. The batch size is set to N=4 for each GPU during training.}

3.4 Ablation study

We study the effectiveness of each component (i.e., RM and BiRef) and practical strategies (i.e., context feature fusion (CFF), image pyramids input (IPT), and regional loss fine-tuning (RLFT)) introduced for our BiRefNet and conduct an investigation about their contributions to improved DIS results. Quantitative results regarding each module and strategy are shown in Tables 2 and 3, respectively.

Baseline. We provide a simple but strong encoder-decoder network as the baseline for the DIS task. To capture better hierarchical features on various scales, we chose the Swin transformer large[19] as our default backbone network. Then, to obtain a better semantic representation in the DIS task, we divided the images in DIS-TR into 219 classes according to their label names and added an auxiliary classification head at the end of the

Table 2 Quantitative ablation studies of the proposed components in the proposed BiRefNet, including reconstruction module (RM), inward reference (InRef), outward reference (OutRef), and their combinations.

Module			$F_\beta^x \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\varphi^m \uparrow$	$\text{HCE}_\gamma \downarrow$
RM	InRef	OutRef						
			0.837	0.785	0.056	0.845	0.887	1204
√			0.855	0.831	0.048	0.865	0.895	1167
	√		0.848	0.825	0.050	0.857	0.903	1152
√	√		0.869	0.834	0.041	0.886	0.912	1093
√		√	0.863	0.831	0.042	0.891	0.918	1106
	√	√	0.861	0.839	0.044	0.881	0.911	1114
√	√	√	0.889	0.851	0.038	0.900	0.924	1065

Table 3 Effectiveness of practical strategies for training high-resolution segmentation, including context feature fusion (CFF), image pyramids input (IPT), regional loss fine-tuning (RLFT), and their combinations. The results are obtained by our final model.

Module			$F_\beta^x \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\varphi^m \uparrow$	$\text{HCE}_\gamma \downarrow$
CFF	IPT	RLFT						
			0.889	0.851	0.038	0.900	0.924	1065
√			0.893	0.856	0.038	0.904	0.928	1054
	√		0.895	0.857	0.037	0.904	0.927	1051
		√	0.890	0.861	0.036	0.899	0.932	1043
√	√	√	0.897	0.863	0.036	0.905	0.937	1039

encoder. In the decoder of the baseline network, each decoder block is made up of two residual blocks[18]. All stages of the encoder and decoder are connected with an 1×1 convolution, except the deepest stage, where an ASPP[43] block is used for connectivity. With this setup, our baseline network has outperformed existing DIS models in most metrics, as shown in Tables 2 and 4.

Reconstruction module. As shown in Table 2, our model gains an overall improvement with the proposed RM. The RM provides multi-scale receptive fields on the HR features for local details and overall semantics. It brings ~2.2% F_β^x relative improvement with little extra computational cost.

Bilateral reference. We separately investigate the effectiveness of the inward reference (InRef, with source images) and the outward reference (OutRef, with gradient labels) in BiRef. InRef supplemented lossless HR information globally, while OutRef drew more attention to the fine-detail parts to achieve higher precision in those areas. As shown in Table 2, they work jointly to bring 2.9% F_β^x relative improvement to BiRefNet. RM and BiRef are combined to achieve 6.2% F_β^x relative improvement.

Training strategies. As shown in Table 3, the proposed strategies improve performance from different perspectives. CCF and IPT improve overall performance, while RLFT specifically improves precision in edge details, which is reflected in metrics such as F_β^ω and HCE_γ.

3.5 State-of-the-art comparison

To validate the general applicability of our method, we conduct extensive experiments on four tasks, i.e., high-resolution DIS, HRSOD, COD, and salient object detection (SOD). We compare our proposed BiRefNet with all the latest task-specific models on existing benchmarks[3,4,14,15,48–51].

Quantitative results. Table 4 shows a quantitative comparison between the proposed BiRefNet and previous state-of-the-art methods. Our BiRefNet outperforms all previous methods in widely used metrics. The complexities of DIS-TE1 to DIS-TE4 are in ascending order. The metrics for structure similarity (e.g., S_α, E_φ^x) focus more on global information. Pixel-level metrics, such as MAE ($\mathcal{M}$), emphasize the precision of details. Metrics based on mean values (e.g., E_φ^m, F_φ^m) better match the requirements of practical applications where maps are thresholded. As seen in Table 4, our BiRefNet outperforms previous methods not only on the accuracy of the global shape but also in the details of the pixels. It is noteworthy that the results are better, especially in metrics that cater more to practical applications.

Additionally, our BiRefNet outperforms existing task-specific models on the HRSOD and COD tasks. As shown in Table 5, BiRefNet achieved much higher accuracy on both high-resolution and low-resolution SOD benchmarks. Compared with the previous state-of-the-art method[15], our BiRefNet achieved an average improvement of 2.0% S_m. Furthermore, as shown in Table 6, in the COD task, BiRefNet also shows a much better performance compared to the previous state-of-the-art models, with an average improvement of 5.6% S_m on the three widely used COD benchmarks. These results show the remarkable generalization ability of our BiRefNet to similar HR tasks.

For a clearer illustration of the generalizability and powerful performance of BiRefNet, we provide a radar picture shown in Fig. 5, where we run our model and the best task-specific models on DIS/HRSOD/COD/SOD tasks. As the results show, our BiRefNet achieves leading results in all four tasks. The other task-specific models show their weakness in similar HR segmentation tasks. For example, FSPNet[24] ranks second in COD benchmarks, while it ranks fourth/third/third in DIS/HRSOD/SOD tasks, respectively.

Qualitative results. Figure 6 shows segmentation maps produced by the most competitive existing DIS models and the proposed BiRefNet. As the results show, we provide samples of all test sets and one validation set. BiRefNet outperforms the previous DIS methods from two perspectives, i.e., the location of target

Table 4 Quantitative comparisons between our BiRefNet and the state-of-the-art methods on DIS5K. We use the results from Ref. [10], where all methods take 1024 × 1024 input. The size of each dataset is marked behind.

Method	DIS-TE1 (500)						DIS-TE2 (500)						DIS-TE3 (500)					
	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$
BASNet$_{19}$ [34]	0.663	0.577	0.105	0.741	0.756	155	0.738	0.653	0.096	0.781	0.808	341	0.790	0.714	0.080	0.816	0.848	681
U^{2}Net$_{20}$ [56]	0.701	0.601	0.085	0.762	0.783	165	0.768	0.676	0.083	0.798	0.825	367	0.813	0.721	0.073	0.823	0.856	738
HRNet$_{20}$ [57]	0.668	0.579	0.088	0.742	0.797	262	0.747	0.664	0.087	0.784	0.840	555	0.784	0.700	0.080	0.805	0.869	1049
PGNet$_{22}$ [15]	0.754	0.680	0.067	0.800	0.848	162	0.807	0.743	0.065	0.833	0.880	375	0.843	0.785	0.056	0.844	0.911	797
IS-Net$_{22}$ [4]	0.740	0.662	0.074	0.787	0.820	149	0.799	0.728	0.070	0.823	0.858	340	0.830	0.758	0.064	0.836	0.883	687
FP-DIS$_{23}$ [9]	0.784	0.713	0.060	0.821	0.860	160	0.827	0.767	0.059	0.845	0.893	373	0.868	0.811	0.049	0.871	0.922	780
UDUN$_{23}$ [10]	0.784	0.720	0.059	0.817	0.864	140	0.829	0.768	0.058	0.843	0.886	325	0.865	0.809	0.050	0.865	0.917	658
BiRefNet	0.860	0.819	0.037	0.885	0.911	106	0.894	0.857	0.036	0.900	0.930	266	0.925	0.893	0.028	0.919	0.955	569
BiRefNet$_{SwinB}$	0.857	0.819	0.038	0.884	0.912	110	0.890	0.854	0.037	0.898	0.930	275	0.919	0.886	0.030	0.915	0.953	597
BiRefNet$_{SwinT}$	0.823	0.774	0.048	0.855	0.887	117	0.862	0.821	0.046	0.877	0.912	290	0.899	0.860	0.036	0.897	0.942	627
BiRefNet$_{PVTv2b2}$	0.839	0.796	0.042	0.870	0.903	111	0.881	0.842	0.040	0.888	0.925	280	0.903	0.866	0.036	0.901	0.941	614

Method	DIS-TE4 (500)						DIS-TE (1-4) (2000)						DIS-VD (470)					
	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\varphi^m\uparrow$	$\text{HCE}_\gamma\downarrow$
BASNet$_{19}$ [34]	0.785	0.713	0.087	0.806	0.844	2852	0.744	0.664	0.092	0.786	0.814	1007	0.737	0.656	0.094	0.781	0.809	1132
U^{2}Net$_{20}$ [56]	0.800	0.707	0.085	0.814	0.837	2898	0.771	0.676	0.082	0.799	0.825	1042	0.753	0.656	0.089	0.785	0.809	1139
HRNet$_{20}$ [57]	0.772	0.687	0.092	0.792	0.854	3864	0.743	0.658	0.087	0.781	0.840	1432	0.726	0.641	0.095	0.767	0.824	1560
PGNet$_{22}$ [15]	0.831	0.774	0.065	0.841	0.899	3361	0.809	0.746	0.063	0.830	0.885	1173	0.798	0.733	0.067	0.824	0.879	1326
IS-Net$_{22}$ [4]	0.827	0.753	0.072	0.830	0.870	2888	0.799	0.726	0.070	0.819	0.858	1016	0.791	0.717	0.074	0.813	0.856	1116
FP-DIS$_{23}$ [9]	0.846	0.788	0.061	0.852	0.906	3347	0.831	0.770	0.047	0.847	0.895	1165	0.823	0.763	0.062	0.843	0.891	1309
UDUN$_{23}$ [10]	0.846	0.792	0.059	0.849	0.901	2785	0.831	0.772	0.057	0.844	0.892	977	0.823	0.763	0.059	0.838	0.892	1097
BiRefNet	0.904	0.864	0.039	0.900	0.939	2723	0.896	0.858	0.035	0.901	0.934	916	0.891	0.854	0.038	0.898	0.931	989
BiRefNet$_{SwinB}$	0.899	0.860	0.040	0.895	0.938	2836	0.891	0.855	0.036	0.898	0.933	954	0.881	0.844	0.039	0.890	0.925	1029
BiRefNet$_{SwinT}$	0.880	0.834	0.049	0.878	0.925	2888	0.866	0.822	0.045	0.877	0.916	980	0.862	0.819	0.045	0.874	0.917	1070
BiRefNet$_{PVTv2b2}$	0.890	0.846	0.045	0.886	0.929	2871	0.878	0.838	0.041	0.886	0.925	969	0.868	0.827	0.044	0.880	0.919	1073

Table 5 Quantitative comparisons between our BiRefNet and the state-of-the-art methods in high-resolution and low-resolution SOD datasets. TR denotes the training set. To provide a fair comparison, we train our BiRefNet with different combinations of training sets, where 1, 2, and 3 represent DUTS[50], HRSOD[17], and UHRSD[18], respectively. The size of each dataset is marked behind.

Methods	TR	High-resolution benchmarks												Low-resolution benchmarks							
		DAVIS-S (92)				HRSOD-TE (400)				UHRSD-TE (988)				DUTS-TE (5019)				DUT-OMRON (5168)			
		$S_m\uparrow$	$F_\beta^x\uparrow$	$E_\varphi^m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^x\uparrow$	$E_\varphi^m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^x\uparrow$	$E_\varphi^m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^x\uparrow$	$E_\varphi^m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^x\uparrow$	$E_\varphi^m\uparrow$	$\mathcal{M}\downarrow$
LDF$_{20}$ [30]	1	0.922	0.911	0.947	0.019	0.904	0.904	0.919	0.032	0.888	0.913	0.891	0.047	0.892	0.898	0.910	0.034	0.838	0.820	0.873	0.051
HRSOD$_{19}$ [17]	1, 2	0.876	0.899	0.955	0.026	0.896	0.905	0.934	0.030	-	-	-	-	0.824	0.835	0.885	0.050	0.762	0.743	0.831	0.065
DHQ$_{21}$ [20]	1, 2	0.920	0.938	0.947	0.012	0.920	0.922	0.947	0.022	0.900	0.911	0.905	0.039	0.894	0.900	0.919	0.031	0.836	0.820	0.873	0.045
PGNet$_{22}$ [18]	1	0.935	0.936	0.947	0.015	0.930	0.931	0.944	0.021	0.912	0.931	0.904	0.037	0.911	0.917	0.922	0.027	0.855	0.835	0.887	0.045
PGNet$_{22}$ [18]	1, 2	0.948	0.950	0.975	0.012	0.935	0.937	0.946	0.020	0.912	0.935	0.905	0.036	0.912	0.919	0.925	0.028	0.858	0.835	0.887	0.046
PGNet$_{22}$ [18]	2, 3	0.954	0.957	0.979	0.010	0.938	0.945	0.946	0.020	0.935	0.949	0.916	0.026	0.859	0.871	0.897	0.038	0.786	0.772	0.884	0.058
BiRefNet	1	0.967	0.966	0.984	0.008	0.957	0.958	0.972	0.014	0.931	0.933	0.943	0.030	0.939	0.937	0.958	0.019	0.868	0.813	0.878	0.040
BiRefNet	1, 2	0.973	0.976	0.990	0.006	0.962	0.963	0.976	0.011	0.937	0.942	0.951	0.024	0.938	0.935	0.960	0.018	0.868	0.818	0.882	0.040
BiRefNet	1, 3	0.975	0.977	0.989	0.006	0.959	0.958	0.972	0.014	0.952	0.960	0.965	0.019	0.942	0.942	0.961	0.018	0.881	0.837	0.896	0.036
BiRefNet	2, 3	0.976	0.980	0.990	0.006	0.956	0.953	0.967	0.016	0.952	0.958	0.964	0.019	0.933	0.928	0.954	0.020	0.864	0.810	0.879	0.040
BiRefNet	1, 2, 3	0.975	0.979	0.989	0.006	0.962	0.961	0.973	0.013	0.957	0.963	0.969	0.016	0.944	0.943	0.962	0.018	0.882	0.839	0.896	0.038

objects and the more accurate segmentation of the details of the objects. For example, in the samples of DIS-TE4 and DIS-TE2, there are neighboring distractors that attract the attention of other models to produce false positives. On the contrary, our BiRefNet eliminates the distractors and accurately segments the target. In samples of DIS-TE3 and DIS-VD, BiRefNet shows much greater

Table 6 Comparison of BiRefNet with recent methods. As seen, BiRefNet performs much better than previous methods. The size of each dataset is marked behind.

Method	CAMO (250)						COD10K (2026)						NC4K (4121)					
	$S_m\uparrow$	$F_\beta^\omega\uparrow$	$F_\beta^m\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^x\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^\omega\uparrow$	$F_\beta^m\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^x\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta^\omega\uparrow$	$F_\beta^m\uparrow$	$E_\varphi^m\uparrow$	$E_\varphi^x\uparrow$	$\mathcal{M}\downarrow$
$SINet_{20}$ [23]	0.751	0.606	0.675	0.771	0.831	0.100	0.771	0.551	0.634	0.806	0.868	0.051	0.808	0.723	0.769	0.871	0.883	0.058
$BGNet_{22}$ [25]	0.812	0.749	0.789	0.870	0.882	0.073	0.831	0.722	0.753	0.901	0.911	0.033	0.851	0.788	0.820	0.907	0.916	0.044
$SegMaR_{22}$ [58]	0.815	0.753	0.795	0.874	0.884	0.071	0.833	0.724	0.757	0.899	0.906	0.034	0.841	0.781	0.820	0.896	0.907	0.046
$ZoomNet_{22}$ [59]	0.820	0.752	0.794	0.878	0.892	0.066	0.838	0.729	0.766	0.888	0.911	0.029	0.853	0.784	0.818	0.896	0.912	0.043
$SINetv2_{22}$ [9]	0.820	0.743	0.782	0.882	0.895	0.070	0.815	0.680	0.718	0.887	0.906	0.037	0.847	0.770	0.805	0.903	0.914	0.048
$FEDER_{23}$ [60]	0.802	0.738	0.781	0.867	0.873	0.071	0.822	0.716	0.751	0.900	0.905	0.032	0.847	0.789	0.824	0.907	0.915	0.044
$HitNet_{23}$ [27]	0.849	0.809	0.831	0.906	0.910	0.055	0.871	0.806	0.823	0.935	0.938	0.023	0.875	0.834	0.853	0.926	0.929	0.037
$FSPNet_{23}$ [28]	0.856	0.799	0.830	0.899	0.928	0.050	0.851	0.735	0.769	0.895	0.930	0.026	0.879	0.816	0.843	0.915	0.937	0.035
BiRefNet	0.904	0.890	0.904	0.954	0.959	0.030	0.913	0.874	0.888	0.960	0.967	0.014	0.914	0.894	0.909	0.953	0.960	0.023

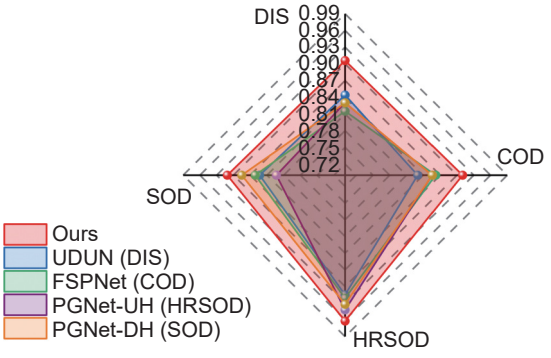

Fig. 5 Quantitative comparisons of the proposed BiRefNet and the best task-specific models. S-measure[13] is used for the comparison here. UDUN[10], FSPNet[24], PGNet-UH[15], and PGNet-DH[15] are currently the best models for the DIS, COD, HRSOD, and SOD tasks, respectively.

performance in precisely segmenting areas where fine details are rich. Compared with previous methods, our BiRefNet can clearly segment slim shapes and curved edges.

We also provide a qualitative comparison on the COD task. Figure 7 shows hard samples with different challenges. For example, in the row of the occluded frog, the area of the frog is divided by the branch that covers it, while our BiRefNet can accurately segment the scattered fragments almost the same as the GT map. In contrast, in the results of the other methods, fragments are difficult to find all, let alone to provide precise segmentation maps. For tiny and slim objects, BiRefNet shows a better ability to find the right target. Our BiRefNet also shows superiority in finding multiple concealed objects.

Efficiency and complexity comparison. We equip our BiRefNet with different backbones to obtain models in different sizes. Runtime, number of parameters, MACs, and performance of them are further tested to provide a comprehensive comparison between them and other methods. First, we provide a quantitative comparison in Table 7. The FPS of the largest BiRefNet can be more than 10, which is acceptable in most practical applications. We also used the compiled version ($BiRefNet_{SwinL\_cp}$) by PyTorch 2.0[55] on $BiRefNet_{SwinL}$ to accelerate its inference by 13%. In

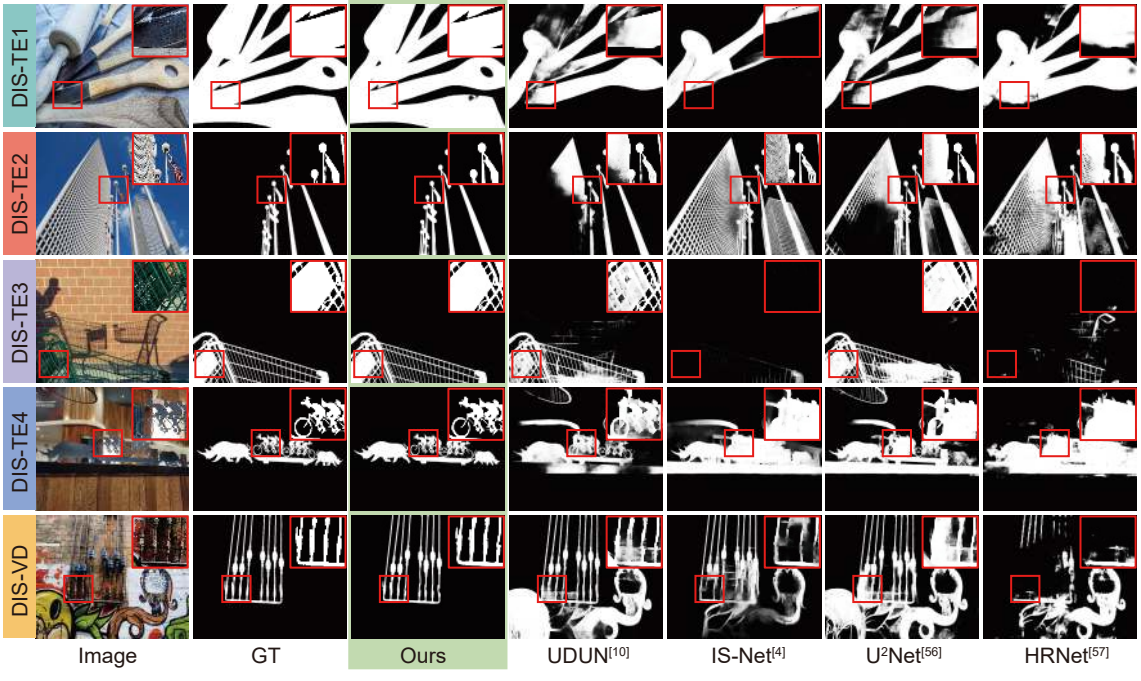

Fig. 6 Qualitative comparisons of the proposed BiRefNet and previous methods on the DIS5K. The results of the previous methods are from Ref. [10], where all models are trained with images in 1024×1024. Zoom in for a better view.

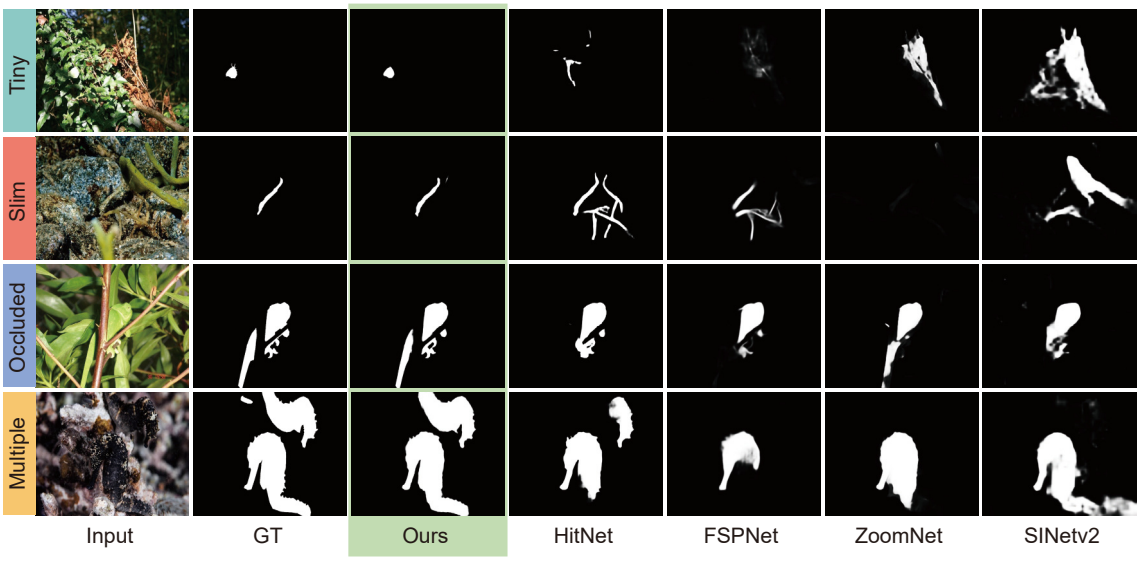

Fig. 7 Visual comparisons of the proposed BiRefNet and other competitors on COD10K benchmark. Samples with different challenges are provided here to show the superiority of BiRefNet from different perspectives.

Table 7 Comparison of different DIS methods on the performance, efficiency, and model complexity. MACs refers to multiply-accumulate operations.

Model	Runtime (ms)	#Params (10^6)	MACs (10^9)	DIS-TEs (HCE, F_β^ω)
BiRefNet$_{SwinL}$	83.3	215	1143	916, 0.858
BiRefNet^{SwinL_cp}	78.3	215	1143	916, 0.858
BiRefNet$_{SwinB}$	61.4	101	561	954, 0.855
BiRefNet$_{SwinT}$	40.9	39	231	980, 0.822
BiRefNet$_{PVTv2b2}$	47.8	35	195	969, 0.838
BiRefNet$_{PVTv2b1}$	36.6	23	147	978, 0.817
BiRefNet$_{PVTv2b0}$	32.9	11	89	1013, 0.806
IS-Net	16.0	44	160	1016, 0.726
UDUN$_{Res50}$	33.5	25	142	977, 0.772

addition, we draw the performance and runtime of each model in Fig. 8 for a clearer display. BiRefNet with different backbones are evaluated and compared with existing DIS methods on DIS-TEs and DIS-VD. Different methods are drawn in different colors and markers. All tests are conducted on a single NVIDIA A100 GPU and an AMD EPYC 7J13 CPU.

4 Potential Application

We envisage that generated fine maps have the potential to be utilized in various practical applications.

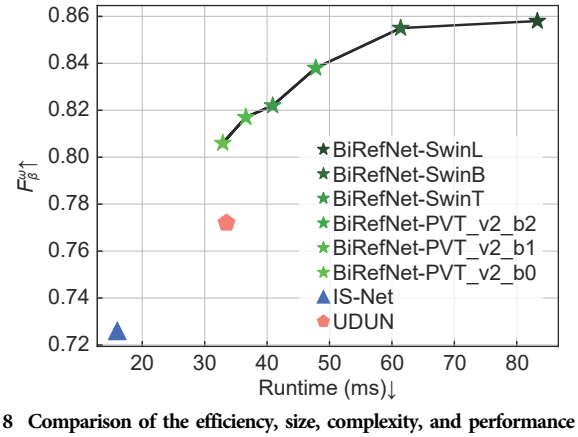

Fig. 8 Comparison of the efficiency, size, complexity, and performance of BiRefNet and existing DIS methods.

Potential application #1: Crack detection. The quality of the walls is important for the health of the architecture[61]. However, segmentation models trained on commonly used datasets (e.g., COCO[62]) can only segment regular foreground objects. The proposed BiRefNet trained on the DIS5K dataset is more aware of the fine details and can also segment targets with higher shape complexities. As shown in Fig. 9a, our BiRefNet can accurately find cracks in the walls and help maintain when to repair them.

Potential application #2: Highly accurate object extraction. Foreground object extraction and background removal have been popular applications in recent years. However, commonly seen methods fail to generate high-quality results when target objects have too high shape complexities[36, 56] or need manual guidance (e.g., scribble, point, and coarse mask) for more accurate segmentation[29, 63]. The proposed BiRefNet trained on DIS5K can generate results with much higher resolution and segment thin threads at the hair level without a mask, as shown in Fig. 9b. On the basis of such refined results, there may be numerous successful downstream applications in the future.

5 Third-Party Creation

Since our project was released on Mar 7, 2024, it has attracted much attention from many researchers and developers in the community to promote it spontaneously. Furthermore, great third-

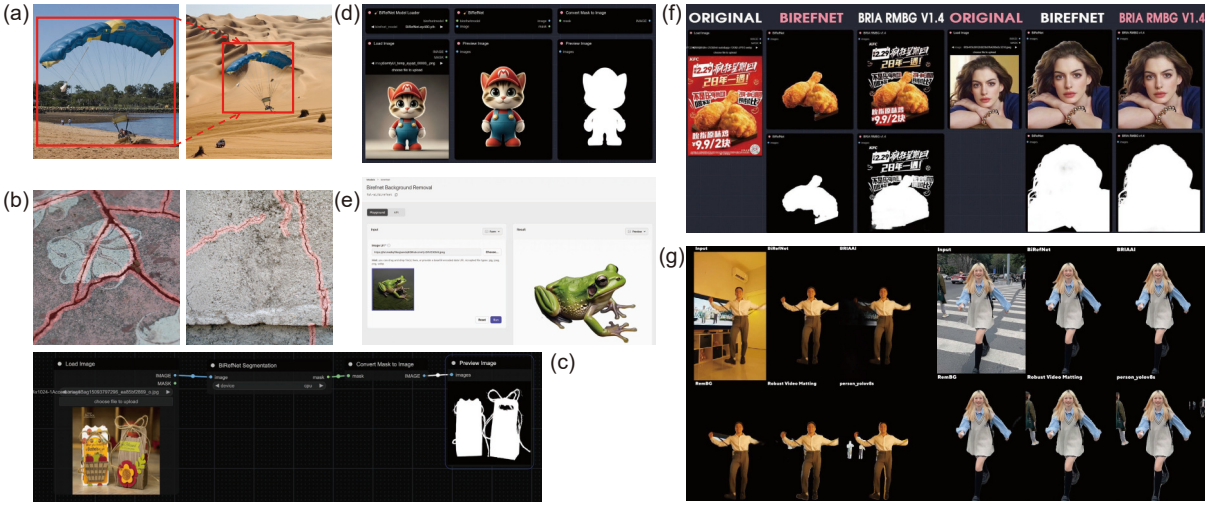

Fig. 9 Potenial applications and selected existing third-party applications based on BiRefNet, and visual comparisons on social media. (a) Potential application #1. Building crack detection for the maintenance of architecture health. (b) Potential application #2. Highly accurate object extraction in high-resolution natural images. (c) A project by viperyl first packs our BiRefNet as a ComfyUI node and makes this state-of-the-art model easier to use for everyone. (d) ZHO also provides a ComfyUI-based project to further improve the UI for our BiRefNet, especially for video data. (e) Fal.AI encapsulates our BiRefNet online with more useful options in UI and API to call the model. (f) ZHO provides a visual comparison between our BiRefNet and previous SOTA method BRIA RMGB v1.4 which has extra training on their private training dataset. (g) Toyxyz conducted a comparison between our BiRefNet and previous competitive human matting methods (e.g., BRIAAI, RemBG, Robust Video Matting, and Person YOLOv8s) with both videos and images.

party applications have also been made based on our BiRefNet. Due to the rapid growth of relevant works, we only list some typical ones.

(1) **Practical applications**. Because of the excellent performance of our BiRefNet, more and more third-party applications have been created by developers in the community[1,2]. As shown in Figs. 9c and 9d, some developers have integrated our BiRefNet into the ComfyUI as a node, which helps a lot in matting foreground segmentation to better processing in the subsequent stable diffusion models. For better online access, Fal.AI has established an online demo of our BiRefNet running on an A6000 GPU[3], as shown in Fig. 9e. In addition to the common prediction of results, this online application also provides an API service for easy use with HTTP requests.

(2) **Social media**. In recent days, our BiRefNet has drawn attention from the community. Many tweets have been posted on the X platform (formerly Twitter)[4]. ZHO provides a visual comparison between our BiRefNet and other methods, as given in Fig. 9f. BiRefNet achieves competitive results with the previous SOTA method BRIA RMGB v1.4 in their tests[5]. It should be noted that our BiRefNet was trained on the training set of the open-source dataset DIS5K[4] under MIT license, while the other one was trained on their carefully selected private data and cannot be used for commercial use. As shown in Fig. 9g, more comparisons on both video and image data have been provided by Toyxyz on Twitter between our BiRefNet and previous great foreground human matting methods[6]. In addition to these posts, PurzBeats has also made an animation with our BiRefNet and uploaded relevant videos[7]. A video tutorial can also be found on YouTube by "AI is in wonderland" in Japanese about how to use our BiRefNet in ComfyUI[8].

6 Conclusion

This work proposes a BiRefNet framework equipped with a bilateral reference, which can perform dichotomous image segmentation, HR salient object detection, and concealed object detection in the same framework. With the comprehensive experiments conducted, we find that unscaled source images and a focus on regions of rich information are vital to generating fine and detailed areas in HR images. To this end, we propose the bilateral reference to fill in the missing information in the fine parts (inward reference) and guide the model to focus more on regions with richer details (outward reference). This significantly improves the model's ability to capture tiny-pixel features. To alleviate the high training cost of HR data training, we also provide various practical tricks to deliver higher-quality prediction and faster convergence. Competitive results on 13 benchmarks demonstrate outstanding performance and strong generalization ability of our BiRefNet. We also show that the techniques of BiRefNet can be transferred and used in many practical applications. We hope that the proposed framework can encourage the development of unified models for various tasks in the academic community and that our model can empower and inspire the developer community to create more great works.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (No. Nankai University, 63243150). We thank Prof. Ming-Ming Cheng for his great support on this project.

1 https://github.com/comfyanonymous/ComfyUI
2 https://github.com/ZHO-ZHO-ZHO/ComfyUI-BiRefNet-ZHO
3 https://fal.ai/models/birefnet
4 https://twitter.com/search?q=birefnet&src=typed_query
5 https://twitter.com/ZHOZHO672070/status/1771026516388041038
6 https://twitter.com/toyxyz3/status/1771413245267746952
7 https://youtu.be/ySGAxxw2fvc
8 https://www.youtube.com/watch?v=o2_nMDUYk6s

Article History

Received: 19 April 2024; Revised: 9 July 2024; Accepted: 23 July 2024

References

[1] D. P. Fan, J. Zhang, G. Xu, M. M. Cheng, and L. Shao, Salient objects in clutter, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2344–2366, 2023.

[2] D. P. Fan, G. P. Ji, P. Xu, M. M. Cheng, C. Sakaridis, and L. Van Gool, Advances in deep concealed scene understanding, *Vis. Intell.*, vol. 1, no. 1, pp. 16, 2023.

[3] D. P. Fan, G. P. Ji, M. M. Cheng, and L. Shao, Concealed object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, 2022.

[4] X. Qin, H. Dai, X. Hu, D. P. Fan, L. Shao, and L. Van Gool, Highly accurate dichotomous image segmentation, in *Proc. 17th European Conf. Computer Vision* (*ECCV*), Tel Aviv, Israel, 2022, pp. 38–56.

[5] W. Lu, W. Sun, X. Min, W. Zhu, Q. Zhou, J. He, Q. Wang, Z. Zhang, T. Wang, and G. Zhai, Deep neural network for blind visual quality assessment of 4K content, *IEEE Trans. Broadcast.*, vol. 69, no. 2, pp. 406–421, 2023.

[6] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, Blind quality assessment for in-the-wild images *via* hierarchical feature fusion and iterative mixed database training, *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 6, pp. 1178–1192, 2023.

[7] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment, *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 1, pp. 64–77, 2020.

[8] W. Sun, X. Min, W. Lu, and G. Zhai, A deep learning based No-reference quality assessment model for UGC videos, in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 856–865.

[9] Y. Zhou, B. Dong, Y. Wu, W. Zhu, G. Chen, and Y. Zhang, Dichotomous image segmentation with frequency priors, in *Proc. 32nd Int. Joint Conf. Artificial Intelligence*, Macao, China, 2023, pp. 1822–1830.

[10] J. Pei, Z. Zhou, Y. Jin, H. Tang, and P. A. Heng, Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 2139–2147.

[11] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim, Revisiting image pyramid structure for high resolution salient object detection, in *Proc. 16th Asian Conf. Computer Vision* (*ACCV*), Macao, China, 2022, pp. 257–273,

[12] X. Li, J. Yang, S. Li, J. Lei, J. Zhang, and D. Chen, Locate, refine and restore: A progressive enhancement network for camouflaged object detection, in *Proc. 32nd Int. Joint Conf. Artificial Intelligence*, Macao, China, 2023, pp. 1116–1124.

[13] D. P. Fan, M. M. Cheng, Y. Liu, T. Li, and A. Borji, Structure-measure: A new way to evaluate foreground maps, in *in Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 4558–4567.

[14] Y. Zeng, P. Zhang, Z. Lin, J. Zhang, and H. Lu, Towards high-resolution salient object detection, in in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 7233–7242.

[15] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, Pyramid grafting network for one-stage high resolution saliency detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 11707–11716.

[16] X. Deng, P. Zhang, W. Liu, and H. Lu, Recurrent multi-scale transformer for high-resolution salient object detection, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 7413–7423.

[17] L. Tang, B. Li, Y. Zhong, S. Ding, and M. Song, Disentangled high quality salient object detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 3560–3570.

[18] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 770–778.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 9992–10002.

[20] D. P. Fan, G. P. Ji, G. Sun, M. M. Cheng, J. Shen, and L. Shao, Camouflaged object detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 2774–2784.

[21] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, Detecting camouflaged object in frequency domain, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, 4494–4503.

[22] Y. Sun, S. Wang, C. Chen, and T. Z. Xiang, Boundary-guided camouflaged object detection, in *Proc. 31st Int. Joint Conf. Artificial Intelligence*, Vienna, Austria, 2022, pp. 1335–1341.

[23] G. P. Ji, D. P. Fan, Y. C. Chou, D. Dai, A. Liniger, and L. Van Gool, Deep gradient learning for efficient camouflaged object detection, *Mach. Intell. Res.*, vol. 20, no. 1, pp. 92–108, 2023.

[24] Z. Huang, H. Dai, T. Z. Xiang, S. Wang, H. X. Chen, J. Qin, and H. Xiong, Feature shrinkage pyramid for camouflaged object detection with transformers, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 5557–5566.

[25] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, High-resolution iterative feedback network for camouflaged object detection, in *Proc. 37th AAAI Conf. Artificial Intelligence*, Washington, DC, USA, 2023, pp. 881–889.

[26] B. Yin, X. Zhang, Q. Hou, B. Y. Sun, D. P. Fan, and L. Van Gool, CamoFormer: Masked separable attention for camouflaged object detection, arXiv preprint arXiv: 2212.06570, 2022.

[27] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, Label decoupling framework for salient object detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 13022–13031.

[28] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu, Multi-view aggregation network for dichotomous image segmentation, arXiv preprint arXiv: 2404.07445, 2024.

[29] J. Li, J. Zhang, S. J. Maybank, and D. Tao, Bridging composite and real: Towards end-to-end deep image matting, *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 246–266, 2022.

[30] N. Xu, B. Price, S. Cohen, and T. Huang, Deep image matting, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 311–320.

[31] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Proc. 18th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI 2015*), Munich, Germany, 2015, pp. 234–241.

[32] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, ICNet for real-time semantic segmentation on high-resolution images, arXiv preprint arXiv: 1704.08545, 2017.

[33] W. I. Grosky and R. Jain, A pyramid-based approach to segmentation applied to region matching, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 5, pp. 639–650, 1986.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid scene parsing network, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 6230–6239.

[35] Q. Yu, J. Zhang, H. Zhang, Y. Wang, Z. Lin, N. Xu, Y. Bai, and A. Yuille, Mask guided matting *via* progressive refinement network, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 1154–1163.

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, BASNet: Boundary-aware salient object detection, in

Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (*CVPR*), Long Beach, CA, USA, 2019, pp. 7471–7481.

[37] T. Shen, Y. Zhang, L. Qi, J. Kuen, X. Xie, J. Wu, Z. Lin, and J. Jia, High quality segmentation for ultra high-resolution images, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 1300–1309.

[38] C. He, K. Li, Y. Zhang, Y. Zhang, Z. Guo, X. Li, M. Danelljan, and F. Yu, Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects, in *Proc. 12th Int. Conf. Learning Representations* (*ICLR*), Vienna, Austria, 2024.

[39] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, and X. Hu, Look closer to segment better: Boundary patch refinement for instance segmentation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 13921–13930.

[40] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, Deep Laplacian pyramid networks for fast and accurate super-resolution, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 5835–5843.

[41] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, Fast and accurate image super-resolution with deep Laplacian pyramid networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, 2019.

[42] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, DenseASPP for semantic segmentation in street scenes, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 3684–3692.

[43] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *in Proc. 15th European Conf. Computer Vision* (*ECCV*), Munich, Germany, 2018, pp. 833–851.

[44] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, Deformable convolutional networks, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 764–773.

[45] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 936–944.

[46] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, Frequency-tuned salient region detection, in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 1597–1604.

[47] Z. Zhang, W. Jin, J. Xu, and M. M. Cheng, Gradient-induced co-saliency detection, in *Proc. 16th European Conf. Computer Vision* (*ECCV*), Glasgow, UK, 2020, pp. 455–472.

[48] T. N. Le, T. V. Nguyen, Z. Nie, M. T. Tran, and A. Sugimoto, Anabranch network for camouflaged object segmentation, *Comput. Vis. Image Underst.*, vol. 184, pp. 45–56, 2019.

[49] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D. P. Fan, Simultaneously localize, segment and rank the camouflaged objects, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 11586–11596.

[50] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, Learning to detect salient objects with image-level supervision, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 11586–11596.

[51] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, Saliency detection *via* graph-based manifold ranking, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 3166–3173.

[52] D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, and A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in *Proc. 27th Int. Joint Conf. Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 698–704.

[53] A. Borji, M. M. Cheng, H. Jiang, and J. Li, Salient object detection: A benchmark, *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[54] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proc. 3rd Int. Conf. Learning Representations* (*ICLR*), San Diego, CA, USA, 2015.

[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. , PyTorch: An imperative style, high-performance deep learning library, in *Proc. 33rd Conf. Neural Information Processing Systems* (*NeurIPS 2019*), Vancouver, Canada, 2019, pp. 8026–8037.

[56] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, U2-Net: Going deeper with nested U-structure for salient object detection, *Pattern Recognit.*, vol. 106, pp. 107404, 2020.

[57] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang et al., Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[58] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, Segment, magnify and reiterate: Detecting camouflaged objects the hard way, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 4703–4712.

[59] Y. Pang, X. Zhao, T. -Z. Xiang, L. Zhang, and H. Lu, Zoom in and out: A mixed-scale triplet network for camouflaged object detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 2150–2160.

[60] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, Camouflaged object detection with feature decomposition and edge reconstruction, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 22046–22055.

[61] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, DeepCrack: learning hierarchical convolutional features for crack detection, *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1498–1512, 2019.

[62] T. Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. Computer Vision* (*ECCV*), Zurich, Switzerland, 2016, pp. 740–755.

[63] L. Dai, X. Song, X. Liu, C. Li, Z. Shi, J. Chen, and M. Brooks, Enabling trimap-free image matting with a frequency-guided saliency-aware network *via* joint learning, *IEEE Trans. Multimedia*, vol. 25, pp. 4868–4879, 2023.