ESNet: Evolution and Succession Network for High-Resolution Salient Object Detection

Hongyu Liu [1] Runmin Cong [2] Hua Li [3] Qianqian Xu [4] Qingming Huang [5] Wei Zhang [2]

Abstract

Preserving details and avoiding high computational costs are the two main challenges for the High-Resolution Salient Object Detection (HRSOD) task. In this paper, we propose a two-stage HRSOD model from the perspective of evolution and succession, including an evolution stage with Low-resolution Location Model (LrLM) and a succession stage with High-resolution Refinement Model (HrRM). The evolution stage achieves detail-preserving salient objects localization on the low-resolution image through the evolution mechanisms on supervision and feature; the succession stage utilizes the shallow high-resolution features to complement and enhance the features inherited from the first stage in a lightweight manner and generate the final high-resolution saliency prediction. Besides, a new metric named Boundary-Detail-aware Mean Absolute Error (MAE_{BD}) is designed to evaluate the ability to detect details in high-resolution scenes. Extensive experiments on five datasets demonstrate that our network achieves superior performance at real-time speed (49 FPS) compared to state-of-the-art methods. Our code is publicly available at: https://github.com/rmcong/ESNet_ICML24.

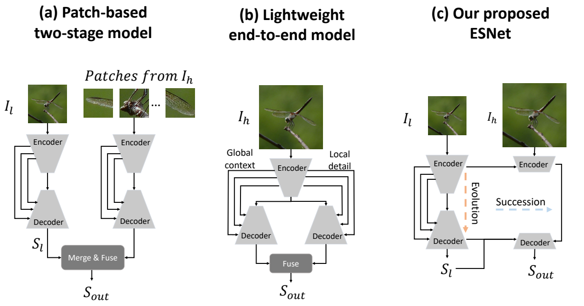


Figure 1: Typical structures for the HRSOD task, where (a) is the patch-based two-stage model, (b) is the lightweight end-to-end model, and (c) is our Evolution and Succession Network (ESNet).

1. Introduction

Salient object detection (SOD) is inspired by the human visual system, aiming at locating the most attractive objects and segmenting them from a given image. Recently, especially in the era of deep learning, SOD task has developed vigorously, and formed a full-scene and multi-source research system. With the rapid development of intelligent shooting devices (*e.g.*, smartphones) and terminal display devices, captured high-resolution (HR) image can offer a more refined viewing experience, preserving greater detail that enhances the quality of perception and understanding.

The task of high-resolution salient object detection (HRSOD) (Zeng et al., 2019) emerged as a means to more effectively adapt to the processing requirements of HR images. Unlike traditional tasks of normal-resolution salient object detection (NRSOD), the HRSOD task addresses two crucial issues. First, from a data source perspective, HR images can capture details more accurately and vividly. This ability to perceive these intricate elements sets HRSOD apart from NRSOD. However, detecting these fine details is challenging. On the one hand, the small proportion of detailed areas within the entire image often gets overlooked during model training. On the other hand, the existing evaluation metrics, such as F-measure and MAE score, fall short in effectively evaluating these critical detail regions. The primary reason for this is that enhancements in detailed

[1] Institute of Information Science, Beijing Jiaotong University & Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China [2] School of Control Science and Engineering, Shandong University & Key Laboratory of Machine Intelligence and System Control, Ministry of Education, Jinan, China [3] School of Computer Science and Technology, Hainan University, Hainan, China [4] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China [5] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Runmin Cong <rmcong@sdu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

areas may not substantially affect the overall performance metrics. Second, since the number of pixels is even tens of times that of normal-resolution images, the amount of computation when processing HR images can be imagined.

To address these issues, existing methods generally fall into two primary categories. One is patch-based two-stage model (Zeng et al., 2019; Tang et al., 2021), as shown in Figure 1(a), which consists of global saliency perception on down-sampled low-resolution image and local saliency perception on divided multiple patches. The other is lightweight end-to-end model (Zhang et al., 2021b; Xie et al., 2022; Wang et al., 2022a), as shown in Figure 1(b), which usually uses a global-local perception structure with some lightweight operations (*e.g.*, atrous convolution (Yu & Koltun, 2016)) to process high-resolution image directly. However, although the above schemes alleviate the challenges posed by HR images to some extent, there is still room for improvement. The cropping operation in the patch-based model may cause discontinuity between patches and also slow down inference. The end-to-end model directly uses HR image as the input of the entire backbone network, which will introduce unnecessary computation, and a large number of atrous convolutions will inevitably lose a certain degree of detail information.

Based on the above analysis, we implement HRSOD task through a two-stage Evolution and Succession Network (ESNet), including an evolution stage with low-resolution localization model (LrLM) and a succession stage with high-resolution refinement model (HrRM), as shown in Figure 1(c). Our core idea is to achieve accurate saliency localization and detail perception on the basis of ensuring computational efficiency with the help of evolution and succession mechanisms. Our ESNet is performed in two stages, but unlike (Zeng et al., 2019) and (Tang et al., 2021), both stages of our network process the entire image, avoiding artifacts caused by the patch-dividing operation. Specifically, our two-stage structure decouples the HRSOD task into saliency localization with details at low resolution and lightweight refinement at high resolution, thereby reducing the computational pressure on the network while ensuring a large perceptual field to extract global saliency information. Moreover, due to the small proportion of detail areas in the whole image, it is difficult to measure the detection effect of these regions by traditional evaluation metrics (*e.g.*, MAE) due to the long-tailed distribution (Xu et al., 2022). Therefore, we design a new Boundary-Detail-aware Mean Absolute Error (MAE_{BD}) metric to achieve a more reasonable detail evaluation for HR scenes, which focuses on the boundary and detail regions of the predicted map in a weighted manner.

Our major contributions can be summarized as: (1) We propose a two-stage ESNet for the HRSOD task with the real-time speed (49 FPS) and competitive performance against 16 SOD methods. (2) We design the evolution mechanisms on supervision and feature are performed to guide the detail learning in an easy-to-hard and coarse-to-fine manner. (3) A high-resolution trigger is designed in the succession stage to achieve local detail supplementation and global saliency enhancement in a lightweight way. (4) Considering that the traditional metrics are not sensitive to the detail region, we propose a metric named MAE_{BD} that is more suitable for HRSOD scenarios to reflect the quality of detail detection.

2. RELATED WORK

In recent years, especially in the era of deep learning, SOD task has developed vigorously, and formed a full-scene and multi-source research system, deriving the RGB SOD task (Chen et al., 2023; Cong et al., 2023b), RGB-D/RGB-T SOD task (Zhou et al., 2021; Cong et al., 2022b; 2023d;a), co-salient object detection (CoSOD) task (Fan et al., 2022; Cong et al., 2023c), 360° omnidirectional image SOD task(Cong et al., 2022a; Li et al., 2020), *etc*.

In order to conform to the trend of the times, (Zeng et al., 2019) first launched the HRSOD task and constructed a corresponding dataset. Methodologically, a two-stage baseline is also given in (Zeng et al., 2019), where the first stage detects salient objects from a global perspective on low-resolution image, then performs local patch refinement on the HR image, and finally fuses and stitches them to obtain HR saliency map. Similarly, (Tang et al., 2021) also adopted this overall structure, and introduced the concept of tri-map which is widely applying in the field of matting(Cai et al., 2019). In the second stage, the high-resolution images are also cropped into some patches to refine the uncertain regions of the tri-map. (Zhang et al., 2021b) and (Wang et al., 2022a) both proposed their end-to-end HRSOD models with the same general idea. They both adopted the structure of detail and context branches, and used atrous convolution and depth-wise convolution to expand the receptive field and reduce the amount of computation. (Xie et al., 2022) proposed another UHRSD dataset for HRSOD task and utilized two different type backbones to extract features from different resolution images for more complementary information.

However, the above approaches all have a certain degree of redundancy. For the two-stage models, the global and local features are considered separately, which leads to redundancy in feature extraction. For the end-to-end models, it is not necessary to use such a large resolution when extracting global information. Moreover, they all pay little attention to those subtle regions that have little impact on quantitative evaluation but are the essence of HRSOD.

Besides, for HR image segmentation models, (Guo et al., 2022) integrated shallow and deep networks well for efficient segmentation. (Dong et al., 2023) leveraged prototypes

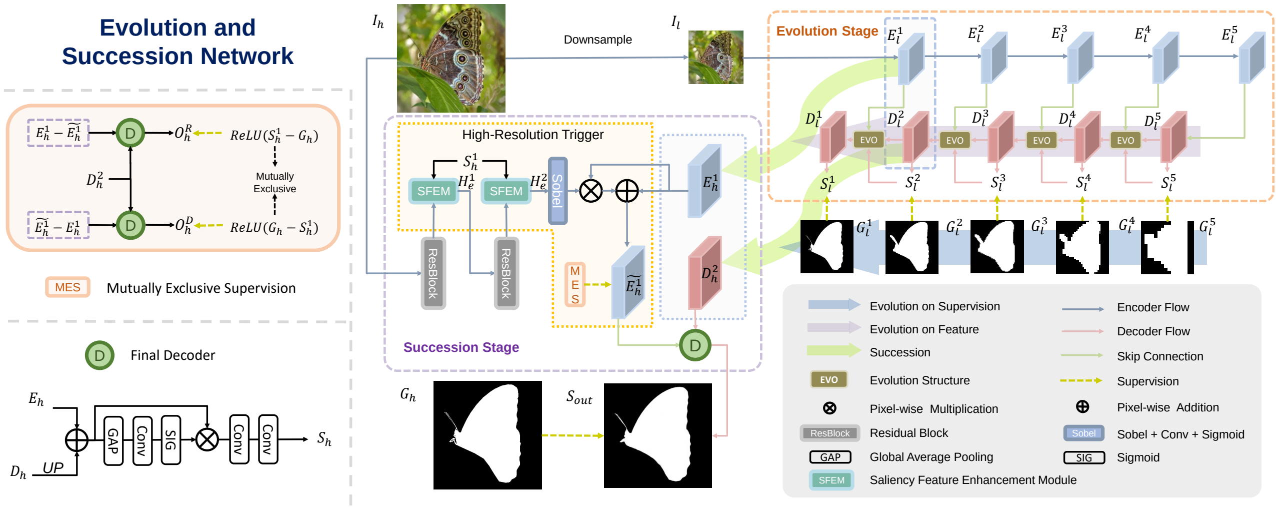


Figure 2: The overall framework of the proposed ESNet with the evolution stage and succession stage.

as learnable local descriptions to achieve a light architecture. Both of them achieve computational lightness by focusing on local context. However, for SOD, a class-agnostic task, whether a pixel is salient or not is not determined by its local representation, but by global contrast, so it is also worthwhile to think about how to achieve lightweighting in the HRSOD task.

3. PROPOSED METHOD

3.1. Overview

In this paper, we achieve HRSOD task from a new perspective and propose a two-stage Evolution and Succession Network (ESNet), as shown in Figure 2, including an evolution stage with low-resolution localization model (LrLM) to enhance attention to detail areas, and a succession stage with high-resolution refinement model (HrRM) to achieve detail refinement in a lightweight way.

The evolution mechanisms on the level of supervision and feature are used in the first stage to guide the model to achieve accurate and detail-preserving salient object localization on the low-resolution images in an easy-to-hard and coarse-to-fine manner. Specifically, considering the computational efficiency, the input high-resolution image $I_h \in \mathbb{R}^{3\times H\times W}$ is firstly down-sampled to $I_l \in \mathbb{R}^{3\times h\times w}$, which is further fed into the pre-trained backbone to extract multi-level encoder features $\left\{E_l^i\right\}_{i=1}^5$. Then, the feature decoder with the evolution structure is used to achieve progressive decoder feature learning and layer-by-layer saliency prediction under the supervision of evolutionay labels, where the decoder features and predicted saliency map are denoted as $\left\{D_l^i\right\}_{i=1}^5$ and $\left\{S_l^i\right\}_{i=1}^5$, respectively. Note that, we also impose an evolutionary variation loss between adjacent-layer features in the evolution structure to strengthen the network's attention to evolutionary change regions.

In succession stage, to supplement high-resolution detail information at low computational cost, we only extract the shallow features from the original high-resolution image I_h, and use the high-resolution trigger to activate valuable detail features and supplement them for the encoder features E_h^1 inherited from the first stage. Finally, the enhanced high-resolution encoder features $\widetilde{E}_h^1$ and the inherited up-sampled decoder features D_h^2 are fed into a simple decoder to generate the final high-resolution saliency map S_{out}.

3.2. The Evolution Stage with Low-resolution Localization Model

As the beginning of the network, the evolution stage forms the foundation for later refinement processes, focusing on accurate and complete localization of salient objects. At the same time, in order to adapt to the requirements for high-quality description of detail in high-resolution scenes, it is also of vital importance to preserve detail regions in the detection results. To achieve these goals, our LrLM adopts an encoder-decoder structure and incorporates evolution mechanisms on supervision and feature. On the one hand, following the rules of easy-to-hard and coarse-to-fine, we upgrade the side supervision in the traditional SOD task to adapt to the HRSOD task, and propose the progressive constraint idea of supervision evolution, providing each layer different supervision labels. The core operation of this part is to gradually inflate the original ground-truth through the max pooling operation to obtain different degrees of supervision information. On the other hand, in order to further model the relationship between the features of different layers, we design an evolution structure in the feature dimension to echo the supervision evolution. In the implementation, we focus on the evolutionary regions corresponding to changes between supervision labels, and then impose an evolutionary variation loss between adjacent-layer features to strengthen the network's attention to evolutionary change regions.

3.2.1. Evolution on supervision

For SOD task and other dense prediction tasks, side supervision has been widely used to constrain the feature learning of each layer in the network (Wu et al., 2019; Chen et al., 2020; Zhao et al., 2021; Liu et al., 2023), especially for the UNet-based segmentation models. Specifically, the existing methods downsample the ground truth or upsample the prediction maps for resolution-matched supervised learning. This is feasible for traditional SOD task, but we believe that it may be overly restrictive to impose the same precise supervision on both high-level and low-level features for the HRSOD task. On the one hand, with the deepening of the network, operations such as pooling and downsampling make the boundary information in the high-level features blurred. If, like the traditional side-supervision methods, using the ground truth with clear boundaries to constrain high-level features would be too harsh, making it difficult to learn. On the other hand, the detail regions, such as the hair and butterfly tentacles, account for a small proportion of the entire ground truth label, and thus are easily blurred or ignored in high-level features. Once this happens, it is difficult to recover in the final output.

To this end, we propose a supervision evolution strategy in the LrLM, aiming to alleviate the learning difficulty of detailed content for high-level features by increasing the proportion of detailed regions in the higher-level supervision. The inspiration and core idea of our design is that we hope to provide each layer different supervision labels, so that the prediction map of each layer can form a gradually evolving shape, and then the features can also achieve a natural evolution from easy to hard and coarse to fine. Notably, our supervision evolution involves not only a simple change in the resolution of supervision, but more importantly, a change in the content of supervision. In this way, the difficulty of network training and learning can be reduced. For all of them, we mainly focus on the detailed regions of the boundaries rather than the object body, since they are the hardest parts to learn but also the most important ones. In the implementation, we gradually dilute the original ground truth through the max pooling operation to obtain different degrees of supervision information, which can be formulated as:

$$G_l^i = \begin{cases} down(G_h) & i = 1 \\ maxpool(G_l^{i-1}) & i = \{2, 3, 4, 5\} \end{cases}, \tag{1}$$

where G_h denotes the full-resolution ground truth, $down$ is the downsampling operation, $maxpool$ is the max pooling operation, G_l^i represents the ground truth of the i^{th} layer, and $G_l^1 \in \mathbb{R}^{1 \times h \times w}$. From Eq. (1) and Figure 2, we can see that the proportion of detail regions in the high-level supervision labels is increased, so that they can be more easily preserved in the final result, making the detail refinement in the second stage possible. The use of these labels facilitates the evolutionary learning of features at each layer. It only needs to capture the approximate scope of salient objects on the high-level features, without paying attention to elaborate boundary outline information, making it more fault-tolerant. The boundaries of salient objects are gradually carved from high to low by progressive decoding. In summary, the supervision evolution can better guide the network for purposeful and planned learning, so as to localize and segment salient objects while preserving detailed regions.

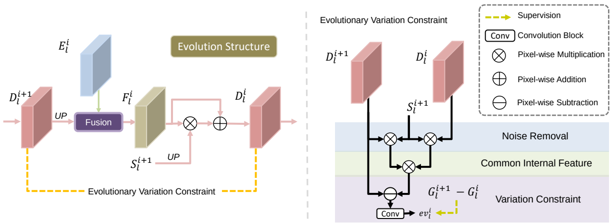


Figure 3: The details of proposed Evolution Structure in LrLM.

3.2.2. Evolution on feature

The design motivation for feature evolution comes from two aspects. First, conventional multi-stage fusion schemes tend to treat cross-layer features equally, but in fact, high-level features aggregate more global information and enable coarse localization of salient objects. In this way, we use higher-level side output as spatial attention map to suppress background noise in shallow-level features during the decoding process.

As shown in Figure 3, we use the saliency map as a medium to construct the relationship between adjacent decoder layers. The upsampled decoder features $D_{l\uparrow}^{i+1}$ and the corresponding encoder features E_l^i are firstly fused, and the upsampled saliency prediction map $S_{l\uparrow}^{i+1}$ of the previous decoder layer is used as a spatial attention map to refine the fusion features and generate the decoder features of the current layer. This process can be described as:

$$F_l^i = Fusion(D_{l\uparrow}^{i+1}, E_l^i), \tag{2}$$

$$D_l^i = F_l^i + F_l^i \otimes S_{l\uparrow}^{i+1}, \tag{3}$$

where $Fusion$ refers to the FFM, CAM and BRM for different levels in CTDNet (Zhao et al., 2021), respectively, and $\otimes$ denotes the pixel-wise multiplication.

Second, the former fusion schemes do not explicitly model the differences between cross-layer features, making the learning direction of the model unclear. Considering that only constraining at the supervision level is not sufficient to make the network learn the nature of evolution, and the high-level features are rich in semantic information and have good consistency within the objects, we hope to specify the direction of evolution by constraining the variation region of each evolution to facilitate the coarse-to-fine learning process and maintain this consistency in the layer-by-layer decoding process. Therefore, for features in two adjacent layers, we activate their internal common parts by multiplication, and then use the subtraction operation to obtain the differences of features and constrain them by the variation of evolutionary labels.

Specifically, considering that the features in each layer have different and noisy representations of the background region, we calculate the evolutionary variation features by excluding the common internal features after using the saliency map output in the previous layer to suppress the noise on the background, and then obtain the final evolutionary variation prediction through the convolution operation. The above process can be formulated as:

$$ev_l^i = conv(D_{l\uparrow}^{i+1} - (D_{l\uparrow}^{i+1} \otimes S_{l\uparrow}^{i+1}) \otimes (D_l^i \otimes S_{l\uparrow}^{i+1})), \quad (4)$$

where ev_l^i denotes the evolutionary variation map of the i^{th} layer, and $conv$ is a convolution layer with the kernel size of 3×3. From this, we can use $(G_{l\uparrow}^{i+1} - G_l^i)$ as the supervision, so as to achieve the purpose of specifying evolutionary direction. This loss can be represented as:

$$\ell_{ev} = \sum_{i=1}^{4} \ell_{bce}\left(ev_l^i, G_{l\uparrow}^{i+1} - G_l^i\right), \quad (5)$$

where ℓ_{bce} is the binary cross-entropy loss.

3.3. The Succession Stage with High-resolution Refinement Model

3.3.1. Succession mechanism

In the first stage, we achieve detail-preserving-focused salient object localization on the low-resolution image via two evolution mechanisms, but the detail representation capability is still insufficient due to the limitation of resolution. Therefore, the succession stage aims to achieve high-quality detail refinement while minimizing computational costs. Its core implementation component is High Resolution Trigger (HRT), which reuses the features generated in the first stage, thereby supplementing and correcting details to obtain high-quality and high-resolution saliency map in a lightweight way. On the one hand, we design the Saliency Feature Enhancement Module (SFEM) to preform global modeling and enhancement on features obtained from the residual block, where the saliency map S_h^1 acts as guidance. On the other hand, in order to ensure that the learning process does not affect the main part of the object, a Mutually Exclusive Supervision (MES) is designed to constrain the supplementation to occur only in the detail area without destroying the main body information.

In fact, the HRSOD task can be achieved by simple upsampling of features or saliency map in LrLM, but this will ignore many details and lead to blurring of boundary regions. Therefore, we cannot just do such a simple succession, but also need to supplement and enhance the features inherited from the LrLM with high-resolution features. Specifically, we take the upsampled encoder features E_h^1 as the basic, and use the high-resolution trigger (HRT) with the help of shallow features extracted from the high-resolution image to update the features of E_h^1 and generate the enhanced high-resolution encoder features $\widetilde{E}_h^1$:

$$\widetilde{E}_h^1 = \mathrm{HRT}(E_h^1, I_h, S_h^1), \quad (6)$$

where HRT is the high-resolution trigger, S_h^1 denotes the upsampled saliency map generated by the first stage. Then, the features $\widetilde{E}_h^1$ and D_h^2 are sent to the final decoder layer as shown in Figure 2 which shares the same structure with LrLM and generate the final saliency prediction:

$$S_{out} = \mathrm{Decoder}_{\mathrm{final}}(\widetilde{E}_h^1, D_h^2). \quad (7)$$

3.3.2. High-resolution trigger

As mentioned earlier, the high-resolution trigger aims to activate valuable detail features from high-resolution images to refine and complement the inherited encoder features E_h^1. To reduce the computational cost, we only use two residual blocks in ResNet (He et al., 2016) to extract the shallow features from the high-resolution image. But such shallow features lack saliency attribute and may contain a lot of redundant and irrelevant information that is unfavorable for our detail supplementation. Furthermore, unlike class-related segmentation tasks (*e.g.*, semantic segmentation), they can be implemented in fine-grained detection to perform corrections to the class attribution of a single pixel based on the features of it (Kirillov et al., 2020). However, whether a pixel belongs to a salient object is not particularly determined by itself but is closely related to the global contrast information. Thus, it is necessary to model cross-pixel relationships in the refinement process. During the low-level feature extraction, especially in high-resolution images processed using regular-scale convolution, the receptive field will be more limited and insufficient to obtain global contextual information. Therefore, we design the SFEM to reuse the saliency map generated by the first stage for attribute reinforcement and global modeling of the features obtained from the residual block, as shown in Figure 4. Considering

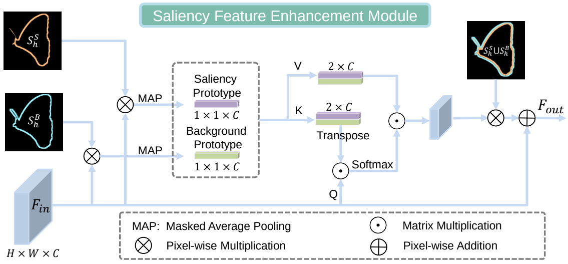
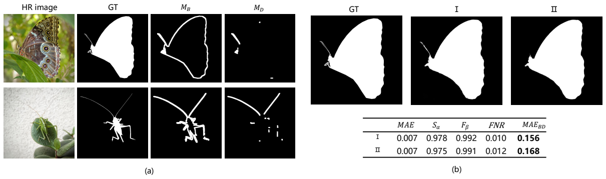


Figure 4: The details of Saliency Feature Enhancement Module in HrRM.



	MAE	S_α	F_β	FNR	MAE_{BD}
I	0.007	0.978	0.992	0.010	**0.156**
II	0.007	0.975	0.991	0.012	**0.168**

Figure 5: (a) Examples of M_B and M_D. (b) Effectiveness and sensitivity of proposed MAE_{BD}.

that both foreground and background have diversity inside, for example, butterfly wings have great variation in color and texture. But in the contour area, there is typically a strong contrast between the foreground and background. So we specifically extract prototypes of the foreground and background in the contour area. Additionally, in the second stage, we focus on refining the detail regions, so that constraining the enhancement operation to the contour region can make the optimization more targeted. First, we determine the saliency boundary regions S_h^S and background boundary regions S_h^B according to the prediction results from the first stage:

$$S_h^B = maxpool(S_h^1) - S_h^1, \tag{8}$$

$$S_h^S = S_h^1 - minpool(S_h^1), \tag{9}$$

where $maxpool$ and $minpool$ are max pooling and min pooling, respectively. Note that, these two regions together constitute the boundary detail part that needs to be refined in the succession stage. Finally, the global prototype vectors of the saliency and background boundary regions are generated by masked average pooling operation, respectively:

$$P_S = MAP(\widetilde{F}_{inS}) = MAP(F_{in} \otimes S_h^S), \tag{10}$$

$$P_B = MAP(\widetilde{F}_{inB}) = MAP(F_{in} \otimes S_h^B), \tag{11}$$

where MAP is the masked average pooling, F_{in} denote the output features of the residual block in the second stage, P_S and P_B are the saliency prototype and background prototype, respectively. Subsequently, the global information is introduced into the local feature extraction process through a prototype-based attention mechanism to facilitate better segmentation in the boundary areas, which is calclulated by:

$$Q = F_{in}, K = V = concat(P_S, P_B), \tag{12}$$

$$F_{out} = F_{in} + \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \otimes (S_h^S \bigcup S_h^B), \tag{13}$$

where F_{out} are the high-resolution features obtained by highlighting the boundary detail regions, d_k denotes the dimension of feature, $concat$ is a spatial-wise concatenation operation, $\bigcup$ is the union operation, and $\mathrm{softmax}$ represents the softmax activation. Finally, we use the Sobel operator to extract the boundary map from the enhanced features H_e^2 of the second residual block, and apply it to the first-stage encoder features in the form of residual connections:

$$\widetilde{E}_h^1 = E_h^1 + \mathrm{Sobel}(H_e^2) \otimes E_h^1. \tag{14}$$

In the above operations, although we restrict the update scope of the first-stage encoder features to the detailed boundary regions, it may also have unpredictable effects to the main part of the object. To this end, a MES is designed to constrain the supplementation to occur only in the detail area without destroying the main body information:

$$\begin{aligned} \ell_{MES} = &\ell_{bce}\left(O_h^D, \mathrm{ReLU}\left(G_h - S_h^1\right)\right) + \\ &\ell_{bce}\left(O_h^R, \mathrm{ReLU}\left(S_h^1 - G_h\right)\right), \end{aligned} \tag{15}$$

where $O_h^R = \mathrm{Decoder}_{\mathrm{final}}((\widetilde{E}_h^1 - E_h^1), D_h^2)$, $O_h^D = \mathrm{Decoder}_{\mathrm{final}}((E_h^1 - \widetilde{E}_h^1), D_h^2)$, and ReLU represents the linear rectification function that can make the supervision of the two ways mutually exclusive in the spatial dimension. In Eq. (15), the first term is used to constrain the missing parts of the first-stage prediction S_h^1 compared to the ground truth G_h, and the second term constrains the redundant parts of S_h^1. These two terms form a mutually exclusive relationship that together promote better feature learning of $\widetilde{E}_h^1$.

3.4. Boundary-Detail-aware MAE Metric

Since the detail regions such as butterfly tentacles are relatively small in the whole image, the detection quality of these regions does not have much impact on the traditional metrics (*e.g.*, MAE). However, these regions are the key to distinguish HRSOD task from ordinary SOD, so we propose a new metric named Boundary-Detail-aware MAE (MAE_{BD}) to measure the ability to detect details in high-resolution scenes, which is defined as:

$$\begin{aligned} MAE_{BD}(S, G) = \theta \cdot MAE\left(S \otimes M_B, G \otimes M_B\right) + \\ (1 - \theta) \cdot MAE\left(S \otimes M_D, G \otimes M_D\right), \end{aligned} \tag{16}$$

$$\theta = \frac{\mathrm{sum}\left(M_D\right)}{\mathrm{sum}\left(M_B\right) + \mathrm{sum}\left(M_D\right)}, \tag{17}$$

where M_B and M_D are the boundary region mask and detail region mask, respectively, and θ is used to balance the weight of errors within the two regions. As shown in Figure 5(a), M_B can be obtained by inflating the salient boundary label. For the calculation of M_D, we first divide the saliency mask G into some local patches (*e.g.*, 80×80 pixels in size), and then calculate the ratio of perimeter to area in each local patch separately. If the ratio is greater than the set threshold (*e.g.*, 0.25), the local patch is regarded as a detail area, and obtain M_D by inflation of these local patches with details.

In Figure 5(b), the difference between the two predictions is mainly represented on the presence or absence of butterfly tentacles, *i.e.*, the quality of detected detail areas. However, due to the relatively small proportion of these regions in the whole image, the evaluation discrimination of the existing traditional indicators (*e.g.*, F-measure, MAE score, FNR, and S-measure) is very small, and it is difficult to represent the detection quality of the detailed regions that are the core of the HRSOD task, while our proposed MAE_{BD} metric goes to intuitively reflect this difference.

4. EXPERIMENT

4.1. Datasets and Evaluation metrics

The **HRSOD** (Zeng et al., 2019), **DAVIS-SOD** (Zhang et al., 2021b), and **UHRSD** (Xie et al., 2022) datasets are used for evaluation. In addition to the new metric Boundary-Detail-aware MAE introduced in Section 3.4, we also adopt four widely used metrics including the F-measure (Niu et al., 2012), MAE score, S-measure (Fan et al., 2017) and FNR (Zhuge et al., 2023). For more details, see A.3.

4.2. Comparisons with the State-of-the-arts

For loss function and implementation details, see A.1&A.2. To prove the effectiveness of our proposed ESNet, we compare with 16 state-of-the-art models, including eleven NRSOD models and five HRSOD models. Among these models, VST (Liu et al., 2021), ICON-S (Zhuge et al., 2023) and PGNet (Xie et al., 2022) are transformer-based SOD models, and the rest SOD models are all based on CNNs.

4.2.1. Quantitative evaluation

For a intuitive performance comparison, Table 1 shows the quantitative results of the proposed ESNet on three high-resolution datasets, where the best performance is marked in bold. Since the NRSOD model cannot use high-resolution input, the results on the three high-resolution datasets are generally far inferior to the HRSOD models. For example, on the HRSOD-TE dataset, our method surpasses the strongest NRSOD method (*i.e.*, PFSNet (Ma et al., 2021)), winning 3.3% performance gain in F-measure and 36.3% performance gain in MAE score. A similar situation occurs on two other high-resolution SOD datasets. Besides, among the HRSOD models, our method also achieves an overall lead. On the HRSOD-TE dataset, compared with the **second best** CNN-based HRSOD method (*i.e.*, HQSOD (Tang et al., 2021)), the percentage gain reaches 2.1% for the F-measure, 12.5% for MAE score, 1.4% for S-measure, and 10.2% for MAE_{BD} score. Also, our Transformer version (*i.e.*, OURS_swin) achieves competitive results compared to the best Transformer-based PGNet (Xie et al., 2022), with the percentage gain of 2.4% for the MAE_{BD} score. For results on NRSOD datasets, see A.5.

4.2.2. Qualitative comparison

To visually demonstrate the advantages of our ESNet, we provide some examples of various SOD models in Figure 6. The results show that our model has obvious advantages in the following aspects. **1) Completeness of the salient objects:** In Figure 6(a) and (d), compared to other models, our ESNet does a better job of guaranteeing the integrity of salient objects. Specifically, it is able to obtain complete object structure, avoiding omissions and loopholes, whether the object is obscured or the object internal structure is complex. **2) High quality detail detection:** Although detail regions make up a relatively small proportion of the whole image, they are at the heart of high-resolution tasks and the basis for a better viewing experience. As shown in the Figure 6(b) and (c), especially the magnified images of local details, our method completely detect the antennae, legs regions of the butterfly benefiting from the introduction of evolutionary mechanisms. **3) Robustness in complex and challenging scenes:** The performance in complex scenarios reflects well the robustness and generalization of the SOD model. As shown in Figure 6(a), the foreground-background contrast in the lower left area of the sloth is very low that all other models except our method fail to detect this region. A similar situation occurs in Figure 6(e). For comparisons between transformer-based models, see A.4.

4.3. Model efficiency

Inference speed has always been a key factor restricting the development and application of HRSOD models. We conduct inference speed testing on the UHRSD-TE dataset using a single NVIDIA 3090 GPU under the same conditions for fair comparison. For the SOTA CNN-based HRSOD model (*i.e.*, HQSOD (Tang et al., 2021)), the inference speed with the size of $1,024 \times 1,024$ is less than 5 FPS. For the latest Transformer-based HRSOD method (*i.e.*, PGNet (Xie et al., 2022)) in 2022, its inference speed is only close to the real-time limit of 30 FPS with the size of $1,024 \times 1,024$. By contrast, our ESNet still maintains a clear inference speed advantage when dealing with a larger resolution image (*i.e.*, $1,280 \times 1,280$), reaching 49 FPS, while also achieving better detection performance.

Table 1: Quantitative results on the high-resolution datasets. The best result is marked in **bold**. 'NRSOD/HRSOD Model' indicates the normal-resolution/high-resolution salient object detection model.

Method	Pub'Year	HRSOD-TE					DAVIS-SOD					UHRSD-TE				
		$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$MAE_{BD}\downarrow$	$FNR\downarrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$MAE_{BD}\downarrow$	$FNR\downarrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$MAE_{BD}\downarrow$	$FNR\downarrow$
CNN-based NRSOD Model																
BASNet (Qin et al., 2019)	CVPR'19	0.038	0.861	0.891	0.235	0.104	0.017	0.926	0.925	0.212	0.101	0.053	0.886	0.883	0.221	0.128
MINet (Pang et al., 2020)	CVPR'20	0.035	0.880	0.900	0.242	0.087	0.017	0.929	0.931	0.220	0.087	0.044	0.892	0.895	0.227	0.119
GateNet (Zhao et al., 2020)	ECCV'20	0.032	0.889	0.911	0.254	0.094	0.018	0.934	0.934	0.235	0.096	0.049	0.893	0.895	0.240	0.135
GCPANet (Chen et al., 2020)	AAAI'20	0.040	0.859	0.892	0.268	0.108	0.016	0.927	0.938	0.237	0.085	0.046	0.893	0.897	0.244	0.118
PFSNet (Ma et al., 2021)	AAAI'21	0.033	0.896	0.907	0.210	0.073	0.012	0.944	0.938	0.178	0.058	0.042	0.902	0.899	0.190	0.109
MSFNet (Zhang et al., 2021a)	MM'21	0.032	0.880	0.895	0.239	0.096	0.015	0.920	0.921	0.222	0.081	0.047	0.889	0.881	0.232	0.134
CTDNet (Zhao et al., 2021)	MM'21	0.031	0.893	0.905	0.233	0.087	0.015	0.936	0.933	0.205	0.072	0.045	0.880	0.885	0.265	0.129
ICON-R (Zhuge et al., 2023)	PAMI'23	0.037	0.887	0.899	0.231	0.068	0.015	0.931	0.929	0.212	0.082	0.048	0.893	0.892	0.219	0.108
EDN (Wu et al., 2022)	TIP'22	0.034	0.885	0.905	0.237	0.097	0.015	0.932	0.933	0.219	0.086	0.045	0.902	0.896	0.230	0.132
CNN-based HRSOD Model																
HRSOD (Zeng et al., 2019)	ICCV'19	0.030	0.889	0.895	0.250	0.160	0.021	0.888	0.913	-	-	-	-	-	-	-
DRFNet (Zhang et al., 2021b)	TIP'21	0.025	0.906	0.913	0.215	0.099	0.012	0.904	0.940	-	-	-	-	-	-	-
HQSOD (Tang et al., 2021)	ICCV'21	0.024	0.907	0.919	0.185	0.071	0.014	0.939	0.937	0.165	0.081	0.040	0.911	0.901	0.174	0.118
DDPNet (Wang et al., 2022a)	AI'22	-	0.906	0.901	-	-	-	-	-	-	-	-	-	-	-	-
OURS	-	**0.021**	**0.926**	**0.932**	**0.166**	**0.053**	**0.011**	**0.949**	**0.945**	**0.157**	**0.055**	**0.038**	**0.915**	**0.909**	**0.164**	**0.102**
Transformer-based NRSOD Model																
VST (Liu et al., 2021)	ICCV'21	0.036	0.891	0.909	0.280	0.083	0.016	0.931	0.934	0.263	0.097	0.041	0.907	0.910	0.259	0.093
ICON-S (Zhuge et al., 2023)	PAMI'23	0.027	0.907	0.918	0.270	0.080	0.016	0.927	0.929	0.254	0.079	0.038	0.910	0.908	0.260	0.106
Transformer-based HRSOD Model																
PGNet (Xie et al., 2022)	CVPR'22	0.020	0.928	0.934	0.161	0.049	0.012	0.954	0.946	0.163	0.052	0.036	0.914	0.911	0.155	0.106
OURS_swin	–	**0.019**	**0.937**	**0.942**	**0.157**	**0.039**	**0.009**	**0.957**	**0.950**	**0.152**	**0.050**	**0.027**	**0.935**	**0.931**	**0.147**	**0.070**

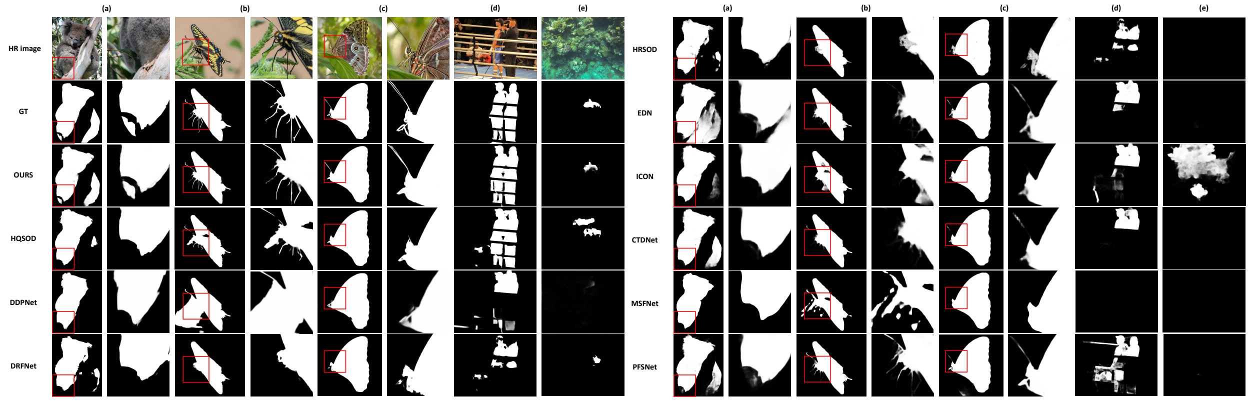


Figure 6: Visual comparisons between our ESNet and SOTA methods under different scenes with obscured objects (a, d), rich detail (b, c), complex backgrounds (d), and low contrast (a, e).

4.4. Ablation Studies

4.4.1. Analysis of the overall architecture

We conduct experiments to verify the role of two-stage design in the overall structure, as shown in Table 2. The specific experimental settings are as follows: FULL (id 0) means our proposed two-stage ESNet with the ResNet50 backbone, *i.e.*, LrLM+HrRM; LrLM (id 1) means that only the first stage LrLM is used here with LR inputs; LrLM+patch (id 2) introduces patch-dividing in the succession stage, while maintaining the evolution stage unchanged.

Although the result of only taking the first stage with the low-resolution input (id 1) drops from full ESNet (id 0), it still outperforms most existing NRSOD methods, which demonstrates the effectiveness of the design of the evolution stage. For example, on the UHRSD-TE dataset, our LrLM achieves a percentage gain of 7.1% for MAE score compared to the SOTA NRSOD PFSNet. As shown in id 2 of Table 2, all metrics become worse compared to our ESNet,

Table 2: Quantitative ablation evaluation of the model structure on the high-resolution datasets. Black bold fonts indicate the best performance.

Method	ID	HRSOD				UHRSD-TE			
		$MAE\downarrow$	$F_\beta\uparrow$	$MAE_{BD}\downarrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$MAE_{BD}\downarrow$	$S_\alpha\uparrow$
FULL	0	**0.021**	**0.926**	**0.166**	**0.932**	**0.038**	**0.915**	**0.164**	**0.909**
LrLM	1	0.023	0.916	0.206	0.926	0.039	0.910	0.199	0.905
LrLM+Patch	2	0.022	0.919	0.190	0.928	0.040	0.911	0.185	0.905
w/o EF&ES	3	0.022	0.925	0.184	0.925	0.040	0.905	0.176	0.902
w/o EF	4	0.026	0.916	0.170	0.920	**0.038**	0.912	0.167	0.908
w/o ev loss	5	0.023	0.919	0.168	0.929	0.039	0.912	0.169	0.907
w/o MES	6	0.022	0.924	0.171	0.929	0.039	0.914	0.168	0.907
whole-att	7	0.022	0.924	0.180	0.928	0.039	0.911	0.176	0.906
standard-att	8	0.023	0.918	0.201	0.926	0.039	0.910	0.196	0.906

particularly the MAE_{BD} metric drops from 0.166 to 0.190 on the HRSOD-TE dataset, attributed to the fact that the patch-dividing operation disrupts the artifacts of edges.

4.4.2. Analysis of the evolution stage

To verify the effectiveness of the design in the evolution stage, we conduct the following experimental settings: w/o ES&EF (id 3) means the evolution designs on feature and supervision are all removed from the full model, *i.e*., the first stage LrLM degenerates to the baseline model CTDNet (Zhao et al., 2021); w/o EF (id 4) means that the evolution on feature is removed, leaving only the supervision evolution; w/o ev loss (id 5) removes the evolutionary variation loss in the evolution process on feature.

In the evolution stage, the evolution mechanisms on supervision and feature are designed to achieve detail-preserving salient object localization. As shown in Table 2, after removing each design separately, the performance is degraded. If we remove all the evolution mechanisms of the design, on the UHRSD-TE dataset, the F-measure is reduced from 0.915 to 0.905, and the MAE_{BD} is worsened from 0.164 to 0.176. Subsequently, when we only add the supervision evolution, there is no overall improvement in performance. Compared with the id 3, the performance of id 4 on the UHRSD-TE dataset improves (*e.g*., the F-measure improves from 0.906 to 0.911), while most metrics are worse on the HRSOD dataset. This is mainly because the remote evolution supervisions only tell the network what the label is at each layer, without explicitly telling the network the relationship between different labels, which is obviously still difficult for the network learning, thereby affecting the robustness of the network. For visual comparisons, see A.7.

4.4.3. Analysis of the succession stage

We conduct ablation experiments in Table 2 to verify the effectiveness of the design and setup in the succession stage. The specific experimental settings are as follows: w/o MES (id 6) means the mutually exclusive supervision is removed; whole-att (id 7) means extracting prototypes and performing feature enhancement using the entire foreground and background regions; standard-att (id 8) means replacing SFEM with a standard form of self-attention by measuring similarity between whole feature maps.

Both MES and SFEM are designed for high-quality detail refinement. Removing or replacing the MES and SFEM results in a obvious decrease in MAE_{BD} metric. This objective change highlights the usefulness of these two designs in improving the quality of detail detection. Besides, for the range of SFEM, as shown in Table 2, whole-att (id 7) approach dose not lead to significant improvements in the quality of details compared to our method of operating exclusively within the contour area, which confirms that our decision to focus on the contour area for prototype extraction and feature enhancement is more effective in enhancing the detail quality. Compared with id 8 of Table 2, our SFEM avoids the extensive computational load associated with per-pixel similarity measurements while improving in the MAE_{BD} metric by 17.4% and 16.3% on both two datasets compared to the standard form self-attention.

5. Conclusion

In this paper, we decouple the HRSOD task into the low-resolution object localization subtask and high-resolution detail refinement subtask, and then propose a two-stage Evolution and Succession Network (ESNet). The evolution stage achieves detail-preserving salient objects localization, while the succession stage realizes supplement and enhancement in a lightweight way. Moreover, we design a new metric named Boundary-Detail-aware MAE (MAE_{BD}) to better evaluate the quality of detail detection in high-resolution scenes. Extensive experiments demonstrate that the effectiveness and efficiency of our ESNet.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61991411, Grant 62122075, Grant 62201179, Grant 62236008, and Grant U21B2038, in part by the Taishan Scholar Project of Shandong Province under Grant tsqn202306079, in part by the National Key Research and Development Plan of China under Grant 2021ZD0112002, in part by Xiaomi Young Talents Program, in part by the Innovation Platform for "New Star of South China Sea" of Hainan Province under Grant NHXXRCXM202306, in part by Project for Self-Developed Innovation Team of Jinan City under Grant 2021GXRC038, and in part by CAAI-Huawei MindSpore Open Fund..

Impact Statement

This work aims at general theoretical issues for the high-resolution salient object detection problem and does not present any foreseeable societal consequence.

References

Cai, S., Zhang, X., Fan, H., Huang, H., Liu, J., Liu, J., Liu, J., Wang, J., and Sun, J. Disentangled image matting. In *Proc. ICCV*, pp. 8818–8827, 2019.

Chen, J., Cong, R., Ip, H. H. S., and Kwong, S. Kepsalinst: Using peripheral points to delineate salient instances. *IEEE Trans. Cybern.*, 2023.

Chen, Z., Xu, Q., Cong, R., and Huang, Q. Global contextaware progressive aggregation network for salient object detection. In *Proc. AAAI*, pp. 10599–10606, 2020.

Cong, R., Huang, K., Lei, J., Zhao, Y., Huang, Q., and Kwong, S. Multi-projection fusion and refinement network for salient object detection in 360° omnidirectional image. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022a.

Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q., and Zhao, Y. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Trans. Image Process.*, 31:6800–6815, 2022b.

Cong, R., Liu, H., Zhang, C., Zhang, W., Zheng, F., Song, R., and Kwong, S. Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection. In *Proc. ACM MM*, pp. 406–416, 2023a.

Cong, R., Qin, Q., Zhang, C., Jiang, Q., Wang, S., Zhao, Y., and Kwong, S. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Trans. Circuits Syst. Video Technol.*, 33(2):534–548, 2023b.

Cong, R., Yang, N., Li, C., Fu, H., Zhao, Y., Huang, Q., and Kwong, S. Global-and-local collaborative learning for co-salient object detection. *IEEE Trans. Cybern.*, 53(3): 1920–1931, 2023c.

Cong, R., Zhang, K., Zhang, C., Zheng, F., Zhao, Y., Huang, Q., and Kwong, S. Does Thermal really always matter for RGB-T salient object detection? *IEEE Trans. Multim.*, 25:6971–6982, 2023d.

Dong, B., Wang, P., and Wang, F. Head-free lightweight semantic segmentation with linear transformer. In Williams, B., Chen, Y., and Neville, J. (eds.), *Proc. AAAI*, pp. 516–524, 2023.

Fan, D., Cheng, M., Liu, Y., Li, T., and Borji, A. Structure-measure: A new way to evaluate foreground maps. In *Proc. ICCV*, pp. 4558–4567, 2017.

Fan, D., Li, T., Lin, Z., Ji, G., Zhang, D., Cheng, M., Fu, H., and Shen, J. Re-thinking co-salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8): 4339–4354, 2022.

Guo, S., Liu, L., Gan, Z., Wang, Y., Zhang, W., Wang, C., Jiang, G., Zhang, W., Yi, R., Ma, L., and Xu, K. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proc. CVPR*, pp. 4351–4360, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778, 2016.

Kirillov, A., Wu, Y., He, K., and Girshick, R. B. Pointrend: Image segmentation as rendering. In *Proc. CVPR*, pp. 9796–9805, 2020.

Li, G. and Yu, Y. Visual saliency based on multiscale deep features. In *Proc. CVPR*, pp. 5455–5463, 2015.

Li, J., Su, J., Xia, C., and Tian, Y. Distortion-adaptive salient object detection in 360° omnidirectional images. *IEEE J. Sel. Top. Signal Process.*, 14(1):38–48, 2020.

Liu, N., Zhang, N., Wan, K., Shao, L., and Han, J. Visual saliency transformer. In *Proc. ICCV*, pp. 4702–4712, 2021.

Liu, Y., Guo, Q., Fu, L., Ke, Z., Xu, K., Feng, W., Tsang, I. W., and Lau, R. W. H. Structure-informed shadow removal networks. *IEEE Trans. Image Process.*, 32:5823–5836, 2023.

Ma, M., Xia, C., and Li, J. Pyramidal feature shrinking for salient object detection. In *Proc. AAAI*, pp. 2311–2318, 2021.

Niu, Y., Geng, Y., Li, X., and Liu, F. Leveraging stereopsis for saliency analysis. In *Proc. CVPR*, pp. 454–461, 2012.

Pang, Y., Zhao, X., Zhang, L., and Lu, H. Multi-scale interactive network for salient object detection. In *Proc. CVPR*, pp. 9410–9419, 2020.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L. V., Gross, M. H., and Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. CVPR*, pp. 724–732, 2016.

Qin, X., Zhang, Z. V., Huang, C., Gao, C., Dehghan, M., and Jägersand, M. Basnet: Boundary-aware salient object detection. In *Proc. CVPR*, pp. 7479–7489, 2019.

Tang, L., Li, B., Zhong, Y., Ding, S., and Song, M. Disentangled high quality salient object detection. In *Proc. ICCV*, pp. 3560–3570, 2021.

Wang, J., Yang, Q., Yang, S., Chai, X., and Zhang, W. Dual-path processing network for high-resolution salient object detection. *Appl. Intell.*, 52(10):12034–12048, 2022a.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. Learning to detect salient objects with image-level supervision. In *Proc. CVPR*, pp. 3796–3805, 2017.

Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. Openauc: Towards auc-oriented open-set recognition. In *Proc. NeurIPS*, 2022b.

Wu, Y., Liu, Y., Zhang, L., Cheng, M., and Ren, B. EDN: salient object detection via extremely-downsampled network. *IEEE Trans. Image Process.*, 31:3125–3136, 2022.

Wu, Z., Su, L., and Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In *Proc. CVPR*, pp. 3907–3916, 2019.

Xie, C., Xia, C., Ma, M., Zhao, Z., Chen, X., and Li, J. Pyramid grafting network for one-stage high resolution saliency detection. In *Proc. CVPR*, pp. 11707–11716, 2022.

Xu, Q., Yang, Z., Jiang, Y., Cao, X., Yao, Y., and Huang, Q. Not all samples are trustworthy: Towards deep robust SVP prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):3154–3169, 2022.

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., and Huang, Q. Optimizing two-way partial AUC with an end-to-end framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (8):10228–10246, 2023.

Yu, F. and Koltun, V. Multi-scale context aggregation by dilated convolutions. In *Proc. ICLR*, 2016.

Zeng, Y., Zhang, P., Lin, Z. L., Zhang, J., and Lu, H. Towards high-resolution salient object detection. In *Proc. ICCV*, pp. 7233–7242, 2019.

Zhang, M., Liu, T., Piao, Y., Yao, S., and Lu, H. Automsfnet: Search multi-scale fusion network for salient object detection. In *Proc. ACM MM*, pp. 667–676, 2021a.

Zhang, P., Liu, W., Zeng, Y., Lei, Y., and Lu, H. Looking for the detail and context devils: High-resolution salient object detection. *IEEE Trans. Image Process.*, 30:3204–3216, 2021b.

Zhao, X., Pang, Y., Zhang, L., Lu, H., and Zhang, L. Suppress and balance: A simple gated network for salient object detection. In *Proc. ECCV*, pp. 35–51, 2020.

Zhao, Z., Xia, C., Xie, C., and Li, J. Complementary trilateral decoder for fast and accurate salient object detection. In *Proc. ACM MM*, pp. 4967–4975, 2021.

Zhou, T., Fan, D., Cheng, M., Shen, J., and Shao, L. RGB-D salient object detection: A survey. *Comput. Vis. Media*, 7 (1):37–69, 2021.

Zhuge, M., Fan, D., Liu, N., Zhang, D., Xu, D., and Shao, L. Salient object detection via integrity learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3738–3752, 2023.

A. appendix

A.1. Loss Function

A.1.1. Loss function of the evolution stage

On the basis of CTDNet (Zhao et al., 2021), the loss function of the evolution stage is defined as follows:

$$\ell_{EVO} = \ell_{sal} + \ell_{bdy} + \varepsilon \cdot \ell_{ev} = \sum_{i=1}^{5} \frac{1}{2^i} \ell_{com} \left(S_l^i, G_l^i \right) + \ell_{bce} \left(S_l^e, G_l^e \right) + \varepsilon \cdot \ell_{ev}, \tag{18}$$

where ℓ_{sal}, ℓ_{bdy} and ℓ_{ev} represent the comprehensive loss of evolution supervision for each layer, the boundary loss, and our proposed evolutionary variation loss, respectively; S_l^e and G_l^e represent the boundary map of saliency prediction from the last decoder layer of the evolution stage and the corresponding saliency boundary label, respectively; the hyperparameter ε is set to 0.1, and ℓ_{com} is a comprehensive loss as defined in CTDNet (Zhao et al., 2021), which includes BCE loss ℓ_{bce}, SSIM loss ℓ_{ssim}, and IOU loss ℓ_{iou}:

$$\ell_{com} \left(S_l^i, G_l^i \right) = \ell_{iou} \left(S_l^i, G_l^i \right) + \gamma \cdot \ell_{bce} \left(S_l^i, G_l^i \right) + \eta \cdot \ell_{ssim} \left(S_l^i, G_l^i \right) \tag{19}$$

where γ is set to 0.6, and η is set to $\delta(i-1)$, and δ is the unit impulse function.

A.1.2. Loss function of the succession stage

The loss function of succession stage consists of the saliency loss and the mutual exclusion loss ℓ_{MES}, which is defined as:

$$\ell_{SUC} = \ell_{com} \left(S_{out}, G_h \right) + \ell_{MES}, \tag{20}$$

where ℓ_{com} is consistent with the constraint on the S_l^1 in the evolution stage, and S_{out} is the final HR saliency map.

A.2. Implementation Details

We implement the proposed ESNet by Pytorch and conduct experiments on a single NVIDIA GeForce RTX 3090 GPU. We also implement our network by using the MindSpore Lite tool[1]. For faster convergence, the evolution stage and the succession stage are trained respectively. Following the setup in (Zeng et al., 2019; Zhang et al., 2021b), the training set of HRSOD and DUTS datasets are used for training. The training samples are further augmented by random cropping and random flipping, then resized to 352×352 and fed to the evolution stage. A SGD optimizer with momentum of 0.9 and weight decay of 0.0005 is used here. The batch size is set to 48 and the training epoch is 80. Warm-up and linear decay learning rate strategy are used with the maximum learning rate of 0.005 for the pre-trained ResNet50 feature extraction backbone and 0.05 for the rest of the network. Then, the trained LrLM will be frozen for inference only without gradient updates, to assist the HrRM training. For the HrRM, it is firstly pre-trained on the DUTS-TR dataset for 30 epochs and then fine-tuned on the HRSOD-TR for 60 epochs to obtain the final fully trained model, with the input size of $1,280 \times 1,280$.

A.3. Dataset

HRSOD dataset (Zeng et al., 2019) contains 2,010 images, where the 1,610/400 images forming the training dataset (HRSOD-TR) and testing dataset (HRSOD-TE), respectively. **DAVIS-SOD** dataset (Zhang et al., 2021b) includes 950 densely annotated images with the resolution of $1,920 \times 1,080$, which are collected from the public video segmentation benchmark DAVIS (Perazzi et al., 2016). **UHRSD** is a new high-resolution SOD dataset released in 2022 (Xie et al., 2022), including 4,932 images for training (UHRSD-TR) and 988 images for testing (UHRSD-TE). For these three high-resolution SOD datasets, the length or width of each image is not shorter than 1,000, which is much higher than traditional SOD datasets. In addition, to verify the generalization of our evolutionary structure, we also evaluate our method on two normal-resolution SOD datasets, including **DUTS** (Wang et al., 2017) and **HKU-IS** (Li & Yu, 2015), which contain 5,019 and 4,447 images, respectively.

A.4. Visual Comparisons with Transformer-based Models

In addition, we also compare our transformer-based version (*i.e.*, OURS_swin) with PGNet(Xie et al., 2022) and VST (Liu et al., 2021). The visual examples are shown in Figure 7. Overall, our model has obvious advantages in irrelevant

[1] https://www.mindspore.cn/

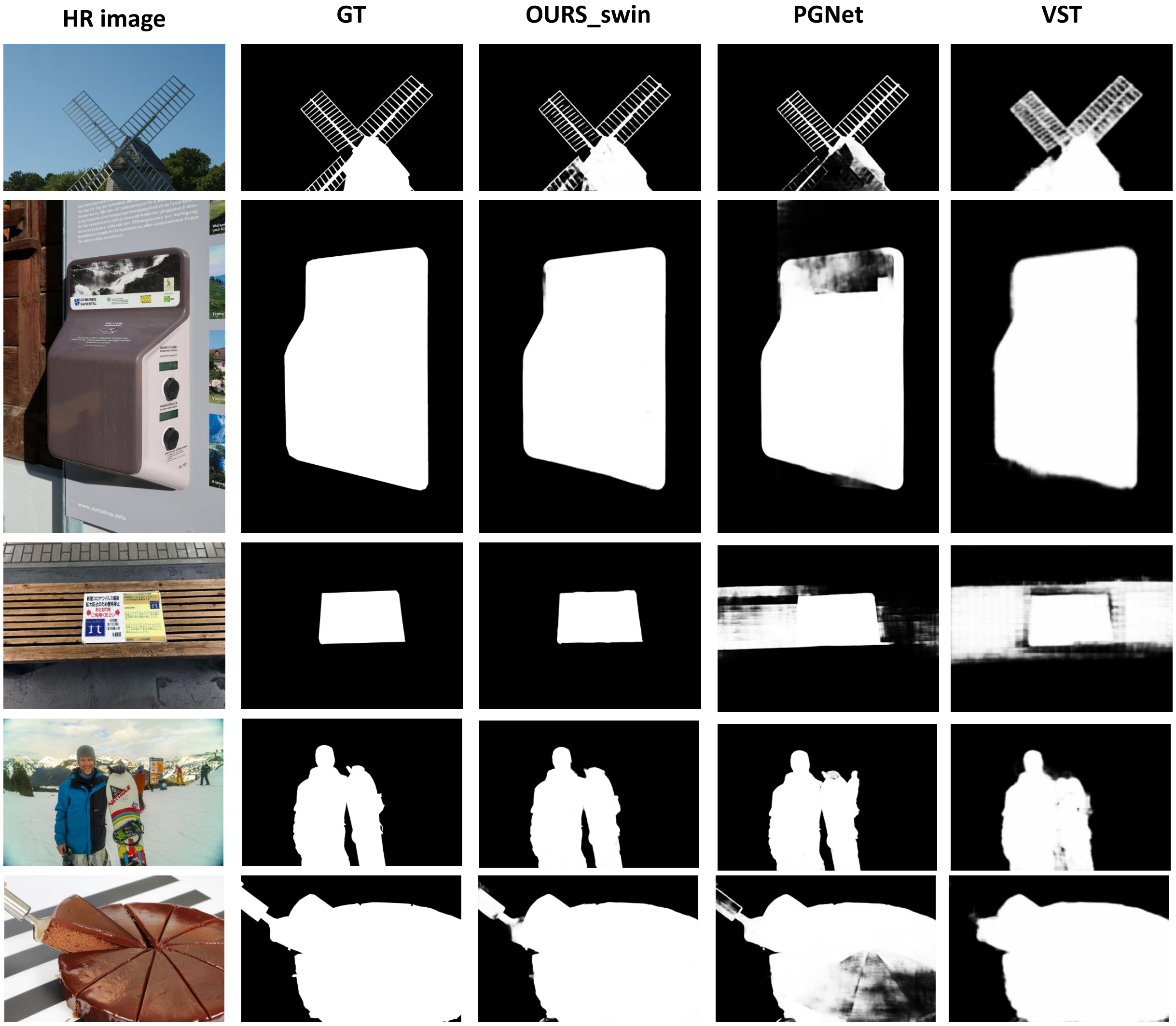


Figure 7: Visual comparison between transformer-based methods, including OURS_swin , PGNet (Xie et al., 2022) and VST (Liu et al., 2021).

background suppression, internal consistency of objects, detail characterization, *etc*. For example, in the first image of Figure 7, the blades of the windmill have many hollow details. Compared with the PGNet (Xie et al., 2022) and VST (Liu et al., 2021), our ESNet not only successfully characterizes these detail regions, but also has better target integrity. In the second and last images, the PGNet method (Xie et al., 2022) cannot fully detect the structure of salient objects, such as the upper region of the machine and the lower right region of the cake. In the third image, the strong interference of chair can not be effectively suppressed by the PGNet (Xie et al., 2022) and VST (Liu et al., 2021). For the above cases, our method wins better results.

A.5. Comparison on NRSOD Dataset

From the perspective of generalization verification, our designed HRSOD model can still perform well on the NRSOD datasets. Table 3 shows the quantitative results on two normal-resolution datasets, where the best performance is marked in bold. For example, on the DUTS-TE dataset, our method outperforms the NRSOD and HRSOD methods on all three metrics, with a percentage gain of 3.1% and 1.0% for the MAE and F-measure compared with the **second best** model,

Table 3: Quantitative results on the normal-resolution DUTS-TE (Wang et al., 2017) and HKU-IS (Li & Yu, 2015) datasets. The best result is marked in **bold**. 'NRSOD/HRSOD Model' indicates the normal/high-resolution SOD model.

Method	Pub'Year	DUTS-TE			HKU-IS		
		$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$
CNN-based NRSOD Model							
BASNet (Qin et al., 2019)	CVPR'19	0.047	0.837	0.866	0.032	0.919	0.909
MINet (Pang et al., 2020)	CVPR'20	0.037	0.863	0.885	0.029	0.926	0.919
GateNet (Zhao et al., 2020)	ECCV'20	0.038	0.872	0.890	0.031	0.926	0.921
GCPANet (Chen et al., 2020)	AAAI'20	0.038	0.866	0.890	0.031	0.928	0.922
PFSNet (Ma et al., 2021)	AAAI'21	0.035	0.879	0.892	0.026	0.934	0.924
MSFNet (Zhang et al., 2021a)	MM'21	0.034	0.863	0.877	0.027	0.922	0.910
CTDnet (Zhao et al., 2021)	MM'21	0.034	0.880	0.893	0.027	0.932	0.922
ICON (Zhuge et al., 2023)	PAMI'23	0.037	0.876	0.889	0.029	0.930	0.920
EDN (Wu et al., 2022)	TIP'22	0.035	0.878	0.893	0.050	0.799	0.850
CNN-based HRSOD Model							
HRSOD (Zeng et al., 2019)	ICCV'19	0.050	0.792	0.823	0.042	0.889	0.877
DRFNet (Zhang et al., 2021b)	TIP'21	0.045	0.786	0.861	0.037	0.891	0.906
HQSOD (Tang et al., 2021)	ICCV'21	0.032	0.881	0.892	0.025	**0.937**	0.923
DDPNet (Wang et al., 2022a)	AI'22	-	0.859	0.869	-	0.935	0.911
OURS	-	**0.031**	**0.890**	**0.899**	**0.024**	**0.937**	**0.926**
Tranformer-based NRSOD Model							
VST(Liu et al., 2021)	ICCV'21	0.037	0.877	0.896	0.030	0.937	0.928
ICON-S(Zhuge et al., 2023)	PAMI'23	0.028	0.895	0.906	0.028	0.933	0.925
Tranformer-based HRSOD Model							
PGNet(Xie et al., 2022)	CVPR'22	0.028	0.903	0.912	0.024	0.939	0.930
OURS_swin	–	**0.022**	**0.921**	**0.924**	**0.019**	**0.951**	**0.939**

respectively.

A.6. Visualization of Features in Evolution Stage

In Figure 8, we provide some visualization results of evolution stage, including the evolutionary labels, the corresponding output maps and decoder features. As can be seen, the higher the level of our evolutionary labels, the more the edges are smoothed, and the proportion of detail regions is significantly increased. Moreover, our design makes it easier to learn higher-level features while paying more attention to the detail areas, and the evolution of the feature level allows the differences between features at each level to be reflected in the evolutionary change areas as much as possible without destroying the main body information.

A.7. Visualization of Ablation Studies

Some visual comparisons are shown in Figure 9. Notably, as depicted in the bottom, the supervision evolution has a noticeable impact on preserving detailed regions like butterfly tentacles, aligning with our original design motivation. As also can be seen, the results of id 3, 4 and 5 are not satisfactory for background suppression and detail characterization, such as the background interference in the left corner of the first image. By contrast, our full model with complete two stage effectively suppresses these background regions and clearly outlines local details.

A.8. Failure Cases

Some typical failure cases are given in Figure 10. The performance of ESNet in scenarios with skeletonized structures inside the object is still falling short, which is related to our evolution strategy of emphasizing the integrity of the object. This is also the direction to be improved in the future utilizing more fine-grained features and discriminations. Meanwhile both this type of area and the detail area are also relatively small, it may help to refer to the long-tailed distribution optimization approach(Wang et al., 2022b; Yang et al., 2023). In addition, the existing HRSOD models, including our ESNet, are dependent on high-resolution, high-quality labeled datasets, which undoubtedly imposes a large labeling cost. In the future,

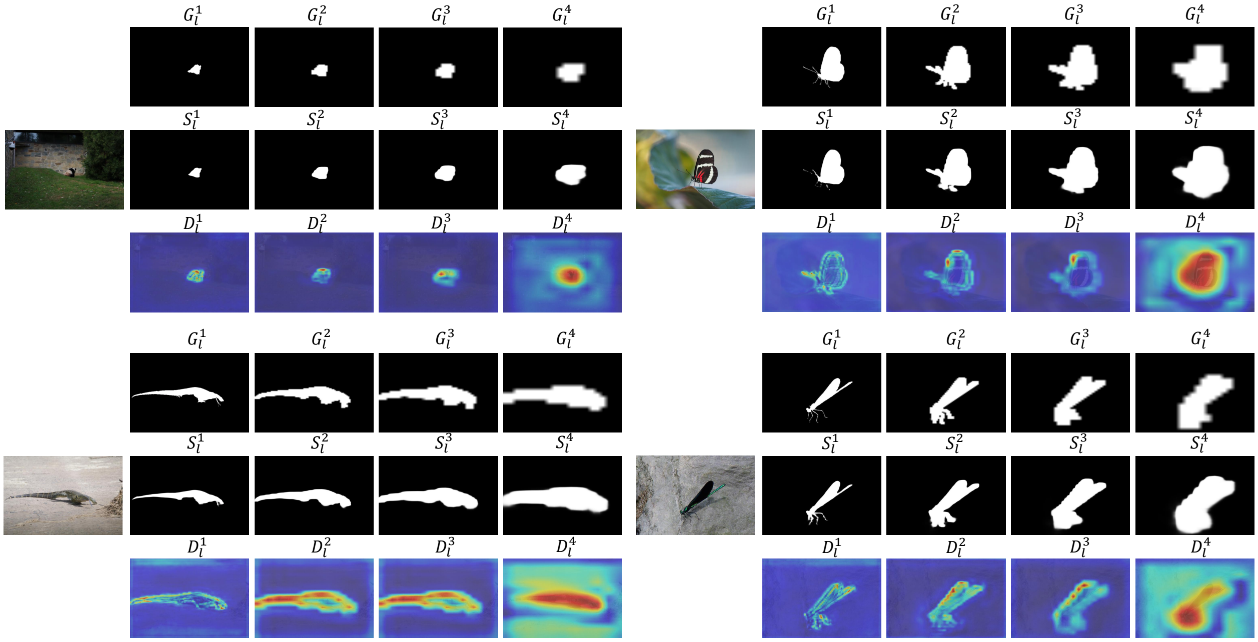

G_l^1 G_l^2 G_l^3 G_l^4
S_l^1 S_l^2 S_l^3 S_l^4
D_l^1 D_l^2 D_l^3 D_l^4
G_l^1 G_l^2 G_l^3 G_l^4
S_l^1 S_l^2 S_l^3 S_l^4
D_l^1 D_l^2 D_l^3 D_l^4
G_l^1 G_l^2 G_l^3 G_l^4
S_l^1 S_l^2 S_l^3 S_l^4
D_l^1 D_l^2 D_l^3 D_l^4
G_l^1 G_l^2 G_l^3 G_l^4
S_l^1 S_l^2 S_l^3 S_l^4
D_l^1 D_l^2 D_l^3 D_l^4

Figure 8: Visualization of evolutionary labels and corresponding output maps and decoder features.

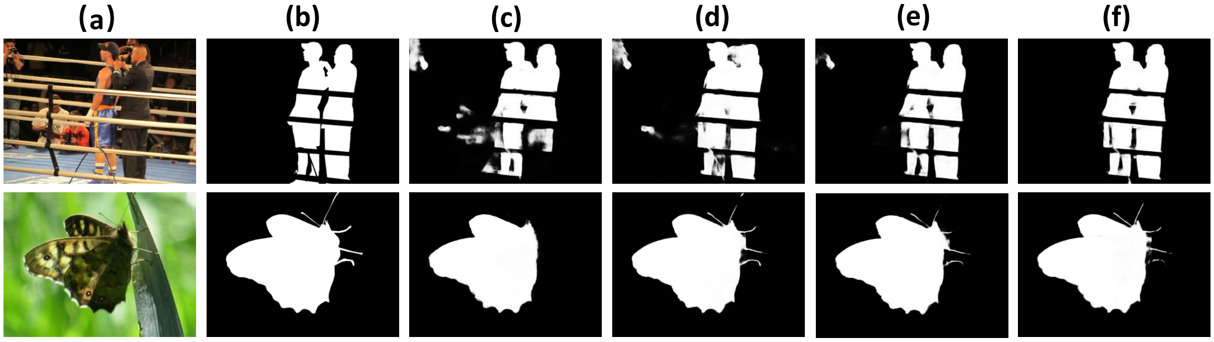


Figure 9: Visual comparison of ablation study on the evolution stage. (a) HR image. (b) GT. (c) w/o EF&ES (id 3). (d) w/o EF (id 4). (e) w/o ev loss (id 5). (f) FULL (id 0).

we hope to explore the possibility of realizing high-quality high-resolution detection in a low-cost form by using a small amount of data or rough annotations.

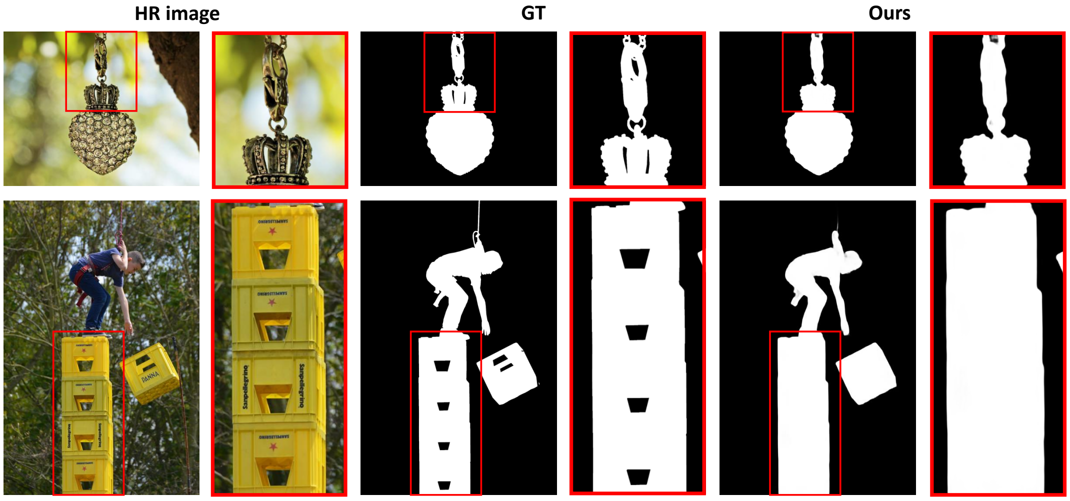


Figure 10: Typical failure cases of our ESNet.