This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

FasterSal: Robust and Real-time Single-Stream Architecture for RGB-D Salient Object Detection

Jin Zhang, Ruiheng Zhang, *Member, IEEE*, Lixin Xu, Xiankai Lu, *Member, IEEE*, Yushu Yu, Min Xu, *Member, IEEE*, and He Zhao, *Member, IEEE*

***Abstract*—RGB-D Salient Object Detection (SOD) aims to segment the most prominent areas and objects in a given pair of RGB and depth images. Most current models adopt a dual-stream structure to extract information from both RGB and depth images. However, this leads to an exponential increase in the number of parameters and computations in the model. Moreover, the discrepancy between RGB pretrained and the 3D geometric relationships in depth maps present a challenge for the encoder in capturing spatial structural details. These issues impact the model's accuracy in locating salient objects and distinguishing edge details. To address these, we propose a novel early feature fusion network, named FasterSal, which enables more efficient RGB-D SOD. FasterSal uses a single stream structure to receive RGB images and depth maps, extracting features based on the 3D geometric relationships in the depth map while fully leveraging the pretrained RGB encoder. This approach effectively avoids the inconsistencies between depth modality and the RGB pretrained encoder. It also significantly reduces the number of network parameters while maintaining efficient feature encoding capabilities. To achieve finer edge learning, the detail-aware loss and texture enhancement module are introduced. These modules are designed to extract latent details in high-frequency component features and to enhance the edge learning capability of the model using distance information. Experimental results on several benchmark datasets confirm the effectiveness and superiority of our method over the state-of-the-art approaches, achieving a good balance between performance and speed with only 3.4 million parameters and a CPU operating speed of 63 FPS. Code and results available at: https://github.com/zhangjinCV/FasterSal.**

***Index Terms*—Saliency detection, RGBD images, single-stream, real-time segmentation, detail awareness.**

I. Introduction

SALIENT object detection (SOD) is a crucial computer vision task that focuses on identifying and segmenting the most prominent object or objects within an image. It disregards subtle associations between similar object classes and instead concentrates on the most visually striking elements, aligning with principles of human visual perception. SOD finds applications in various domains, including semantic segmentation [2, 3], image enhancement [4, 5], video compression [6], and visual tracking [7–9].

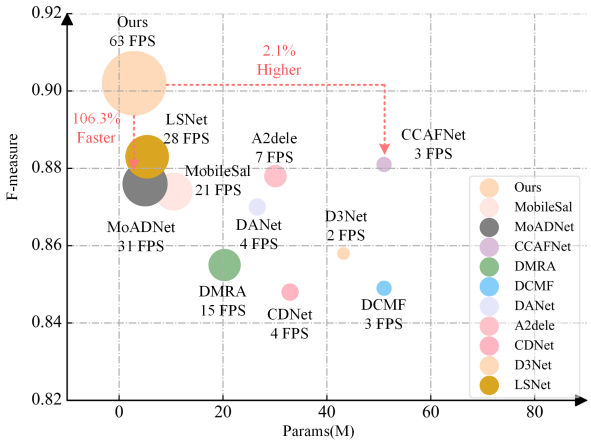

Fig. 1. The comparison between our methods **FasterSal** and existing methods in terms of parameters, accuracy and speed. The metrics F-measure are calculated in the NLPR dataset [1]. Different colors indicate different methods, and the size of the circle represents the speed of detection on CPU environment.

Recent advancements in RGB-D salient object detection methods [10–15] have shown superior performance over traditional RGB-based approaches. Chen et al. [16] proposed an innovative method for RGB-D salient object detection, which effectively integrates depth cues into the saliency detection process. Their novel neural network architecture uses depth information to distinctly enhance the separation of salient objects from their backgrounds, particularly in complex scenes. Similarly, Pang et al. [17] introduced CAVER, a framework that leverages context-aware analysis and entropy-based refinement in RGB-D detection. This method excels in differentiating salient objects from backgrounds in challenging environments, where they often share similar features.

Despite the impressive performance of these models, the development and application of RGB-D SOD technology still face significant challenges. A major issue is how to strike the best balance between detection effectiveness and speed. Heavyweight models [14, 16, 18–24], though powerful, are often impractical for real-world deployment due to their large parameter sizes and high computational demands. This limitation is evident, as illustrated in Fig. 1, where these models perform well on high-end devices but struggle to deliver real-

This work was funded by the STI 2030—Major Projects under grant 2022ZD0209600, the National Natural Science Foundation of China under grant 62201058, 62475016 and the Beijing Institute of Technology Research Fund Program for Young Scholars under grant 6120210047. *(Corresponding author: Ruiheng Zhang, email: ruiheng.zhang@bit.edu.cn)*

J. Zhang, R. Zhang, L. Xu, and Y. Yu are with the State Key Laboratory of Electromechanical Dynamic Control, the School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, 100081, China.

X. Lu is with the School of Software, Shandong University, Jinan 250101, China

M. Xu is with the School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia

H. Zhao is with the Department of Eye and Vision Sciences, University of Liverpool, U.K.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

time performance on edge and mobile devices.

To achieve faster inference, researchers have proposed more efficient RGB-D SOD methods. Zhou et al.'s "LSNet" [25] employs a lightweight network design while introducing spatial enhancement mechanisms to improve the detection performance of salient objects. Wu et al.'s "MobileSal" [26] notably simplified the process by fusing RGB with depth information at the coarsest level, thereby bypassing the intensive computations of low-level fusion. Pursuing efficiency, Jin et al.'s "MoADNet" [27] involved using a compact encoder for depth map, leading to a network that is both leaner and less demanding computationally. However, these methods also have some noteworthy issues: i) They use RGB pretrained parameters to extract depth features directly, which is different from the image input in RGB pretrained and results in the destruction of the 3D geometry in the depth map; ii) During the training process, mismatched pretrained parameters can bring about significant changes in the model representation distribution, thereby affecting refined details predictions; iii) Compared to standard RGB methods, their dual-stream structure is effective, but it generates additional computational costs and still impacts overall efficiency (shown in Fig. 1). In Fig. 2, we use RGB pretrained weights for feature extraction of both RGB and depth images at low- and high-level. From the figure, it is observed that the low-level features still retain rich detail textures, while the high-level features from the depth images completely fail to extract effective image semantics. As mentioned in DFormer [28], the depth branches of existing dual-branch RGBD models all adopt RGB pre-trained weights, which severely affects the representation of depth information, especially in lightweight model architectures, where there is a lack of effective utilization of bimodal information. Moreover, we posit that employing RGB pretrained weights for the extraction of low-level features is effective and enhances the capture of texture details, with specific experiments documented in Table V.

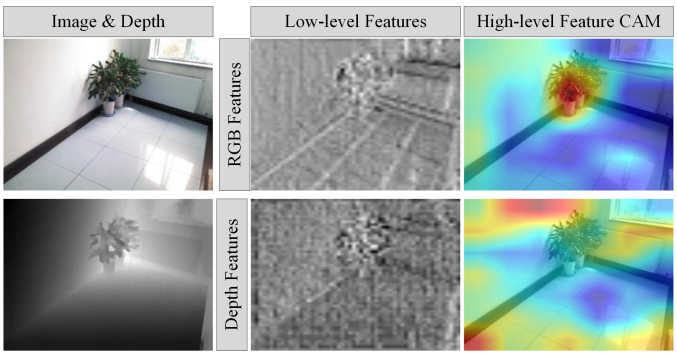

Fig. 2. Feature extraction of RGB and depth maps using RGB pretrained weights. Redder CAMs indicate areas of greater concern to the model.

To address these issues, we propose an innovative architecture named **FasterSal**, aimed at overcoming the challenges faced by RGB-D salient object detection. Considering the minimal contribution of RGB pretrained encoders to depth information and the parameter overhead of the dual-stream structure, we design a single stream encoder with 4-channel input to simultaneously receive RGB images and depth maps. Unlike previous dual-stream methods [26, 27], the interaction between RGB and depth images is implemented within the same encoder. This allows the RGB images to fully use the pretrained parameters as well as the 3D geometric relationships in the depth maps, and also avoids the distribution fluctuation problems caused by modality inconsistencies. Furthermore, we introduce a Texture Enhancement Module (TEM) and Detail-Aware Loss (DA Loss) to capture the high-frequency fine textures and edges of objects. TEM enables the model to focus more on high-frequency details. During training, DA Loss improves the model's edge learning capabilities, thereby enhancing overall prediction quality. Finally, we introduce the Object Awareness Module (OAM) and Attention-Based Decoder (ABD). These modules provide valuable contextual clues for smaller objects and seamlessly integrate structural information extracted from depth features, achieving seamless integration of information at different abstraction levels. Our main contributions include:

- We present FasterSal, a powerful single-stream architecture designed for RGB-D SOD. FasterSal uses a more efficient interaction method to fuse RGB and depth information, avoiding the problem of mismatch in the encoding of 3D geometric relations in depth maps.
- We introduce a texture enhancement module and detail-aware loss in FasterSal, which effectively balances the model's modeling of low-frequency semantics and high-frequency details, and enhances the model's handling of object details from a distance-aware perspective, thus enhancing the model's edge learning capability.
- We evaluate the proposed FasterSal on 5 RGB-D SOD datasets to demonstrate its comparable accuracy to heavyweight methods, along with more than a 2× efficiency improvement compared to lightweight methods (63 FPS vs. 31 FPS) and a smaller network size (3.4M vs. 5.0M).

II. Related Work

A. RGB-D Saliency Detection

As a multimodal learning task, most existing RGB-D SOD models [27, 29–34] primarily focus on efficient fusion of multimodal features, achievable through implicit multimodal feature aggregation or explicit modality contribution assessment. With the advancement of RGB-D SOD, research efforts in this domain can broadly be categorized into the early fusion methods [35–37] and the cross-level middle-fusion methods [26, 27, 31]. Furthermore, the domain of pure late fusion methods [38, 39] remains underdeveloped due to inherent noise in depth information. Compare to cross-level middle-fusion methods, the early fusion one exhibits deficiencies in feature interaction, often resulting in suboptimal performance. However, early fusion methods are able to better leverage the advantages of ImageNet pretrained backbones to extract color and depth features effectively, thereby compensating for deficiencies in individual grouping cues within color and depth spaces. Importantly, the early fusion methods significantly reduce the computational demands of the overall structure, allowing the development of more lightweight models.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

Moreover, the utilization of depth information as supervised feedback signifies an innovative approach [26, 40]. This technique harnesses the feedback from depth information to stimulate the model's comprehension of the saliency of distinct objects. In doing so, it indirectly tackles the formidable challenge of proficient foreground-background segmentation in pure RGB mode, a challenge exacerbated by low contrast.

B. Early Fusion RGB-D Saliency Detection

The early fusion approach in RGB-D saliency detection begins by merging RGB images with depth maps. Subsequently, these fused bimodal images are fed into an encoding-decoding architecture to accomplish feature extraction and fusion. Zhang et al. [35] introduced a singular streaming architecture that exploits uncertainty for RGB-D saliency detection. Fu et al. [36] adopted a dual-modality early fusion approach wherein RGB images and depth maps are concatenated into a novel dimension, deviating from a simplistic channel-wise concatenation. Chen et al. [41] treated the dual-modal RGB-D task as a 3D vision issue and harnessed 3D convolutions to extract features from the amalgamated RGB and depth images. Meanwhile, Zhao et al. [42] investigated the utilization of pretrained ImageNet models for comprehensive feature extraction from both RGB and depth modalities, incorporating depth maps for intermediate feature supervision. Unlike the aforementioned methods, FasterSal fully uses 3D geometric relations while also thoroughly modeling the texture details of objects. In generating accurate objects, it is able to better outline their detailed textures.

C. Cross-level Middle-fusion RGB-D Saliency Detection

Cross-level middle-fusion approach remains the prevailing methodology in current RGB-D SOD tasks. Piao et al. [43] devised an effective depth refinement block, which employs residual connections to comprehensively extract and fuse multi-level complementary cues from both RGB and depth images. Taking into account the intrinsic disparities between RGB and depth data, Zhang et al. [44] proposed an asymmetric dual-stream architecture for extracting RGB and depth information. Pang et al. [17] integrated and enhanced information from different modalities through an innovative Transformer architecture for cross-modal perspective blending to improve the performance of salient object detection. Jin et al. [45] considered the impact of noisy depth information on the final results and constructed an innovative complementary depth network to fully harness the salient depth features within RGB-D SOD. To further obtain detailed structural features, Vision Transformers [46] have been extensively applied in RGB-D SOD tasks. Tang et al. [47] used high-resolution Transformers to extract richly detailed and globally structured features from both RGB and depth data. Cong et al. [48] combined CNN with Transformers to complement bimodal details and global information.

D. Lightweight RGB-D Saliency Detection

The substantial parameter count, high computational demands, and significant inference latency of heavyweight RGB-D SOD models have constrained their practical applicability in real-world settings, particularly on resource-constrained edge devices. In response to this challenge, several approaches have been dedicated to constructing lightweight models to enable genuine practical applications. Wu et al. [26] leveraged MobileNet V2 as the backbone network and achieved lightweight detection by fusing deep features from both RGB and depth modalities. Their proposed MobileSal, when running on a 2080TI GPU, achieves a remarkable frame rate of 450 FPS. Jin et al. [27] took into account the complexity of dual-stream structures and employed a simpler encoder for depth information extraction, followed by multi-scale decoders for the fusion and decoding of bimodal features. Zhou et al. [25] introduced a spatial enhancement mechanism to improve the pairs so that they can effectively distinguish salient objects from the background when dealing with thermal imaging images with complex backgrounds and variable temperature distributions.

Despite the preliminary developments in lightweight RGB-D models, several challenging issues persist. Firstly, the lack of sufficiently robust lightweight architectures hinders their ability to match the performance of heavyweight models. Secondly, these models still rely on a dual-stream structure, making it difficult to further compress the model's size and achieve faster inference speeds. Lastly, these models treat deep and low-level features in the same manner, disregarding the characteristics of different features, which makes it challenging to extract object textures from low-level features and object structures from deep-level features.

III. Proposed Method

In this section, we present the FasterSal architecture, a tailored solution for RGB-D SOD. FasterSal employs an encoding-decoding architecture with a single stream, designed to deliver superior performance in salient object detection. As depicted in Fig. 3, our method combines RGB and depth modalities to a 4-channel bimodal input to generate multi-level feature representations from the backbone. These features are then refined by the FasterSal middleware, enhancing both deep and low-level characteristics. Subsequently, the refined features undergo multi-level fusion in the attention-based decoder, resulting in accurate salient object segmentation with rich detail and precise localization. Additionally, FasterSal incorporates selective reinforcement mechanisms during training to improve object edge pixel and small object detection. In the following sections, we provide a detailed exposition of each component within our structure.

A. Encoder

To effectively accommodate the 4-channel bimodal input data, we have adapted the MobileNetV3 input layer, ensuring seamless integration of all four channels. Furthermore, leveraging pretrained parameters from ImageNet, we initialize the MobileNetV3 backbone and the extra fourth channel. This initialization strategy harnesses the knowledge embedded in the pretrained model, facilitating the early fusion of RGB and depth features in a high-dimensional space. With input dimensions represented as $\mathbb{R}^{3\times H\times W}$, we extract five hierarchical features denoted as $\{F^i|i=1,2,3,4,5\}$ with corresponding sizes

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

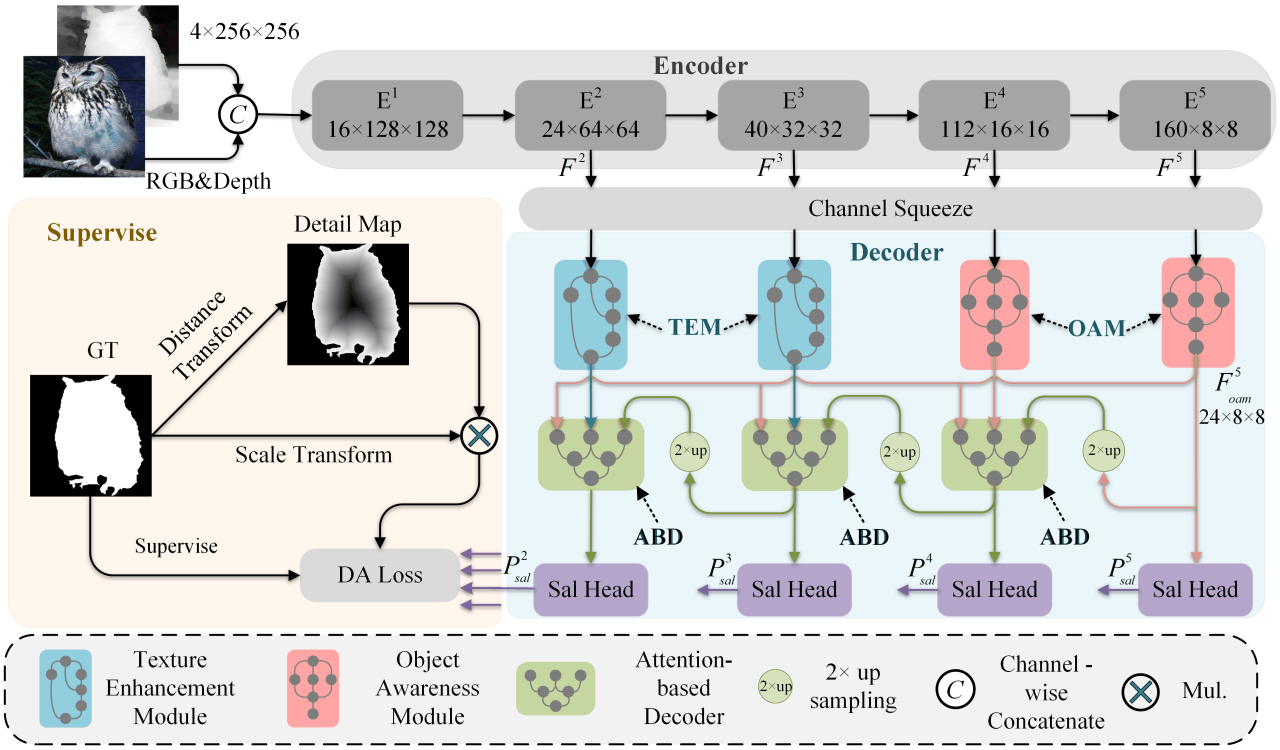

Fig. 3. Overall pipeline of the proposed FasterSal. This diagram illustrates an end-to-end encoder-decoder architecture, with the encoder tasked with feature extraction across four hierarchical levels. Following channel compression, these features are directed into the refinement middleware modules TEM and OAM. Ultimately, they are decoded by the ABD modules to produce the ultimate saliency map. The custom-designed DA loss function is specifically crafted to amplify feedback related to object specifics and small-scale objects.

$[\frac{H}{2^i}, \frac{W}{2^i}]$. Notably, features F^1, F^2, and F^3 are categorized as low-level features due to their larger dimensions, encapsulating a wealth of texture details. Conversely, features F^4 and F^5 are distinguished as deep-level features for their ability to encompass advanced global class information within smaller dimensions. In alignment with established practices in the field [26, 43], feature F^1 is omitted due to its computational overhead and marginal performance contribution. Furthermore, all encoded features are adjusted to 32 channels using 3×3 convolution to ensure uniformity across multi-level features.

B. Textural Enhancement Module

In the context of pixel-wise segmentation tasks, the role of rich textures within low-level features is paramount. These textures are not only instrumental in assisting the model in foreground-background discrimination during decoding but also in the precise recovery of intricate object details. Nonetheless, the presence of complex backgrounds and the interference of low-frequency noise can disrupt the model's ability to discern object texture details accurately. Taking inspiration from denoising techniques like mean filtering, we introduce the texture enhancement module (TEM) to amplify sharp noise, thereby accentuating textures within low-level features.

Our approach commences with a 7×7 pooling operation that serves to smoothen the features while highlighting their low-frequency components. Subsequently, we perform a pixel-wise subtraction operation between the smoothed features and the original ones. This operation effectively exposes the sharp high-frequency information concealed within complex low-frequency components, bringing out object texture details, as described in Equ. 1:

$$F^i_{hf} = |F^i - Up(Pool(F^i))|, \tag{1}$$

where F^i means the i-th feature extracted from the backbone and $i \in \{2, 3\}$. $Up(\cdot)$ is the upsampling operation. $Pool(\cdot)$ represents the average pooling operation with size 7×7.

To mitigate any potential information loss arising from the subtraction operation, we reintegrate the original features and the high-frequency information using an additive approach:

$$F^i_{fuse} = F^i_{hf} + Conv(Conv(F^i) + F^i), \tag{2}$$

where $Conv(\cdot)$ means a convolution layer with 3×3 filters.

The fused features are then input into the AFF module [49], where spatial and channel-wise weights are thoughtfully applied:

$$F^i_{tem} = AFF(F^i_{fuse}), \tag{3}$$

This culminates in the generation of texture-enhanced features, processed by TEM. The entire process can be seen in Fig. 4.

C. Object Awareness Module

In the SOD task, we focus on enhancing precise object localization and abstract semantic understanding through an efficient Object Awareness Module (OAM). Our approach

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

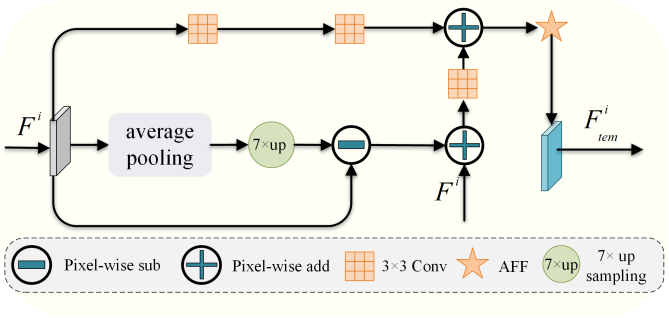

Fig. 4. The processing flow of input features by the texture enhancement module. For input feature F^i, we enhance high-frequency features through 7×7 average pooling and pixel-level subtraction, extract features via convolutions with small receptive fields, and finally enhance low-level feature details through pixel-level addition and AFF attention.

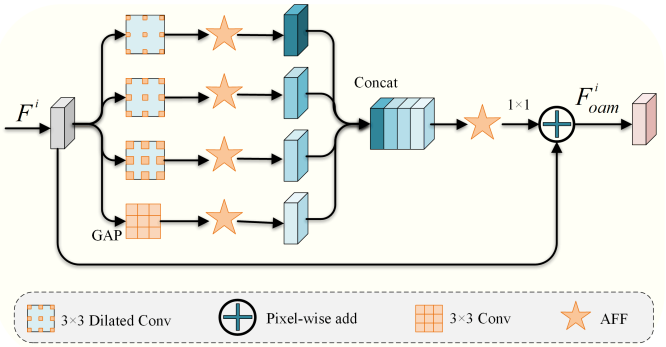

Fig. 5. The processing flow of input features by the object awareness module. For input feature F^i, we obtain target information under different receptive fields using convolutions with various dilation rates and global average pooling, and perform channel-level filtering through the AFF module. Finally, we retain the details of the original features without loss through 1×1 convolution and residual connections.

involves a novel structural design for deep-level feature processing. It combines convolutional layers with varying dilation rates and integrates the AFF module to analyze multi-level receptive fields, extracting scale and rotation-invariant features for improved object-specific information extraction.

The AFF module effectively addresses information loss that can occur with larger dilation rates by ensuring selective feature extraction within extensive receptive fields. We then synthesize information from diverse receptive fields and perform channel-wise concatenation to obtain comprehensive multi-scale features. This step enriches our understanding of images by offering insights from different scales and object perspectives:

$$\begin{aligned} F^i_{ms} = concat(&AFF(D_jConv(F^i)),\\ &AFF(Up(GAP(F^i)))) \qquad for\ j = 2,3,4 \end{aligned} \tag{4}$$

where F^i means the i-th feature extracted from the backbone and $i \in \{4,5\}$. $D_jConv(\cdot)$ means a 3×3 dilation convolution layer with the dilation rate of j. $GAP(\cdot)$ is the global average pooling operation. $Concat(\cdot)$ means the channel-wise concatenation.

Additionally, we use the AFF module to filter and retain the most relevant features. Finally, we employ a residual connection strategy to merge the original input features with the multi-scale features. This fusion ensures that the Object Awareness Module learns rich information and preserves essential details:

$$F^i_{ms} = Conv(AFF(F^i_{ms})), \tag{5}$$

$$F^i_{oam} = F^i + F^i_{ms}. \tag{6}$$

This design empowers our model to fully perceive objects in images and improve the precision of localizing salient objects within their context. The entire process can be seen in Fig. 5.

D. Attention-based Decoder

Our framework employs two key modules, TEM and OAM, to capture fine details in low-level features and deep semantic information in high-level features, enhancing image understanding. TEM focuses on texture and minute details, while OAM extracts deep semantics. To fully and effectively leverage these features at different levels, our decoder, referred to as ABD (as illustrated in Fig. 6), plays a pivotal role in decoding tasks. Let's take the first ABD in the decode branch as an example to describe its working principle in detail. In this ABD, we initially perform convolutional operations between the deepest feature representation (denoted F^5_{oam}) and low-level contextual features (including F^2_{tem} and F^3_{abd}) to obtain higher-level feature representations. Furthermore, we transform the deepest feature representation into a single-channel pseudo-salient map, which is then multiplied and added to the other two contextual features. This clever feature fusion strategy enables us to fully exploit the rich semantic information from deep-level features during decoding while enhancing the texture details of the decoded features:

$$F^5_{sal} = Conv(F^5_{oam}), \tag{7}$$

$$F^2_{mul} = F^5_{sal} \times Conv(F^2_{tem}) + Conv(F^2_{tem}), \tag{8}$$

$$F^3_{mul} = F^5_{sal} \times Conv(F^3_{abd}) + Conv(F^3_{abd}). \tag{9}$$

Next, we concatenate the fused contextual features at the channel-wise and output them through the AFF module. This selectively emphasizes relevant spatial locations and channel dependencies, thereby enhancing the discriminative power of the decoded features. Finally, we subject the features passed through the AFF module to convolutional operations and generate the ultimate comprehensive representation:

$$F^2_{abd} = Conv(AFF(Concat(F^2_{mul}, F^3_{mul}))) \tag{10}$$

This representation captures detailed texture information or deep semantic content, providing robust support for decoding tasks.

E. Supervision

In recent SOD tasks, the precise delineation of object edges plays a pivotal role in achieving better performance.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

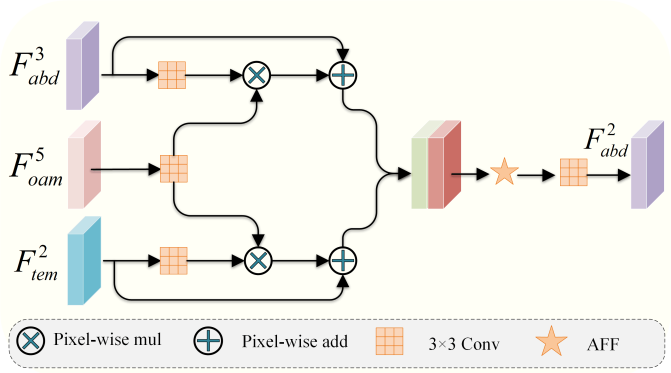

Fig. 6. The processing flow of input features by the attention-based decoder. We take the first ABD in the decode branch as an example, and the input features are deepest feature F^5_{oam}, contextual features F^2_{tem} and F^3_{abd}.

Recognizing this, Wei et al. [50] introduced a novel approach to enhance the model's ability to learn object edges effectively. Their technique utilizes a distance transform function to partition the ground truth into two distinct maps: the detail map and the body map. This division allows the model to focus separately on acquiring knowledge related to object details and major structural components, thereby fortifying the learning of object edge pixels. Specifically, detail maps are generated by performing a distance transform operation that computes the distance from each foreground pixel to the nearest background pixel. Notably, pixels located further from the center of the foreground object are assigned higher weights in these detail maps. This weight assignment prioritizes the learning of fine-grained object characteristics, a crucial aspect of object edge definition. However, one of the challenges in utilizing detail maps with continuous values lies in their integration into classification models effectively. To address this challenge and further optimize model performance, we propose a different approach: employing detail maps with continuous values as weight factors and subsequently incorporating these weights into common loss functions.

Algorithm 1: Distance Transform

input : Binary GT
output: Distance Weight (DW)

$GT_{fg}, GT_{bg} \leftarrow GT$;
$DW = [GT.shape]$;
for p_{xy} *in* GT_{fg} **do**
  $k = 1$;
  while
    $GT[x-k : x+k+1, y-k : y+k+1] == k^2$ **do**
    $k = k + 2$;
  $DW[x, y] = \sqrt{k^2 \times 2}$;
$DW = GT - Norm(DW)$;
return DW

In this work, we introduce the concept of a detail-aware (DA) loss function, designed to refine the learning process further. We begin with the intersection over union (IoU) loss function as our baseline. The first step involves performing a distance transform on ground truth (GT) using Algorithm 1, resulting in the derivation of detail weights (DW).

In Algorithm 1, $p_{(x,y)}$ represents the pixel in foreground of GT. $Norm(\cdot)$ normalises the values in DW to [0-1]. The size of k indicates the distance of the current pixel from the nearest background pixel.

Additionally, to provide enhanced supervision for learning smaller objects, we perform a scale transformation tailored to the size of the foreground region within each input batch of GTs. This process yields scale weights (SW) for individual ground truth samples.

$$SW^n = 1 + \frac{1}{sum(G^n_{xy})/max(sum(G^n_{xy}))}, \quad for\ n = 1, 2, ..., N \tag{11}$$

where n represents the sample in a batch, $sum(\cdot)$ and $max(\cdot)$ are functions that sum and find the maximum value of all batch samples, respectively, and N is the total number of samples in a batch.

Finally, we combine the detail weights with the scale weights to obtain the ultimate weight, denoted as $w_{xy} = SW \times DW$. This weight factor is integrated into the computation of the IoU loss, forming an integral part of the DA loss. The DA Loss formula is expressed as follows:

$$L_{da} = 1 - \frac{\sum^{H,W}_{x,y=1} (G_{xy} * P_{xy}) * (\alpha + \beta w_{xy})}{\sum^{H,W}_{x,y=1} \left(G_{xy} + P_{xy} - G^s_{xy} * P_{xy}\right) * (\alpha + \beta w_{xy})}, \tag{12}$$

where G represents the ground truth. P signifies the predicted saliency map. Hyperparameters α and β balance the contribution of IoU and the weight factor, and the choice of suitable values for these parameters is discussed in Section IV-E1. The weight factor w_{xy} is a product of DW and SW, with SW determined based on the total sum of foreground pixels for each sample within an input batch of data. Smaller foreground surfaces, which correspond to smaller salient objects, result in larger weights. Furthermore, these SWs are normalized to the [1-2] range.

Following [27, 50–52] and most existing methods, we use multi-level supervision to guide the model. Specifically, features $F^i_{abd}, i \in \{2, 3, 4\}$ and F^5_{oam} are fed into the 'Sal Head' consisting of a 3×3 convolution and Sigmoid activation, to obtain saliency maps $P^i_{sal}, i \in \{2, 3, 4, 5\}$ for each layer. The saliency map generated from feature F^2_{abd} serves as the primary output, while the saliency maps obtained from other features assist in the calculation of auxiliary losses. Since the value of the auxiliary loss is larger than the dominant loss, we assign them a smaller weight. The total loss is defined as:

$$L_{total} = \sum_{i=2}^{5} \frac{1}{2^{i-2}} L_{da}(P^i, G). \tag{13}$$

IV. Experiments

A. Datasets

We conduct experiments on five RGB-D datasets: NJU2K (1985 image pairs) [53], NLPR (1000 image pairs) [1], SIP

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

(929 image pairs) [32], DUT-RGBD (1200 image pairs) [43], and STERE (1000 image pairs) [51]. These datasets encompass a wide range of scenes and challenges, including diverse indoor and outdoor environments, complex lighting conditions, and varied human body postures, facilitating comprehensive evaluations of salient object detection methods.

B. Evaluation Metrics

Following most existing methods [27, 54–59], we use four standard evaluation metrics to comprehensively evaluate the model's performance: mean absolute error ($\mathcal{M}$), E-measure (E_ϕ), F-measure (F_β), and S-measure (S_α). $\mathcal{M}$ evaluates the mean absolute error between the ground truth and the prediction, but it is more paranoid about small objects. S_α evaluates the structural similarity between ground truth and prediction. F_β is the harmonic mean of precision and recall, and we calculate the average F-measure across different thresholds. E_ϕ leverages both image-level statistics and local pixel-level statistics to assess binary saliency maps.

In addition to quantitative metrics on different datasets, we conduct testing of each model against criteria to compare their practical applicability. These criteria include performance metrics such as FPS and inference latency on both CPU and GPU, as well as model's parameters and computational requirements (FLOPs). To ensure a fair comparison, we conduct tests under identical conditions in both the CPU and GPU environments. The CPU used is an Intel(R) Xeon(R) Gold 6240R, and the GPU used is an NVIDIA GeForce RTX 3090. We use the same code to evaluate all models, and the input dimensions matched those specified in the original paper.

C. Implementation Details

In our experimental setup, we harness the computational prowess of four NVIDIA Tesla V100 GPUs, each equipped with 32GB of video memory. To enhance the robustness of our model, we incorporate a variety of data enhancement techniques during the training phase, which are in line with established practices employed by existing models [40, 41]. These augmentation strategies encompass operations such as random cropping, blurring, brightness adjustments, and image flipping. Subsequently, these augmented images are resized to dimensions of 256×256 pixels, making them suitable for input into our model architecture. Throughout the training process, we employ the Adam optimizer to iteratively update the model parameters, configuring the weight decay with a value set at 0.01. Our training regimen spans a total of 400 epochs, with a mini-batch size of 128. Remarkably, each round of training only demands a mere 7 seconds. We initiate training with an initial learning rate of 1.6e-2 and gradually reduce it using a cosine annealing schedule until it converges to 1.6e-4 upon completion of training.

D. Comparison with State-of-the-art Methods

We compare FasterSal with 16 state-of-the-art methods, including the lightweight models MobileSal [26], MoADNet [27], LSNet [25], and the heavyweight models CCAFNet [31], DCMF [13], DQSD [60], CDNet [45], D3Net [32], A2dele [39], DANet [42], S2MA [61], DMRA [43], MMCI [62], TANet [63], CPFP [33], CAVER [17]. To ensure the fairness of the test, we employ an identical set of verification codes, which has been supplied by Jin et al. [27], to calculate the metrics based on the result maps provided by the authors of these methods.

1) Quantitative Comparison: Table I serves as a comprehensive quantitative comparison that sheds light on the prowess of FasterSal when pitted against existing models, encompassing both lightweight and heavyweight contenders, across a spectrum of five diverse datasets. Compared to heavyweight models such as CCAFNet, DCMF and CAVER, FasterSal emerges as a formidable competitor, demonstrating comparable performance on several datasets while notably outperforming them on the NJU2K and NLPR datasets. A standout accomplishment for FasterSal lies in its superior performance in the F_β metric, signifying heightened precision in foreground and background detection. Furthermore, FasterSal achieves commendable results in the M, S_α, and E_ϕ metrics. When juxtaposed with lightweight models like MoADNet, MobileSal and LSNet, FasterSal not only excels in terms of model performance but also manages to maintain an edge in parameters, FLOPs, and resource efficiency. For the sake of convenient comparison of the overall performance of the model on the dataset, we rank the model based on four evaluation metrics. From the ranking, it can be observed that, compared to the existing heavyweight model CAVER, FasterSal still maintains a very high comparability, for example, on the NJU2K and DUT-RGBD datasets.

Additionally, we present a comprehensive evaluation of existing RGB-D SOD models, offering insights into the trade-off between model parameters and performance. In Fig. 7, we provide visualizations depicting the relationship between model parameters and key performance metrics, specifically M, F_β, S_α and E_ϕ. These metrics are computed as averages across the five datasets employed in our study. An intriguing observation emerges from these visualizations: heavyweight models, characterized by a high number of parameters, tend to excel in terms of performance metrics but often sacrifice practical applicability due to their computational demands. Conversely, lightweight models, designed for faster inference speed, frequently compromise accuracy. However, our proposed FasterSal takes a nuanced approach by striking a harmonious balance between these two extremes. It manages to achieve both high-speed inference and exceptional precision in the realm of RGB-D SOD.

2) Efficiency Comparison: A comparison of these models in terms of efficiency is given in Table II to further demonstrate the significant advantages of FasterSal in different hardware environments. We conduct rigorous tests on these models using an Intel(R) Xeon(R) Gold 6240R CPU and an NVIDIA GeForce RTX 3090 GPU separately, while adhering to the recommended input dimensions outlined in the paper. It is evident that the proposed FasterSal achieves the highest FPS (64 and 1900) and the lowest latency in CPU and GPU settings. Compared to the runner-up MoADNet, FasterSal demonstrates a remarkable performance improvement of 106.3% on GPU

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

TABLE I

PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD FASTERSAL AND THE STATE-OF-THE-ART LIGHTWEIGHT AND HEAVYWEIGHT METHODS. ↓ (↑) MEANS THAT THE HIGHER (LOWER) IS BETTER. 'RANK' INDICATES THE AVERAGE RANKING ACROSS FOUR EVALUATION METRICS, WITH OUR METHOD HIGHLIGHTED IN **BOLD**.

Method Pub &Year		CPFP CVPR19	TANet TIP19	MMCI PR19	DMRA ICCV19	S2MA CVPR20	DANet ECCV20	A2dele CVPR20	D3Net TNNLS21	CDNet TIP21	DQSD TIP21	DCMF TIP22	CCAFNet TMM22	MoADNet TCSVT22	MobileSal TPAMI22	LSNet TIP23	CAVER TIP23	Ours
Params		69.5M	232.4M	241.7M	20.3M	86.6M	26.6M	30.1M	43.2M	32.9M	396.8M	51.0M	41.8M	5.0M	6.5M	5.39M	55.79M	**3.4M**
FLOPs		101.1G	372.9G	412.0G	25.6G	141.0G	66.1G	41.7G	55.1G	72.0G	812.6G	102.3G	76.6G	1.3G	1.58G	1.21G	21.86G	**0.9G**
Size		256.3M	885.4M	922.1M	77.5M	330.6M	101.8M	115.8M	164.6M	125.6M	1515.7M	224.8M	159.5M	19.3M	39.4M	32.7M	213.30M	**13.5M**
NJU2K	$M\downarrow$	0.053	0.061	0.079	0.051	0.054	0.047	0.051	0.047	0.048	0.050	0.043	0.037	0.041	0.045	0.038	0.032	**0.034**
	$F_\beta\uparrow$	0.837	0.844	0.813	0.872	0.838	0.859	0.874	0.863	0.866	0.860	0.847	0.897	0.892	0.848	0.899	0.874	**0.906**
	$S_\alpha\uparrow$	0.878	0.878	0.859	0.886	0.894	0.897	0.869	0.900	0.885	0.899	0.910	0.910	0.906	0.896	0.911	0.920	**0.908**
	$E_\phi\uparrow$	0.900	0.893	0.882	0.908	0.899	0.916	0.916	0.914	0.908	0.913	0.907	0.942	0.935	0.909	0.940	0.922	**0.949**
	Rank↓	15	16	17	12	14	7	110	6	12	9	8	2	5	10	2	4	**1**
NLPR	$M\downarrow$	0.038	0.041	0.059	0.031	0.030	0.029	0.028	0.030	0.032	0.029	0.029	0.026	0.027	0.025	0.024	0.022	**0.022**
	$F_\beta\uparrow$	0.818	0.795	0.729	0.855	0.848	0.870	0.878	0.858	0.848	0.841	0.849	0.881	0.874	0.874	0.883	0.895	**0.902**
	$S_\alpha\uparrow$	0.884	0.886	0.855	0.899	0.915	0.915	0.896	0.911	0.902	0.916	0.922	0.922	0.915	0.919	0.918	0.929	**0.920**
	$E_\phi\uparrow$	0.920	0.916	0.871	0.942	0.940	0.949	0.945	0.944	0.935	0.934	0.938	0.953	0.947	0.953	0.956	0.959	**0.960**
	Rank↓	15	16	17	13	11	7	8	10	14	12	8	4	6	5	3	1	**2**
SIP	$M\downarrow$	0.064	0.075	0.086	0.088	0.058	0.054	0.070	0.063	0.076	0.065	-	0.054	0.058	0.058	0.049	0.043	**0.049**
	$F_\beta\uparrow$	0.819	0.809	0.795	0.815	0.850	0.862	0.827	0.835	0.805	0.843	-	0.864	0.850	0.855	0.883	0.884	**0.870**
	$S_\alpha\uparrow$	0.850	0.835	0.833	0.800	0.872	0.878	0.826	0.860	0.823	0.863	-	0.876	0.865	0.866	0.885	0.893	**0.870**
	$E_\phi\uparrow$	0.899	0.894	0.886	0.858	0.911	0.916	0.887	0.902	0.880	0.900	-	0.916	0.911	0.908	0.927	0.927	**0.929**
	Rank↓	11	13	14	16	6	4	12	9	15	10	-	4	7	7	2	1	**3**
DUT-RGBD	$M\downarrow$	0.099	0.093	0.113	0.048	0.043	0.047	0.042	0.097	0.048	0.072	0.036	0.038	0.031	0.041	-	0.029	**0.030**
	$F_\beta\uparrow$	0.736	0.779	0.753	0.883	0.886	0.877	0.892	0.752	0.874	0.817	0.896	0.904	0.923	0.912	-	0.919	**0.925**
	$S_\alpha\uparrow$	0.749	0.808	0.791	0.888	0.903	0.889	0.884	0.775	0.880	0.845	0.927	0.903	0.927	0.896	-	0.930	**0.918**
	$E_\phi\uparrow$	0.814	0.866	0.855	0.927	0.935	0.925	0.929	0.847	0.918	0.889	0.944	0.944	0.959	0.940	-	0.955	**0.958**
	Rank↓	16	13	14	9	7	9	8	15	11	12	4	5	1	6	-	1	**1**
STERE	$M\downarrow$	0.051	0.060	0.068	0.047	0.051	0.048	0.070	0.046	0.042	0.051	0.043	0.044	0.042	0.042	0.054	0.034	**0.040**
	$F_\beta\uparrow$	0.830	0.835	0.829	0.867	0.831	0.841	0.825	0.856	0.873	0.839	0.839	0.869	0.868	0.851	0.854	0.872	**0.875**
	$S_\alpha\uparrow$	0.879	0.871	0.873	0.886	0.890	0.892	0.826	0.899	0.896	0.892	0.910	0.891	0.898	0.901	0.871	0.914	**0.888**
	$E_\phi\uparrow$	0.912	0.893	0.873	0.920	0.910	0.915	0.892	0.921	0.922	0.912	0.913	0.933	0.935	0.919	0.919	0.931	**0.939**
	Rank↓	14	15	16	9	13	11	17	7	4	11	8	5	3	5	12	1	**1**

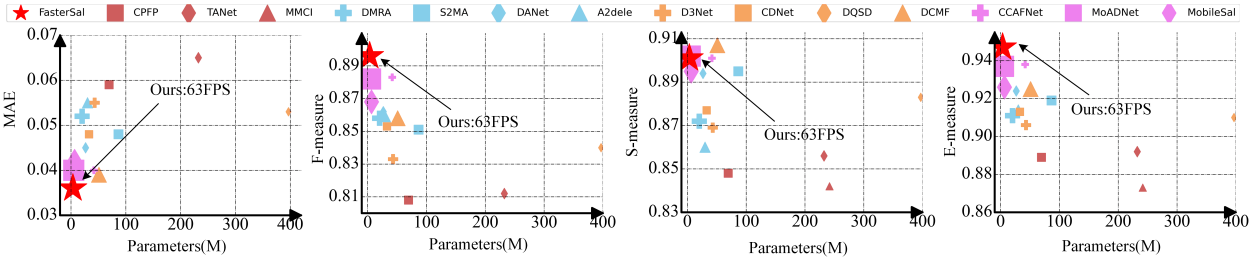

Fig. 7. Illustration of the trade-off between the accuracy and efficiency of existing models and proposed FasterSal. Metrics $\mathcal{M}$, E_ϕ, F_β and S_α are the average score of the five testing datasets.

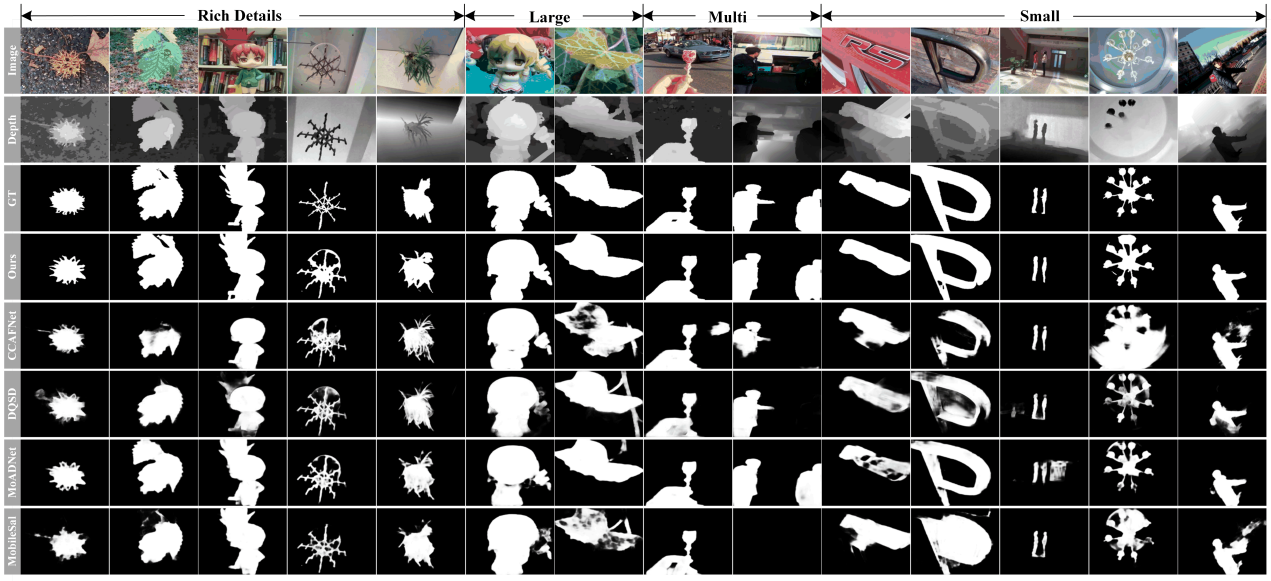

Fig. 8. Visual comparison results between FasterSal and other state-of-the-art models in different difficult scenarios. The third row is GT (ground truth), the fourth row is our result, and rows 5-8 are the comparative results. The first five columns have richer image details, the sixth and seventh columns represent large-size objects, the eighth and ninth columns show cases with multiple objects, and the last five columns are cases with small objects.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

TABLE II
EFFICIENCY COMPARISON BETWEEN FASTERSAL AND EXISTING MODELS UNDER DIFFERENT HARDWARE CONDITIONS. WE USED THE SAME CPU AND GPU ENVIRONMENTS TO TEST THE EFFICIENCY OF THESE MODELS, AND THE INPUT SIZE IS THE ONE RECOMMENDED BY THE ORIGINAL PAPER. OUR APPROACH IS HIGHLIGHTED.

Method	Backbone (type)	Params (M)	FLOPs↓ (G)	FPS(CPU)↑ (images/s)	Latency(CPU)↓ (ms)	FPS(GPU)↑ (images/s)	Latency(GPU)↓ (ms)	Resolution (RGB&Depth)
DMRA$_{ICCV19}$	VGG16	20.3	25.6	18	55.6	458	2.2	256×256
DANet$_{ECCV20}$	VGG16	26.6	66.1	4	242.4	121	8.2	384×384
A2dele$_{CVPR20}$	VGG16	30.1	41.7	8	122.1	238	4.2	256×256
CDNet$_{TIP21}$	VGG16	32.9	72.0	4	205.4	126	7.9	224×224
CCAFNet$_{TMM22}$	VGG16	41.8	76.6	4	218.5	107	9.3	224×224
CAVER$_{TIP23}$	ResNet50	55.8	21.9	21	52.6	511	2.0	256×256
MoADNet$_{TCSVT22}$	MobileNetV3	5.0	1.3	31	31.7	921	1.1	256×256
MobileSal$_{TPAMI22}$	MobileNetV2	6.5	1.6	20	49.7	632	1.6	320×320
LSNet$_{TIP23}$	MobileNetV2	5.4	1.2	28	35.2	899	1.2	224×224
Ours	**MobileNetV3**	**3.4**	**0.9**	**63**	**15.9**	**1900**	**0.5**	**256×256**

and 103.2% on CPU, achieving more than a twofold increase in efficiency while reducing the model size by 32% (3.4M vs. 5M). When compared to heavyweight models, FasterSal exhibits a speed increase of more than threefold. These results unequivocally affirm the outstanding effectiveness and reliability of single-stream RGB-D models in enabling faster real-world applications.

3) Visual Comparison: We present a visual comparison of FasterSal against state-of-the-art methods using representative RGB-D SOD scenes, as illustrated in Fig. 8. These scenes encompass a diverse array of scenarios, ranging from scenes featuring salient objects replete with intricate details (columns 1^{st} to 5^{th}) to those with larger-sized salient objects (columns 6^{th} and 7^{th}). We also include scenarios with multiple salient objects (columns 8^{th} and 9^{th}), as well as those with smaller-sized salient objects (columns 10^{th} to 14^{th}). Upon careful examination of these results, it becomes evident that FasterSal excels in several critical aspects. In particular, it demonstrates a remarkable ability to preserve semantic accuracy, ensuring that the salient objects are identified. Moreover, FasterSal exhibits a commendable proficiency in preserving the integrity of fine edges, a particularly challenging task when dealing with small objects and complex scenes.

E. Ablation Experiments

To demonstrate the effectiveness of the proposed single-stream structure FasterSal, proposed modules, especially the middlewares as well as the weighted loss function, we design the following experiments, and all experiments are based on NJU2K and SIP datasets in terms of the mean absolute error ($\mathcal{M}$), E-measure (E_ϕ) , F-measure (F_β) , and S-measure (S_α)

1) Effectiveness of the DA Loss: To comprehensively assess the impact of hyperparameters α and β on our model's performance and to gauge the effectiveness of the proposed DA Loss function, we conducted a series of ablation experiments. By keeping other parameter values constant, we incrementally varied the settings of α and β to explore their influence on model performance. Notably, parameter α plays a crucial role in emphasizing fine details and small objects during the model's learning process. Therefore, we validate on the boundary maps, which are extracted from the predicted maps using the method provided by CAVER [17]. Table III presents our experimental findings, highlighting that when both α and β are set to 1, the model's performance peaks. This achievement underscores the efficacy of our proposed weighting factors, as they significantly enhance the model's learning capability compared to the baseline scenario outlined in the first row, which solely uses IoU loss. However, overemphasizing fine details by increasing α may decrease the model's ability to capture the overall structure, potentially degrading performance. Therefore, our meticulous experimentation leads us to conclude that setting both α and β to 1 represents the optimal parameterization.

TABLE III
THE PERFORMANCE OF THE MODEL UNDER DIFFERENT HYPERPARAMETER COMBINATIONS IN DA LOSS.

Loss	NJU2K				SIP			
	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$
α=1,β=0	0.026	0.631	0.775	0.947	0.033	0.610	0.759	0.928
α=1,β=1	**0.026**	**0.643**	**0.790**	**0.952**	**0.031**	**0.622**	**0.763**	**0.939**
α=2,β=1	0.027	0.629	0.770	0.941	0.034	0.610	0.755	0.923
α=3,β=1	0.027	0.612	0.768	0.935	0.035	0.608	0.753	0.920
α=4,β=1	0.030	0.610	0.761	0.933	0.037	0.600	0.740	0.915

2) Effectiveness of Single-Stream Structure: The core foundation of FasterSal lies in its single-stream architecture. To thoroughly assess the performance differences between the single-stream and dual-stream versions of FasterSal, results are listed in Table IV. In the dual-stream FasterSal (rows 2^{nd} and 3^{rd} of the table), RGB and depth features are independently extracted using MobileNet V3, and then two different methods of multi-level feature fusion are explored: channel-wise concatenation (row 2^{nd}) and pixel-wise addition (row 3^{rd}). Additionally, the performance of using only the RGB modality with our single-stream structure is measured (row 1^{st}), to validate the efficacy of early fusion of the two modal images. A key observation from the table is row 2^{nd}, where the channel concatenation fusion method, due to the destruction of depth information by RGB pretraining, introduces more noise information, leading to unimproved performance. In contrast, row 4^{th} (our result) clearly demonstrates the significant gap between the two approaches.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

TABLE IV
RESULTS USING FASTERSAL WITH DIFFERENT STRUCTURES. THE BEST RESULT VALUES ARE MARKED IN **BOLD**.

Architecture	Params	FLOPs	NJU2K				SIP				NLPR				DUT-RGBD				STERE			
			$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$
RGB Input	3.4	0.9	0.042	0.884	0.889	0.934	0.058	0.853	0.851	0.913	0.024	0.894	0.912	0.956	0.033	0.920	0.913	0.955	0.046	0.861	0.873	0.924
Dual-Stream(concat)	6.2	2.1	0.038	0.891	0.899	0.942	0.049	0.879	0.874	0.924	0.025	0.895	0.916	0.958	0.037	0.905	0.907	0.950	0.042	0.870	0.885	0.936
Dual-Stream(add)	6.0	1.8	0.034	0.899	0.906	0.947	0.049	**0.881**	**0.874**	0.928	0.022	0.894	0.917	0.960	0.033	0.915	0.913	0.955	0.042	0.870	0.884	0.939
Single-Stream (FasterSal)	**3.4**	**0.9**	**0.034**	**0.906**	**0.908**	**0.949**	**0.049**	0.870	0.870	**0.929**	**0.022**	**0.902**	**0.920**	**0.960**	**0.030**	**0.925**	**0.918**	**0.958**	**0.040**	**0.875**	**0.888**	**0.939**
Single-Stream (MoADNet)	*3.7*	*1.0*	*0.040*	*0.888*	*0.909*	*0.947*	*0.053*	*0.868*	*0.871*	*0.918*	*0.024*	*0.895*	*0.913*	*0.955*	*0.030*	*0.923*	*0.911*	*0.949*	*0.041*	*0.871*	*0.883*	*0.930*

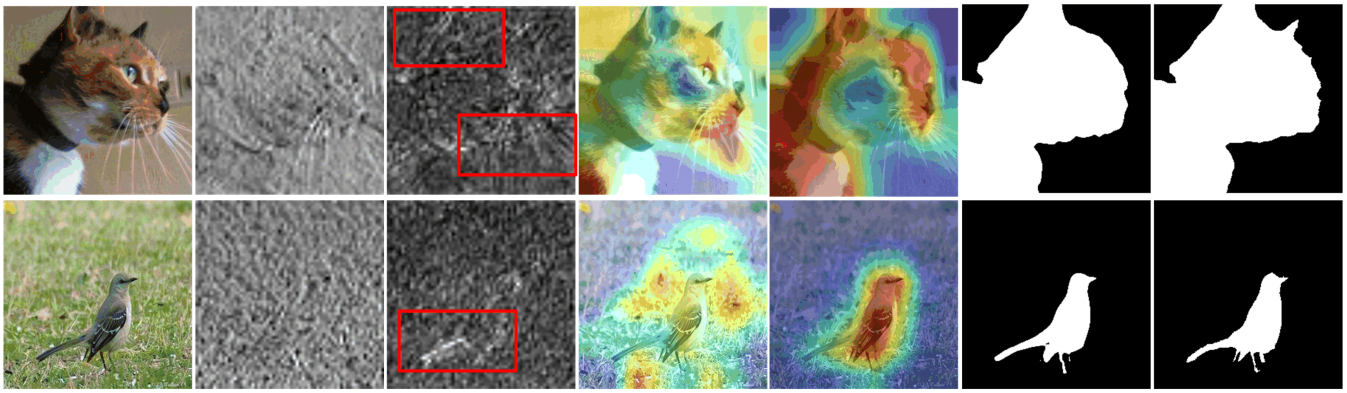

Fig. 9. Performance of the proposed module.

The last row of the Table IV shows the effects of applying our method on MoADNet. As can be seen, our method still outperforms the original model on five datasets, which further demonstrates the superiority of our approach.

3) Impact of Pretrained Backbone on FasterSal: Within the FasterSal framework, we expand the input of MobileNet V3 to include four channels and use pretrained parameters from ImageNet to initialize the fourth channel. To ensure the effectiveness of this modification within the single-stream architecture, we explore various initialization methods for the newly introduced fourth channel. These methods include zero initialization, one initialization, and Kaiming random initialization. The experimental results listed in Table V illustrate the impact of different initialization strategies. The findings suggest that weights pretrained on ImageNet are more advantageous for our FasterSal framework. Specifically, compared to the other three initialization methods, using ImageNet pretrained initialization leads to a 0.97% reduction in M and achieves improvements of 2.15% in F_β, 1.12% in S_α, and 1.45% in E_ϕ across two datasets.

TABLE V
MODEL PERFORMANCE WITH DIFFERENT INITIALIZATIONS ON THE EXPANDED FOURTH CHANNEL. THE BEST RESULT VALUES ARE MARKED IN **BOLD**.

Initialization	NJU2K				SIP			
	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$
One Init	0.044	0.882	0.894	0.932	0.066	0.835	0.863	0.918
Zero Init	0.038	0.896	0.903	0.941	0.052	0.858	0.864	0.922
Random Init	0.040	0.888	0.898	0.938	0.067	0.839	0.845	0.896
ImageNet Pretrained	**0.034**	**0.906**	**0.908**	**0.949**	**0.049**	**0.870**	**0.870**	**0.929**

4) Effectiveness of Modules: Within the FasterSal framework, we introduce three innovative modules: TEM, OAM, and ABD, to assess their contributions through a series of ablation experiments. For these experiments, we use MobileNetV3 as the backbone architecture, building upon a basic UNet-based structure, referred to as "B." Our exploration begins by replacing the UNet decoder with ABD and then systematically introducing OAM and TEM to discern the impact of each module on the overall model performance. Additionally, we incorporate ASPP [64] and PPM [65] into the model for an ablation analysis of OAM. The comprehensive experimental results listed in Table VI demonstrate that each module significantly contributes to the overall performance. With the addition of ABD and OAM, the model's Mean Absolute Error (MAE) on the NJU2K dataset improved from 0.041 to 0.037 and then to 0.035, respectively. Notably, compared to ASPP and PPM, OAM achieved a more substantial performance improvement, reducing the average error (M) by 3% and increasing the precision by more than 5% in both datasets. This indicates that the proposed OAM outperforms other feature enhancement modules.

In Fig. 9, we provide visualizations in the form of class activation maps, offering insights into the interplay between encoded features and TEM and OAM. These visualizations serve to highlight the pivotal role of these intermediary modules in our study. From these visual outcomes, we can clearly observe the efficacy of TEM in attenuating low-frequency information within low-level features, thereby enhancing the representation of object textures and complex details. Simultaneously, OAM effectively discerns the position and edges of prominent objects by applying different exploration strategies

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

TABLE VI
THE EFFICACY OF THE PROPOSED MODULES. THE BEST RESULT VALUES ARE MARKED IN **BOLD**.

Module	Params	FLOPs	NJU2K				SIP			
			$M \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$
B	2.99	0.5	0.041	0.882	0.895	0.936	0.058	0.847	0.857	0.916
B+ABD	3.09	0.63	0.037	0.896	0.900	0.943	0.054	0.860	0.860	0.919
B+ADB+OAM	3.26	0.66	0.035	0.903	0.907	0.945	0.049	0.864	0.864	0.927
B+ADB+ASPP	3.28	0.69	0.036	0.893	0.902	0.937	0.050	**0.873**	0.868	0.921
B+ADB+PPM	3.25	0.63	0.039	0.889	0.900	0.931	0.052	0.870	0.869	0.920
FasterSal	**3.36**	**0.88**	**0.034**	**0.906**	**0.908**	**0.949**	**0.049**	**0.870**	**0.870**	**0.929**

across various spatial contexts. In summary, the introduction of TEM and OAM significantly enhances the model's capability in detecting object textures and precisely locating objects.

V. CONCLUSION

In this study, we present FasterSal, a cutting-edge single-stream architecture designed for RGB-D salient object detection. This innovative approach effectively tackles the challenges of modal inconsistency and excessive parameter counts common in dual-stream systems. Employing a single stream encoder capable of handling both RGB and depth image inputs, FasterSal skillfully uses pretrained RGB encoders while integrating the intricate 3D geometric details from depth maps. The incorporation of a texture enhancement module and a detail-aware loss significantly refines the model's ability to discern edges, emphasizing high-frequency details. Extensive tests across diverse datasets validate FasterSal's exceptional balance between performance and computational efficiency. With a modest parameter count of just 3.4 million, FasterSal stands out for its remarkable speed and precision, positioning it as a highly effective solution for real-world applications.

REFERENCES

[1] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pp. 92–109, Springer, 2014.

[2] X. Xu, S. Wang, Z. Wang, X. Zhang, and R. Hu, "Exploring image enhancement for salient object detection in low light images," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 17, no. 1s, pp. 1–19, 2021.

[3] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[4] B. Ding, R. Zhang, L. Xu, G. Liu, S. Yang, Y. Liu, and Q. Zhang, "U 2 d 2 net: Unsupervised unified image dehazing and denoising network for single hazy image enhancement," *IEEE Transactions on Multimedia*, 2023.

[5] Z. Pan, F. Yuan, J. Lei, W. Li, N. Ling, and S. Kwong, "Miegan: Mobile image enhancement via a multi-module cascade neural network," *IEEE Transactions on Multimedia*, vol. 24, pp. 519–533, 2021.

[6] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3454–3463, 2019.

[7] X.-F. Zhu, X.-J. Wu, T. Xu, Z.-H. Feng, and J. Kittler, "Robust visual object tracking via adaptive attribute-aware discriminative correlation filters," *IEEE transactions on multimedia*, vol. 24, pp. 301–312, 2021.

[8] R. Zhang, J. Tan, Z. Cao, L. Xu, Y. Liu, L. Si, and F. Sun, "Part-aware correlation networks for few-shot learning," *IEEE Transactions on Multimedia*, 2024.

[9] R. Zhang, Z. Cao, S. Yang, L. Si, H. Sun, L. Xu, and F. Sun, "Cognition-driven structural prior for instance-dependent label transition matrix estimation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[10] X. Cheng, X. Zheng, J. Pei, H. Tang, Z. Lyu, and C. Chen, "Depth-induced gap-reducing network for rgb-d salient object detection: an interaction, guidance and refinement approach," *IEEE Transactions on Multimedia*, 2022.

[11] S. Yao, M. Zhang, Y. Piao, C. Qiu, and H. Lu, "Depth injection framework for rgbd salient object detection," *IEEE Transactions on Image Processing*, 2023.

[12] X. Wang, L. Zhu, S. Tang, H. Fu, P. Li, F. Wu, Y. Yang, and Y. Zhuang, "Boosting rgb-d saliency detection by leveraging unlabeled rgb images," *IEEE Transactions on Image Processing*, vol. 31, pp. 1107–1119, 2022.

[13] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for rgb-d saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1285–1297, 2022.

[14] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 235–252, Springer, 2020.

[15] K. Song, H. Wang, Y. Zhao, L. Huang, H. Dong, and Y. Yan, "Lightweight multi-level feature difference fusion network for rgb-dt salient object detection," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101702, 2023.

[16] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *IEEE*

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

Transactions on Image Processing, vol. 29, pp. 4296–4307, 2020.

[17] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 892–904, 2023.

[18] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "Rgbd salient object detection via disentangled cross-modal fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[19] R. Zhang, L. Li, Q. Zhang, J. Zhang, L. Xu, B. Zhang, and B. Wang, "Differential feature awareness network within antagonistic learning for infrared-visible object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[20] S. Chen and Y. Fu, "Progressively guided alternate refinement network for rgb-d salient object detection," in *European conference on computer vision*, pp. 520–538, Springer, 2020.

[21] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "Rgb-d salient object detection with cross-modality modulation and selection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 225–241, Springer, 2020.

[22] W. Liu, P. Zhang, Y. Lei, X. Huang, J. Yang, and M. Ng, "A generalized framework for edge-preserving and structure-preserving image smoothing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6631–6648, 2021.

[23] P. Zhang, W. Liu, Y. Zeng, Y. Lei, and H. Lu, "Looking for the detail and context devils: High-resolution salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3204–3216, 2021.

[24] W. Liu, P. Zhang, X. Huang, J. Yang, C. Shen, and I. Reid, "Real-time image smoothing via iterative least squares," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 3, pp. 1–24, 2020.

[25] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Transactions on Image Processing*, vol. 32, pp. 1329–1340, 2023.

[26] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10261–10269, 2021.

[27] X. Jin, K. Yi, and J. Xu, "Moadnet: Mobile asymmetric dual-stream networks for real-time and lightweight rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7632–7645, 2022.

[28] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," *arXiv preprint arXiv:2309.09668*, 2023.

[29] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the f-measure for threshold-free salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8849–8857, 2019.

[30] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1407–1417, 2021.

[31] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Ccafnet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in rgb-d images," *IEEE Transactions on Multimedia*, vol. 24, pp. 2192–2204, 2021.

[32] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[33] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3927–3936, 2019.

[34] L. Zhang, Q. Zhang, and R. Zhao, "Progressive dual-attention residual network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5902–5915, 2022.

[35] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8582–8591, 2020.

[36] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3052–3062, 2020.

[37] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, 2022.

[38] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[39] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9060–9069, 2020.

[40] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 1745–1754, 2020.

[41] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1063–1071, 2021.

[42] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pp. 646–662, Springer, 2020.

[43] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7254–7263, 2019.

[44] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 374–390, Springer, 2020.

[45] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[47] B. Tang, Z. Liu, Y. Tan, and Q. He, "Hrtransnet: Hrformer-driven two-modality salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 728–742, 2022.

[48] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, "Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection," *arXiv preprint arXiv:2308.08930*, 2023.

[49] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3560–3569, 2021.

[50] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13025–13034, 2020.

[51] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, IEEE, 2012.

[52] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-irtarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Transactions on Multimedia*, vol. 24, pp. 1735–1749, 2021.

[53] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*, pp. 1115–1119, IEEE, 2014.

[54] Y. Yao, T. Chen, G.-S. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, and J. Zhang, "Non-salient region object mining for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2623–2632, 2021.

[55] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 367–383, 2018.

[56] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[57] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[58] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "Cmsalgan: Rgb-d salient object detection with cross-view generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1343–1353, 2020.

[59] W. Zhou, J. Wu, J. Lei, J.-N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3d images using a deep convolutional residual autoencoder," *IEEE Transactions on Multimedia*, vol. 23, pp. 3388–3399, 2020.

[60] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 2350–2363, 2021.

[61] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13756–13765, 2020.

[62] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[63] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[64] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Jin Zhang received his M.S. degree in Control Science and Engineering from the Shanghai Institute of Technology, China, in 2023. He is currently pursuing his Ph.D. at Beijing Institute of Technology. His research interests include computer vision, image segmentation, and weakly supervised learning. Since 2020, he has made significant contributions to the field of Salient Object Detection.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3521699

Ruiheng Zhang (M'19) received the B.E. degree in 2014 from Beijing Institute of Technology, China. He was a dual-Ph.D. from University of Technology Sydney, Australia and Beijing Institute of Technology, China. He is an Associate Professor in Beijing Institute of Technology. He is the author of more than 40 research papers, including Remote Sensing of Environment, IEEE TMM, IEEE TCSVT, ISPRS, Pattern Recognition, ICLR, IJCAI and so on. He is involved as a member of the Editorial Board of Frontiers in Robotics and AI, Artificial Intelligence and Applications. He has served as TPC of ASIP, SPML, ICDIP, VSIP. His current research interests include deep learning, object understanding, and multi-modal remote sensing.

Lixin Xu is a Professor in Beijing Institute of Technology. He has published 100 journal and conference papers. He received his PhD degree in information engineering from Harbin Institute of Technology. His current research interests include deep learning, MEMS, and target detection. He has served as the Editor and Reviewer for several international journals and conferences. He is the editorial board of Journal of Detection and Control.

Xiankai Lu (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. He is currently a Research Professor with the School of Software, Shandong University. From 2018 to 2020, he was a Research Associate with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, object tracking, video object segmentation, and deep learning.

Yushu Yu received the B.S., M.S., and Ph.D. degrees in mechanical engineering from the Beijing University of Aeronautics and Astronautics (BUAA), Beijing, China, in 2007, 2010, and 2013, respectively. From 2013 to 2014, he was an Engineer with China Aerospace Science and Industry Corporation. He conducted research at the BUAA, Nanyang Technological University, Singapore, and the Chalmers University of Technology, Sweden, from 2014 to 2019. He is currently an Associate Professor with the Beijing Institute of Technology (BIT), China. His research interest includes control and robotics. He received the 2012 IEEE International Conference on Mechatronics and Automation Toshio Fukuda Best Award in Mechatronics, and the Excellent Doctoral Dissertation of BUAA, in 2014.

Min Xu (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, the M.S. degree from the National University of Singapore, Singapore, and the Ph.D. degree from the University of Newcastle, Callaghan, NSW, Australia. She is an Associate Professor with the School of Electrical and Data Engineering (SEDE), Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), and also the Leader of Visual and Aural Intelligence Laboratory with the Global Big Data Technologies Center (GBDTC), UTS. She has published 170+ research papers in prestigious international journals and conference proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (T-PAMI), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON MOBILE COMPUTING (T-MC), PR, ICLR, ICML, CVPR, ICCV, ACM MM, AAAI. Her research interests include multimedia, computer vision, and machine learning. Dr. Xu is an Editorial Board Member for Elsevier Journal of Neurocomputing and served as a program chair/area chair for many major conferences.

He Zhao (Member, IEEE) received the B.E. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 2020 and 2014, respectively. He was a Research Intern with Tencent, Shenzhen, China, in 2019. His research interests include medical image processing, deep learning, and computer vision.

Authorized licensed use limited to: University of Liverpool. Downloaded on March 03,2025 at 16:16:35 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.