Efficient Mirror Detection via Multi-Level Heterogeneous Learning

Ruozhen He, Jiaying Lin*, Rynson W.H. Lau

Department of Computer Science, City University of Hong Kong
ruozhenhe2-c@my.cityu.edu.hk, jiayinlin5-c@my.cityu.edu.hk, rynson.lau@cityu.edu.hk

Abstract

We present HetNet (Multi-level **Het**erogeneous **Net**work), a highly efficient mirror detection network. Current mirror detection methods focus more on performance than efficiency, limiting the real-time applications (such as drones). Their lack of efficiency is aroused by the common design of adopting homogeneous modules at different levels, which ignores the difference between different levels of features. In contrast, HetNet detects potential mirror regions initially through low-level understandings (*e.g.*, intensity contrasts) and then combines with high-level understandings (contextual discontinuity for instance) to finalize the predictions. To perform accurate yet efficient mirror detection, HetNet follows an effective architecture that obtains specific information at different stages to detect mirrors. We further propose a multi-orientation intensity-based contrasted module (MIC) and a reflection semantic logical module (RSL), equipped on HetNet, to predict potential mirror regions by low-level understandings and analyze semantic logic in scenarios by high-level understandings, respectively. Compared to the state-of-the-art method, HetNet runs 664% faster and draws an average performance gain of 8.9% on MAE, 3.1% on IoU, and 2.0% on F-measure on two mirror detection benchmarks. The code is available at https://github.com/Catherine-R-He/HetNet.

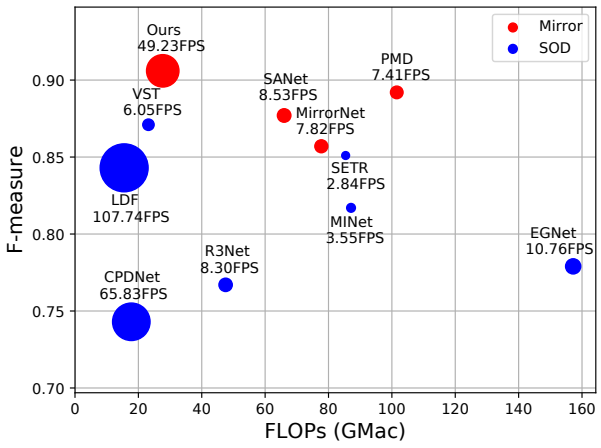

Figure 1: Comparison of our proposed method with Mirror Detection and SOD models on F-measure, FLOPs (GMac), and FPS using the MSD dataset. Our method achieves a new SOTA result with considerable efficiency.

Introduction

Mirrors are common objects in our daily lives. The reflection of mirrors may cause depth prediction errors and confusion about reality and virtuality. Ignoring them in computer vision tasks may cause severe safety issues in situation such as drone and robotic navigation. In addition, owing to the limited computation resources, application scenarios may heavily depend on the model efficiency while efficient mirror detection is essential for real-time computer vision applications.

Recently, Yang *et al.* (Yang et al. 2019) propose MirrorNet based on contextual contrasted features. Lin *et al.* (Lin, Wang, and Lau 2020) propose PMD, which considers content similarity. Guan *et al.* (Guan, Lin, and Lau 2022) propose SANet, which focuses on semantic associations. Though experimental results show their superior performances on mirror detection, these methods suffer from huge computation costs since they apply the same modules for both low-level features with large spatial resolutions and high-level features with small spatial resolutions at every stage. Besides, they heavily rely on post-processing algorithms, *e.g.,* CRF (Krähenbühl and Koltun 2011), which heavily limit the usage of these existing methods to real-world scenarios with the demand for real-time processing, as demonstrated in Figure 1. For more comparisons, we also include some methods from a related task, salient object detection (SOD). Although some of them (*e.g.,* LDF (Wei et al. 2020), CPDNet (Wu, Su, and Huang 2019)) contain few FLOPs with high FPS, they are not able to achieve competitive performances. Thus, it is challenging and significant to propose a method that meets the trade-off between accuracy and efficiency.

Figure 2 compares different network architectures for mirror detection. As shown in Figure 2(a-c), existing mirror detection network architectures use the same module for both low-level and high-level features (Yang et al. 2019; Lin,

*Corresponding authors: Jiaying Lin and Rynson W.H. Lau

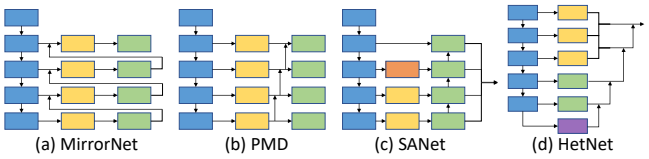

Figure 2: Illustrations of different network architectures used in the existing mirror detection methods. (a) to (d) is the network architecture of MirrorNet (Yang et al. 2019), PMD (Lin, Wang, and Lau 2020), SANet (Guan, Lin, and Lau 2022), and HetNet, respectively. We denote different modules in different colors. Both MirrorNet and PMD share the same high-level design, which uses the same type of modules to learn from the backbone features at different levels, while SANet aggregates all-level features simultaneously. Different from these designs, we split the backbone features as low- and high-level features, and adopt different modules to learn them. For example, in (d), the low-level features are fed into multi-orientation intensity-based contrasted modules (in yellow), and the high-level features are forwarded into reflection semantic logical modules (in green). Such design can fully exploit features at different levels effectively and efficiently.

Wang, and Lau 2020; Guan, Lin, and Lau 2022). However, although low-level features contain the representation of colors, shapes, and textures, learning these features require a higher computational cost due to their larger spatial resolutions compared with high-level features. On the other hand, high-level features involve more semantic information, and it is more difficult for models to directly extract precise mirror boundaries as high-level features contain only rough spatial information. On account of the representation gap between low-level and high-level features, it is inappropriate to adopt the same module for both types of features in model design.

To overcome these limitations, we propose in this paper a highly efficient HetNet (Multi-level **Het**erogeneous **Net**work). We observe that both low-level and high-level understandings assist mirror detection. We also notice that cells on the retina convert the received light signals into electrical signals (Wald 1935), which are then processed and transmitted to the visual cortex for further information processing, integration and abstraction (Hubel and Wiesel 1962), thereby forming vision. HetNet mimics this process in mirror detection. Specifically, human eyes initially accept low-level information (*e.g.*, intensity contrast, colors) from the scene, and after transmission and abstraction (*e.g.*, objects' edges and shapes) in our brains, we confirm the reflection semantic high-level understandings (*e.g.*, content similarity, contextual contrast) of objects to finally determine mirrors. Considering these observations, we propose our new model HetNet to take advantage of the characteristics of low-level and high-level features individually. HetNet includes multi-orientation intensity-based contrasted (MIC) modules to learn low-level features at shallow stages to help localize mirrors, and reflection semantic logical (RSL) modules to help extract high-level understandings at deep stages and then aggregate them with low-level understandings to output the final mirror mask. In addition, we fully use the backbone network for better learning of low-level features without a huge computational cost. With the benefit of the disentangled learning on low-level and high-level features, our model exploits low-level and high-level features effectively and efficiently. Experimental results show that with the proper heterogeneous design for features of different levels, our model outperforms the state-of-the-art mirror detection method PMD (Lin, Wang, and Lau 2020) with 72.73% fewer model FLOPs and 664% faster.

Our main contributions are summarized as follows:

- We propose the first highly efficient mirror detection model HetNet which learns specific understandings via heterogeneous modules at different levels. Unlike existing mirror detection methods that adopt the same modules in all stages, our heterogeneous network can reduce the computational cost via a proper design for different levels of features.
- We propose a novel model that consists of multi-orientation intensity-based contrasted (MIC) modules for initial localization by low-level features in multi-orientation, and reflection semantic logical (RSL) modules for semantic analysis through high-level features.
- Experiments demonstrate that HetNet outperforms all relevant baselines, achieving outstanding efficiency (664% faster) and accuracy (an average enhancement of 8.9% on MAE, 3.1% on IoU, and 2.0% on F-measure on two benchmarks) compared with the SOTA method PMD.

Related Works

Mirror Detection. The initial work of automatic mirror detection was proposed by Yang *et al.* (Yang et al. 2019). They localize mirrors through multi-scale contextual contrasting features. Thus, this method has limitations if the mirror and non-mirror contents are similar. To solve the problem, Lin *et al.* (Lin, Wang, and Lau 2020) propose a method focused on the relationship between features of mirror and non-mirror areas. Recently, two concurrent works (Tan et al. 2022) and (Huang et al. 2023) adopt heavy structures (e.g., transformer) for mirror detection despite low efficiency. However, these methods fail when an area is likely to be a mirror from a context aspect. They are also not efficient for real-time mirror detection. To overcome the limitation, our method preliminary localizes mirror regions based on the intensity contrast. Then it integrates contextual relationships inside and outside mirrors to refine the final predicted regions. Experiments results prove our approach has better performance on both MSD (Yang et al. 2019) and PMD (Lin, Wang, and Lau 2020) datasets.

Salient Object Detection. It detects the most salient objects in images. Early methods are mainly based on low-level features such as color and contrast (Achanta et al. 2009) and spectral residual (Hou and Zhang 2007). Recently, most approaches have depended on deep learning. Qin *et al.* (Qin et al. 2019) propose a densely supervised encoder-decoder with a residual refine module to generate and refine saliency

maps. Chen *et al.* (Chen et al. 2020) propose a global context-aware progressive aggregation network to aggregate multi-level features. Pang *et al.* (Pang et al. 2020) design an aggregate interaction module to extract useful inter-layer features via interactive learning. Ma *et al.* (Ma, Xia, and Li 2021) extract effective features and denoise through an adjacent fusion module. Liu *et al.* (Liu et al. 2021) propose a transformer-based model from a sequence-to-sequence perspective. Nevertheless, the mirror reflects a part of a scene, including both salient and non-salient objects. Hence, salient object detection may not detect mirrors precisely.

Shadow Detection. It identifies or removes shadow areas of images. Nguyen *et al.* (Nguyen et al. 2017) propose a conditional generative adversarial network supporting multi-sensitivity level shadow generation. Hu *et al.* (Hu et al. 2018) propose a direction-aware spatial context module based on spatial RNN to learn spatial contexts. Zhu *et al.* (Zhu et al. 2018) refine context features recurrently through a recurrent attention residual module. Zheng *et al.* (Zheng et al. 2019) design a distraction-aware shadow module to solve indistinguishable area problems. Zhu *et al.* (Zhu et al. 2021) propose a feature decomposition and reweighting to adjust the significance of intensity and other features. Han *et al.* (Han et al. 2022) further incorporate shadow detection in a blind image decomposition setting. The main factor of shadow detection is the distinct intensity contrast between non-shadow and shadow areas. However, there are usually no strong intensity contrasts in the mirror detection scene but many weak ones. Therefore, it is hard to detect mirrors by shadow detection methods.

Methodology

HetNet is based on two observations. We observe that humans are easily attracted to regions with distinctive low-level features (*e.g.*, intensity contrast) first, and then pay attention to high-level information (*e.g.*, content similarity, contextual contrast) to check for object details to detect mirrors. These observations motivate us to learn low-level features at shallow stages and extract high-level features at deep stages with heterogeneous modules. Figure 3 illustrates the pipeline.

Overall Structure

After obtaining the multi-scale image features from the backbone network (Xie et al. 2017), multi-orientation intensity-based contrasted (MIC) modules are used in the first three stages, while reflection semantic logical (RSL) modules and a global extractor (GE) are used in the last three, as shown in Figure 3. The global extractor (GE) extracts multi-scale image features following the pyramid pooling module (Zhao et al. 2017). Outputs after MIC, RSL and GE are denoted as f_i, where i is the stage number starting from 1 to 6. Learning intensity-based low-level understandings, f_1 and f_2, f_2 and f_3 are then fused to f_{21}, f_{22}, respectively. To integrate reflection semantic understandings, we fuse f_6 with f_5 first and then with f_4 to produce f_{23}. Finally, after f_{23} is aggregated with f_{22} to f_{31}, the output feature map is a fusion of f_{31} and f_{21}. The fusion strategy is multiplication and two 3×3 convolution layers with BatchNorm, where the low-level features f_{low} and high-level features f_{high} are fused in a cross aggregation strategy (Zhao et al. 2021; Cai et al. 2020; Chen et al. 2018). First, the interim low-level features are computed from f_{low} multiplies upsampled f_{high}. The interim high-level features are the product of f_{high} and f_{low} processed by a 3×3 convolution layer. The interim low-level and high-level features are fused after applying a 3×3 convolution layer with BatchNorm and ReLU.

The MIC Module

Only learning contextual information is insufficient, especially when contextual information is limited or complex. Thus, utilizing an additional strong cue to facilitate mirror detection is necessary. Gestalt psychology (Koffka 2013) believes that most people see the whole scene first, and then pay attention to individual elements of the scene. In addition, the whole is not equivalent to the sum of the individual elements. Instead, it takes into account the degree of association of these elements (*e.g.*, shape, position, size, color). Based on this, we believe that observing the same scene from different orientations may obtain different information. We use ICFEs to imitate orientation-selective preference visual cortex cells (Hubel and Wiesel 1962) to be proficient in learning features in one orientation. Strengthened contrast information is acquired by combining information learned from two single orientations. Considering low-level contrasts between mirror and non-mirror regions, we first design a MIC module to focus on two-orientation low-level contrasts to localize possible mirror regions. In addition, to reduce computational costs, ICFEs process input features as two parallel 1D features in two directions separately.

A MIC module consists of two Intensity-based Contrasted Feature Extractor (ICFE) modules, as shown in Figure 4. Given the input image features after a 1×1 convolution f_{in}^{low}, we first extract one orientation contrast features f_1^{low} with ICFE directly and then extract contrast features f_2^{low} at another orientation after rotating 90 degrees counterclockwise. To combine two-orientation contrasts into original orientation, we get f_3^{low} by element-wise multiplication of f_1^{low} and f_2^{low} rotated back. Finally, we use a 3×3 convolution and a 1×1 convolution layer to extract the intensity-based contrasted low-level features f_{out}^{low}. Each of the convolution layers is followed by BatchNorm and ReLU.

$$f_1^{low} = \textbf{ICFE}(f_{in}^{low}), \quad f_2^{low} = \textbf{ICFE}(\textit{Rot}(f_{in}^{low}, 1)), \tag{1}$$

$$f_3^{low} = f_1^{low} \odot \textit{Rot}(f_2^{low}, -1), \tag{2}$$

$$f_{out}^{low} = \textbf{BConv}_{1\times1}(\textbf{BConv}_{3\times3}(f_3^{low})), \tag{3}$$

where *Rot(f, d)* denotes rotating f 90° for d times. $\odot$ represents element-wise multiplication. $\textbf{BConv}_{k\times k}(\cdot)$ refers to a k×k convolution with BatchNorm and ReLU activation function.

Inspired by the direction-aware strategy (Hou, Zhou, and Feng 2021), which embeds spatial information by a pair of 1D feature encoders instead of 2D global pooling, the input f_{in}^{low} of ICFE first pools horizontally and vertically.

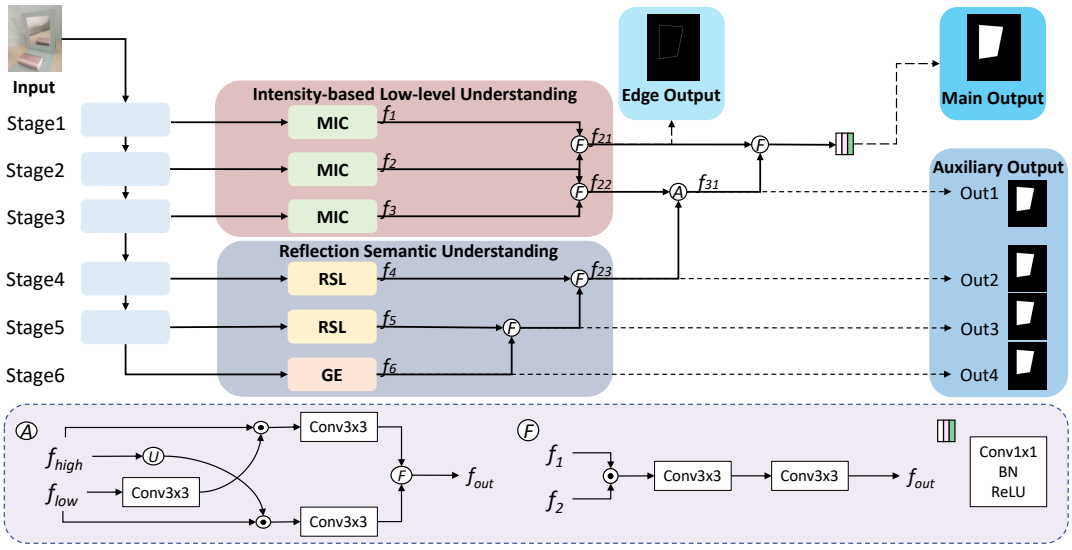

Figure 3: An overview of our proposed method. We first use ResNeXt-101 (Xie et al. 2017) as a backbone and a global extractor (GE) to extract multi-scale image features. We then apply multi-orientation intensity-based contrasted (MIC) modules at the first three stages and reflection semantic logical (RSL) modules at the remaining stages. We use edge and auxiliary output supervisions alongside the main output during the training process.

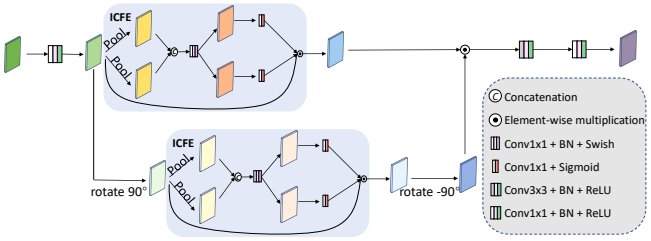

Figure 4: The architecture of the multi-orientation intensity-based contrasted (MIC) module. We first rotate the input features to obtain the other orientation and then use two Intensity-based Contrasted Feature Extractors (ICFEs) to extract low-level features in two orientations. After rotating the features back to the original orientation, we multiply them and take the product into convolution layers to compute low-level contrasts.

The concatenated pooling results go through a 1×1 convolution layer with BatchNorm and Swish function. After processing by two split branches, $f_h^{mid} \in \mathbb{R}^{C\times H\times 1}$ and $f_w^{mid} \in \mathbb{R}^{C\times 1\times W}$ are applied a 1×1 convolution layer and sigmoid function before multiplying together with f_{in}^{low}. Formally, we have:

$$f_1^{mid} = \textbf{SConv}_{1\times1}(\mathcal{P}_h(f_{in}^{low}) \copyright permute(\mathcal{P}_v(f_{in}^{low}))), \quad (4)$$

$$f_h^{mid}, f_w^{mid} = split(f_1^{mid}), \quad (5)$$

$$f_{out}^{mid} = \sigma(\textbf{Conv}_{1\times1}(f_h^{mid})) \odot \sigma(\textbf{Conv}_{1\times1}(f_w^{mid})) \odot f_{in}^{low}, \quad (6)$$

where $P_{h,v}(\cdot)$, $\copyright$, σ denote *h*orizontal or *v*ertical average pooling, concatenation, and Sigmoid, respectively. $\textbf{Conv}_{k\times k}(\cdot)$ represents a $k\times k$ convolution, and $\textbf{SConv}_{k\times k}(\cdot)$ refers to a $\textbf{Conv}_{k\times k}(\cdot)$ with BatchNorm and Swish activation function.

The RSL Module

As there are usually many low-level contrasts in the scene, it could be difficult to detect mirrors simply by low-level understandings. Owing to the reflection, parts outside the mirrors are similar to the contents inside mirrors. Besides, reflected contents may be distinctive to objects around mirrors, making content discontinuity a clue. Based on the above observations, we use a RSL module to learn high-level understandings to assist in finalizing mirror detection combined with the initial localization.

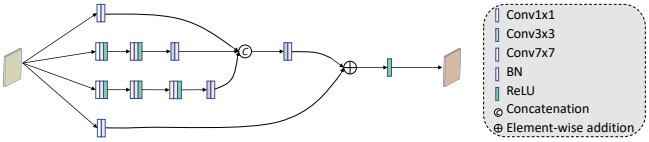

Figure 5: The architecture of the reflection semantic logical (RSL) module.

RSL extracts information in four branches of different numbers of convolution layers with different kernel sizes and dilation rates. Each branch extracts semantic features from a receptive field. Thus, integrating all branches expands a wider receptive field. To obtain the output features f_i^s of branch i from the input high-level features f_{in}^{high} in

our RSL, we have:

$$f_1^s = \mathcal{N}(\mathbf{Conv}_{1\times1}(f_{in}^{high})), \tag{7}$$

$$f_2^s = \mathcal{N}(\mathbf{Conv}_{3\times3}^{p,d=7}(\mathbf{BConv}_{7\times7}^{p=3}(\mathbf{BConv}_{1\times1}(f_{in}^{high})))), \tag{8}$$

$$f_{mid}^s = \mathbf{BConv}_{1\times1}(f_{in}^{high}) \tag{9}$$

$$f_3^s = \mathcal{N}(\mathbf{Conv}_{3\times3}^{p,d=7}(\mathbf{BConv}_{7\times7}^{p=3}(\mathbf{BConv}_{7\times7}^{p=3}(f_{mid}^s))), \tag{10}$$

$$f_4^s = \mathcal{N}(\mathbf{Conv}_{1\times1}(f_{in}^{high})), \tag{11}$$

where *N* denotes BatchNorm, and *R* denotes ReLU. Superscripts *p*, *d* of **Conv**$(\cdot)$ related modules represent padding and dilation rate, respectively. The default padding and dilation rate are set as 1.

We then combine the output features from all branches:

$$f_{mid}^s = \mathcal{N}(\mathbf{Conv}_{3\times3}(f_1^s © f_2^s © f_3^s)), \tag{12}$$

$$f_{out}^s = \mathcal{R}(f_{mid}^s + f_4^s). \tag{13}$$

Under such design, our RSL can acquire rich reflection semantic logical information to determine real mirrors from potential regions predicted by the previous modules.

Loss Function

During the training process, we apply multi-scale supervision for mirror maps and also supervise edge extraction for initial localization. We use the pixel position aware (PPA) loss (Wei, Wang, and Huang 2020) for multi-scale mirror map supervision, and binary cross entropy (BCE) loss for mirror edge supervision. The PPA loss is the sum of weighted BCE (wBCE) loss and weighted IoU (wIoU) loss. wBCE loss concentrates more on hard pixels (*e.g.*, holes) than the BCE loss (De Boer et al. 2005). wIoU measures global structure and pays more attention to important pixels than the IoU loss (Máttyus, Luo, and Urtasun 2017). The final loss function is therefore:

$$Loss = L_{bce} + \sum_{i=0}^{4} \frac{1}{2^i} L_{ppa}^i, \tag{14}$$

where L_{ppa} is the pixel position aware (PPA) loss between the i-th mirror map and the ground truth mirror map, while L_{bce} is the binary cross-entropy (BCE) loss.

Experiments

Datasets and Evaluation Metrics

We conduct experiments on two datasets: MSD (Yang et al. 2019) and PMD (Lin, Wang, and Lau 2020). MSD focuses more on similar indoor scenes, but PMD contains diverse scenes. MSD includes 3,063 images for training and 955 for testing, while PMD has 5,096 images for training and 571 for testing. We train our method on each training set and then test it separately. We adopt three evaluation metrics: Mean Absolute Error (MAE), Intersection over union (IoU), and F-measure to evaluate the performances of the models quantitatively. MAE represents average pixel-wise error between the prediction mask and ground truth. F-measure (F_β) is a trade-off between precision and recall. It is computed as:

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

Method	MSD			PMD		
	MAE↓	IoU↑	F_β ↑	MAE↓	IoU↑	F_β ↑
R^3Net	0.111	0.554	0.767	0.045	0.496	0.713
CPDNet	0.116	0.576	0.743	0.041	0.600	0.734
EGNet	0.096	0.630	0.779	0.088	0.210	0.590
LDF	0.068	0.729	0.843	0.038	0.633	0.783
MINet	0.088	0.664	0.817	0.038	0.608	0.765
SETR	0.071	0.690	0.851	0.035	0.564	0.797
VST	0.054	0.791	0.871	0.036	0.591	0.736
MirrorNet	0.065	0.790	0.857	0.043	0.585	0.741
PMD	0.047	0.815	0.892	0.032	0.660	0.794
SANet	0.054	0.798	0.877	0.032	0.668	0.795
HetNet	0.043	0.828	0.906	0.029	0.690	0.814

Table 1: Quantitative comparison with the state-of-the-art methods on two benchmarks with evaluation metrics MAE, IoU, and F_β.

where β^2 is set to 0.3, for precision is more important (Achanta et al. 2009). Larger F_β is better.

Implementation Details

We implement our model by PyTorch and conduct experiments on a GeForce RTX2080Ti GPU. We use ResNeXt-101 (Xie et al. 2017) pretrained on ImageNet as our backbone network. Input images are resized to multi-scales with random crops and horizontal flips during the training process. We use the stochastic gradient descent (SGD) optimizer with a momentum value of 0.9 and a weight decay of 5e-4. In the training phase, the maximum learning rate is 1e-2, the batch size is 12, and the training epoch is 150. It takes around 5 hours to train. As for the inference process, input images are only resized to 352×352 and then directly predict final maps without any post-processing.

Comparison to the State-of-the-art Methods

To prove the effectiveness and efficiency of our method, we compare it with 10 state-of-the-art methods, including salient object detection methods (R^3Net (Deng et al. 2018), CPDNet (Wu, Su, and Huang 2019), EGNet (Zhao et al. 2019), LDF (Wei et al. 2020), MINet (Pang et al. 2020), SETR (Zheng et al. 2021), VST (Liu et al. 2021)), and mirror detection methods MirrorNet(Yang et al. 2019), PMD(Lin, Wang, and Lau 2020), SANet(Guan, Lin, and Lau 2022). Table 1 illustrates the quantitative comparison on the three metrics. Our method achieves the best performances on all metrics. Besides, as shown in Table 2, we compare Parameters, FLOPs, and FPS with relevant methods. As the hardware environment influences FPS, we conduct all the experiments on the same PC to ensure fairness. Quantitative results show that our method meets the balance between efficiency and accuracy.

We provide some visual comparisons with state-of-the-art methods. As shown in Figure 6, our method can generate more precise segmentation than other counterparts. It performs excellently in various challenging scenarios, such as high-intrinsic similar surroundings (rows 1, 2), mirrors split by objects (row 3), ambiguous regions outside mirrors (row 4), tiny mirrors (rows 5, 6), and partially hidden mirrors

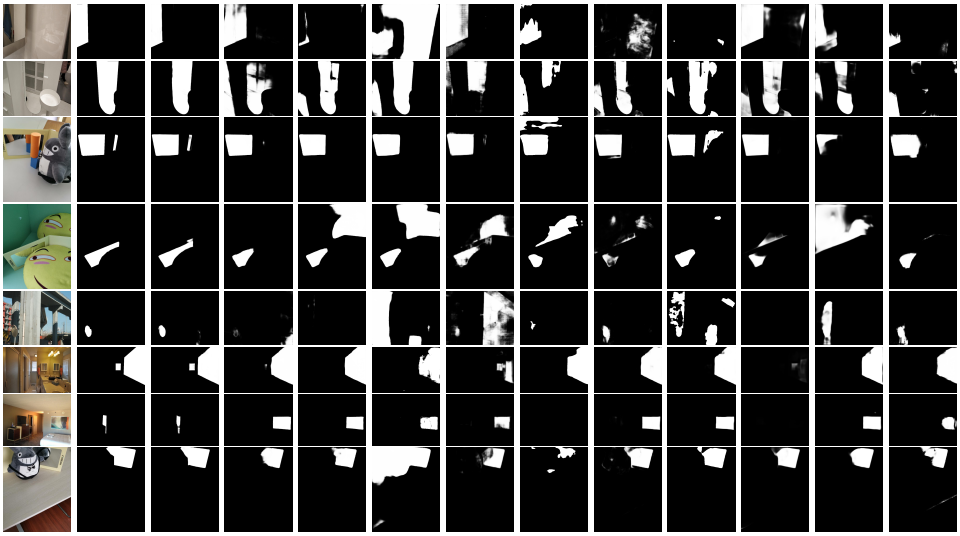

Figure 6: Qualitative comparison of our model with relevant state-of-the-arts in challenging scenarios.

Method	Input Size	Para.	FLOPs	FPS
R^3Net	300×300	56.16	47.53	8.30
CPDNet	352×352	47.85	17.77	65.83
EGNet	256×256	111.64	157.21	10.76
LDF	352×352	25.15	15.51	107.74
MINet	320×320	162.38	87.11	3.55
SETR	480×480	91.76	85.41	2.84
VST	224×224	44.48	23.18	6.05
MirrorNet	384×384	121.77	77.73	7.82
PMD	384×384	147.66	101.54	7.41
SANet	384×384	104.80	66.00	8.53
HetNet	352×352	49.92	27.69	49.23

Table 2: Quantitative comparison on efficiency. We compare our model with relevant state-of-the-art models on Parameters(M), FLOPs(GMAC), and FPS.

(rows 7, 8). Note that we do not use any post-processing to generate our maps. This shows that our method is effective and robust in processing complex images.

Ablation Study

To better analyze the architecture and effectiveness of our network, we conduct ablation studies of each component in our proposed network on the MSD dataset.

Network Architecture Analysis. In this section, we focus on analyzing the network structure to demonstrate the rationality and necessity of learning specific information with the heterogeneous modules at different stages. As shown in Table 3, using MICs at all stages (A_a) performs the worst. Models with RSLs at shallow stages and MICs at deep stages (A_{ba}) or RSLs (A_b) at all stages have similar performances, which are not as effective as our HetNet. Low-level features focus more on colors, shapes, texture and contain more precise spatial information, while high-level features involve more semantics with rough spatial information. MIC aims to learn low-level contrasts, but RSL extracts high-level understandings. Hence, it is more reasonable to learn low-level information to roughly localize mirrors with MICs at shallow (1-3) stages and learn high-level information by RSLs at deep (4-5) stages.

Architecture	MAE↓	IoU↑	F_β ↑	Para.	FLOPs	FPS
A_{ba}	0.046	0.821	0.897	50.13	60.89	37.73
A_a	0.049	0.811	0.889	47.58	26.90	43.34
A_b	0.046	0.817	0.897	52.50	61.67	39.00
HetNet	0.043	0.828	0.906	49.92	27.69	49.23

Table 3: The ablation study results of network architecture. A_{ba}, A_a, A_b denote applying RSL at the shallow stages and MIC at the deep stages, MIC at all stages, and RSL at all stages, respectively.

Component Analysis. To verify the component effectiveness, we conduct ablation experiments by gradually adding them to the network. We simply run the original backbone network (Xie et al. 2017) (I) with the remaining HetNet architecture without the 6th stage as a baseline, and then insert GE (II) into it to complete 6 stages. We gradually add three alternative components to it. The first one adds one ICFE (III) instead of MIC. The second includes the entire MIC component (IV). Both of the above two methods exclude RSLs. The third one applies RSLs without MICs (V).

Table 4 illustrates the experimental results. The ablated model (I) performs the worst of all. We may also observe that adding MICs (IV) or RSLs (V) is generally better than the other alternative (*i.e.*, III). As MICs learn low-level information in two orientations instead of a single orientation, while (IV) performs better than (III). In addition, "basic + GE + MICs" (IV) shows a slight overall advantage over "basic + GE + RSLs" (V), but when MICs and RSLs are

Ablation	Base	GE	ICFEs	MICs	RSLs	MAE↓	IoU↑	F_β ↑	Para.	FLOPs	FPS
I	√					0.056	0.773	0.872	47.53	26.61	60.31
II	√	√				0.050	0.805	0.882	47.53	26.68	58.05
III	√	√	√			0.053	0.781	0.878	47.55	26.68	53.87
IV	√	√		√		0.049	0.811	0.892	47.56	26.68	51.45
V	√	√			√	0.049	0.809	0.890	49.92	27.47	54.20
HetNet	√	√		√	√	0.043	0.828	0.906	49.92	27.69	49.23

Table 4: The ablation study results of components. By adding each component gradually, our model achieves the best performance.

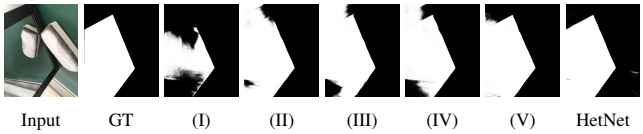

Figure 7: A visual example of the ablation study. (I) to (V) correspond to the prediction maps from five ablated models: "basic", "basic + GE", "basic + GE + ICFEs", "basic + GE + MICs", "basic + GE + RSLs", respectively.

combined (HetNet or "basic + GE + MICs + RSLs"), it outperforms all the other ablated models. It proves that the cooperation of low- and high-level cues is more effective than single-level understandings in mirror detection. Figure 7 shows a visual example where MICs and RSLs play an important role together.

Effectiveness of the Rotation Strategy in the MIC Module. In Table 5, we compare five rotation strategies in MIC and show the effectiveness of our multi-orientation strategy. MIC performing better than 1 ICFE ("ICFE") or 2 parallel ICFEs ("ICFE+ICFE"). This shows that rotation helps learn more comprehensive low-level information in two orientations. To obtain as much distinct information as possible, we adopt an equal division strategy for orientations. If we use lines to denote orientations on a 2D plane, we expect intersecting lines to divide 360° equally, *e.g.*, 1 line into 2*180°, 2 lines into 4*90°, and 3 lines into 6*60°. However, rotating tensors for non-(90*k)° induces information loss. For example, if a tensor rotates 60°, we need to add paddings or crop it. Hence, 90*k° are better options. The failure of ICFE*3 is possibly caused by observing one orientation (0°,180°) twice, which makes information on this orientation stronger than the other (90°). ICFE*4 performs better than ICFE*3 for its balanced observation on two orientations. However, it may introduce more noise so that it is worse than our HetNet (MIC).

Figure 8: Failure cases. Our model may fail in scenarios with complex reflection mirrors, like distorting mirrors.

Strategy	MAE↓	IoU↑	F_β ↑	Para.	FLOPs	FPS
ICFE	0.049	0.802	0.887	49.92	27.68	49.52
ICFE+ICFE	0.048	0.806	0.891	49.92	27.69	45.58
ICFE*3	0.047	0.821	0.895	49.93	27.69	40.73
ICFE*4	0.046	0.820	0.901	49.93	27.70	36.08
MIC	0.043	0.828	0.906	49.92	27.69	49.23

Table 5: The ablation study results of MIC rotation strategies. ICFE, ICFE+ICFE, ICFE*3, ICFE*4, and MIC denote a single ICFE, 2 same-orientation ICFEs, 3 ICFEs with $(0°, 90°, 180°)$ orientations, 4 ICFEs with $(0°, 90°, 180°, 270°)$, and 2 ICFEs with $(0°, 180°)$ orientations, respectively.

Conclusions

In this paper, we propose a highly efficient mirror detection model HetNet. To meet the trade-off of efficiency and accuracy, we adopt heterogeneous modules at different stages to provide the benefits of feature characteristics at both low and high levels. Additionally, considering the low-level contrasts inside and outside mirrors, we propose a multi-orientation intensity-based contrasted (MIC) module to learn low-level understanding in two orientations to select likely mirror regions. To further confirm mirrors, we propose a reflection semantic logical (RSL) module to extract high-level information. Overall, HetNet has outstanding and efficient feature extraction performances, making it effective and robust in challenging scenarios. Experimental results on two benchmarks illustrate that our method outperforms SOTA methods on three evaluation metrics with an average enhancement of 8.9% on MAE, 3.1% on IoU, 2.0% on F-measure, as well as 72.73% fewer model FLOPs and 664% faster than the SOTA mirror detection method PMD (Lin, Wang, and Lau 2020).

Our method does have limitations. Since both MSD and PMD datasets collect mostly regular mirrors, our method may fail in some mirrors with special reflection property occasions. In Figure 8, the three mirrors reflect the same man with three different statuses. In the right image, the mirrors show complex intensity contrast. Hence, the distortion of high-level or low-level information is even more challenging. For future work, we are currently considering additional information to help detect different kinds of mirrors.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, 1597–1604. IEEE.

Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; and Sun, J. 2020. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, 455–472. Springer.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Chen, Z.; Xu, Q.; Cong, R.; and Huang, Q. 2020. Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10599–10606.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1): 19–67.

Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; and Heng, P.-A. 2018. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 684–690. AAAI Press Menlo Park, CA, USA.

Guan, H.; Lin, J.; and Lau, R. W. 2022. Learning Semantic Associations for Mirror Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5941–5950.

Han, J.; Li, W.; Fang, P.; Sun, C.; Hong, J.; Armin, M. A.; Petersson, L.; and Li, H. 2022. Blind Image Decomposition. In *European Conference on Computer Vision (ECCV)*.

Hou, Q.; Zhou, D.; and Feng, J. 2021. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713–13722.

Hou, X.; and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, 1–8. Ieee.

Hu, X.; Zhu, L.; Fu, C.-W.; Qin, J.; and Heng, P.-A. 2018. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7454–7462.

Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R. W.; and Zuo, W. 2023. Symmetry-Aware Transformer-based Mirror Detection. *AAAI*.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106.

Koffka, K. 2013. *Principles of Gestalt psychology*. Routledge.

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3697–3705.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4722–4732.

Ma, M.; Xia, C.; and Li, J. 2021. Pyramidal feature shrinking for salient object detection. In *AAAI*, volume 35, 2311–2318.

Máttyus, G.; Luo, W.; and Urtasun, R. 2017. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE international conference on computer vision*, 3438–3446.

Nguyen, V.; Yago Vicente, T. F.; Zhao, M.; Hoai, M.; and Samaras, D. 2017. Shadow detection with conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4510–4518.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9413–9422.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7479–7489.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2022. Mirror Detection With the Visual Chirality Cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wald, G. 1935. Carotenoids and the visual cycle. *The Journal of general physiology*, 19(2): 351–371.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12321–12328.

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

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3907–3916.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. 2019. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8809–8818.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for

salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8779–8788.

Zhao, Z.; Xia, C.; Xie, C.; and Li, J. 2021. Complementary Trilateral Decoder for Fast and Accurate Salient Object Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4967–4975.

Zheng, Q.; Qiao, X.; Cao, Y.; and Lau, R. W. 2019. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5167–5176.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhu, L.; Deng, Z.; Hu, X.; Fu, C.-W.; Xu, X.; Qin, J.; and Heng, P.-A. 2018. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 121–136.

Zhu, L.; Xu, K.; Ke, Z.; and Lau, R. W. 2021. Mitigating Intensity Bias in Shadow Detection via Feature Decomposition and Reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4702–4711.