# Progressive Mirror Detection

Jiaying Lin Guodong Wang Rynson W.H. Lau†
Department of Computer Science, City University of Hong Kong

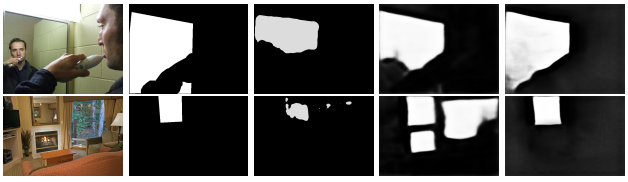
## Abstract

*The mirror detection problem is important as mirrors can affect the performances of many vision tasks. It is a difficult problem since it requires an understanding of global scene semantics. Recently, a method was proposed to detect mirrors by learning multi-level contextual contrasts between inside and outside of mirrors, which helps locate mirror edges implicitly. We observe that the content of a mirror reflects the content of its surrounding, separated by the edge of the mirror. Hence, we propose a model in this paper to progressively learn the content similarity between the inside and outside of the mirror while explicitly detecting the mirror edges. Our work has two main contributions. First, we propose a new relational contextual contrasted local (RCCL) module to extract and compare the mirror features with its corresponding context features, and an edge detection and fusion (EDF) module to learn the features of mirror edges in complex scenes via explicit supervision. Second, we construct a challenging benchmark dataset of 6,461 mirror images. Unlike the existing MSD dataset, which has limited diversity, our dataset covers a variety of scenes and is much larger in scale. Experimental results show that our model outperforms relevant state-of-the-art methods.*

## 1. Introduction

Mirrors are ubiquitous in our daily lives. As they can affect the performances of a lot of vision tasks, such as depth prediction and object detection, they are beginning to receive some discussions [5, 2]. For example, Anderson *et al*. [2] note that mirrors are potential obstacles in vision-and-language navigation (VLN) tasks and existing methods for VLN tend to ignore mirrors. Braun *et al*. [5] find that the error caused by the reflections from mirror-like surfaces is one of the six major types of error in the person detection problem. Zendel *et al*. [30] conduct a safety analysis of existing datasets for computer vision tasks, and find that the existence of mirrors is a hazardous factor to these tasks.

Recently, Yang *et al*. [29] make the first attempt to automatically detect mirrors. They propose a model, called MirrorNet, to segment mirrors from a single image based on extracting multi-level contextual contrasted features. Through extracting the contextual contrast information, their method implicitly learns to detect mirror edges for segmenting the mirrors. However, this method may fail when the contextual contrasts between the inside and outside of a mirror are not obvious. The first row of Figure 1 shows a very common scenario, where a person is looking at a mirror in (a). Since he and his mirrored image have almost identical appearances and overlap with each other (i.e., the arm of the person and his cloth in the reflection), the visual contrast between them is small. If we detect the mirror using MirrorNet, it gets confused by the mixture and fails to separate the two correctly, as shown in (c). In addition, since MirrorNet only considers contextual contrast, it is more prone to overpredict the mirror regions. In the second row of Figure 1, the mirror, window and fireplace all exhibit contextual contrasts with their surroundings. Hence, MirrorNet fails to distinguish them, and detects all of them as mirrors.



(a) Input (b) GT (c) PSPNet (d) MirrorNet (e) Ours

Figure 1: Two popular scenarios where existing methods [32, 29] fail. While PSPNet [32] is a segmentation model, MirrorNet [29] is designed for mirror detection. As MirrorNet is based on extracting contrasted features, it wrongly segments the mirror region in the top image, and wrongly detects the window and fireplace in the bottom image as mirrors. In contrast, our model extracts correspoding context features between inside and outside of the mirror to accurately identify mirror regions.

To tackle the above problem, we propose in this paper a novel approach to detect mirrors. We note that the content of a mirror reflects its surrounding context. This means that there is often a corresponding relationship between objects inside a mirror and those outside of the mirror. In this work, we explore this corresponding relationship in a progressive manner. Figure 2 explains our idea. We can see from (b)

[0†] Rynson W.H. Lau is the corresponding author. He leads this project.

that it is difficult even for human to recognize the mirror simply by its content. However, we can infer the potential mirror region once we can establish a relation between some objects inside the mirror with those of the outside, as in (c). Finally, we can filter and refine our inferred mirror region through explicitly detecting mirror edges, as in (d).

Our method for mirror detection is based on two novel modules. First, we propose a Relational Contextual Contrasted Local (RCCL) module to extract all contextual contrasted and relational features to find out all potential mirror regions. Second, we propose an edge detection and fusion (EDF) module to explicitly detect mirror edges in multi-scale. Using a refinement network, mirror regions are extracted based on the relational contextual contrasted features from RCCL and edge information from EDF.

To train our model, we also propose a dataset of mirror images. Although Yang *et al*. [29] propose the MSD mirror dataset of 4,018 images with ground truth annotations, as shown in Figure 3, a lot of their images are very similar to each other, and therefore have similar contexts. In addition, majority of them are zoom-in images of indoor scenes. This can significantly reduce the robustness of the trained model. Hence, we construct a more challenging benchmark dataset that includes a variety of mirrors and contexts. It is derived from six public image datasets developed for different problems. Our dataset contains a total of 6,461 mirror images and corresponding annotated masks. We have conducted extensive experiments to evaluate our model, in comparison with the state-of-the-art methods, and show that the proposed model outperforms existing methods on both the MSD dataset and our dataset.

Our main contributions can be summarized as:

- We propose a novel progressive method for mirror detection. It is based on two new modules, a *RCCL module* for extracting and comparing mirror features and contextual features for correspondences and a *EDF module* to extract multi-scale mirror edge features.
- We propose a more challenging benchmark dataset, which consists of 6,461 mirror images and corresponding masks from diverse scenes.
- We have conducted extensive experiments to evaluate the performance of our method on both MSD and our own dataset, to demonstrate its effectiveness.

## 2. Related Work

In this section, we briefly summarize recent works that are relevant to the mirror detection problem.

**Mirror Detection.** In [29], Yang *et al*. propose the first model for automatic mirror detection. It focuses on extracting multi-scale contextual contrasted features between regions inside and outside of the mirror, which help implicitly locate the mirror edges. However, the assumption of having contrasted features between the content inside a mirror and that of the outside may fail when the two contents are very similar. To address this limitation, we propose in this work to explicitly consider the relationship between the features inside and outside of the mirrors as well as explicitly detect mirror edges in a multi-scale manner. Results show that our approach can detect mirrors more accurately.

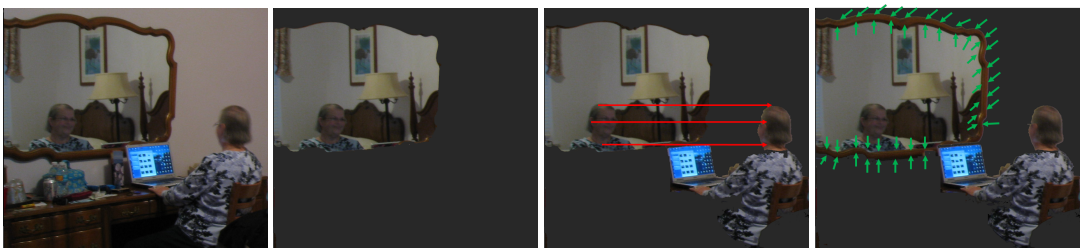
**Salient Object Detection.** This is a popular research problem, and has attracted much attention. Earlier methods are mostly based on low-level features, such as priors [28] and region contrast [21, 9]. Recent methods are mostly CNN-based. Deng *et al*. [11] propose a residual learning method to improve feature refinement. Wu *et al*. [26] adopt a cascaded partial decoder framework to refine the saliency map. Some recent works also address the importance of salient edges in saliency object detection. Qin *et al*. [22] propose a boundary-aware method and a new hybrid loss to learn salient features in pixel, patch and map levels. Unlike salient object detection, which assumes the objects to be detected as salient, mirrors are not always distinctive. Hence, this kind of methods cannot address our problem.

**Semantic Segmentation.** This is a very popular research problem in recent years. It aims to assign pixel-level category labels to an input image. Current state-of-the-arts semantic segmentation approaches, *e.g*., [31, 13, 17], have extensively exploited the popular deep CNNs to extract discriminative features for pixel-level classification. Due to the limited receptive fields of a single convolutional layer, semantic segmentation methods also leverage multi-scale features to encode contextual information across different layers for accurate and dense prediction. For example, PSPNet [35] and DeepLab [8] propose Pyramid Pooling Module (PPM) and ASPP (atrous spatial pyramid pooling), respectively, to extract pyramid contextual representations in an efficient and effective way. Ding *et al*. [12] propose to aggregate context contrasted local features and gated multi-scale features to improve performance. Zhang *et al*. [31] explore the fine-grained representation using co-occurrent features for semantic segmentation.

Semantic segmentation methods rely on object appearances for predictions. However, the appearance of a mirror primarily reflects the appearances of its surrounding objects. Thus, using a segmentation method for mirror detection may end up detecting the objects inside the mirror, instead of the mirror itself. Hence, we focus in this work to develop a more powerful detector specifically for mirrors.

## 3. Our Method

In this paper, we propose a novel progressive mirror detection method. Figure 4 illustrates the pipeline. We first feed the input image to the backbone feature extraction network [27] to extract multi-scale image features. For each



(a) Image (b) Mirror content only (c) Relation (d) Relation edges

Figure 2: Visualization of our progressive approach to recognizing mirrors from a single image. By finding correspondences between objects inside and outside of the mirror and then explicitly locating the miror edges, we can detect the mirror region more reliably.

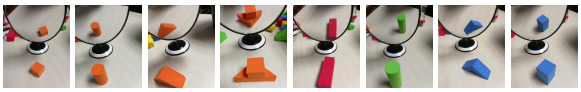


Figure 3: Some example images from the MSD dataset indicating the high similarity of the dataset.

level of image features, we then extract the relational contextual contrasted (RCC) features using the proposed Relational Contextual Contrasted Local (RCCL) module. A decoder is then used to decode the extracted RCC features into a mirror map. In addition, we also propose an edge detection and fusion (EDF) module to explicitly detect mirror edge features, given both low-level image features and high-level RCC features as input, to output a boundary map. Finally, we feed the predicted mirror maps of different scales from the decoders and the boundary map from the EDF module to a refinement module to produce the final output mirror map.

### 3.1. RCCL Module

Our relational contextual contrasted local (RCCL) module is designed to extract relational contextual contrasted features. Unlike the original context contrasted local (CCL) block in [12] and the contextual contrasted feature extraction (CCFE) module in [29], which focus only on contextual contrasted information, our module also tries to consider the relation between the contextual contrast and contextual similarity. The reason for us to take contextual similarity into account is that we notice the content of a mirror can sometimes be similar to the content around the outside of the mirror, *e.g.*, a mirror in front of a white wall while reflecting another white wall opposite to it.

Our RCCL module consists of two blocks: a global relation (GR) block and a contextual contrasted local (CCL) block. Given the input image features $f_{in}$, we first extract global features $f_G$ using the global feature extractor (GFE), local features $f_L$ using the local feature extractor (LFE), and context features $f_C$ using the context feature extractor (CFE). The relational feature extractor (RFE) in the GR block takes the global features $f_G$ as input to extract global relational features $f_{GR}$. Specifically, for each pixel $\mathbf{x}_i$ in $f_G$, the RFE computes the relation score $R$ as:

$$R(\mathbf{x}_i, \mathbf{x}_k) = \theta(\mathbf{x}_i^T \mathbf{x}_k), \tag{1}$$

where $\mathbf{x}_k$ represents the corresponding pixels of $\mathbf{x}_i$, and $\theta$ is a linear transformation. Unlike the non-local method [24], which considers all the other pixels of the image (except $\mathbf{x}_i$) as the corresponding pixels of $\mathbf{x}_i$, we select the corresponding pixels $\mathbf{x}_k$ based on the characteristic of mirror-reflection invariant [14], which points out that the real object and its mirror reflection may have a spatial similarity relation. To fully cover all possible corresponding pixels for pixel $\mathbf{x}_i$ in our searching phase as well as to reduce redundant computations, our method considers all pixels along the eight directions from $\mathbf{x}_i$, i.e., all pixels to the right, to the left, to the top, to the bottom, to the upper left, to the upper right, to the bottom left and to the bottom right, as corresponding pixels. Compared with the original non-local method, which suffers from a huge computational burden in the segmentation process, our RFE has a much smaller set of corresponding pixels to enable efficient mirror detection.

In summary, each pixel $\mathbf{z}_i$ in $f_{GR}$ is computed from pixel $\mathbf{x}_i$ in $f_G$ by:

$$\mathbf{z}_i = S(\sum_{j}^{8} \gamma_j (\sum_{k \in D_j} \theta(\mathbf{x}_i^T \mathbf{x}_k))), \tag{2}$$

where $D_j$ is the set of indices of pixels along direction $j$, and $\mathbf{x}_k$ refers to the corresponding pixels of $\mathbf{x}_i$ along a given direction. $\theta$ is a linear transformation. $\gamma_j$ is a learnable factor. $S$ is a sigmoid function. We enumerate the eight directions around $\mathbf{x}_i$ to obtain its spatial corresponding relation.

For $f_L$ and $f_C$ in the CCL block, we extract the contextual contrasted map by subtracting $f_C$ from $f_L$, so that potential mirror regions can be extracted. We then multiply the subtracted features with $f_{GR}$ to form the final relational context features $f_{RC}$.

The global feature extractor is a 1×1 convolution layer with batch normalization. The local feature extractor is a 3×3 convolution layer with 1 stride, dilation rate of 1 and 1 padding. The context feature extractor is similar to the

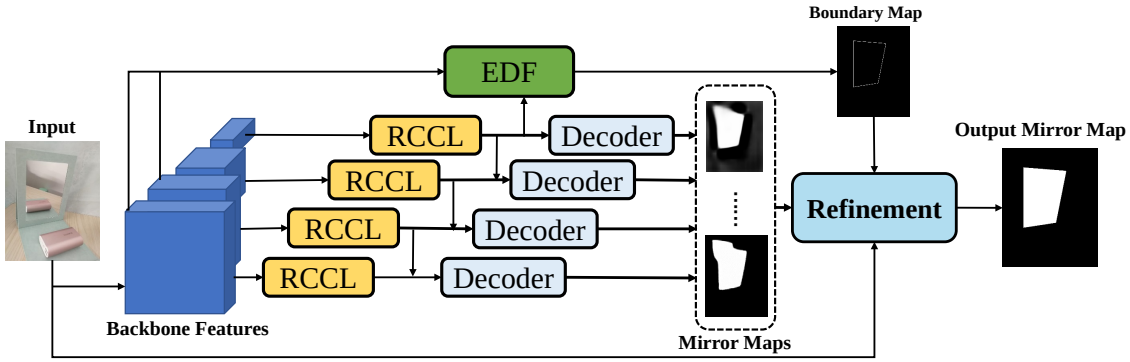



Figure 4: Overview of our network. The backbone network [27] first extracts multi-scale image features, which are used by the RCCL modules to extract relation contextual features. Each decoder then takes the relational contextual features as input and outputs a mirror map. The EDF module extracts mirror edge features to produce a boundary map, given the input low-level image features and high-level relational contextual features. Finally, the refinement module takes all mirror maps and the boundary map to output a final mirror map.

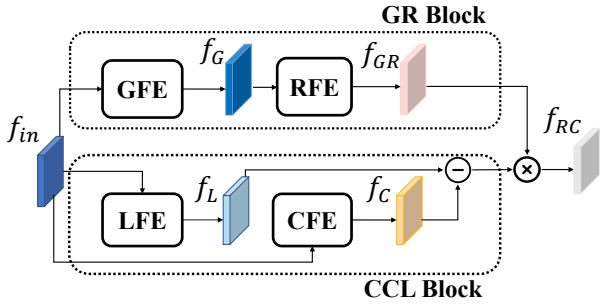



Figure 5: Architecture of the Relational Contextual Contrasted Local (RCCL) module. GFE is a global feature extractor. RFE is a relational feature extractor. LFE is a local feature extractor. CFE is a context feature extractor. All these extractors together help extract relational contextual contrasted features.

local feature extractor but with different dilation rates and padding. In our implementation, we set the dilation rates to 2, 4, 8 and 8 for the highest-level RCCL to the lowest-level RCCL, respectively. The decoders used after the RCCL modules consist of a $1\times1$ convolution layer followed by an upsampling layer to output intermediate mirror maps.

### 3.2. EDF Module

Our edge detection and fusion (EDF) module is designed to extract multi-scale mirror edge features to produce a boundary map. Unlike a recent edge extraction module [34], which only uses low-level features to help a CNN detect edge information, our EDF module considers both low-level and high-level image features in extracting mirror edges. The reason for us to take high-level image features to extract edge features is that a mirror may sometimes have an ambiguous region boundary caused by real objects overlapping in front of their reflections, e.g., the top image in Figure 1, which would require high-level semantics to help detect the boundary of the mirror region more accurately.

Figure 6 shows the architecture of the EDF module. We first take the lowest and the second-lowest levels of backbone features, $f_{in}^{L1}$ and $f_{in}^{L2}$, together with the highest-level relational contextual features, $f_{RC}^{L4}$, as inputs to the EDF module and resize them to the same size as the input image. We then use a low-level edge extractor to extract low-level edge features $E_L$ from $f_{in}^{L1}$ and $f_{in}^{L2}$, and a high-level edge extractor to extract high-level edge features $E_H$ from $f_{RC}^{L4}$. Finally, we use an edge fusion and prediction network to fuse the low-level edge features $E_L$ and the high-level edge features $E_H$ to output a predicted boundary map. The reason for us to extract the low-level and high-level edge features separately and then fuse them together, instead of using a single edge extractor, is that from our experiments, we find that using only a single edge extractor tends to produce sparse edges. By using separate edge extractors, one can focus on the low-level edge features and the other can focus on the high-level edge features. A much finer boundary map can be obtained after fusing the two results.

To supervise our EDF module, we need to have ground truth edges. We use the Canny edge detector [7] to extract the mirror edges from the ground truth masks in our dataset to produce the ground truth edge maps.

The low-level edge extractor consists of three convolution layers, including 256, 128 and 64 filters with a kernel size of $3\times3$ and 1 padding. The high-level edge extractor consists of a convolution layer, including 512 filters with a kernel size of $1\times1$. The fusion layer and the prediction layer are each of a convolution layer with a kernel size of $1\times1$.

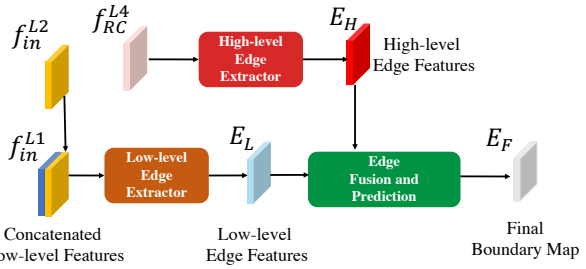



Figure 6: Architecture of the Edge Detection and Fusion (EDF) module. It contains two edge extractors for low-level and high-level feature extraction, and one edge fusion layer together with a prediction layer for fusing the two sets of features to produce the final boundary map.

### 3.3. Refinement Module

To combine the predicted boundary map with the multi-scale mirror maps to form the output mirror map, we add a refinement module to learn to fuse all these maps with reference to the input image. Our refinement module is composed of two convolution layers, with a kernel size of 3×3, and 1 padding with batch normalization. We first concatenate the multi-scale mirror maps together with the input image as input features to the refinement module to obtain refine mirror features. We then feed the refine mirror features to a convolution layer with a kernel size of 1×1 to obtain the final mirror map.

### 3.4. Loss Function

We use the Lovász-Softmax loss [4] to supervise the training of the multi-scale mirror maps. For the EDF module, we use the binary cross-entropy (BCE) loss to supervise the extraction of the boundary map.

The final loss function is:

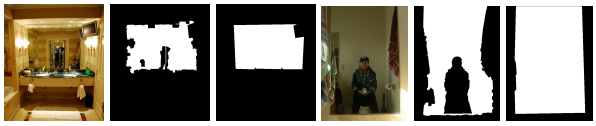
$$Loss = \sum_{s=1}^{S} w_s L_s + w_b L_b + w_f L_f, \tag{3}$$

where $L_s$ is the lovasz-hinge loss between the $s$-th mirror map and the ground truth mirror map, while $L_f$ is the lovasz-hinge loss between the final output mirror map and the ground truth mirror. $L_b$ is the binary cross-entropy (BCE) loss. We empirically set the weight balanced factors $w_s$, $w_b$, $w_f$ to 1, 5, 2, respectively.

## 4. Experiments

### 4.1. Datasets

**Datasets.** Currently, there is only one mirror dataset available, i.e., MSD [29], which has 4,018 mirror images with corresponding masks. However, we note that MSD includes mostly indoor scenes, with small mirrors. In particular, a lot of images in MSD are very similar to each other,



(a) (b) (c) (d) (e) (f)

Figure 7: Two wrongly annotated examples (a) and (d) from the original datasets that form our benchmark. While the mirror masks provided by the original dataset (b) and (e) are very coarse, the regions representing the reflected persons are not considered as part of the mirror. Our corresponding ground truth masks are shown in (c) and (f), respectively.

| Dataset | MSD | Ours |
|---|---|---|
| Similarity | 34.73% | 21.85% |

Table 1: The similarity scores of MSD and of our benchmark.

which can significantly reduce the robustness of mirror detection methods. Figure 3 shows eight images that are very similar to each other. We also use SSIM [25] to study the similarity of the images in MSD. We first resize all of them to the same size, and compute the sum of the image similarity of every pair of images in MSD. We then divide the sum by the number of image pairs to obtain the average similarity score of the whole dataset as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4}$$

$$Similarity = \frac{\sum_{k=1}^{N} SSIM(x_k, y_k)}{N}, \tag{5}$$

where $\mu_x, \mu_y$ and $\sigma_x, \sigma_y$ are the means and standard deviations of images $x$ and $y$. $\sigma_{xy}$ is the covariance of images $x$ and $y$. $C_1$ and $C_2$ are to avoid division by zero, and are set to $0.01^2$ and $0.03^2$, respectively. $N$ is the total number of image pairs. $k$ is the index of the $k^{th}$ image pair, containing two different images $(x_k, y_k)$. The similarity score ranges from 0 to 1. The second column of Table. 1 shows that MSD has a rather high similarity score.

To address the limitations of MSD as discussed above, we propose a large-scale benchmark here, which contains a total of 6,461 mirror images with ground truth annotations. All these images are obtained from six public datasets: ADE20K [35, 36], NYUD-V2 [19], MINC [3], Pascal-Context [18], SUNRGBD [23], and COCO-Stuff [6]. We select all images from these six datasets that contain mirrors in them. As such, our benchmark contains very diverse images, covering a variety of scenes. To evaluate the diversity of the images, we have also computed the average similarity score of our benchmark, as shown in the third column of Table. 1. We can see that our benchmark has a much lower similarity score than MSD.

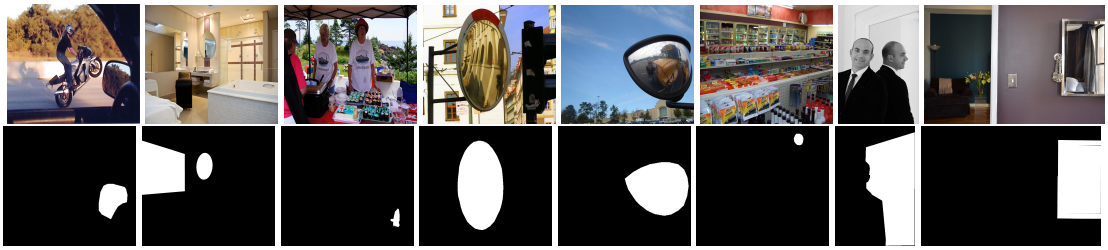


Figure 8: Images in our benchmark show high diversity and low similarity. They cover a varieity of daily scenes containing planar mirrors or concave mirrors.

| Dataset | Total number of images selected | Error rate |
|---|---|---|
| ADE20K [35, 36] | 1352 | 3.03% |
| COCO-Stuff [6] | 3766 | 91.72% |
| MINC [3] | 387 | 16.02% |
| NYUD-V2 [19] | 159 | 18.87% |
| SUNRGBD [23] | 716 | 31.15% |
| Pascal-Context [18] | 81 | 6.17% |
| Total | 6461 | 59.04% |

Table 2: Composition of our benchmark for mirror detection. We collect mirror images from six existing datasets. The second column shows the total number of mirror images obtained from each of the six datasets, while the third column shows the error rate of the their mirror annotations in each of the datasets.

In addition, we have noted that although all six datasets include the mirror label, there are a lot of labelling problems in their mirror annotations. For example, as demonstrated in Figure 7, a lot of them have very coarse or even incorrect mirror masks (left example), and some consider the regions of the reflected objects as non-mirror regions (right example). Table 2 summarizes the error rates of the mirror annotations of individual datasets. Each error rate indicates the percentage of wrongly labeled images in the original dataset. We have corrected all incorrect annotations in our benchmark. Figure 8 shows some example images/masks from our benchmark. We can see that they have much finer annotations.

To evaluate on our benchmark, we adopt the leave-one-out cross-validation strategy. Therefore, we test a model on all mirror images from one of the six datasets, but train it on all mirror images from the remaining five datasets. We perform this test six times on the six datasets in a similar way, to obtain six sets of results.

### 4.2. Evaluation Metrics

We employ two popular metrics to quantitatively evaluate the performance of our model. As the mirror detection problem is similar to the salient object detection problem, we use maximum F-measure ($F_\beta$) and mean absolute error ($MAE$) to evaluate the performance. $F_\beta$ is computed as:

$$F_\beta = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{6}$$

where $\beta^2 = 0.3$ as suggested in [1].

Mean absolute error ($MAE$) is computed as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y)|, \tag{7}$$

where $Y$ is the ground truth. $W$ and $H$ are width and height of the test image. $P$ is the predicted output.

### 4.3. Implementation Details

We use ResNeXt101 [27] pretrained on ImageNet [10] as the backbone feature extraction network. We have implemented the proposed model on PyTorch [20] and train it on a PC with a GeForce RTX2080Ti card. We use stochastic gradient descent as the optimizer with a momentum value of 0.9 and a weight decay of $5e-4$. The learning rate in the training phase is initialized as $1e-3$. We use the "poly" learning rate policy with a power of 0.9, which is the same as PSPNet [32]. We set the batch size to 10 and the number of training epochs to 150. We apply CRF [16] to our predicted map for final output. The parameters in all the layers, except the backend network, are randomly initialized. Training our model takes 16 hours and testing takes 0.13s per image, on a single GTX 2080Ti.

### 4.4. Comparison with the State-of-the-arts Methods

In this experiment, we compare our method with state-of-the-art methods from relevant fields.

Table 3 shows the mirror detection performance on the MSD dataset and our proposed benchmark. Our method achieves the best performances on both metrics, $F_\beta$ and MAE, compared with all the other methods. Figure 9 provide visual comparisons. The first three rows of images contain some ambiguous regions that look similar to mirrors. While MirrorNet tends to detect these regions as mirrors, our method can differentiate them well and accurately identify the mirror regions. In the third row, MirrorNet even

| Method | MSD | | ADE20K | | COCO-Stuff | | MINC | | NYUD-V2 | | Pascal-Context | | SUNRGBD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ | $F_\beta \uparrow$ | MAE$\downarrow$ |
| DSC [15] | 0.812 | 0.087 | 0.169 | 0.176 | 0.528 | 0.158 | 0.341 | 0.145 | 0.094 | 0.161 | 0.074 | 0.225 | 0.141 | 0.149 |
| BDRAR [37] | 0.792 | 0.093 | 0.683 | 0.137 | 0.624 | 0.219 | 0.354 | 0.134 | 0.686 | 0.132 | 0.465 | 0.219 | 0.624 | 0.204 |
| PSPNet [32] | 0.746 | 0.117 | 0.351 | 0.106 | 0.344 | 0.126 | 0.289 | 0.185 | 0.237 | 0.070 | 0.336 | 0.048 | 0.371 | 0.100 |
| $R^3$Net [11] | 0.846 | 0.068 | 0.722 | 0.034 | 0.651 | 0.076 | 0.560 | 0.089 | 0.631 | 0.035 | 0.557 | 0.034 | 0.630 | 0.039 |
| CPDNet [29] | 0.769 | 0.111 | 0.695 | 0.030 | 0.602 | 0.081 | 0.531 | 0.072 | 0.655 | 0.038 | 0.550 | 0.039 | 0.606 | 0.042 |
| BASNet [22] | 0.791 | 0.082 | 0.614 | 0.047 | 0.525 | 0.103 | 0.585 | 0.079 | 0.439 | 0.076 | 0.354 | 0.056 | 0.483 | 0.064 |
| EGNet [33] | 0.802 | 0.086 | 0.632 | 0.044 | 0.578 | 0.099 | 0.555 | 0.070 | 0.440 | 0.054 | 0.549 | 0.041 | 0.641 | 0.038 |
| MirrorNet [29] | 0.857 | 0.065 | 0.704 | 0.126 | 0.624 | 0.136 | 0.608 | 0.069 | 0.706 | 0.127 | 0.432 | 0.151 | 0.620 | 0.105 |
| Ours | **0.898** | **0.045** | **0.743** | **0.029** | **0.659** | **0.074** | **0.621** | **0.063** | **0.726** | **0.034** | **0.560** | **0.030** | **0.657** | **0.032** |

Table 3: Quantitative results on the MSD dataset (second column) and on our benchmark (third to eighth columns). We compare our model with relevant state-of-the-art methods: shadow detection methods DSC [15] and BDRAR [37]; semantic segmentation method PSPNet [32]; salient object detection methods $R^3$Net, CPDNet [26], BASNet [22] and EGNet [33]; and mirror detection method MirrorNet [29]. The best results are shown in bold.

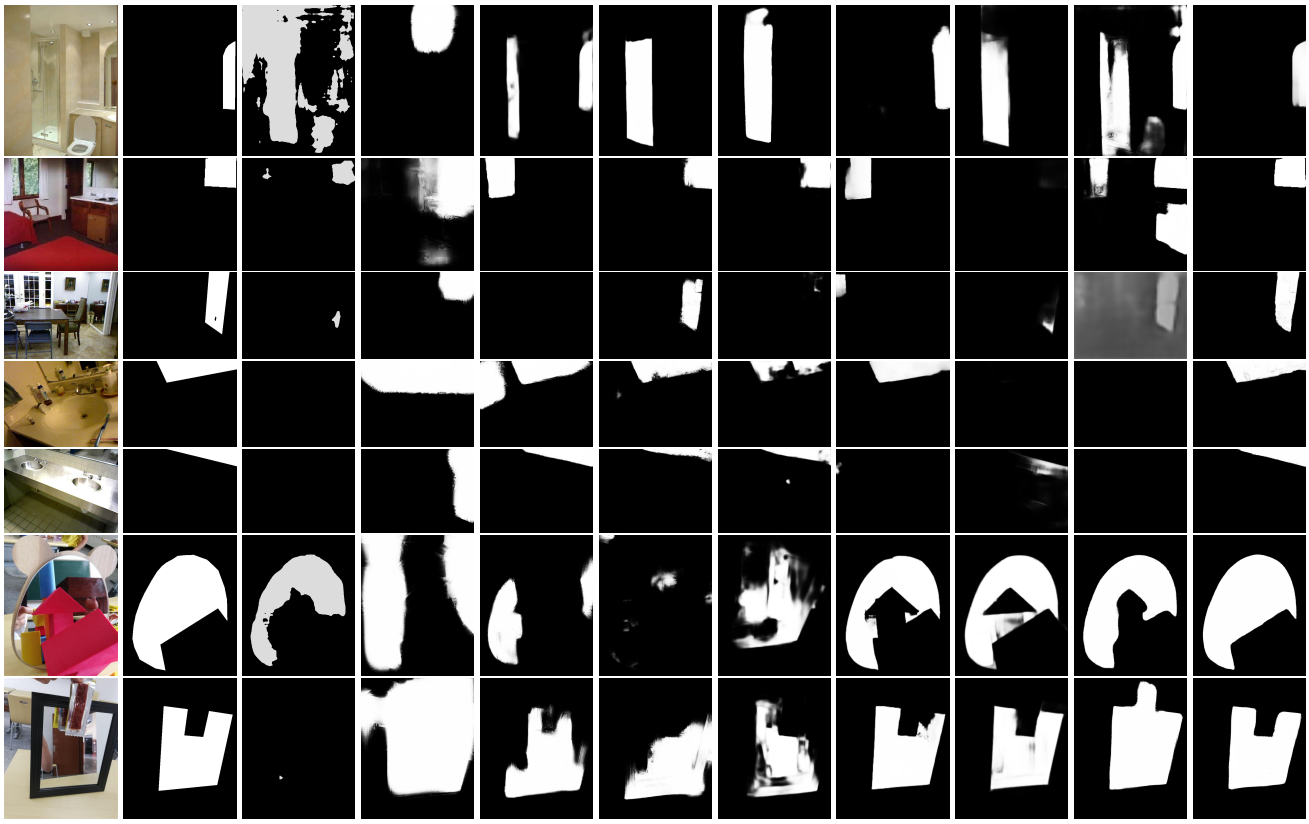



Figure 9: Qualitative results of our model, compared with relevant state-of-the-art methods.

considers the whole image as potentially covered by a mirror. The fourth and fifth rows of images contain mirrors that can easily be missed. While MirrorNet fails to detect them, our method can detect them accurately. The sixth and seven rows of images contain mirrors that are partially overlapped by the real objects. As a result, the real objects can be easily mixed up with their reflections. While MirrorNet fails to differentiate them correctly, our method can separate them accuractly and outperform all baseline methods.

We have also observed that although BASNet [22] and EGNet [33] are boundary-aware networks for salient object detection and have shown to perform well among the SOD methods, our method can still significantly outperforms these two methods on mirror detection.

| Ablation | $F_\beta \uparrow$ | MAE$\downarrow$ |
|---|---|---|
| Basic | 0.859 | 0.061 |
| Basic + s-ED | 0.858 | 0.068 |
| Basic + EDF | 0.864 | 0.062 |
| Basic + RCCL | 0.866 | 0.059 |
| Basic + EDF + GR | 0.874 | 0.052 |
| Basic + EDF + CCL | 0.876 | 0.049 |
| Basic + EDF + RCCL | 0.889 | 0.047 |
| Ours | **0.898** | **0.045** |

Table 4: Ablation study results, trained and tested on the MSD dataset. "Basic" denotes our network without the RCCL, EDF and refinement modules. "s-ED" is the EDF module with only a single low-level edge extractor, instead of both low-level and high-level edge extractors in our "EDF". "Basic+EDF+RCCL" is the full model but without the refinement module. "Our" is the proposed full model. The best results are shown in bold.

### 4.5. Ablation Study

Table 4 demonstrates the effectiveness of each component in our model. As shown in the last row, our final proposed network with the RCCL module, EDF module and refinement module outperforms other baselines on all metrics. We can see that the EDF module with only a single edge extractor (s-ED) does not help improve the performance. However, when the basic network containing both EDF and RCCL modules (using the top-level mirror map as output), it can greatly outperform the other ablated models, especially on $F_\beta$. We attribute it to the effect of our contextual relation extraction process carried out by the RCCL modul, which significantly benefits the mirror detection task from a global view. Figure 10 shows a visual example of the component analysis. We can see that the refinement module can help improve the performance by removing the over-predicted region.

To investigate the effectiveness of our EDF module, we visualize the edge maps extracted, as shown in Figure 11. We can find that our edge map extracted from high-level features (d) can differentiate between the red arrow and its reflection well, while the edge map extracted from low-level features (c) fails.

## 5. Conclusion

In this paper, we have proposed a progressive method for detecting mirrors in a single image. The method includes two novel modules, the relational contextual contrasted local module (RCCL) for extracting and comparing mirror and contextual features for correspondence, and the edge detection and fusion (EDF) module for extracting multi-scale mirror edge features. In addition, we have constructed a challenging large-scale benchmark with diverse scenes from six public datasets. It includes 6,461 images covering daily scenes with mirrors. Our experimental results demonstrate that the proposed model achieves state-of-the-art performances on both our benchmark and the existing dataset.

(a) (b) (c) (d) (e) (f) (g)

Figure 10: A visual example of the ablation study. (a) is the input image and (b) is the ground truth. (c) to (g) correspond to the predictions from the five ablated models: "Basic", "Basic + s-ED", "Basic + EDF", "Basic + EDF + RCCL", and ours, respectively.

(a) Input (b) GT (c) LL map (d) HL map

Figure 11: Visual comparison of edge maps extracted by low-level and high-level features.

Figure 12: Failure cases. Our model may fail on some images with very little relational and contextual contrasted information.

Our method does have limitations. Since our method relies on detecting and correlating features inside and outside of the mirrors, it may fail if a region appears like a mirror even from a human point of view. In Figure 12, the wall behind the man in the left image appears like some of the mirrors in our benchmark and is detected as a mirror region. On the right image, the wooden frame causes the background of the two persons to appear like a mirror. Hence, the background is also detected as a mirror region. As a future work, we are currently considering additional information, such as depth and light-field images, to help detect mirrors beyond human-like visual perception.

## References


[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and

Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018.

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. *CVPR*, 2015.

[4] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018.

[5] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE TPAMI*, 41(8):1844–1861, Aug 2019.

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018.

[7] John Canny. A computational approach to edge detection. *IEEE TPAMI*, (6):679–698, 1986.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE TPAMI*, 40(4), 2017.

[9] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[11] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690. AAAI Press, 2018.

[12] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018.

[13] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, June 2019.

[14] Xiaojie Guo, Xiaochun Cao, Jiawan Zhang, and Xuewei Li. MIFT: A mirror reflection invariant feature descriptor. In *ACCV*, pages 536–545. Springer-Verlag, 2010.

[15] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, pages 7454–7462, 2018.

[16] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, pages 109–117. 2011.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[18] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[19] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[21] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[22] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, June 2019.

[23] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, June 2015.

[24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[25] Zhou Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, Apr. 2004.

[26] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, June 2019.

[27] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, July 2017.

[28] Chuan Yang, Lihe Zhang, and Huchuan Lu. Graph-regularized saliency detection with convex-hull-based center prior. *IEEE Signal Processing Letters*, 20(7):637–640, 2013.

[29] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019.

[30] Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernandez Dominguez. Analyzing computer vision data - the good, the bad and the ugly. In *CVPR*, July 2017.

[31] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *CVPR*, June 2019.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[33] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, Oct 2019.

[34] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-class part parsing with joint boundary-semantic awareness. In *ICCV*, October 2019.

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[36] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, Mar 2019.

[37] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018.