Where Is My Mirror?

Xin Yang[1,*], Haiyang Mei[1,*], Ke Xu[1,3], Xiaopeng Wei[1], Baocai Yin[1,2], Rynson W.H. Lau[3,†]
[1] Dalian University of Technology, [2] Peng Cheng Laboratory, [3] City University of Hong Kong

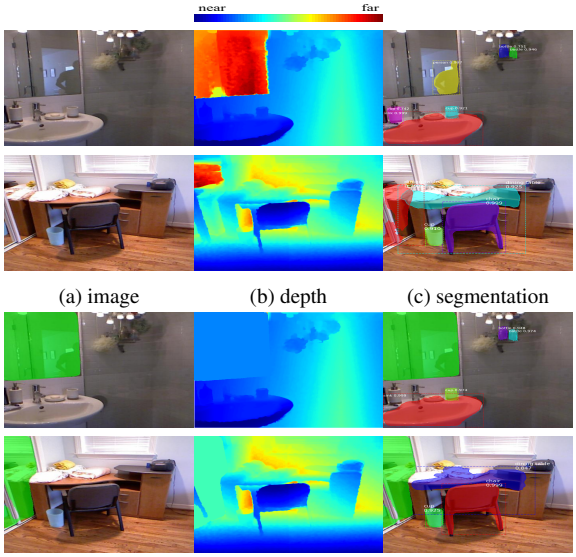
Abstract

Mirrors are everywhere in our daily lives. Existing computer vision systems do not consider mirrors, and hence may get confused by the reflected content inside a mirror, resulting in a severe performance degradation. However, separating the real content outside a mirror from the reflected content inside it is non-trivial. The key challenge is that mirrors typically reflect contents similar to their surroundings, making it very difficult to differentiate the two. In this paper, we present a novel method to segment mirrors from an input image. To the best of our knowledge, this is the first work to address the mirror segmentation problem with a computational approach. We make the following contributions. First, we construct a large-scale mirror dataset that contains mirror images with corresponding manually annotated masks. This dataset covers a variety of daily life scenes, and will be made publicly available for future research. Second, we propose a novel network, called MirrorNet, for mirror segmentation, by modeling both semantical and low-level color/texture discontinuities between the contents inside and outside of the mirrors. Third, we conduct extensive experiments to evaluate the proposed method, and show that it outperforms the carefully chosen baselines from the state-of-the-art detection and segmentation methods.

1. Introduction

Mirrors are very common and important in our daily lives. The presence of mirrors may, however, severely degrades the performance of existing computer vision tasks, *e.g.*, by producing wrong depth predictions (Figure 1(b)) or falsely detecting the reflected objects as real ones (Figure 1(c)). Hence, it is essential to these systems to be able to detect and segment mirrors from the input images.

Automatically segmenting mirrors from the background is extremely challenging, due to the fact that the contents reflected by the mirrors are very similar to those outside them (*i.e.*, their surroundings). This makes them fundamentally different from other objects that have been addressed well by the state-of-the-art segmentation methods [47, 12]. Meanwhile, as the contents reflected by the mirrors may not necessarily be salient, directly applying state-of-the-art saliency detection methods [8, 21] for detecting mirrors is also not appropriate.



(a) image (b) depth (c) segmentation

(d) MirrorNet (e) our depth (f) our segmentation

Figure 1: Problems with mirrors in existing vision tasks. In depth prediction, NYU-v2 dataset [32] uses a Kinect to capture depth as ground truth. It wrongly predicts the depths of the reflected contents, instead of the mirror depths (b). In instance semantic segmentation, Mask RCNN [12] wrongly detects objects inside the mirrors (c). With MirrorNet, we first detect and mask out the mirrors (d). We then obtain the correct depths (e), by interpolating the depths from surrounding pixels of the mirrors, and segmentation maps (f).

In this work, we aim to address the mirror segmentation problem. We note that humans can generally detect the existence of mirrors well. To do this, we observe that humans typically try to identify content discontinuity at the mirror boundaries in order to differentiate if some content belong to the reflection of a mirror. Hence, a straightforward solution to this problem is to apply low-level features to detect mirror boundaries. Unfortunately, this may fail if an object partially appears in front of a mirror, *e.g.*, the second

*Joint first authors. †Rynson Lau is the corresponding author, and he led this project. Project page: https://mhaiyang.github.io/ICCV2019_MirrorNet/index

example in Figure 1. In this case, separating the reflection of the object from the object itself may not be straightforward. Instead, this discontinuity includes both low-level color/texture changes as well as high-level semantics. This observation inspires us to leverage the contextual contrasted information for mirror segmentation.

In this paper, we address the mirror segmentation problem in two ways. First, we have constructed a large-scale mirror segmentation dataset (MSD), which contains $4,018$ pairs of images with mirrors and their corresponding segmentation masks, covering a variety of daily life scenes. Second, we propose a novel network, called MirrorNet, with a Contextual Contrasted Feature Extraction (CCFE) module, to segment mirrors of different sizes, by learning the contextual contrast inside and outside of the mirrors.

We have the following main contributions:

- We construct the first large-scale mirror dataset, which consists of $4,018$ images containing mirrors and their corresponding manually annotated mirror masks, taken from diverse daily life scenes.
- We propose a novel network that incorporates a novel contextual contrasted feature extraction module for mirror segmentation, by learning to model the contextual contrast inside and outside of the mirrors.
- Through extensive experiments, we show that the proposed network outperforms many baselines derived from state-of-the-art segmentation/detection methods.

2. Related Work

In this section, we briefly review state-of-the-art methods from relevant fields, including semantic/instance segmentation, saliency/shadow detection, as well as mirror detection works from the 3D community.

Semantic segmentation. It aims to assign per-pixel predictions of object categories to the input image. Based on the fully convolutional encoder-decoder structure [25], state-of-the-art semantic segmentation approaches typically leverage multi-scale (level) context aggregation to learn discriminative features for recognizing the objects and delineating their boundaries. Specifically, low-level encoder features are combined with their corresponding decoder features by feeding recorded pooling indices [3] or concatenation [31]. Dilated convolutions are used in [7, 42] to expand the receptive fields to compensate for the lost details in the encoder part. PSPNet [48] leverages pyramid pooling to obtain multi-scale representations in order to differentiate objects of similar appearances. Zhang *et al*. [46] propose to fuse the low-/high-level features so as to take advantages of both high resolution spatial and rich semantic information in the encoder part. Zhang *et al*. [43] propose to explicitly predict the objects in the scene and use this prediction to selectively highlight the semantic features. Ding *et al*. [10] propose to learn contextual contrasted features to boost the segmentation performance of small objects.

However, applying existing segmentation methods for mirror segmentation (*i.e*., treating mirrors as one of the object categories) cannot solve the fundamental problem of mirror segmentation, which is that the reflected content of a mirror can also be segmented too. In this paper, we focus on the mirror segmentation problem and formulate it as a binary classification problem (*i.e*., mirror or non-mirror).

Instance segmentation. It aims to simultaneously recognize, localize and segment out objects while differentiating individual instances of the same category. State-of-the-art detection based instance segmentation methods extend object detection methods, *e.g*., Faster-RCNN [30] and FPN [20], to obtain instance maps. Mask RCNN [12] uses one additional branch to predict instance segmentation masks from the box predictions of Faster-RCNN [30]. PANet [23] further proposes to add bottom-up paths to facilitate feature propagation in Mask RCNN [12] and aggregates multi-level features for detection and segmentation. MaskLab [6] adopts Faster-RCNN [30] to locate objects and combines semantic segmentation with pixel-direction (to its instance center) prediction for instance segmentation. Another line of works first use a segmentation method to obtain per-pixel labels, and then a clustering method to group the pixels into instances, via depth estimation [45], spectral clustering [19], and neural networks [38, 22].

Similar to semantic segmentation, instance segmentation methods cannot differentiate between the content inside a mirror and that of outside. As a result, they would segment objects inside the mirror too.

Salient object detection (SOD). It aims to identify the most conspicuous object(s) in an image. While conventional SOD methods rely on low-level hand-crafted features (*e.g*., color and contrast), deep learning based SOD methods consider either or both bottom-up and top-down saliency inferences. Wang *et al*. [35] propose to integrate local pixel-wise saliency estimation and global object proposal search for salient object detection. Multi-level feature aggregation from deep networks is also explored for detecting and refining the detection [17, 44, 13]. Recent works apply attention mechanisms for learning global and local contexts [21] or learning foreground/background attention maps [8] to help detect salient objects and eliminate non-salient objects.

The content reflected by a mirror, however, may or may not be salient. Even though if it is salient, it is likely that only part of it is salient. Hence, applying existing SOD methods to detect mirrors may not address the mirror segmentation problem.

Shadow detection. It aims to detect/remove shadows from the input images. Hu *et al*. [14] propose to use direction-aware features to analyze the contrasts between

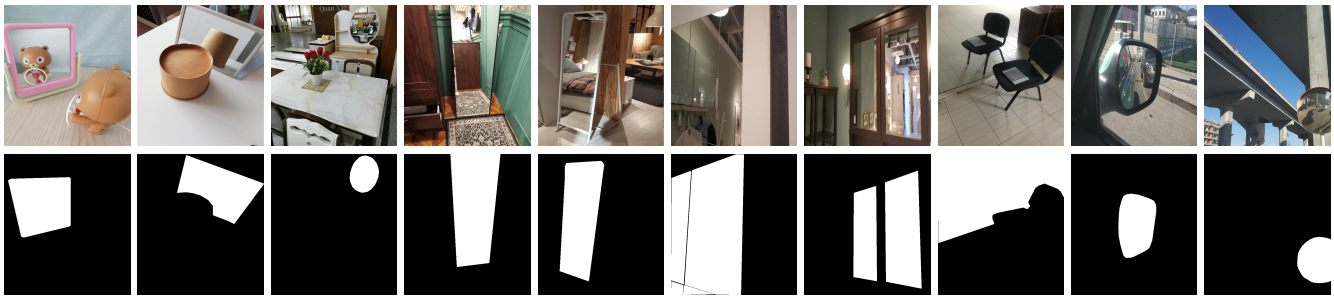


Figure 2: Example mirror image/mask pairs in our mirror segmentation dataset (MSD). It shows that our MSD covers a variety of our daily life scenes that contain mirrors.

shadow/non-shadow regions for shadow detection. Le *et al*. [16] propose to train a shadow detection network with augmented adversarial shadow examples generated from a shadow attenuation network. Zhu *et al*. [50] propose a bidirectional feature pyramid to leverage the spatial contexts from shallow and deep CNN layers. A conditional GAN [26] is also applied to model both local features and global image semantics for shadow detection [27] and removal [34]. Qu *et al*. [29] propose a multi-context network, together with a new dataset, for shadow removal.

In general, shadow detection methods are largely based on detecting the intensity contrast between shadow and non-shadow regions. In contrast, the contents inside and outside of a mirror typically have very similar intensity, making the mirror segmentation problem more difficult to address.

Mirror detection in the 3D community. To our knowledge, there are only two works that consider mirror segmentation in 3D reconstruction. Matterport3D [5] proposes the user to manually segment the mirrors on an iPad during scanning. Whelan *et al*. [36] attach a hardware tag (based on the AprilTag [28]) to the scanner. If a tag is detected in the captured image, it signals the presence of a mirror. A total variation-based segmentation method is then used to segment the mirror based on a set of hand-crafted features (*e.g*., depth discontinuity and intensity variance).

Instead of using any special hardware, in this paper, we propose the first automatic method for mirror segmentation and the first mirror dataset with mirror annotations.

3. Mirror Segmentation Dataset

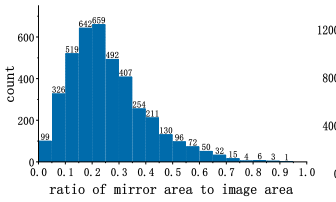
To address the mirror segmentation problem, we construct the first large-scale mirror dataset, named MSD. It includes $4,018$ pairs of images containing mirrors and their corresponding manually annotated masks.

Dataset construction. We use several latest smartphones for capturing images and Labelme[1] for manual labeling of mirrors. While capturing the images, we consider common types of mirrors (including cosmetic, dressing, decorative, bathroom, and road mirrors) that are often used in our daily life scenes (*e.g*., bedroom, living room, office, garden, street, and parking lot). Some example mirror images in our MSD dataset are shown in Figure 2. The dataset contains 3,677 images taken from indoor scenes and 341 images taken from outdoor scenes. The reason that we have many more indoor images than outdoor ones is that we want to focus on indoor scenes in this work. The outdoor images are mainly to provide more diverse mirror shapes and scenarios. For splitting the dataset into training and test sets in a fair way, we first divide the images into different groups based on the mirror types. Since we may have taken several images using each specific mirror with different combinations of foreground/background objects and camera orientations, to make sure that mirrors appearing in the training set do not appear in the test set, we split the images by randomly splitting the mirror types. Finally, we have 3,063 images for training and 955 images for testing.

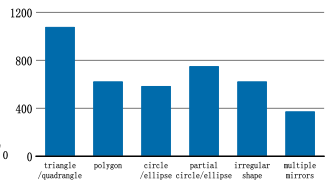
Dataset analysis. Figure 3 shows statistical analysis on the mirror properties in our captured images (including mirror area, shape, location in the image, and global color contrast between inside/outside of the mirror) for a comprehensive understanding of the proposed MSD dataset.

- **Mirror area distribution.** We define it as a ratio between the mirror area and image area. As shown in Figure 3(a), majority of the mirrors fall in the range of $(0.0, 0.7]$. Mirrors falling in the range of $(0.0, 0.1]$ are small mirrors that can easily be cluttered with other background objects. Mirrors falling in the range of $[0.5, 0.95]$ are typically located close to the camera. Foreground object occlusion often happens in this situation. Mirrors falling in the range of $[0.95, 1.0]$ are not included in MSD, as the images may not provide sufficient contextual information even for humans to determine whether there is a mirror in them.
- **Mirror shape distribution.** There are some popular mirror shapes (*e.g*., elliptic and rectangular). However, if a mirror is partially occluded by an object in front of it, the resulting shape of the mirror becomes irregular. Figure 3(b) shows that MSD includes images of different mirror shapes and multiple mirrors.

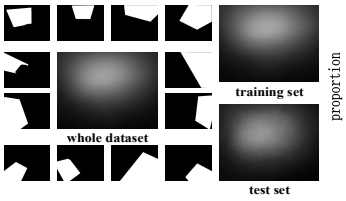
[1] https://github.com/wkentaro/labelme



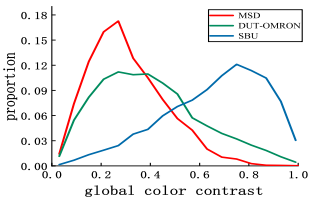
(a) mirror area distribution



(b) mirror shape distribution



(c) mirror location distribution



(d) color contrast distribution

Figure 3: Statistics of the MSD dataset. We show that MSD has mirrors with reasonable property distributions, including mirror area, silhouette, location and color contrast.

- **Mirror location distribution.** To analyze the spatial distribution of mirrors in MSD, we compute probability maps to show how likely each pixel belongs to a mirror, as in Figure 3(c). Although our MSD has mirrors covering different locations, the mirrors tend to cluster around the upper part of the image. This is reasonable as mirrors are usually placed approximately around the human eyesight. We can also see that the mirror location distributions for the training/test splits are consistent to that of the whole dataset.
- **Color contrast distribution.** As mirrors can reflect unpredictable contents, we analyze the global color contrast between the contents inside/outside of the mirrors, to check if mirror contents in our dataset are salient and can easily be detected. We use χ^2 distance to measure the contrasts between two RGB histograms computed separately from mirror and non-mirror regions, similar to [18, 11]. We further compare this distribution to two existing datasets, *i.e.*, the DUT-OMRON saliency dataset [41] and SBU shadow dataset [33], as shown in Figure 3(d). We can see that MSD has the lowest global color contrast, making the mirror segmentation task more challenging.

4. Proposed Network

We observe that in order for humans to know if we are looking at a mirror, we typically look for content discontinuity, in terms of low-level color/texture changes as well as high-level semantic information. This inspires us to leverage the contrast between the mirror and non-mirror regions. To this end, we propose a novel Contextual Contrasted Feature Extraction (CCFE) block to extract multi-scale contextual contrasted features for mirror localization. Building upon the CCFE block, a novel CCFE module is designed to hierarchically aggregate long-range contextual contrasted information to effectively detect mirrors of different sizes.

4.1. Overview

Figure 4 illustrates the proposed mirror segmentation network, called MirrorNet. It takes a single image as input and extracts multi-level features by the feature extraction network (FEN). The deepest features, which are full of semantics, are then fed to the proposed CCFE module to learn contextual contrasted features for locating the mirrors with the coarsest mirror map, by detecting the dividing boundaries where the contrasts appear. This mirror map functions as an attention map to suppress the feature noise of the next-upper FEN features in the non-mirror regions, so that the next-upper layer can focus on learning discriminative features in the candidate mirror regions. In this way, MirrorNet progressively leverages contextual contrasted information to refine the mirror region in a coarse-to-fine manner. Finally, we upsample the coarsest network output to obtain the original image resolutions as the output.

4.2. Contextual Contrasted Feature Extraction

Figure 5 shows the structure of the proposed CCFE module. Given the features extracted by the Feature Extraction Network, the CCFE aims to produce multi-scale contextual contrasted features for detecting mirrors of different sizes.

CCFE block. To effectively detect mirror boundaries (where contents may change significantly), we design the CCFE block to learn contextual contrasted features between a local region and its surrounding, as:

$$\mathbb{CCF} = f_{local}(\mathbf{F}, \Theta_{local}) - f_{context}(\mathbf{F}, \Theta_{context}), \quad (1)$$

where $\mathbf{F}$ is the input features. f_{local} represents a local convolution with a 3×3 kernel (dilation rate = 1). $f_{context}$ represents a context convolution with a 3×3 kernel (dilation rate = x). Θ_{local} and $\Theta_{context}$ are parameters. $\mathbb{CCF}$ is the desired contextual contrasted features.

We further propose to learn multi-scale contextual contrasted features to avoid the ambiguities caused by nearby real objects and their reflections in the mirror, by considering non-local contextual contrast. Hence, we set the dilation rate x to 2, 4, 8, and 16, such that long-range spatial contextual contrast can be obtained. The multi-scale contextual contrasted features are then concatenated and refined via the attention module [37], to produce feature maps that highlight the dividing boundaries.

CCFE module. A large mirror can easily cause under-segmentation, as the content inside it may exhibit high contrast within itself. To address this problem, global image contexts should be considered. Hence, we propose to leverage the global contextual contrast by cascading the CCFE blocks to form a deep CCFE module with larger receptive

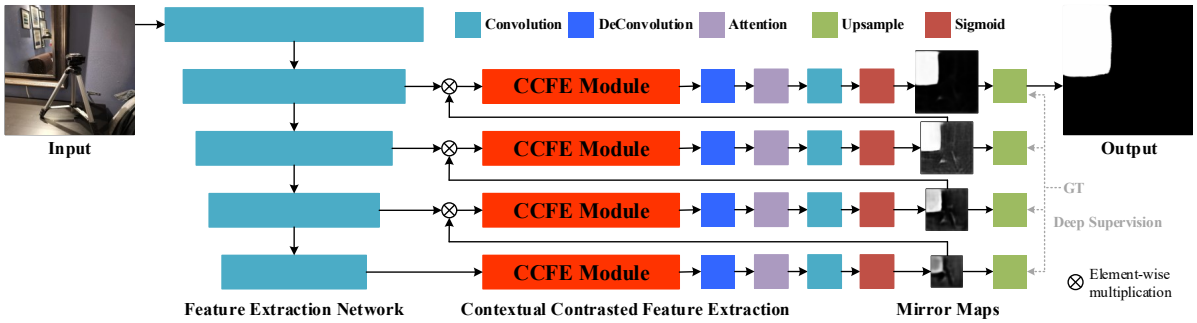


Figure 4: Overview of MirrorNet. First, a pre-trained Feature Extraction Network is used to extract multi-scale feature maps. Second, CCFE modules are embedded to different layers of the Feature Extraction Network to learn different scales of contextual contrasted features. Third, MirrorNet leverages these different scales of features in a coarse to fine manner to produce mirror maps, which function as attention maps to help the upper layers focus on learning contextual contrasted features in the candidate mirror regions. Fourth, the coarsest mirror map is progressively refined and increased in spatial resolution as it propagates from the bottom layers up to the upper layers.

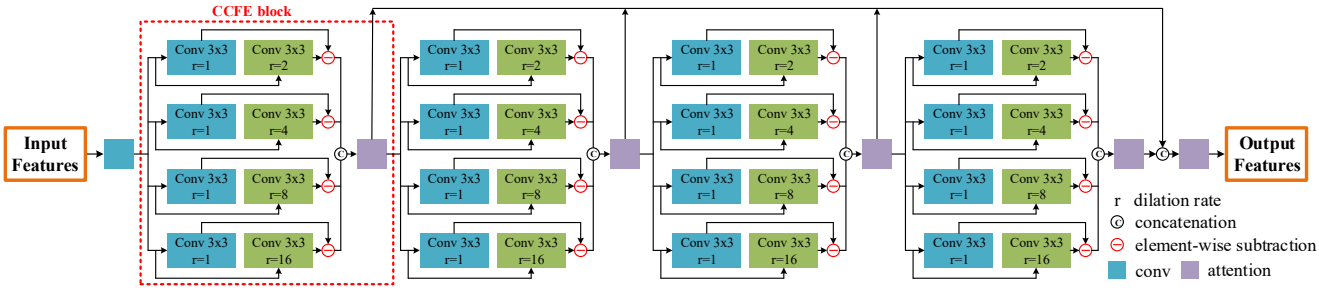


Figure 5: The Contextual Contrasted Feature Extraction (CCFE) module. The input features are passed through four chained CCFE blocks and the output of each CCFE block are fused via an attention module to generate multi-level contextual contrasted features. In each CCFE block (red dashed box), we first compute the contextual contrasts between local information (extracted by standard convolutions) and their surrounding contexts (extracted by dilated convolutions with different dilation rates) in parallel, and then adaptively select useful ones from these concatenated multi-scale contextual contrasted features via an attention module.

fields, such that the global image contexts are captured in deeper blocks of the CCFE module. We also adopt the attention module [37] to highlight the candidate mirror regions of the concatenated multi-level features from different blocks in the CCFE module.

Discussion. Although we have drawn some inspiration from the Context Contrast Local (CCL) block in [10] in our network design, our network is different from the CCL block in both motivations and implementations. First, while the CCL block aims to detect small objects, our CCFE modules are used to locate mirrors by detecting the dividing boundaries. They also serve as attention modules to enhance the feature responses in mirror regions and suppress the feature noise in the non-mirror regions. Second, the CCL block has only one scale of contrast and is only embedded in the deepest layer for their purpose of small objects detection using semantical contrast. We extend the CCL block to our CCFE module by incorporating multi-scale contextual contrasted feature extraction, to provide sufficient contextual information for locating mirrors in different sizes. We also embed our CCFE modules to all side-outputs of the feature extraction network, such that our network takes advantages of both rich semantical contrasted contexts from deeper layers and low-level contrasted contexts from upper layers, for mirror segmentation.

4.3. Loss Function

Per-pixel cross entropy is commonly used as the loss function in semantic segmentation, salient object detection and shadow detection problems. However, it is not sensitive to small objects, and can easily be dominated by large objects. Hence, we choose the lovász-hinge loss [4] for optimizing our network. It is a surrogate for the non-differentiable intersection over union (IoU) metric, which preserves the scale invariance property of the IoU metric. Deep supervision [40] is also adopted to facilitate the learning process. The loss function is:

$$Loss = \sum_{s=1}^{S} w_s L_s, \tag{2}$$

where w_s represents the balancing parameters. L_s is the lovász-hinge loss between the s-th upsampled mirror map and the ground truth.

4.4. Implementation Details

We have implemented MirrorNet on the PyTorch framework [1]. For training, input images are resized to a resolution of 384×384 and are augmented by horizontally random flipping. We use the pre-trained ResNeXt101 network [39] as the feature extraction network. The remaining parts of our network are randomly initialized. For loss optimization, we use the stochastic gradient descent (SGD) optimizer with momentum of 0.9 and a weight decay of 5×10^{-4}. Batch size is set to 10. The learning rate is initialized to 0.001 and decayed by the poly strategy [24] with the power of 0.9, for 160 epochs. There are $S = 4$ loss terms in Eq. 2, and the balancing parameters w_s are empirically set to 1. It takes about 12 hours for the network to converge on an NVIDIA Titan V graphics card. For testing, images are also resized to a resolution of 384×384 for network inferences. We then use the fully connected conditional random field (CRF) [15] to further enhance the network outputs by optimizing the spatial coherence of pixels as the final mirror segmentation results.

5. Experiments

5.1. Experimental Settings

Evaluation metrics. For a comprehensive evaluation, we adopt five metrics that are commonly used in the related fields (*i.e.*, semantic segmentation, salient object detection and shadow detection), for quantitatively evaluating the mirror segmentation performance. Specifically, we use the intersection over union (IoU) and pixel accuracy metrics from the semantic segmentation field as our first and second metrics. We also use the F-measure and mean absolute error (MAE) metrics from the salient object detection field. F-measure is defined as the weighted harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{3}$$

where β^2 is set to be 0.3 to emphasize more on precision over recall, as suggested in [2].

Finally, we adopt the balance error rate (BER) from the shadow detection field, to evaluate the mirror segmentation performance. It considers the unbalanced areas of mirror and non-mirror regions, and is computed as:

$$BER = 100 \times (1 - \frac{1}{2}(\frac{TP}{N_p} + \frac{TN}{N_n})), \tag{4}$$

where TP, TN, N_p and N_n represent the numbers of true positives, true negatives, mirror pixels, and non-mirror pixels, respectively.

Compared methods. We select the state-of-the-art methods from the related fields for comparison. Specifically, we choose PSPNet [48] and ICNet [47] from the semantic segmentation field, Mask RCNN [12] from the instance segmentation field, DSS [13], PiCANet [21], RAS [8] and R^3Net [9] from the salient object detection field, DSC [14] and BDRAR [50] from the shadow detection field. We use their publicly available codes and train them on our proposed training set for a fair comparison.

5.2. Comparison to the State-of-the-arts

Evaluation on the MSD test set. Table 1 reports the mirror segmentation performance on the proposed MSD test set. We can see that our method achieves the best performance with a large margin on all five metrics: intersection over union (IoU), pixel accuracy (Acc), F-measure (F_β), mean absolute error (MAE), and balance error rate (BER). Figure 6 shows visual comparisons. We can see that our method can effectively locate and segment small mirrors (4*th*, 5*th* and 7*th* rows). While the state-of-the-arts typically under-segment the large mirrors with high contrasts among their contents, our method successfully detects the mirror regions as a whole (*e.g.*, 1*st* and 3*rd* rows). Our method can also accurately delineate the mirror region boundaries, where there are ambiguities caused by nearby objects and their reflections in the mirror (2*nd* row). In general, our method can segment mirrors of different sizes with accurate boundaries. This is mainly contributed by the proposed multi-scale contextual contrasted feature learning.

More mirror segmentation results. Figure 7 shows some mirror segmentation results from our MirrorNet on

method	CRF	IoU↑	Acc↑	F_β↑	MAE↓	BER↓
Statistics	-	30.76	0.595	0.436	0.360	32.94
PSPNet [48]	-	63.18	0.750	0.746	0.117	15.82
ICNet [47]	-	57.18	0.694	0.709	0.125	18.78
Mask RCNN [12]	-	63.10	0.820	0.756	0.095	14.38
DSS [13]	-	59.08	0.665	0.743	0.125	18.82
PiCANet [21]	-	71.69	0.844	0.808	0.088	11.02
RAS [8]	-	60.46	0.695	0.758	0.111	17.61
R^3Net [9] w/o C	-	72.66	0.805	0.840	0.080	11.47
R^3Net [9]	√	73.19	0.805	0.845	0.068	11.40
DSC [14]	-	69.68	0.816	0.811	0.087	11.79
BDRAR [50] w/o C	-	66.97	0.821	0.798	0.099	12.48
BDRAR [50]	√	67.39	0.820	0.792	0.093	12.43
MirrorNet w/o C	-	78.38	0.932	0.841	0.066	6.50
MirrorNet	√	**78.88**	**0.932**	**0.856**	**0.066**	**6.43**

Table 1: Comparison to state-of-the-arts on MSD test set. All methods are trained on MSD training set. "w/o C" is without using CRF [15] for post-processing. "Statistics" refers to thresholding mirror location statistics from our training set as a mirror mask for detection. The best and second best results are marked in **bold** and red, respectively.

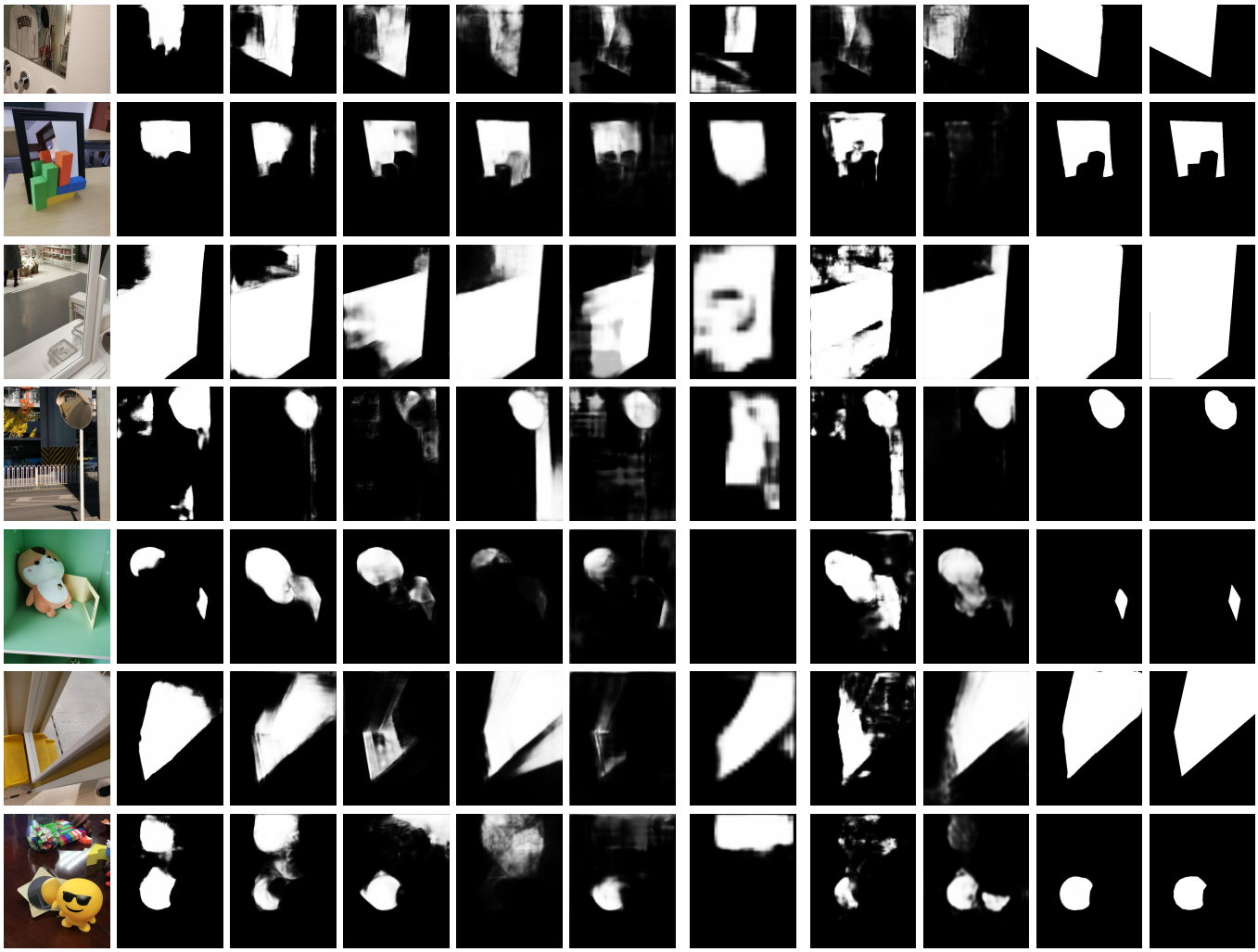


Figure 6: Visual comparison of MirrorNet to the state-of-the-art methods on the proposed MSD test set.

Figure 7: Some mirror segmentation results of MirrorNet on the ADE20K dataset [49].

the ADE20K dataset [49], which demonstrate the effectiveness of MirrorNet. Figure 8 shows mirror segmentation results of some challenging images downloaded from the Internet. These images contain not only mirrors but also other mirror-like objects, such as paintings (2*nd*, 3*rd* and 6*th* rows), windows (5*th* row), and door (4*th* row). We can see that the existing methods are distracted by these mirror-like objects. However, MirrorNet can distinguish mirrors from paintings/windows (*e.g.*, 2*nd*, 3*rd* and 5*th* rows), as the content within a mirror region is usually semantically consistent with the rest of the image while the content within a painting/window region is often different. MirrorNet is designed to learn different levels of contextual contrast features between the mirror region and outside. For example, the mirror region in the 2*nd* row reflects the indoor scene, which is similar to the surroundings of the mirror, while the paintings contain very different scenes. Such differences can be learned by the CCFE module. We understand that there are limitations with this assumption, which can be an interesting future work. In addition, MirrorNet can distinguish the mirror from the door in the 4*th* row. A possible reason is that the bottom of the door region is continuous and thus the door region is not considered as a mirror.

5.3. Component Analysis

Table 2 demonstrates the effectiveness of the lovász-hinge loss [4] and the proposed CCFE module. We can see that the lovász-hinge loss [4] performs better than the

Networks	IoU↑	Acc↑	BER↓
basic + BCE loss	74.00	0.821	10.61
basic + lovász-hinge loss [4]	75.32	0.820	10.46
basic + CCFE w/o contrasts	78.54	0.851	8.56
basic + CCFE w/ 1B4C	76.31	0.882	8.02
basic + CCFE w/ 4B1C	78.50	0.853	9.08
MirrorNet	**78.88**	**0.932**	**6.43**

Table 2: Component analysis. "basic" denotes our network with all CCFE modules removed, "CCFE w/o contrasts" denotes using multi-scale dilated convolutions without computing their feature contrasts. "1B4C" denotes using 1 CCFE block with 4 parallel scales of contrasts, while "4B1C" denotes using 4 CCFE blocks with 1 scale of contrasts. Our proposed CCFE module contains 4 blocks and each of them contains 4 scales of contrast extraction.

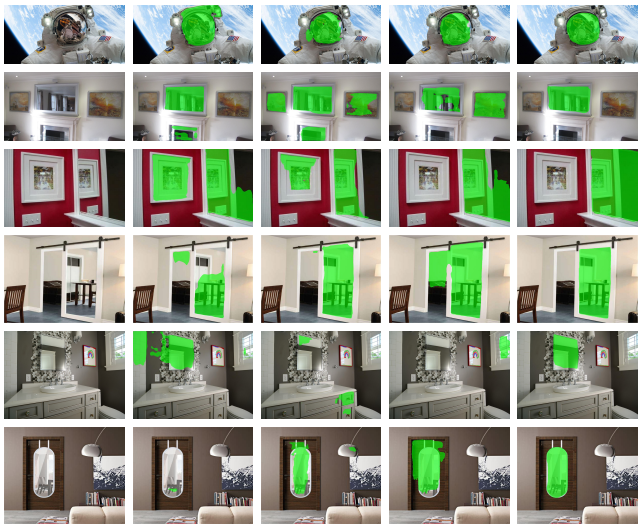


Image PSPNet[48] DSC[14] PiCANet[21] MirrorNet

Figure 8: More mirror segmentation results on challenging images obtained from the Internet.

binary cross entropy (BCE) loss in our task, due to its scale-invariant property. In addition, while multi-scale dilated convolutions (*i.e.*, "CCFE w/o contrasts") benefit the segmentation performance, we can see that using only one CCFE block with 4 parallel scales of contrast extraction ("basic + CCFE w/ 1B4C") can improve both the pixel accuracy and BER. In contrast, using four CCFE blocks with one single scale of contrast extraction mainly improves the IoU. Our proposed multi-scale contextual contrasted feature learning takes advantage of both. Figure 9 shows a visual example, in which we can see that our method successfully learns the global contextual contrasted features for addressing the mirror under-segmentation problem.

6. Conclusion and Future Work

In this paper, we have presented a novel method to segment mirrors from an input image. Specifically, we have constructed the first large-scale mirror dataset (MSD). It contains 4,018 images with mirrors and corresponding masks. We have also proposed a novel network to leverage multi-scale contextual contrasts for mirror detection. We have conducted extensive experiments to verify the superiority of the proposed network against state-of-the-art methods developed for other relevant problems, on both the proposed MSD test set, the ADE20K dataset [49], and some challenging images obtained from the Internet.

image basic basic + CCFE w/o contrasts basic + CCFE w/ contrasts

Figure 9: Visual example of the component analysis.

input images our results

Figure 10: Failure cases. Our mirror segmentation method can fail in extreme scenarios where insufficient contextual contrasts can be extracted.

Our method does have limitations. As it relies on modeling the contextual contrasts presented in the input images, it tends to fail in some extreme scenes where insufficient contextual contrasts between the mirrors and their surroundings can be perceived, as shown in Figure 10.

As a first attempt to address the automatic mirror segmentation problem, we focus in this paper on segmenting mirrors that appear in our daily life scenes. However, in some cities, the glass walls of skyscrapers may often exhibit mirror-like effects and reflect the surrounding objects/scenes. There are also very large mirrors that may appear outside some stores. As a future work, we are interested to extend our method to detect this kind of mirrors that appear in city streets, which may benefit outdoor vision tasks such as autonomous driving and drone navigation.

Acknowledgements. This work was supported in part by the NNSF of China under Grants 91748104, U1811463, 61632006, 61425002, and 61751203, by the Open Project Program of the State Key Lab of CAD&CG (Grant A1901), Zhejiang University, the Open Research Fund of Beijing Key Laboratory of Big Data Technology for Food Safety (No. BTBD-2018KF), and a SRG grant from City University of Hong Kong (Ref: 7004889).

References

[1] Pytorch. https://pytorch.org/.

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE TPAMI*, 2017.

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

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, 2017.

[6] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.

[8] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018.

[9] 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*, 2018.

[10] 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.

[11] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

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

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011.

[16] Hieu Le, Tomas Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+D Net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, 2018.

[17] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.

[18] Yin Li, Xiaodi Hou, Christof Koch, James Rehg, and Alan Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[19] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *IEEE TPAMI*, 2018.

[20] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[21] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[22] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *ICCV*, 2017.

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[24] Wei Liu, Andrew Rabinovich, and Alexander Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[27] Vu Nguyen, Tomas Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017.

[28] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *ICRA*, 2011.

[29] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *CVPR*, 2017.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 2017.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

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

[33] Tomás Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, 2016.

[34] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018.

[35] Lijun Wang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015.

[36] Thomas Whelan, Michael Goesele, Steven Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM TOG*, 2018.

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[38] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv:1605.06885*, 2016.

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

[40] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[41] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

[43] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.

[44] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Ruan Xiang. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[45] Ziyu Zhang, Alexander Schwing, Sanja Fidler, and Raquel Urtasun. Monocular object instance segmentation and depth ordering with cnns. In *ICCV*, 2015.

[46] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, 2018.

[47] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.

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

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

[50] 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.