Mirror Detection with the Visual Chirality Cue

Xin Tan, Jiaying Lin, Ke Xu,† Pan Chen, Lizhuang Ma†, and Rynson W.H. Lau†

Abstract—Mirror detection is challenging because the visual appearances of mirrors change depending on those of their surroundings. As existing mirror detection methods are mainly based on extracting contextual contrast and relational similarity between mirror and non-mirror regions, they may fail to identify a mirror region if these assumptions are violated. Inspired by a recent study of applying a CNN to help distinguish whether an image is flipped or not based on the visual chirality property, in this paper, we rethink this image-level visual chirality property and reformulate it as a learnable pixel level cue for mirror detection. Specifically, we first propose a novel flipping-convolution-flipping (FCF) transformation to model visual chirality as learnable commutative residual. We then propose a novel visual chirality embedding (VCE) module to exploit this commutative residual in multi-scale feature maps, to embed the visual chirality features into our mirror detection model. Besides, we also propose a visual chirality-guided edge detection (CED) module to integrate the visual chirality features with contextual features for detection refinement. Extensive experiments show that the proposed method outperforms state-of-the-art methods on three benchmark datasets.

Index Terms—Mirror Detection, Visual Chirality, Intrinsic Mirror Property, Salient Object Detection.

1 Introduction

Mirrors are very common objects in our daily lives. However, they are very challenging to detect, as they typically do not have their own visual appearances but reflect those of their surrounding objects. This property makes them very difficult to be distinguished from their surrounding objects, and fails many existing computer vision tasks, *e.g.*, depth prediction [20] and instance detection [39]. Hence, while detecting mirrors from an input image is an important task, it is also very challenging.

Although a few methods [44], [21], [25] have been proposed for mirror detection, they all rely on learning the relations (*i.e.*, contrast [44], correspondence [21], and depth discontinuity [25]) between mirror/non-mirror regions. However, these relation assumptions can be easily violated in real world scenes, causing these mirror detection methods to fail. For example, both MirrorNet [44] and PMDNet [21] under-detect the mirror region in the top image of Figure 1, as the mirror region has weak context contrast to the non-mirror region while correspondences do not exist. In contrast, both of them over-detect the mirror region in the bottom image of Figure 1, as they heavily rely on looking for relations between the two regions. We can see that MirrorNet [44] correctly locates the mirror region but extends it to the upper-left region of the image, as the contrast between the upper-left region and the mirror region is lower compared to the contrast between the lower-left region (the desk) and the mirror region. PMDNet [21] mistakenly establishes correspondences between the left side and the right side of the image, and hence identifies the left side (including the upper-left region) as the mirror region. This kind of ambiguities actually appear everywhere in real world scenes, and they cannot be addressed simply by modeling the contrast/correspondence relations between mirror and non-mirror regions. Hence, there is a need to incorporate some intrinsic mirror properties in the detection.

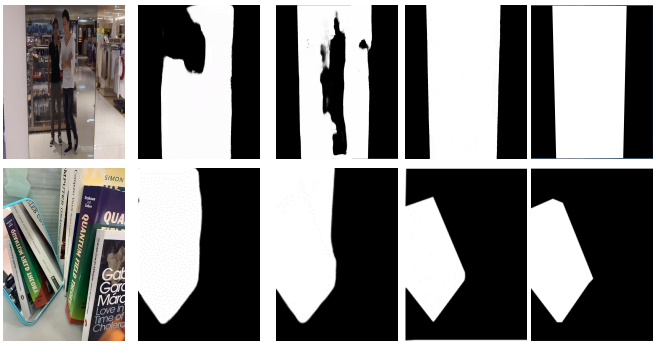


Fig. 1. Existing single image based mirror detection methods [44] [21], which are based on modeling contrasts/correspondences between mirror and non-mirror regions, may fail when these relations are not reliable. For example, MirrorNet [44] would fail if the contrasts between mirror/non-mirror regions are weak (top row) or have multiple degrees (bottom row). PMDNet [21] would fail if correspondences do not exist (top row) or are incorrectly detected (bottom row). Our method (Ours) leverages the visual chirality cue, which is an intrinsic property of mirrors reflecting real-world scenes, to accurately differentiate mirror and non-mirror regions.

Recently, Lin *et al.* [22] studied the visual chirality problem. Visual chirality refers to the statistical change of visual data caused by image flipping (*i.e.*, reflection) in data augmentation, and is defined on the image distribution *(at*

- *X. Tan and L. Ma are with the School of Computer Science and Technology, East China Normal University, Shanghai, China and the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. X. Tan is also with the Department of Computer Science, City University of Hong Kong, HKSAR, China. E-mail: tanxin2017@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn*
- *J. Lin, K. Xu and Rynson W.H. Lau are with the Department of Computer Science, City University of Hong Kong, HKSAR, China. E-mail: jiayinlin5-c@my.cityu.edu.hk, kkangwing@gmail.com, Rynson.Lau@cityu.edu.hk.*
- *P. Chen is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. E-mail: prilochen@sjtu.edu.cn.*

† *Ke Xu, Lizhuang Ma, and Rynson W.H. Lau are the corresponding authors. Rynson Lau leads this project.*
Manuscript received xx xx, 2021; revised xx xx, 2021.

the image-level). They use a ResNet network [13] to classify whether an image is flipped or not. A high classification accuracy indicates the existence of visual chirality on this distribution. Inspired by this visual chirality property, we propose in this work to apply it for mirror detection, as the effect of mirror reflection on the content is similar to that of flipping. However, as their work [22] classifies visual chirality at image-level, the classification may only provide coarse localization of mirrors when applied to mirror detection. Although they propose a computational method, called "commutative residual" to indicate the existence of visual chirality at the pixel-level, this method contains several non-differentiable operations that they are not able to embed it into their classification network. Our preliminary experiments show that directly applying their classification network on the mirror dataset MSD [44] to determine if each image contains mirrors or not achieves only 58.5% accuracy, just slightly higher than random guess.

To address the above problems, we rethink the definition of "visual chirality" in [22], and propose in this paper to reformulate the image-level visual chirality property as a pixel-level cue for mirror detection. Our key insight is that if we can define visual chirality on the content distribution *(at the pixel-level)*, the change of the visual chirality property over an image can be learned by a deep model and then used to detect the presence of mirrors and their locations.

To this end, we first propose a novel flipping-convolution-flipping (FCF) transformation to compute the commutative residual, which represents the visual chirality cue in the feature domain. It allows learning mirror features with the help of the visual chirality cue. We refer to these features as the *visual chirality features* for the rest of this paper, We then propose to integrate the reformulated commutative residual into a novel visual chirality embedding (VCE) module, which leverages the visual chirality features to locate the mirror regions. In addition, we also propose a novel visual chirality-guided edge detection (CED) module to enrich the mirror features with the visual chirality features to facilitate mirror boundary detection. Finally, we propose a neural network model, named *VCNet*, that integrates the VCE and CED modules for mirror detection. From Figure 1, we can see that VCNet with the visual chirality cue (Ours) correctly detects the mirror regions, without being confused by the surrounding contents. We perform extensive experiments on three standard benchmark datasets to evaluate the performances of our method, and show that VCNet outperforms state-of-the-art methods on mirror detection.

In summary, this paper has three main contributions:

1) We consider "visual chirality" as an intrinsic mirror property for mirror detection. We reformulate it from its original image-level definition to a pixel-level definition for segmenting mirrors. We represent it via a novel flipping-convolution-flipping (FCF) transformation.
2) We propose the *VCNet* model for mirror detection, with a novel visual chirality embedding (VCE) module to leverage visual chirality features to locate the mirror regions and a novel visual chirality-guided edge detection (CED) module to leverage visual chirality features to guide the detection of mirror boundaries.
3) We conduct extensive experiments to evaluate the proposed method and show that it outperforms state-of-the-art methods on three mirror detection benchmarks.

2 Related Work

In this section, we briefly introduce the mirror detection problem, and discuss chirality and visual chirality. Since the mirror detection problem is a rather new problem, we also review two closely related problems, salient object detection and shadow detection.

Mirror Detection. Recently, two methods are proposed to address the mirror detection problem from a single RGB image. Yang *et al.* [44] propose the first mirror detection network, MirrorNet, to detect mirrors by modeling the content discontinuity between mirror/non-mirror regions. They mainly leverage the multi-level contextual contrasted information between mirror and non-mirror regions to detect mirrors in a coarse-to-fine manner. Lin *et al.* [21] further propose to consider content correspondences between the contents inside and outside of the mirror via a relational contextual contrasted local module, with an additional module to explicitly learn mirror boundaries. Despite the success, these two methods can easily over- or under-detect mirror regions in real world scenes, where most images likely contain some ambiguous contents. This is because these methods mainly focus on learning the relations between mirror/non-mirror regions, without considering any intrinsic mirror properties, resulting in incorrect detection of mirror/non-mirror regions.

Most recently, Mei *et al.* [25] proposed PDNet, which jointly exploits RGB and depth information to leverage the depth discontinuity between mirror and non-mirror regions for mirror detection. However, there are three fundamental limitations with this work. First, capturing the depth information requires additional equipment, which may not always be possible in practice. Second, the captured depth information is typically unreliable. Third, the contents in both mirror and non-mirror regions may also exhibit depth discontinuities, causing their model to fail.

In this paper, we propose to address the mirror detection problem by exploiting the intrinsic visual chirality property of mirrors to learn discriminative mirror features for mirror detection. Extensive experiments show that our method outperforms the above state-of-the-art methods, which rely on mirror/non-mirror relation learning.

Chirality / Visual Chirality. Chirality (or geometry chirality) describes an asymmetry property in many sciences, including Mathematics, Physics, Chemistry and Biology. As quoted from [18], where the term *chirality* was first introduced, "any geometrical figure, or group of points" is *chiral* if "its image in a plane mirror, ideally realized, cannot be brought to coincide with itself." This term is then applied to different sciences. For example, in Mathematics, it is defined as a figure that cannot be mapped to its mirror image by rotations and translations alone [30].

Recently, Lin *et al.* [22] introduce *visual chirality*, which is defined as a measure of the approximation error caused by

assuming that visual distributions are symmetrical under flipping (or reflection), for the purpose of relating visual chirality to data augmentation for analysis. They further demonstrate that such an error can be experimentally determined by applying a CNN to classify if an image is flipped or not.

The following example may help differentiate these two concepts. A human left hand is always geometrically chiral, as it cannot be exactly mapped to its mirror image by rotation and translation. However, if the visual distribution includes samples of both left and right hands, the distribution is defined as visual achiral, as these samples form reflective pairs and can occur with similar frequency. In contrast, if the visual distribution includes only one hand, the distribution is visual chiral as the distribution will be different due to flipping.

In [49], a direct application of visual chirality [22] for freehand sketch recognition is proposed. It aims to assign image-level labels for recognizing sketches. In this paper, we discuss how to reformulate the visual chirality property for mirror detection. Our method aims to assign pixel-level labels for locating mirrors. Our experiments show that the image-level method in [49] does not work on our problem.

Salient Object Detection / Shadow Detection. While salient object detection (SOD) is to detect visually distinctive objects in an image, shadow detection is to detect shadow regions in an image.

In the SOD task, early works mainly detect salient objects based on prior knowledge, such as center prior [12], background prior [40] and color prior [33]. Recent deep learning based methods propose to learn saliency features for SOD. He *et al.* [14] formulate a superpixel-wise CNN to address the SOD problem. Zhao *et al.* [47] propose to guide the SOD task by jointly modelling complementary salient object information and salient edge information. Qin *et al.* [32] further propose a boundary-aware network to exploit boundary information for SOD. Pang *et al.* [28] propose to integrate deep features from adjacent network layers to exploit contextual information for SOD. Most recently, Siris *et al.* [35] propose a context-aware learning approach to explicitly exploit semantic scene contexts for SOD.

In the shadow detection task, recent methods also exploit deep representations of shadow features. Hu *et al.* [15] propose to detect shadows by exploiting the spatial contextual information in a direction-aware manner. Zhu *et al.* [53] propose to aggregate spatial contextual information of shadows by fusing deep features of every two adjacent layers in both top-down and bottom-up directions. Zheng *et al.* [48] propose to explicitly learn the distraction-aware shadow features, by formulating distractions as false predictions of existing methods. Most recently, Zhu *et al.* [54] observe that intensity plays an important role in existing shadow detection methods, and propose to decompose the input into intensity-variant and intensity-invariant features in order to learn the shadow features better.

While SOD and shadow detection aim to learn the salient objects or shadow features, mirrors do not have a consistent visual appearance to learn. Hence, these SOD and shadow detection methods may not be straightforwardly adapted to address the mirror detection problem.

3 Visual Chirality

In [22], Lin *et al.* introduce visual chirality to explain the potential asymmetric problem of the flipping operation used in data augmentation. In this paper, we apply this concept for mirror detection. In this section, we first summarize the key idea of visual chirality as described in [22]. We then rethink the definition of visual chirality in order to apply it in our mirror detection problem.

3.1 The Original Visual Chirality

Given an image dataset and its distribution denoted as $\mathbf{D}$, if we flip the images in this dataset (*e.g.*, for the purpose of data augmentation), we typically assume that the augmented new distribution $\mathbf{D}'$ (*i.e.*, the distribution of the original dataset plus the flipped dataset) is symmetrical with respect to the flipping. However, Lin *et al.* [22] find that a deep network can easily distinguish if an image from $\mathbf{D}'$ is the original or flipped version of the image. This means that $\mathbf{D}'$ is not symmetrical, and Lin *et al.* [22] define the error between the real asymmetric $\mathbf{D}'$ and the assumed symmetric $\mathbf{D}'$ as *visual chirality*. In other words, there is at least one sample x_1 in $\mathbf{D}'$ that satisfies $D'(x_1) \neq D'(T(x_1))$, where $\mathbf{T}$ is the image flipping operation.

To measure the visual chirality value, Lin *et al.* [22] use a ResNet network to classify if an image is the original or flipped version. They apply the Class Activation Maps (CAM) [50], which is a visualization method, on the classification features to highlight the discriminative regions where visual chirality dominates.

A straightforward idea to incorporate the visual chirality cue for mirror detection is to use the visual chirality dominated regions to guide mirror detection, by combining the classification features in [22] with the mirror detection features of an existing mirror detection method. To do this, we first augment the training images with horizontal flipping and train a ResNet network to classify if an input image is flipped or not, as suggested by [22]. However, from our experiment, the accuracy of such a classification is only 58.5% on MSD [44]. The main reason for the low accuracy is that each image in MSD contains both mirror region (with flipped content) and non-mirror region (with non-flipped content). Hence, the classification result is then significantly disturbed by the relative size between the mirror and non-mirror regions of each image. Nonetheless, after training the classification network, we then incorporate its classification features into MirrorNet [44] and PMDnet [21] for mirror detection. Note that the classification network has the same backbone network as those used in MirrorNet and PMDnet, and we can fuse the classification features with the mirror detection features from MirrorNet or PMDnet via element-wise summation.

The second and fifth rows of Table 1 show the performances of this straightforward design for mirror detection. We are surprised to see that such a strategy not only fails to improve the performance, but instead significantly lowers it. There are two main reasons for the performance degradation. First, the flipping of the training samples is at the image-level, which cannot help locate the mirror inside an image which is a pixel-level problem. Second, according to [22], the visual chirality dominated regions (visualized

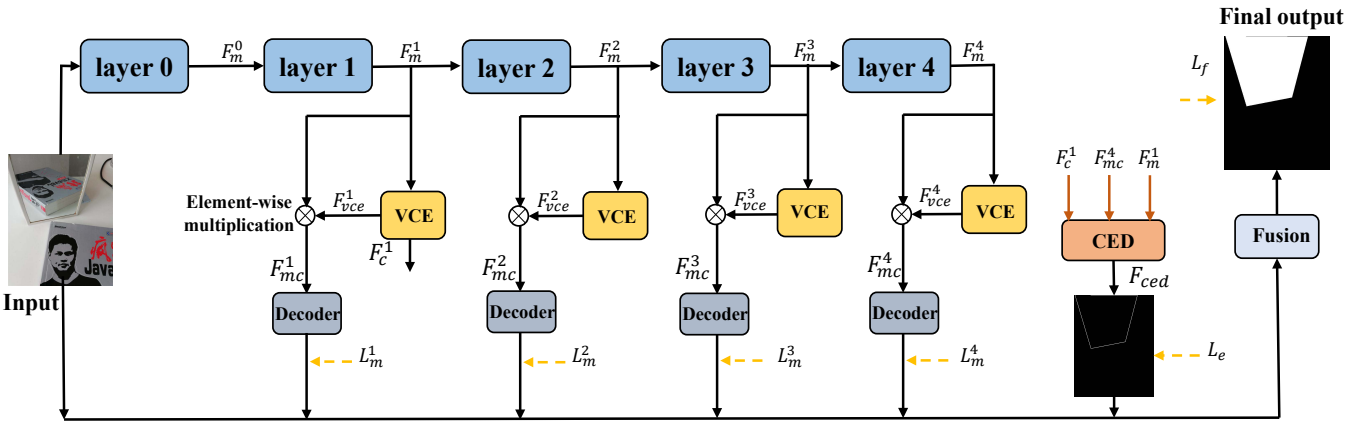


Fig. 2. The pipeline of VCNet. Given an input image, it first extracts multi-scale backbone features, and then transforms these features into multi-level visual chirality features via the proposed (VCE) modules. These features guide the backbone features to produce discriminative mirror features, which are then used by the decoders to detect mirrors at different scales. In addition, we also use the visual chirality features at different levels, *i.e.*, F_m^1, F_c^1 and F_{mc}^4, to learn mirror boundary information via the proposed CED module. Finally, we fuse the predicted mirror maps at different scales, the predicted mirror boundary map, and the input image using a fusion layer to obtain the final mirror map.

by CAM) are very coarse, which cannot help delineate the mirror boundaries.

In addition, we conduct an experiment by applying [49] to the mirror detection task. As [49] proposes the VCA module to apply the image-level visual chirality [22] for recognizing freehand sketches, we incorporate the VCA module from [49] in the existing networks (*i.e.*, MirrorNet [44] and PMDNet [21]). The third and sixth rows of Table 1 demonstrate that [49] does not work for our problem, as it is based on the original image-level visual chirality.

Nevertheless, this preliminary investigation inspires us to reformulate visual chirality for mirror detection.

TABLE 1
Comparing different ways of applying visual chirality [22] for mirror detection, on the MSD dataset [44] and the PMD dataset [21]. Note that VCA [49] converts visual chirality into attention.

	MSD		PMD	
Methods	$F_\omega \uparrow$	MAE $\downarrow$	$F_\omega \uparrow$	MAE $\downarrow$
MirrorNet [44]	0.857	0.065	0.748	0.061
MirrorNet + Lin *et al.* [22]	0.763	0.081	0.612	0.082
MirrorNet + VCA [49]	0.833	0.075	0.706	0.079
PMDNet [21]	0.892	0.047	0.790	0.032
PMDNet + Lin *et al.* [22]	0.772	0.061	0.692	0.051
PMDNet + VCA [49]	0.856	0.067	0.765	0.042

3.2 Rethinking Visual Chirality in Mirror Detection

As mirrors reflect their surrounding contents, the contents inside and outside of a mirror naturally form the distribution in $\mathbf{D}'$. This inspires us to use visual chirality to help differentiate between mirror regions (with flipped contents) and non-mirror regions (with non-flipped contents) of an image. To apply it to our pixel-level mirror detection problem, we need to redefine visual chirality on the content distribution at the pixel-level. Due to the existence of visual chirality, such content distribution is also asymmetrical, and the content distribution of mirror regions is expected to be different from that of non-mirror regions. Our aim is to learn such a difference. To do this, we need to reformulate the computation of "commutative residual" to represent the visual chirality cue at pixel-level in the feature domain.

FCF Transformation. As demonstrated in [22], the existence of visual chirality in an image X can be examined by the *commutative residual* $E(X)$ as:

$$E(X) = |\mathbf{J}(\mathbf{T}(X)) - \mathbf{T}(\mathbf{J}(X))|, \tag{1}$$

where $\mathbf{T}$ is a flipping transformation operation, and $\mathbf{J}$ is an image processing operation. A non-zero *commutative residual* map indicates the existence of visual chirality, as its existence makes these two operations not reversible.

Such a formulation, however, cannot be directly applied to our mirror detection network for two reasons. First (and also the main reason), this formulation is defined at image-level and is conditioned on the image processing operation $\mathbf{J}$ (*e.g.*, demosaicing or JPEG compression), *i.e.*, the detected chirality information changes with respect to the image processing operation $\mathbf{J}$ used. The presence of $\mathbf{J}$ here also makes this formulation non-differentiable for end-to-end learning of mirror features. Second, this formulation changes the spatial contextual information of mirrors, as the flipping operation $\mathbf{T}$ is applied to both terms in Eq. 1.

In order to apply the visual chirality information to our mirror detection problem, we propose to re-formulate Eq. 1. We note that the image processing operation $\mathbf{J}$ should satisfy two criteria. First, it should work as a pixel-level operator. Second, it should be learnable so that it could be end-to-end optimized to provide reliable mirror localization. Hence, we define $\mathbf{J}$ as the convolution operation. We also need to re-consider the usage of the flipping operation $\mathbf{T}$ to keep the spatial contexts consistent, as we want the commutative map to capture only the changes in the visual chirality property. To do this, we propose to add an additional flipping operation to align the spatial contexts such that the commutative residual $E(X)$ represents only the change in visual chirality. Hence, given the input features F, we define the proposed FCF transformation as:

$$E(F) = |\mathbf{C}(F) - \mathbf{T}(\mathbf{C}(\mathbf{T}((F))))|, \tag{2}$$

where $\mathbf{C}$ represents the convolution operation. Eq. 2 allows our model to learn mirror features with the help of the visual chirality cue.

Note that computing the commutative residual map does not impose any assumptions on operation $\mathbf{J}$, according to the three Propositions in [22]. Briefly, Propositions 1 and 2 in the Supplemental of [22] show that $\mathbf{J}$ preserves the symmetry of $\mathbf{D}$ if $\mathbf{J}$ is commutative. Proposition 3 further shows that even if $\mathbf{J}$ and $\mathbf{T}$ are not commutative, $\mathbf{J}$ will still preserve the symmetry property of $\mathbf{D}$ with respect to $\mathbf{T}$. Hence, replacing image processing operations with a differentiable convolutional operation does not affect whether $\mathbf{D_J}$ is symmetrical with respect to $\mathbf{T}$ or not. This allows us to use a convolutional operation to compute the commutative residual map in our FCF Transformation.

4 Our Model

Figure 2 shows our proposed mirror detection network. The key idea of our network design is to leverage visual chirality as an intrinsic mirror property for mirror detection. We first apply a backbone network [43] to extract multi-scale (5 scales in total) image features F_m^i $(i = 0, 1, 2, 3, 4)$ from the input image. We follow previous mirror detection methods [44], [21] to detect mirrors at different scales. Since the coarsest backbone features usually contain noisy information, we detect mirrors in the 2^{nd} to 5^{th} scales. For the backbone features from each of these scales, we assign a VCE module to learn visual chirality features F_{vce}^i $(i = 1, 2, 3, 4)$ for mirror detection and a decoder to predict the mirror map. In addition, we propose a CED module to leverage both image features and visual chirality features to predict mirror boundaries. Finally, we use a fusion module to fuse the results from mirror detection and mirror boundary detection to produce the final mirror mask.

4.1 Visual Chirality Embedding (VCE) Module

Our VCE module aims to extract the visual chirality based features for mirror detection. Based on the proposed FCF Transformation in Eq. 2, we design the VCE module to learn visual chirality features from the mirror features. As shown in Figure 3, the VCE module consists of two extractors: the dilated feature extractor and the visual chirality feature extractor. The dilated feature extractor aggregates multi-scale spatial information of the input backbone features F_m to compute the visual chirality features, which are then applied as an attention to help locate the mirror regions. To this end, the dilated feature extractor outputs three kinds of features, basic mirror features F_{bsc}, local features F_{loc}, and context features F_{con}, which are then sent to the visual chirality feature extractor to mine the visual chirality cue to generate visual chirality features F_c. Finally, F_{bsc} is combined with F_c to obtain the final output F_{vce}.

For the dilated feature extractor, given the backbone features F_m, it uses convolutional layers with different dilation rates in a pyramidal structure to learn richer information. As a convolution with a larger dilation rate provides a larger receptive field for learning more global information, we employ four convolution layers with dilation rates of 1, 2, 4 and 6. We then concatenate and process the output

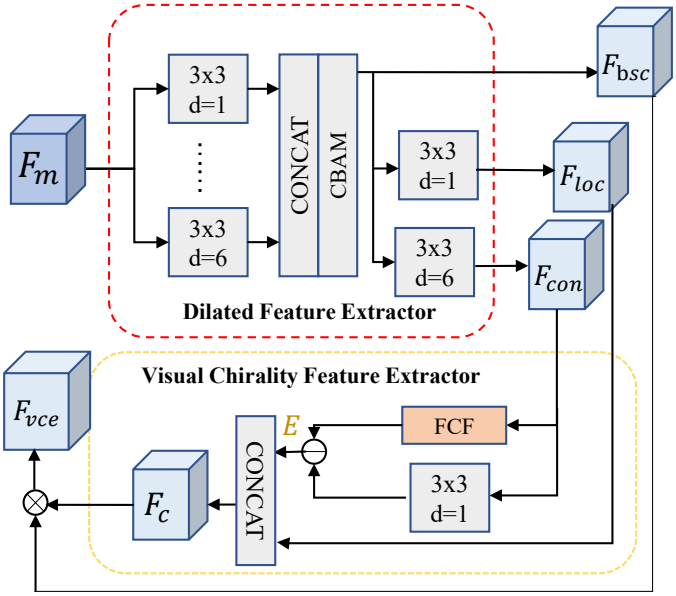


Fig. 3. The proposed visual chirality embedding (VCE) module. It aims to transform the input backbone features F_m into visual chirality features F_{vce} via a dilated feature extractor, followed by a visual chirality feature extractor. The proposed FCF transformation is embedded in the visual chirality feature extractor to extract the visual chirality cue.

multi-level features by a batch normalization layer with ReLU activation, and use a CBAM [41] to refine the feature maps, as in [44], to obtain the basic mirror features F_{bsc}. We further process F_{bsc} using two different convolutional layers. One is a 3×3 convolutional layer with a dilation rate of 1 to generate the local features F_{loc}. Another is a 3×3 convolutional layer with a dilation rate of 6 to generate the context features F_{con}.

The inputs to the visual chirality feature extractor are F_{loc} and F_{con}. Since capturing the visual chirality information typically requires a good understanding of the context information, we extract visual chirality E from context features F_{con} with Eq. 2 as:

$$E = |\mathbf{C}(F_{con}) - \mathbf{T}(\mathbf{C}(\mathbf{T}(F_{con})))|. \quad (3)$$

We then condition the commutative residual E on the local features F_{loc} via concatenation to produce F_c, to enhance its per-pixel localization capability. Finally, we use F_c to spatially re-weight the basic mirror features F_{bsc} to produce the visual chirality features F_{vce} for locating mirror regions.

4.2 Visual Chirality-guided Edge Detection (CED) Module

Edge detection has shown to be useful in mirror detection by [21]. However, their work focuses only on learning the geometric edge features. We note that along a mirror boundary, the visual chirality property of the distribution over the mirror region is also different from that of the distribution over the non-mirror region. Based on this observation, we propose the CED module to leverage the visual chirality cue to help detect mirror boundaries.

Figure 4 shows the architecture of the CED module. It takes three types of features at different levels as inputs. First, as low-level image features are helpful to boundary

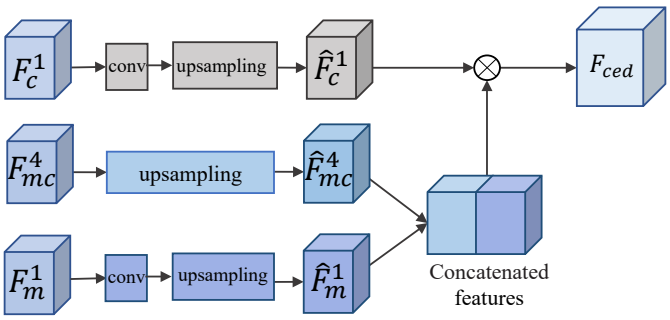


Fig. 4. The proposed visual chirality-guided edge detection (CED) module. It uses the extracted visual chirality based features (*i.e.*, F_c^1 and F_{mc}^4) to guide the detection of mirror edge features (*i.e.*, F_{ced}) from the mirror features (*i.e.*, F_m^1).

detection [21], we take F_m^1, extracted by the second backbone layer, as the low-level image features. We do not use the features from the first backbone layer as they are noisy. Second, high-level visual chirality cues are often identified by humans as mirrored texts or objects. Hence, we take F_{mc}^4 as the high-level visual chirality features, since it models both high-level semantics and high-level visual chirality cues Third, we take F_c^1 as the low-level visual chirality features, since it models low-level visual chirality cues.

We use three branches to process the three types of input features. The low-level visual chirality features F_c^1 are convolved by two convolutional layers, followed by an up-sampling operation to produce $\hat{F}_c^1$. The high-level visual chirality features F_{mc}^4 are simply processed by an up-sampling operation to align its spatial resolution with $\hat{F}_c^1$, denoted as $\hat{F}_{mc}^4$. The low-level mirror features F_m^1 are also convolved by two convolutional layers to produce $\hat{F}_m^1$.

To leverage the visual chirality features to guide the edge detection step, we first use the high-level visual chirality features $\hat{F}_{mc}^4$ to enrich the mirror features via concatenation, and then use the low-level visual chirality features $\hat{F}_c^1$ to further enrich the concatenated features through element-wise multiplication, to produce the output features F_{ced}.

To supervise our CED module, we need to prepare the ground truth mirror edge maps. To do this, we apply the Canny edge detector to extract mirror edges from the ground truth mirror masks.

4.3 Details of Other Modules

Two other modules are used in our network: decoder and fusion module.

Decoder. Each decoder in Figure 2 consists of a deconvolutional layer, a CBAM block and a convolutional layer with up-sampling operation to convert the feature maps into a 1-channel map.

Fusion Module. Our fusion module in Figure 2 is an 8-channel convolutional layer with a 1×1 kernel. We concatenate the input image (3 channels), the outputs of four decoders (1 channel per-decoder output), and the output of CED (1 channel), and then feed it into the Fusion module to produce the final output.

4.4 Loss Function

We adopt the Lovász-Softmax loss function [3] to compute the difference between the mirror prediction and ground truth mirror mask. We use the binary cross-entropy (BCE) loss function to measure the accuracy of the predicted edges from the CED module. The loss function is formulated as:

$$\mathcal{L} = \sum_{i=1}^{4} w_i L_m^i + w_e L_e + w_f L_f, \tag{4}$$

where L_m^i represents the lovasz-hinge loss for supervising the multi-level mirror maps predicted by these decoders. L_f is the lovasz-hinge loss for supervising the final mirror prediction. L_e is the BCE loss for edge prediction. w_i, w_e and w_f are the balancing parameters, which are empirically set to $w_i = 1$, $w_e = 100$ and $w_f = 1$.

5 Experiments

5.1 Experimental Settings

Implementation Details. We have implemented our model on the Pytorch framework and trained it on a PC with two RTX1080Ti cards. For a fair comparison with the existing mirror detection methods [44], [21], [25], we use ResNeXt101 [43] pretrained on ImageNet [7] as the backbone network to extract multi-level features. To train the proposed network, we initialize the learning rate as $1e-3$, and decay it with the poly learning rate policy. We use the SGD optimizer with a momentum of 0.9 and a weight decay of $5e-4$ for loss optimization. All input images are rescaled to a resolution of 384×384. The batch size is set to 10 for training. We also apply CRF [19] as post-processing to refine the mirror detection results. When training on the MSD dataset [44], we set the number of training epochs to 100. When training on the PMD dataset [21] and RGBD-Mirror [25], we set the number of training epochs to 160. During inference, we resize the network predictions to the same resolutions as the original images. From our experiments, the average inference time of our model is 0.13s for an input image of resolution 384×384 on a single GTX1080Ti card.

Evaluation Metrics. We apply two popular metrics, the F-measure and mean absolute error (MAE), to evaluate the performance of our model. The F-measure is defined as:

$$F_w = \frac{(1+\omega^2) * precision * recall}{\omega^2 * precision + recall}, \tag{5}$$

where w^2 is set to 0.3 as suggested in [21]. The MAE is defined as:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |\hat{M}(i,j) - \hat{G}(i,j)|, \tag{6}$$

where $\hat{M}$ and $\hat{G}$ are the predicted mirror map and the ground truth mirror mask, respectively. W and H are the image width and height.

5.2 Comparison on the MSD and PMD Datasets

Evaluation Datasets. We conduct our experiments on two available mirror datasets with RGB images, the MSD dataset [44] and the PMD dataset [21]. MSD contains 3,063 training images and 955 test images, which are mainly collected from indoor scenes and then manually annotated. PMD contains 5,096 training images and 571 test images. This

TABLE 2
Quantitative comparison between the proposed VCNet and 11 state-of-the-art methods, on the MSD dataset [44] and the PMD dataset [21]. Best results are marked in **bold**.

	MSD		PMD	
Methods	$F_\omega \uparrow$	MAE $\downarrow$	$F_\omega \uparrow$	MAE $\downarrow$
DSC [15]	0.812	0.087	0.511	0.123
BDRAR [53]	0.792	0.093	0.616	0.101
CCNet [16]	0.749	0.112	0.614	0.092
R^3Net [8]	0.846	0.068	0.646	0.057
CPDNet [42]	0.769	0.111	0.741	0.041
BASNet [32]	0.791	0.082	0.609	0.066
PoolNet [23]	0.785	0.099	0.580	0.090
EGNet [47]	0.802	0.086	0.672	0.087
MINet [29]	0.812	0.082	0.760	0.037
MirrorNet [44]	0.857	0.065	0.748	0.061
PMDNet [21]	0.892	0.047	0.790	0.032
Ours	**0.898**	**0.044**	**0.812**	**0.028**

dataset covers more diverse scenes and is more challenging, as it is collected from six different public datasets, including ADE20K [51], [52], NYUv2 [34], MINC [2], Pascal-Context [26], SUNRGBD [36], and COCO-stuff [4].

Methods for Comparisons. To study how well our model performs, we compare it with a variety of state-of-the-art models, including MirrorNet [44] and PMDNet [21] for mirror detection, DSC [15] and BDRAR [53] for shadow detection, R^3Net [8], CPDNet [42], BASNet [32], EGNet [47], PoolNet [23] and MINet [29] for saliency object detection, and CCNet [16] for semantic segmentation. We report the performances of MirrorNet [44] and PMDNet [21] on MSD by running their released pre-trained models. For the performances of other methods, we either obtain their performances from [21], if available, or train their models on the two evaluation datasets using their released codes.

Quantitative Evaluation. We first quantitatively compare the proposed method with the state-of-the-art methods. Table 2 shows the results. We can see that the proposed method achieves consistently better performances across the two datasets on both metrics. Particularly, our model achieves a notable performance gain on the more challenging PMD dataset [21]. The relatively small performance gain on MSD because MSD contains mostly simple indoor scenes that existing methods already perform well on them. The relatively large performance gain on PMD shows that the proposed model is particularly effective when applied on complex real-world scenes.

Qualitative Evaluation. We then qualitatively compare the proposed VCNet with the state-of-the-art methods. Figure 5 shows some visual results. The first five rows show some challenging examples, with objects that have visual patterns similar to the mirrors. In the first row, most methods fail to differentiate the mirror on the right from the window on the left, as these two regions have similar properties (*e.g.*, relatively higher intensity and regular boundary). They either miss both or detect both as mirrors. In the second row, the jacket inside the mirror confuses existing methods in detecting the mirror. Note that as there are no correspondences, PMDNet [21] also fails to locate the mirror. The third row shows an image with complex illuminations. Existing methods falsely detect the floor with reflected light as a mirror. The fourth image shows a mirror on the left and a confusing wooden frame on the right. Most existing methods are confused by this combination and are unable to identify the mirror region correctly. The fifth image shows a confusing setting. Most existing methods wrongly identify the left hand wall as the mirror. In contrast, our method can successfully detect the mirror regions correctly in all these confusing scenarios, with the help of the visual chirality cue. The last two rows show mirrors that are partially occluded by another objects. The sixth image shows a mirror with a photo frame in front of it. Due to similar appearances, most existing methods detect both of them as mirrors. The seventh image shows a mirror with an occluder of a distinct appearance in front of it. Existing methods consider the occluder as part of the mirror. In contrast, our method can detect the mirror regions well in both cases.

To sum up, these qualitative results show that our method has obvious advantages over state-of-the-arts in distinguishing mirrors from non-mirror regions in challenging scenarios, benefited from the mining of the visual chirality cue.

We further compare our method with the latest single-image mirror detection method, PMDNet [21], in both image and feature domains, on three images as shown in Figure 6. In (c), we visualize the relation contextual features learned by PMDNet, and in (d), we show the detection results by PMDNet. In (e), we visualize our learned visual chirality features F_c^1, and in (f), we show our detection results. In the first row, although PMDNet can correctly identify the correspondences between the cabinet on the right and the reflection (the center vertical color strip) in the mirror, it mistakenly considers the cabinet as the mirror, instead of the reflection. In contrast, our visual chirality features can help correctly locate the mirror region, allowing our model to detect the correct mirror region. Likewise, in the second row, PMDNet can also detect correspondences between the contents inside and outside of the mirrors, but mistakenly recognizes the outside region as the mirror. In contrast, our model can correctly detect the mirror region via the visual chirality features. In the third row, PMDNet is not able to detect the whole mirror as it fails to detect correspondences in the input image, while our model can segment the whole mirror region.

These results show that our learned visual chirality features can help distinguish the reflected objects from the real ones, resulting in correct detection of the mirror regions.

5.3 Comparison on RGBD-Mirror Dataset

Although PDNet [25] uses more expensive input data than ours (*i.e.*, RGBD images instead of RGB images), it would be interesting to compare our results with theirs.

Evaluation Dataset. We train all compared models on the RGBD-Mirror dataset [25], which is the latest RGBD-based mirror dataset with $3,049$ RGBD images and corresponding mirror masks. There are $2,000$ images for training

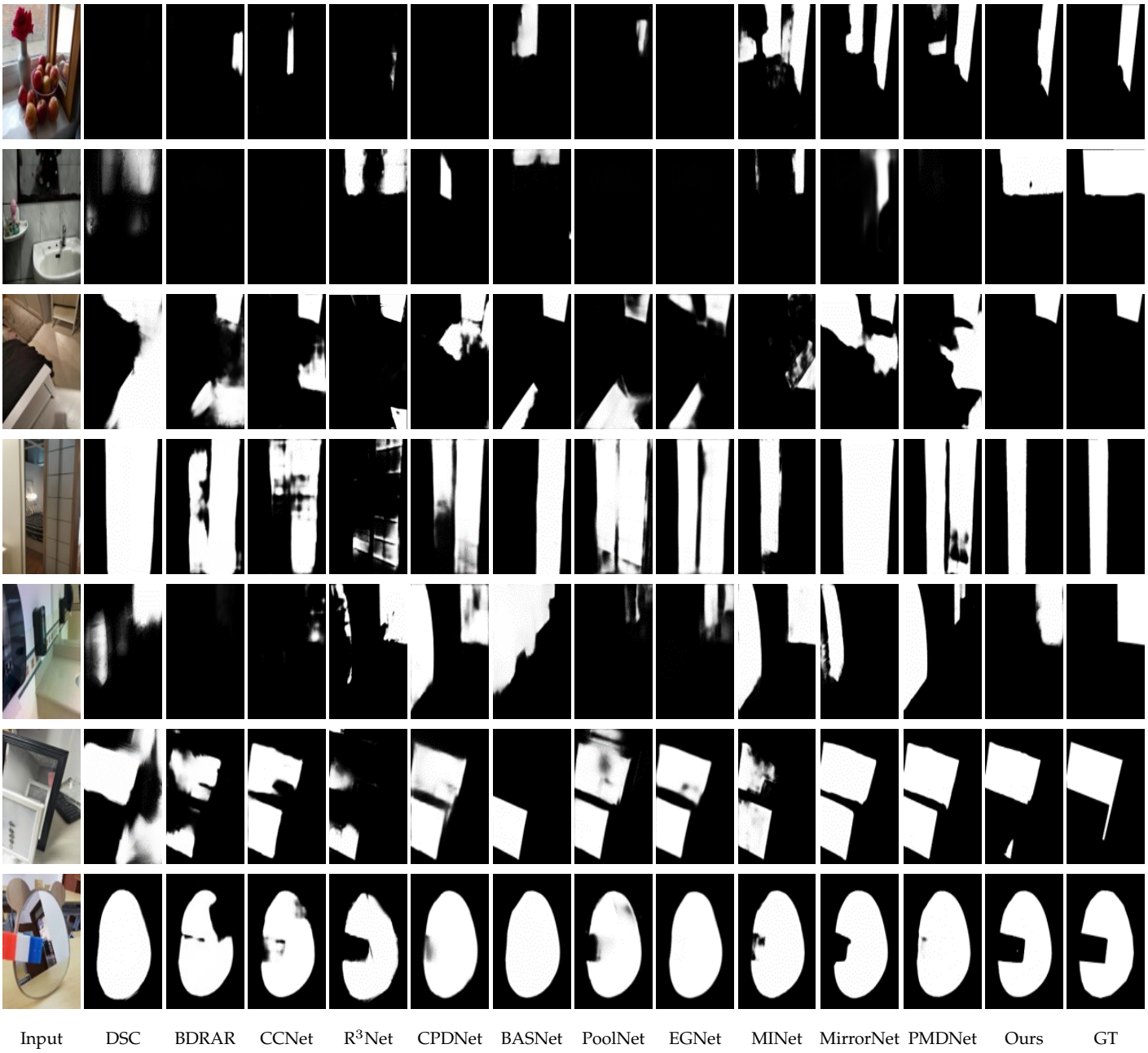
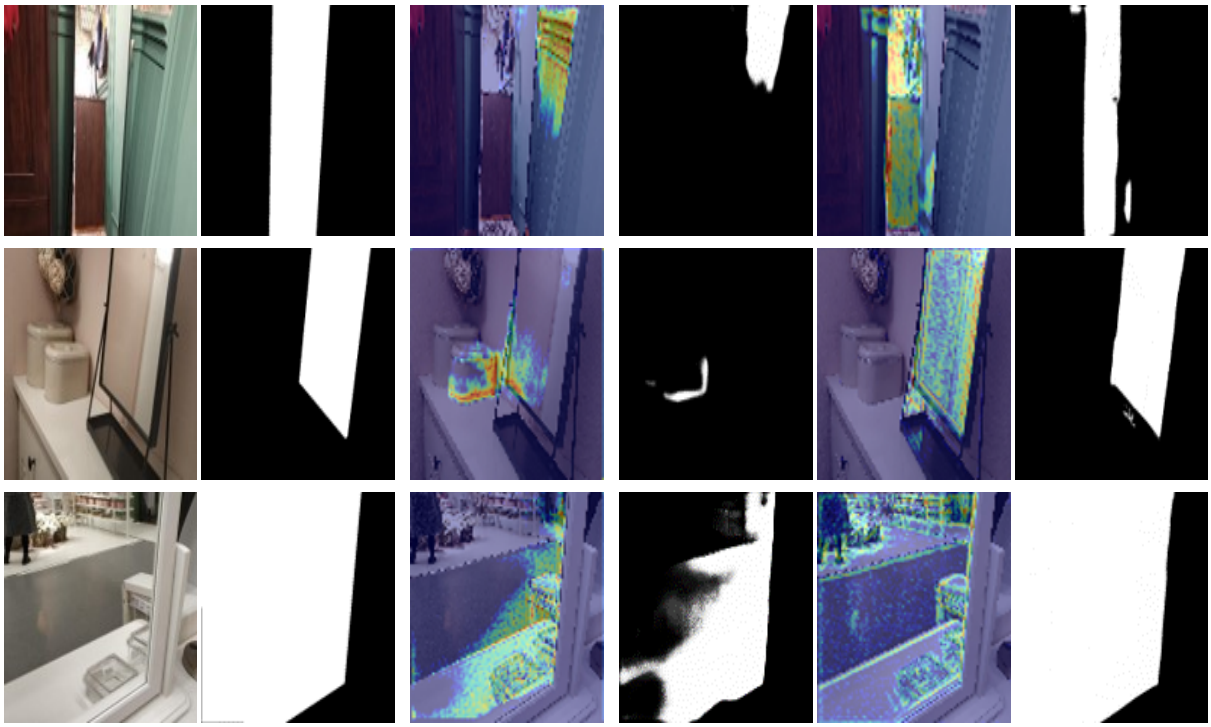


Fig. 5. Visual comparison between the proposed VCNet and 11 state-of-the-art methods on the MSD and PMD datasets. From left to right are the input images, results from DSC [15] and BDRAR [53] for shadow detection, CCNet [16] for semantic segmentation, R^3Net [8], CPDNet [42], BASNet [32], PoolNet [23], EGNet [47] and MINet [29] for salient object detection, MirrorNet [44] and PMDNet [21] for mirror detection, the proposed VCNet and the ground truth. Our VCNet can detect the mirror regions accurately.

and $1,049$ images for evaluation. The authors selected images with mirrors and their corresponding depth maps from 4 existing datasets, including Matterport3D [5], SUNRGBD [36], ScanNet [6], and 2D3DS [1]. Finally, there is one more dataset Mirror3D [37] proposed for refining the depth in mirror regions. However, the images used in this dataset [37] are already covered in the other datasets that we use for comparisons in our paper, as Mirror3D is constructed from images taken from Matterpot3D [5], NYUv2 [34] and ScanNet [6].

Methods for Comparisons. We compare our method with PDNet [25], which is the only RGBD-based mirror detection method. We also compare our model with eight state-of-the-art RGBD saliency detection methods, including S2MA [24], SSF [46], A2dele [31], CoNet [17], JL-DCF [11], HDFNet [27], ATSA [45], and BBS-Net [10]. For a comprehensive evaluation, we also include MirrorNet [44], PMD-Net [21] and the RGB image based PDNet [25] (w/o Depth).

Quantitative Evaluation. Table 3 shows the performance comparison on the RGBD-Mirror dataset. We can see that our model achieves the best performance on the F score, and the second best performance on MAE. Note that our method does not use the depth information in the dataset, while PDNet is specifically designed to take both RGB and depth information as inputs. For a fair comparison, we have also trained a version of PDNet without using the depth



(a) Input (b) GT (c) Features (PMDNet) (d) PMDNet (e) Features (Ours) (f) Ours

Fig. 6. Visual comparison between the proposed VCNet and PMDNet [21]. Although PMDNet can correctly detect the correspondences between inside and outside of the mirrors in the top two images, it wrongly identifies the outside regions as mirrors. In contrast, VCNet uses the visual chirality features to correctly identify the mirror regions in both cases. For the bottom image, PMDNet fails to detect the whole mirror region, due to the lack of correspondences, while VCNet can accurately locate the mirror region.

TABLE 3
Quantitative comparison between the proposed VCNet and other methods on the RGBD-Mirror dataset [25]. Note that all methods with a ✔were trained on the RGBD data, while methods with a ✗were trained on the RGB data only. Best results are marked in **bold** and the second best results are marked by an underline.

Methods	Depth	$F_\omega \uparrow$	MAE $\downarrow$
S2MA [24]	✔	0.677	0.071
SSF [46]	✔	0.599	0.097
A2dele [31]	✔	0.614	0.087
CoNet [17]	✔	0.576	0.120
JL-DCF [11]	✔	0.727	0.065
HDFNet [27]	✔	0.549	0.095
ATSA [45]	✔	0.664	0.090
BBS-Net [10]	✔	0.736	0.059
MirrorNet [44]	✗	0.723	0.062
PMDNet [21]	✗	0.775	0.054
PDNet [25] (w/o Depth)	✗	0.783	0.053
PDNet [25]	✔	<u>0.825</u>	**0.042**
Ours	✗	**0.849**	<u>0.052</u>

information, *i.e.*, PDNet (w/o Depth). We can see that our model now outperforms it on both metrics. These results demonstrate the importance of our visual chirality features on mirror detection.

Qualitative Evaluation. Figure 7 shows the visual comparison between our results and the results from PDNet [25] (trained and tested with RGBD images). In the first row, PDNet misses the right mirror region as the depth information exhibits huge discontinuity within the mirror region. In the second row, PDNet mis-recognizes the door as one of the mirrors since the door region also has a large depth discontinuity to its surroundings. In the third row, the real desk has very similar depth information as its reflection inside the mirror, which fails PDNet. Our model successfully detects the correct mirror regions in all these challenging scenes, showing that the visual chirality cue is more robust than the depth cue for detecting mirrors in real world scenes.

5.4 Model Analysis

Ablation Study. We have conducted ablated experiments to verify the effectiveness of our design. First, we remove the proposed VCE and CED modules to form the baseline for comparison. We then add the proposed VCE module as an ablated version, denoted as "Baseline+VCE". To evaluate the proposed FCF formulation, we include an ablated version by removing the FCF transformation from the VCE module, denoted as "Baseline+VCE (w/o FCF)". To demonstrate the effectiveness of conditioning the commutative residual E on the local features F_{loc}, we have conducted an additional ablation study by removing the local features F_{loc} from the VCE module, denoted as "Baseline+VCE (w/o F_{loc})". To study the proposed CED module, we add different levels of mirror and visual chirality features one by one. We first separately apply the low level

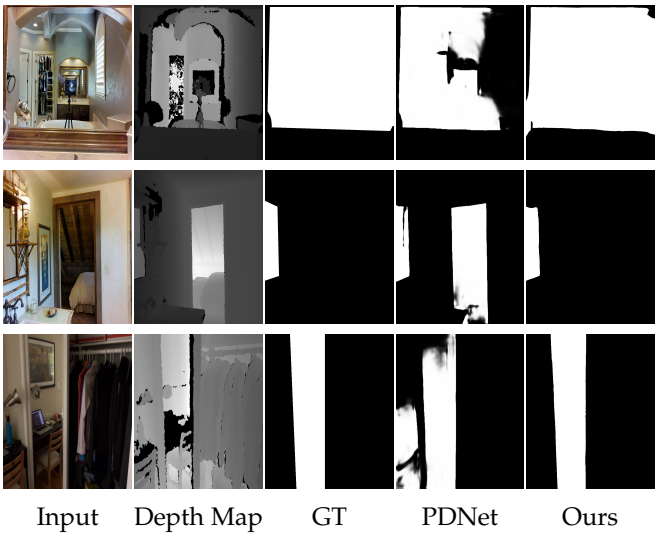


Fig. 7. Visual comparison between our model and PDNet [25] on RGBD-Mirror dataset. Note that our model does not use the depth information while PDNet uses it.

mirror features, high-level mirror/visual chirality features, and low-level visual chirality features, denoted as "Baseline+VCE+lm", "Baseline+VCE+hmc", "Baseline+VCE+lc", respectively. We then combine the low-level mirror features with high-level mirror/visual chirality features, denoted as Baseline+VCE+lm+hmc, to study the removal of the low-level visual chirality features on the detection performance. Finally, we compare with our full model, *i.e.*, "Ours".

Table 4 shows the performances of different ablated models. We can see that adding the VCE module (4^{th} row) introduces obvious improvements over the Baseline model (1^{st} row) on both F_ω and MAE, which demonstrates the effectiveness of mining the visual chirality cue for mirror detection. However, if we remove the proposed FCF Transformation (2^{nd} row), the performance of the model is similar to that of the Baseline model, indicating the importance of learning visual chirality. In addition, if we only remove the local feature F_{loc} (3^{rd} row), the performance is better than that of "Baseline+VCE (w/o FCF)", which indicates the contribution of FCF. Further, we can see that using only a single level of mirror and visual chirality features (5^{th}-7^{th} rows) in the CED module degrades the performance, as predicting the mirror boundaries typically requires a comprehensive understanding of multi-level information. By adding the low-level mirror features to the CED module (8^{th} row), the performance increases. Finally, by further adding the low-level visual chirality information to the CED module to become our full model (9^{th} row), there is a more significant jump in performance, showing that the multi-level learning of the visual chirality cue is important to mirror detection.

In Table 4, we observe that our baseline outperforms most of the compared methods, except PMDNet [21]. In particular, compared to MirrorNet [44], our baseline method fuses multi-scale features with the input image via a light-weight fusion module, while MirrorNet directly fuses the features from the shallow layers to those of the deep layers without the input image. This straightforward adaption

TABLE 4
Ablation study of the proposed FCF transformation, VCE module and CED module, on the MSD dataset. Best results are marked in **bold**.

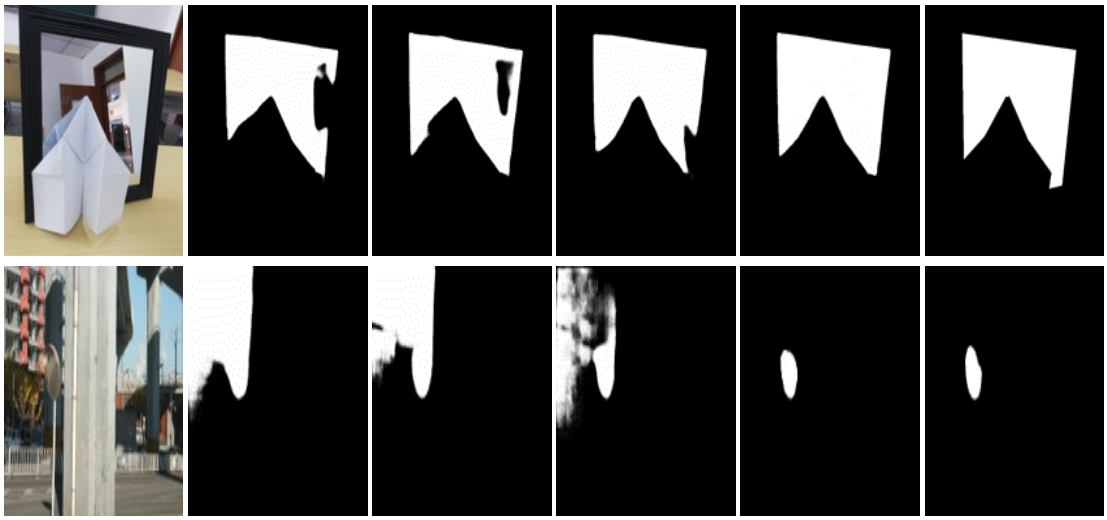
	Ablation	$F_\omega \uparrow$	MAE $\downarrow$
1	Baseline	0.873	0.062
2	Baseline+VCE (w/o FCF)	0.875	0.060
3	Baseline+VCE (w/o F_{loc})	0.878	0.058
4	Baseline+VCE	0.880	0.053
5	Baseline+VCE+lm	0.874	0.061
6	Baseline+VCE+hmc	0.875	0.060
7	Baseline+VCE+lc	0.880	0.057
8	Baseline+VCE+lm+hmc	0.887	0.050
9	Ours	**0.898**	**0.044**

results in a notable performance gain. By replacing our fusion strategy with the one used in MirrorNet, the F_ω performance of our baseline method drops from 0.873 to 0.847, and the MAE performance drops from 0.062 to 0.072.

Figure 8 shows the visual results of different ablated models on two challenging images. In the first row, we can see that the Baseline model only locates the rough mirror region. Introducing the VCE module can help recover the missing part, with the help of the visual chirality cue. By further incorporating the proposed CED module, our full model is able to correctly detect the mirror region. In the second row, we can see that the Baseline model produces incorrect predictions. Gradually incorporating the proposed modules can help correct the prediction errors and remove the non-mirror regions. These visual results again demonstrate the effectiveness of the proposed modules.

Study of the CED module. To demonstrate the motivation of using special features in our CED module, *i.e.*, the input features used in the CED module, to detect mirror boundaries, we conduct another two experiments with different feature sets. Since low-level image features are helpful to boundary detection [21], we take F_m^1, extracted by the second backbone layer, as one of the inputs to the CED module. As Table 5 shows, if we replace F_m^1 with F_m^0, the F_ω performance for the best result drops from 0.898 to 0.891, and the MAE performance drops from 0.044 to 0.049. This is because F_m^0 from the first backbone layer is noisy. F_{mc}^4 is a combination of the backbone features from layer 4 and the corresponding visual chirality features. Since high-level visual chirality cues are often identified by humans as mirrored texts or objects, we take F_{mc}^4 as the high-level visual chirality features since it models both high-level semantics and high-level visual chirality cues. As shown in Table 5, if we replace F_{mc}^4 with F_c^4, the F_ω performance of the best result drops from 0.898 to 0.885, and the MAE performance drops from 0.044 to 0.058. This indicates that both high-level image features and visual chirality features are important.

To summarize, we use F_m^1, F_c^1, and F_{mc}^4 to represent low-level image features, low-level visual chirality cues, and high-level semantics and visual chirality cues, respectively, for detecting mirror boundaries. The ablation study in Table 4 shows that the multi-level inputs to the CED module (ablated models 5, 6, 7 and 8) continuously improve the mir-



(a) Input (b) Baseline (c) Base.+VCE (w/o FCF) (d) Base.+VCE (e) Ours (f) GT

Fig. 8. Visual comparison of different ablated models on two challenging images. (b) is the baseline model, without the VCE and CED modules. (c) and (d) add the VCE module to the baseline model, but without and with the FCF Transformation, respectively. (e) is our full model, while (f) is the ground truth.

TABLE 5
The motivation of using special features in our CED module.

	Feature Sets	$F_\omega \uparrow$	MAE $\downarrow$
lm feature	$F_m^1 + F_{mc}^4 + F_c^1$	0.898	0.044
	$F_m^0 + F_{mc}^4 + F_c^1$	0.891	0.049
hmc feature	$F_m^1 + F_{mc}^4 + F_c^1$	0.898	0.044
	$F_m^1 + F_c^4 + F_c^1$	0.885	0.058

ror detection performance. In addition, the experiments in Tables 2 and 3 also demonstrate the generalization capability of the module on all existing mirror detection datasets.

(a) Input (b) GT (c) Ours

Fig. 9. Failure cases. From left to right are the input images, ground truths and our results. Our method may fail to accurately delineate the mirror boundaries if the mirror is partially occluded by a complex object (1^{st} row) or by a very small object (2^{nd} row).

6 Conclusion

In this paper, we have investigated the visual chirality cue for mirror detection. We have proposed a novel flipping-convolution-flipping (FCF) transformation to model the visual chirality cue in the feature domain. Based on this formulation, we have proposed a novel visual chirality embedding VCE module for locating mirror regions, and a novel visual chirality-guided edge detection (CED) module for refining the mirror boundaries. Extensive experiments have shown that our model outperforms state-of-the-art relevant methods on two RGB mirror benchmark datasets. We have further compared our model with the latest RGBD mirror detection model. Even though the RGBD mirror detection model uses additional depth information, our model performs comparable to it. Ablation studies have also verified the effectiveness of model design.

Our method does have some limitations. As shown in Figure 9, although our method can locate the mirror regions by learning the visual chirality features, if the mirror is partially occluded by an object with a complex structure (top row), it may not have sufficient confidence in every location to determine the mirror boundary accurately. For the same reason, if the occluding object is too small, it may not have sufficient confidence in every location to determine the occluded region as a non-mirror region.

As a future work, our main goal is to address the current failure cases. Since our model may fail to delineate boundaries due to the occlusion of mirrors by fine/tiny objects, we are interested in developing a mechanism that can zoom in the local regions with occlusions, and apply the visual chirality cue to localize the mirror regions. In the long run, we are interested in exploring the possibilities of applying the visual chirality property in other related computer vision tasks, for example, the reflection removal [9] task and the night-time semantic segmentation [38] task.

Acknowledgments

This work is partially supported by the National Key Research and Development Program of China (No.

2019YFC1521104), National Natural Science Foundation of China (No. 61972157, 72192821, 62106268), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), a GRF (RGC Ref: 11205620) from the Hong Kong Research Grants Council, a SRG (Ref: 7005674) from City University of Hong Kong, Shanghai Sailing Program (22YF1420300) and the High-Level Talent Program for Innovation and Entrepreneurship (ShuangChuang Doctor) of Jiangsu Province (No. JSSCBS20211220). Xin Tan is supported by the Postgraduate Studentship (Mainland Schemes) from the City University of Hong Kong.

REFERENCES

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.
[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, pages 3479–3487, 2015.
[3] 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.
[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 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 *International Conference on 3D Vision (3DV)*, pages 667–676, 2017.
[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.
[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
[8] 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 *AAAI*, pages 684–690, 2018.
[9] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson Lau. Location-aware single image reflection removal. In *ICCV*, pages 5017–5026, 2021.
[10] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292, 2020.
[11] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, pages 3052–3062, 2020.
[12] shan Gao, Vijay Mahadevan, and Nuno Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NeurIPS*, pages 497–504, 2008.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
[14] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 115(3):330–344, 2015.
[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] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, 2019.
[17] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020.
[18] William Thomson Baron Kelvin. *The molecular tactics of a crystal*. Clarendon Press, 1894.
[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011.
[20] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *CVPR*, pages 2186–2196, 2020.
[21] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *CVPR*, pages 3697–3705, 2020.
[22] Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *CVPR*, pages 12295–12303, 2020.
[23] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3912–3921, 2019.
[24] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *CVPR*, pages 13756–13765, 2020.
[25] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021.
[26] 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*, pages 891–898, 2014.
[27] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pages 235–252, 2020.
[28] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.
[29] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9410–9419, 2020.
[30] Michel Petitjean. Chirality in metric spaces. in memoriam michel deza. *Optimization Letters*, 14(2):329–338, 2020.
[31] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, pages 9060–9069, 2020.
[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.
[33] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, pages 853–860, 2012.
[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012.
[35] Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, and Rynson W.H. Lau. Scene context-aware salient object detection. In *ICCV*, 2021.
[36] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.
[37] Jiaqi Tan, Weijie Lin, Angel X Chang, and Manolis Savva. Mirror3d: Depth refinement for mirror surfaces. In *CVPR*, pages 15990–15999, 2021.
[38] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021.
[39] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson Lau. Learning to detect instance-level salient objects using complementary image labels. *International Journal of Computer Vision*, 130:729–746, 2022.
[40] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42, 2012.
[41] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.
[42] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.
[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.
[44] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *ICCV*, pages 8809–8818, 2019.
[45] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *ECCV*, pages 374–390, 2020.
[46] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *CVPR*, pages 3472–3481, 2020.
[47] 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*, pages 8779–8788, 2019.
[48] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *CVPR*, pages 5167–5176, 2019.
[49] Ying Zheng, Yiyi Zhang, Xiaogang Xu, Jun Wang, and Hongxun Yao. Visual chirality meets freehand sketches. In *ICIP*, pages 1544–1548. IEEE, 2021.
[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.
[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.
[52] 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, 2019.
[53] 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*, pages 121–136, 2018.
[54] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W.H. Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *ICCV*, 2021.

Xin Tan is currently with East China Normal University, China. He received dual Ph.D. degrees in Computer Science from Shanghai Jiao Tong University and City University of Hong Kong. He received his B.Eng. degree in Automation from the Chongqing University, China in 2017. His research interests lie in computer vision and deep learning. He serves as a program committee member/reviewer for CVPR, ICCV, AAAI, IJCAI and International Journal of Computer Vision (IJCV).

Jiaying Lin is currently a PhD student in Computer Science at City University of Hong Kong. He received the B.Eng. degree in Computer Science and Technology from South China University and Technology in 2019. His research interests include computer vision and computer graphics. He serves as a program committee member/reviewer for CVPR, ECCV, IEEE Transactions on Image Processing (TIP) and IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).

Ke Xu is currently with the City University of Hong Kong. He obtains the dual Ph.D. degrees from Dalian University of Technology and City University of Hong Kong. He has served as a program committee member/reviewer for several CV and AI conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, IJCV and TIP. His research interests include deep learning and image enhancement.

Pan Chen is now a Master student at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He received his B.Eng. degree in Software Engineering from the East China Normal University, China in 2019. His research interests lie in saliency detection, self-supervised learning and action recognition.

Lizhuang Ma received his B.S. and Ph.D. degrees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a Distinguished Professor, at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China and the School of Computer Science and Technology, East China Normal University, China. He was a Visiting Professor at the Frounhofer IGD, Darmstadt, Germany in 1998, and a Visiting Professor at the Center for Advanced Media Technology, Nanyang Technological University, Singapore from 1999 to 2000. His research interests include computer vision, computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He serves as the reviewer of IEEE TPAMI, IEEE TIP, IEEE TMM, CVPR, AAAI etc.

Rynson W.H. Lau received his Ph.D. degree from University of Cambridge. He has been on the faculty of Durham University, Hong Kong Polytechnic University, and City University of Hong Kong.

Rynson serves on the Editorial Board of International Journal of Computer Vision (IJCV) and IET Computer Vision. He has served as the Guest Editor of a number of journal special issues, including ACM Trans. on Internet Technology, IEEE Trans. on Multimedia, IEEE Trans. on Visualization and Computer Graphics, and IEEE Computer Graphics & Applications. He has also served in the committee of a number of conferences, including Program Co-chair of ACM VRST 2004, ACM MTDL 2009, IEEE U-Media 2010, and Conference Co-chair of CASA 2005, ACM VRST 2005, ACM MDI 2009, ACM VRST 2014. Rynson's research interests include computer graphics and computer vision.