Designing a Lightweight Edge-Guided Convolutional Neural Network for Segmenting Mirrors and Reflective Surfaces

Mark Edward M. Gonzales
De La Salle University
Taft Avenue, Malate
Manila 1004, Philippines
mark_gonzales@dlsu.edu.ph

Lorene C. Uy
De La Salle University
Taft Avenue, Malate
Manila 1004, Philippines
lorene_c_uy@dlsu.edu.ph

Joel P. Ilao
De La Salle University
Taft Avenue, Malate
Manila 1004, Philippines
joel.ilao@dlsu.edu.ph

ABSTRACT

The detection of mirrors is a challenging task due to their lack of a distinctive appearance and the visual similarity of reflections with their surroundings. While existing systems have achieved some success in mirror segmentation, the design of lightweight models remains unexplored, and datasets are mostly limited to clear mirrors in indoor scenes. In this paper, we propose a new dataset consisting of 454 images of outdoor mirrors and reflective surfaces. We also present a lightweight edge-guided convolutional neural network based on PMDNet. Our model uses EfficientNetV2-Medium as its backbone and employs parallel convolutional layers and a lightweight convolutional block attention module to capture both low-level and high-level features for edge extraction. It registered F_β scores of 0.8483, 0.8117, and 0.8388 on the Mirror Segmentation Dataset (MSD), Progressive Mirror Detection (PMD) dataset, and our proposed dataset, respectively. Applying filter pruning via geometric median resulted in F_β scores of 0.8498, 0.7902, and 0.8456, respectively, performing competitively with the state-of-the-art PMDNet but with 78.20× fewer floating-point operations per second and 238.16× fewer parameters. The code and dataset are available at `https://github.com/memgonzales/mirror-segmentation`.

Keywords

Mirror segmentation, object detection, convolutional neural network (CNN), CNN filter pruning

1 INTRODUCTION

Despite the ubiquitous presence of mirrors and reflective surfaces in everyday scenes — from indoor rooms to outdoor buildings — existing computer vision systems have difficulty detecting them due to their lack of a consistent distinguishing appearance and the visual similarity of reflections with their surroundings [Par21]. This results in complications in tasks such as robot navigation [And18] and three-dimensional scene reconstruction [Zha18], where approaches to accommodate the presence of mirrors entail having to augment visual information from cameras with cues from specialized hardware, including ultrasonic sensors and dedicated illumination devices [Tin16].

Mirrors and reflective surfaces also pose potential hazards to autonomous driving and driver assistance systems that rely on stereo vision since they can cause glare spots, irregularly distorted reflections, and infinite reflections [Zen17]. These challenges are pronounced given the presence of safety mirrors in road and parking space junctions, as well as large reflective glass surfaces in the façades of several high-rise buildings. Hence, developing systems that can reliably recognize and localize them is critical to autonomous navigation.

While general object detection and segmentation frameworks have achieved success in various applications [He17, Zha17], they are unable to satisfactorily distinguish reflections from the actual objects. Consequently, directly applying them to mirror detection has yielded subpar results, as the reflections also tend to get segmented [Yan19]. Meanwhile, salient object detection techniques may not necessarily tag mirrors as salient [Yan19, Lin20a].

In this regard, the segmentation of mirrors and reflective surfaces posits itself as a challenging task that necessitates tailored approaches. Early works focused on exploiting contrasts and relationships between the contents inside and outside the mirror [Yan19, Lin20a]. Recently, depth [Mei21], semantic association with surrounding objects [Gua22], and visual chirality [Tan22] have also been explored to enrich the set of cues.

However, despite their success, designing lightweight mirror segmentation models remains an unexplored direction. Most systems have over 100 million parame-

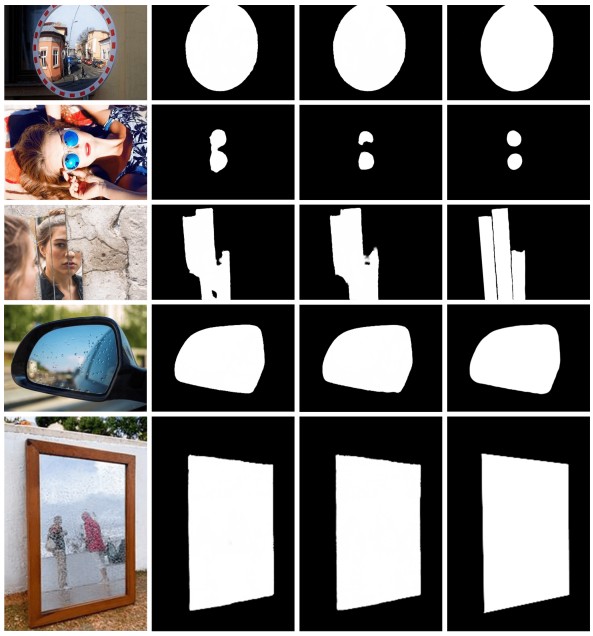


Figure 1: Existing datasets consist mostly of clear indoor mirrors. Our proposed dataset focuses on outdoor mirrors and reflective surfaces of varying shapes and sizes (first column). Our edge-guided CNN and its pruned version perform competitively with the state-of-the-art. This pruned version is also lightweight and can be deployed to resource-constrained devices.

ters, with MirrorNet [Yan19], PMDNet [Lin20a], and SANet [Gua22] having 121.77, 147.66, and 105.84 million parameters, respectively. Existing datasets are also mostly limited to clear mirrors in indoor scenes; outdoor mirrors and reflective surfaces (e.g., tinted car windows and building façades) are not well represented. These may be prohibitive to the integration of models into resource-constrained devices, such as drones and autonomous navigation vehicles.

In an attempt to address these gaps, our study seeks to contribute the following:

- We propose a dataset of outdoor mirrors and reflective surfaces with 454 images and their corresponding ground-truth masks.
- We modified the architecture of PMDNet [Lin20a] and extensively tested different feature extraction backbones and edge-related modules to guide the segmentation.
- We pruned our best-performing edge-guided convolutional neural network, resulting in a lightweight model with 1.52 billion floating-point operations per second (FLOPS) and 0.62 million parameters. It performs competitively with the state-of-the-art PMDNet but with 78.20× fewer FLOPS and 238.16× fewer parameters.

2 RELATED WORKS

Early attempts to detect and segment mirrors require the assistance of specialized hardware [Whe18] or user interaction [Cha17]. The first model to perform the task given solely an RGB image input is MirrorNet [Yan19]. Using ResNeXt-101 [Xie17] as its multi-scale feature extraction backbone, content discontinuities inside and outside the mirror are captured via a dedicated contextual contrasted feature extraction module.

PMDNet [Lin20a] extends this by considering not only discontinuities but also similarities between the reflection and the surroundings via a dedicated module connected to the side-outputs of a ResNeXt-101 backbone. Moreover, an edge detection and fusion module captures both high-level and low-level features from the feature maps generated by the backbone. However, MirrorNet and PMDNet may have some difficulty handling cases where there are insufficient correlational features or contextual contrast, such as when the reflection occupies most of the image.

Recent studies have also investigated the integration of various cues. Adopting ResNet-50 [He16] as its backbone, PDNet [Mei21] captures not only RGB features but also depth. Aside from the limitations posed by the need for specialized hardware to capture depth, objects such as doorways may confuse its depth-aware module.

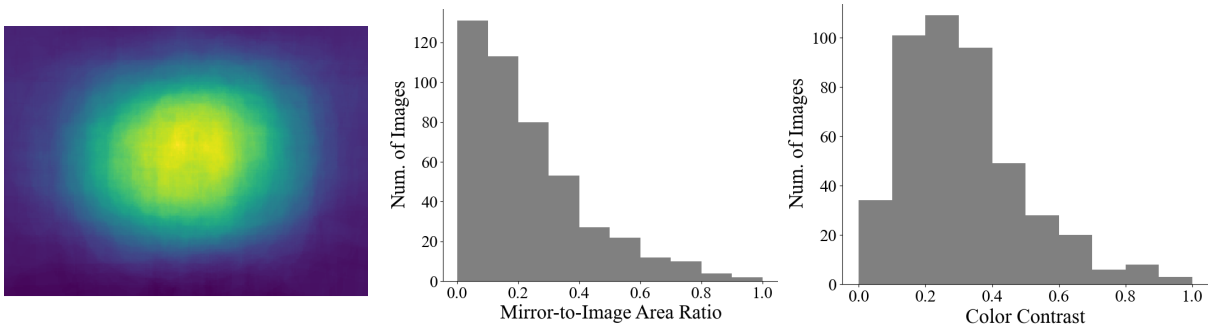
The scene-aware SANet [Gua22] capitalizes on semantic associations, i.e., the observed placement of mirrors together with certain objects for functional purposes. Since this approach relies on annotations, low-quality labels may affect performance. Annotated datasets may also be expensive to construct and may thus not be readily available for most real-world use cases.

VCNet [Tan22] frames visual chirality [Lin20b], the change in image statistics upon reflection, as a commutative residual. Similar to MirrorNet and PMDNet, it utilizes a ResNeXt-101 backbone. While its use of a visual chirality cue allows its edge detection module to learn features other than the conventional geometric properties, it has difficulty excluding small occluding objects and handling boundaries with complex shapes.

Our work builds on insights from these previous works and explores another direction by focusing on the construction of a lightweight model that is capable of performing competitively with the state-of-the-art. We also demonstrate the effectiveness of using EfficientNet [Tan19] as a promising and less computationally expensive alternative to the usual ResNeXt backbone used in existing mirror detection and segmentation models.

3 OUTDOOR MIRRORS AND REFLECTIVE SURFACES DATASET

Following previous works [Yan19, Lin20a, Mei21, Gua22, Tan22], we used two publicly available mirror datasets in our study: MSD [Yan19] and PMD



(a) (b) (c)

Figure 2: Dataset Statistics. (a) Distribution of the mirror location, with yellow corresponding to higher frequencies and blue corresponding to lower frequencies. (b) Mirror-to-image area ratio. (c) Color contrast between the mirror and the surrounding area, as measured by taking the χ^2 distance between their RGB histograms, following [Yan19].

[Lin20a]. MSD consists of 4018 images; however, most are zoomed-in images of indoor scenes that exhibit high similarity. PMD aggregates 6016 images from multiple datasets including ADE20K [Zho17] and NYUD-V2 [Sil12]. Although the images in PMD are more varied than those in MSD, outdoor mirrors and reflective surfaces remain underrepresented.

To help address this limitation, we propose the De La Salle University – Outdoor Mirrors and Reflective Surfaces (DLSU-OMRS) dataset. The images were scraped from Shutterstock using the key phrases *outdoor mirror* and *street mirror* and manually filtered to remove duplicates and heavily manipulated photos. Ground-truth masks were produced through manual segmentation. The DLSU-OMRS dataset contains 454 images, with an average structural similarity index of 28.67%. As characterized in Figure 2 and Table 1, most mirrors are located near the center and occupy up to 20% of the image. The color contrast [Yan19] of most images is also below 40%, which suggests that the contents inside the mirrors are visually similar to their surroundings, making our dataset more challenging.

	Num. of Images
One Mirror	338
Multiple Mirrors	116
	Num. of Mirrors
By Shape	
Triangle	4
Quadrilateral	258
Polygonal (≥ 5 straight edges)	9
Round/Elliptical	160
Irregular	355
By Presence of Occlusion	
Present	192
Not Present	594

Table 1: Mirror Shape and Occlusion Statistics. For images with multiple mirrors, each mirror is categorized separately by shape and by the presence of an occluding object. In total, our DLSU-OMRS dataset has 454 images and 786 mirrors within those images.

4 MODEL CONSTRUCTION

4.1 Model Architecture

Using PMDNet as the base model (Figure 3a), we introduced two modifications in an attempt to improve performance and lower computational costs.

First, we explored seven feature extraction backbones that were pretrained on ImageNet [Den09]: ResNet-50 [He16], Xception-65 [Cho17], VoVNet-39 [Lee19], MobileNetV3 [How19], EfficientNetLite4 [Tan19], EfficientNet-Edge-Large (pruned following the lottery ticket hypothesis) [Tan19], and EfficientNetV2-Medium [Tan19]. These were selected in light of their application in object segmentation [Cha22, Lin22].

Second, we modified PMDNet's edge detection and fusion module. While PMDNet extracts low-level edge features by connecting the side-output of the lowest-level backbone to a sequence of three convolutional layers (Figure 3b), our proposed design (Figure 3c) connects it to a boundary extraction module with four parallel convolutional layers of varying kernel sizes and dilation rates, adapted from GDNet [Mei22]; suppose this module's output is denoted by f_{low}.

To extract high-level edge features, our design shares PMDNet's approach of feeding the highest-level relational contextual contrasted local module's output to a convolutional block attention module [Woo18], a lightweight module that infers spatial and channel attention maps; suppose its output is denoted by f_{high}.

The intermediate output maps f_{low} and f_{high} are then concatenated and passed to an edge prediction block. Our edge prediction block expands that of PMDNet, changing it from a single 3×3 convolutional layer to a 1×1 convolutional layer (with batch normalization and ReLU) connected to a 3×3 convolutional layer.

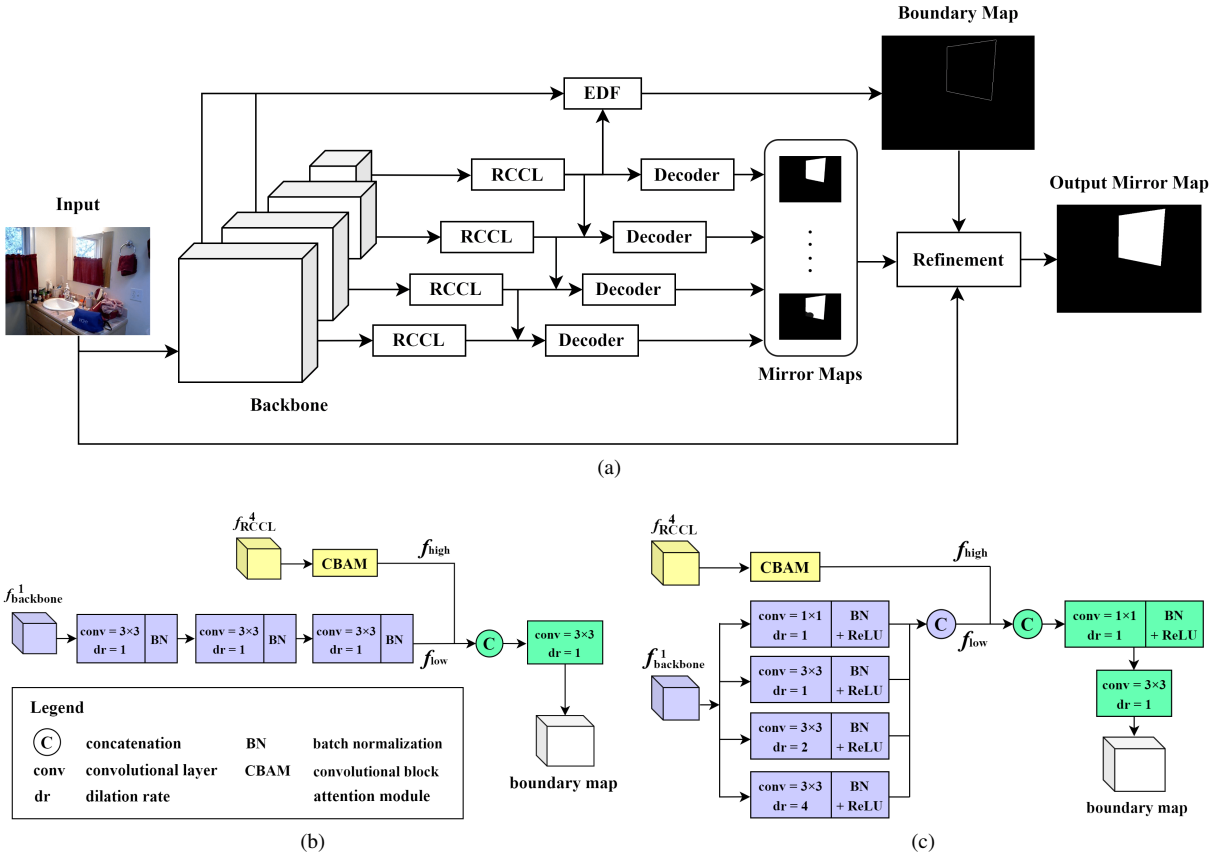


Figure 3: Model Architecture. (a) Overall architecture (the diagram is adapted from [Lin20a]). Its main components are the relational contextual contrasted local (RCCL) module, which is designed to extract contrasts and similarities inside and outside the mirror, and the edge extraction and fusion (EDF) module. (b) Original EDF module of PMDNet. (c) Our proposed modification to the EDF module. The blue blocks in (b) and (c) are for low-level edge feature extraction; f^1_{backbone} is the side-output of the lowest-level backbone, and f_{low} is the low-level edge feature map. The yellow blocks are for high-level edge feature extraction; f^4_{RCCL} is the output of the highest-level RCCL module, and f_{high} is the high-level edge feature map. The green blocks combine and process f_{low} and f_{high} for edge prediction.

These aforementioned convolutional layers both have a dilation rate of 1. Our modified architecture (Figure 3c) aims to exploit richer edge semantics without adding significant overhead to the model's complexity.

4.2 Model Training

Our models were built using PyTorch and trained on the training partition of the split PMD dataset, which consists of 5096 images. The input images were then resized to 352×352 and augmented through random horizontal flipping and jittering the brightness, contrast, saturation, and hue by a random value in the interval $[0.9, 1.1]$. They were normalized following the mean and standard deviation of the images in ImageNet [Den09]. The batch size was set to 10.

The learning rate was initialized to 1×10^{-3} and updated via a polynomial strategy with 0.9 as the power. The loss function was minimized using stochastic gradient descent with a weight decay of 5×10^{-4} and momentum of 0.9. The models were trained for 150 epochs, with the exception of those with ResNet (200 epochs) and EfficientNet (140 epochs) backbones.

4.3 Loss Function

We combined three loss functions to supervise the training of our model. First, intersection-over-union (IoU) loss was used for the multi-scale mirror maps (i.e., excluding the final mirror map). Second, a Laplacian-based loss [Zha19] for emphasizing edges was used for the boundary map. Third, an additive loss that combines the weighted IoU and the weighted binary cross-entropy (BCE) loss proposed by [Wei20] was used for the final (output) mirror map.

Our choice of loss functions differs from the usual approach in existing mirror segmentation models [Yan19, Lin20a, Mei21, Gua22, Tan22], which mostly employ Lovász-Softmax [Ber18] for the mirror maps and BCE loss for the boundary map.

A drawback of Lovász-Softmax is its high computational cost, as noted in our initial experiments and in related studies [Alo19]. Our use of IoU loss for the multi-scale mirror maps is more efficient, albeit generally outperformed by Lovász-Softmax. To compensate for this while maintaining efficiency, we employed an additive loss that combines weighted IoU and BCE for the final mirror map. Unlike ordinary IoU and BCE loss, which focus only on individual pixels, their weighted variants draw the model to a larger receptive field [Wei20].

In addition, our use of a Laplacian-based loss function tailored for emphasizing edges is an alternative strategy to BCE, which is sensitive to imbalanced edge/non-edge distribution [Den18], a problem that is more pronounced since our edge extraction module is concerned only with the edges of the mirrors.

To formalize, let $\mathscr{L}_{\text{mirror}}(\hat{M}_i, M)$ denote the IoU loss between the i^{th} predicted mirror map $\hat{M}_i$ and the ground truth M; $\mathscr{L}_{\text{edge}}(\hat{E}, E)$, the Laplacian-based loss between the predicted boundary map $\hat{E}$ and the ground truth E; and $\mathscr{L}_{\text{output}}(\hat{M}, M)$, the additive loss between the predicted output mirror map $\hat{M}$ and the ground truth M. Note that the ground-truth boundary maps were obtained by applying Canny edge detection [Can86] on the ground-truth mirror maps.

Our final loss function $\mathscr{L}$ is given by Equation 1.

$$\begin{aligned}\mathscr{L} = \sum_{i=1}^{4} w_{\text{mirror}} \cdot \mathscr{L}_{\text{mirror}}(\hat{M}_i, M) \\ + w_{\text{edge}} \cdot \mathscr{L}_{\text{edge}}(\hat{E}, E) \\ + w_{\text{output}} \cdot \mathscr{L}_{\text{output}}(\hat{M}, M)\end{aligned} \tag{1}$$

The weighting coefficients w_{mirror} (for $i = 1$ to 4), w_{edge}, and w_{output} were set to 1, 5, and 2, respectively, following [Lin20a]. These values were empirically found to yield the best performance from a parameter space of $\{(1,1,1),(1,2,2),(1,5,2),(1,5,5),(1,5,7)\}$.

4.4 Model Compression

To further decrease its complexity, we subjected our best-performing model to filter pruning via geometric median (FPGM), a one-shot pruning technique that reduces redundant filters by leveraging the geometric median as a data centrality estimator to capture the mutual information shared by filters in the same layer [He19]. FPGM has also been applied in previous studies on object detection and segmentation [Hao22]. In our work, we applied FPGM on the convolutional and linear layers at a sparsity level of 10%.

After pruning, we performed retraining for 20 epochs to recover lost accuracy; to this end, we adopted a learning rate rewinding policy [Ren20], which uses the original learning rate schedule to retrain unpruned weights from their final values.

4.5 Model Evaluation

We evaluated the performance of our built models on MSD, the test partition of the split PMD dataset, and our proposed DLSU-OMRS dataset, which contain 955, 571, and 454 images, respectively.

We employed two evaluation metrics: maximum F-measure (F_β) and mean absolute error (MAE). Given the ground truth $Y(\cdot,\cdot)$, the predicted output $\hat{Y}(\cdot,\cdot)$, and an image of width w and height h, the formal definitions of these measures are given in Equations 2 and 3; β^2 was set to 0.3, as suggested by [Ach09].

$$F_\beta = \frac{(1+\beta^2)\cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \tag{2}$$

$$MAE = \frac{1}{w \cdot h} \sum_{x=1}^{w} \sum_{y=1}^{h} |\hat{Y}(x,y) - Y(x,y)| \tag{3}$$

Moreover, the number of floating-point operations per second (FLOPS) and the number of parameters were identified to measure our models' complexity.

5 RESULTS AND ANALYSIS

5.1 Model Performance

Table 2 compares the performance of our models with two relevant state-of-the-art systems. VST [Liu21] is a transformer-based salient object detection model that can handle scenarios with similar foreground and background, as is the case for most images with mirrors. PMDNet is the base model of our work.

Our model that uses an EfficientNetV2-Medium backbone and employs our compound loss function and edge extraction and prediction module (second to last row of Table 2) registered the top performance across both metrics on the PMD dataset, as well as the lowest MAE on MSD. It performed competitively with PMDNet, achieving a slight edge on MSD and PMD. While it was slightly outperformed on DLSU-OMRS, our model has the advantage of having 4.79× fewer FLOPS and 2.77× fewer parameters.

The pruned version of this model (last row of Table 2) also performed competitively with PMDNet and registered the highest F_β on both MSD and DLSU-OMRS, slightly outperforming the said baseline by 0.0148 and 0.0033 points, respectively. It also achieved the second-lowest MAE on both of these datasets. Among our models, this pruned version has the least computational complexity, clocking in 78.20× fewer FLOPS and 238.16× fewer parameters compared to PMDNet.

On another note, although our model with an EfficientNet-Lite backbone was not able to outperform PMDNet, its F_β scores across all three benchmark datasets were consistently within 0.06 points of the

Model	Computational Complexity		MSD		PMD		DLSU-OMRS	
	GFLOPS ↓	# of Params ↓	F_β ↑	MAE ↓	F_β ↑	MAE ↓	F_β ↑	MAE ↓
VST [Liu21]	46.36	44.48M	0.4290	0.2739	0.1317	0.261	0.5730	0.2274
PMDNet [Lin20a]	118.86	147.66M	0.8350	0.0816	0.8011	0.0324	0.8423	**0.0878**
Ours (Compound Loss)								
ResNet-50	105.47	129.04M	0.7548	0.1119	0.7650	0.0403	0.7874	0.1011
Ours (Compound Loss + Edge Extraction)								
ResNet-50	116.46	130.12M	0.7695	0.1098	0.7524	0.0409	0.8042	0.1025
Xception-65	75.28	129.12M	0.7800	0.0973	0.7566	0.0401	0.7643	0.1164
VoVNet-39	98.25	61.90M	0.7014	0.1196	0.7578	0.0412	0.7868	0.1088
MobileNetV3	6.61	20.76M	0.7515	0.1153	0.7508	0.0427	0.8256	0.1006
EfficientNet-Lite	6.99	15.54M	0.7909	0.1027	0.7769	0.0387	0.8178	0.1048
EfficientNet-Edge-Large (Pruned)	17.02	10.42M	0.7682	0.1082	0.7831	0.0349	0.8035	0.1044
EfficientNetV2-Medium	24.79	53.35M	0.8483	**0.0800**	**0.8117**	**0.0313**	0.8388	0.1032
Ours (Compound Loss + Edge Extraction + FPGM Pruning)								
EfficientNetV2-Medium	**1.52**	**0.62M**	**0.8498**	0.0813	0.7902	0.0364	**0.8456**	0.0955

Table 2: Performance of the Models. The row labels for our models denote the backbone. Higher F_β and lower MAE correspond to better performance. The best scores are given in bold; the second-best scores are underlined.

highest scores. Moreover, it has 17.00× fewer FLOPS and 9.50× fewer parameters compared to PMDNet. These results suggest the applicability of the EfficientNet family of networks as a promising and less computationally expensive alternative to the ResNeXt backbone used in existing mirror segmentation models.

Figure 4 provides a qualitative comparison of how the different models handle some challenging cases.

5.2 Performance of the Pruned Model

To quantify the extent to which pruning can be applied without overly compromising the model's performance, we applied FPGM pruning to the best-performing unpruned model at different sparsity levels and retrained the pruned model for 20 epochs following a learning rate rewinding policy.

As seen in Tables 3 and 4, raising the sparsity from 10% to 20% decreased the F_β score by around 0.02 to 0.04 points; further increasing it to 40% already resulted in a significant drop of around 0.16 to 0.22 points. A visual example is provided in Figure 5.

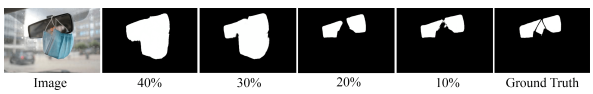


Figure 5: Visual Example of Performance Under Different Sparsity Levels. In this image taken from our proposed dataset, the performance of the pruned model noticeably degrades at 30% sparsity and above, as it already fails to properly distinguish the hung face mask from the mirror.

Sparsity Level	MSD	PMD	DLSU-OMRS
40%	0.6267	0.6006	0.6876
30%	0.7695	0.7566	0.7963
20%	0.8073	0.7795	0.8211
10%	**0.8498**	0.7902	**0.8456**
Unpruned	0.8483	**0.8117**	0.8388

Table 3: F_β Under Different Sparsity Levels

Sparsity Level	MSD	PMD	DLSU-OMRS
40%	0.4633	0.4790	0.1485
30%	0.0970	0.0410	0.1039
20%	0.0905	0.0352	**0.0940**
10%	0.0813	0.0364	0.0955
Unpruned	**0.0800**	**0.0313**	0.1032

Table 4: MAE Under Different Sparsity Levels

	MSD	PMD	DLSU-OMRS
Unpruned	0.8483	**0.8117**	0.8388
Not Retrained	**0.8505**	0.7858	0.8432
Retrained	0.8498	0.7902	**0.8456**

Table 5: F_β of Pruned Model (Sparsity = 10%) Before and After Retraining

	MSD	PMD	DLSU-OMRS
Unpruned	**0.0800**	**0.0313**	0.1032
Not Retrained	0.4185	0.4585	0.4407
Retrained	0.0813	0.0364	**0.0955**

Table 6: MAE of Pruned Model (Sparsity = 10%) Before and After Retraining

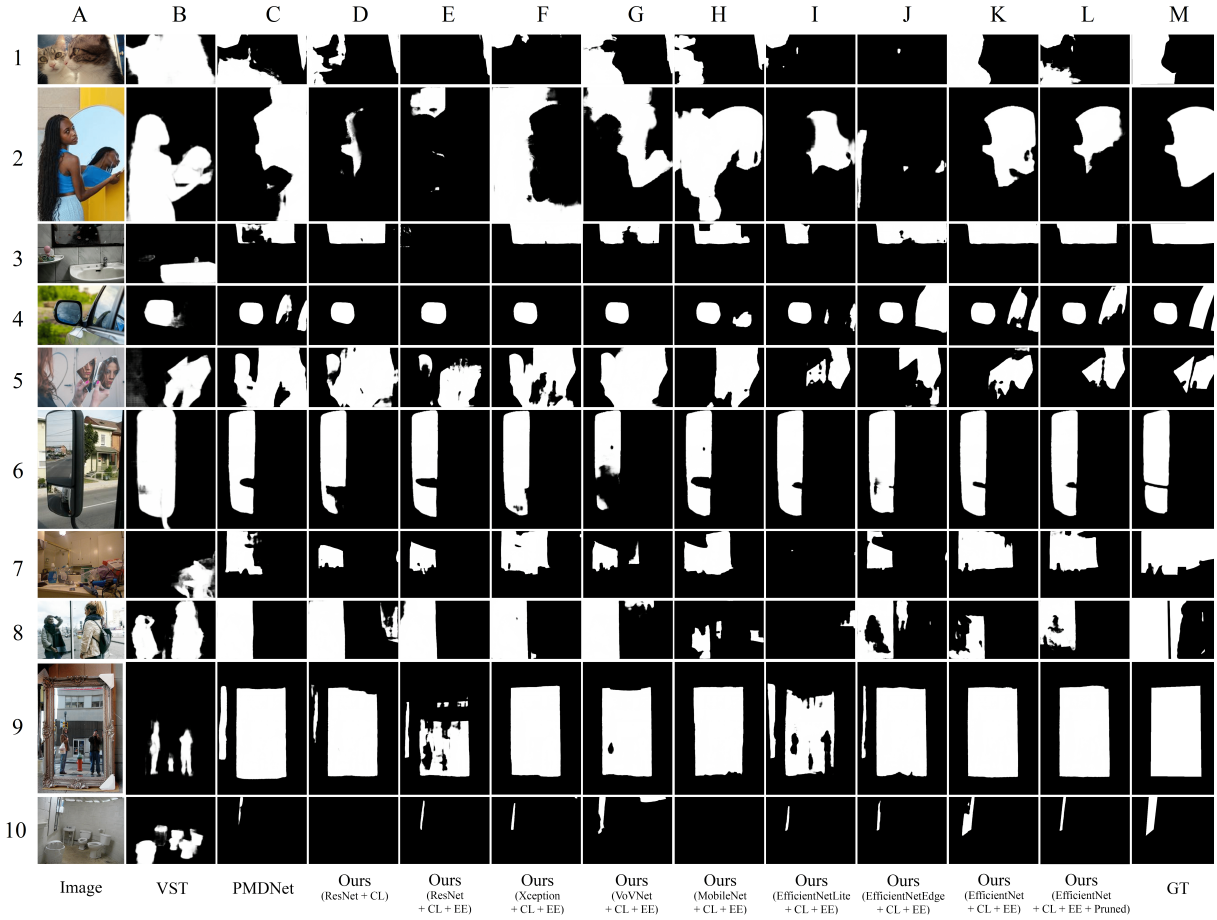


Figure 4: Qualitative Comparison on Challenging Cases. *CL* and *EE* indicate that the model uses our proposed *compound loss* and *edge extraction and prediction module*, respectively. *GT* pertains to the *ground truth*. Salient object detection models (column B) may not necessarily tag mirrors as salient. Our best-performing model (column K) can handle some cases that may be challenging even for a state-of-the-art model (column C). These include images where *(i)* the object occludes the mirror and, alongside its reflection, occupies a large portion of the image (rows 1 and 2), *(ii)* the reflection has a similar color to the mirror's frame (row 3), and *(iii)* multiple mirrors and reflective surfaces are present (row 4). Our pruned version (column L) was able to segment irregularly shaped mirror shards (row 5), although, in general, it seems to have some difficulty handling cases where mirrors are separated by only a thin divider (row 6) and where the object and reflection occupy the majority of the image (rows 1 and 2). Although our best-performing model and its pruned version captured the largest fraction of the ground-truth mask in row 7, it remains challenging to handle cases where the contextual features inside and outside the mirror appear continuous (row 8).

Tables 5 and 6 report the performance after pruning the model at 10% sparsity but prior to retraining. Although the F_β score was comparable, there was a significant increase in MAE prior to retraining. This increased MAE can be attributed to the resulting output maps emphasizing the mirrors but failing to completely mask out the surroundings, as seen in Figure 6.

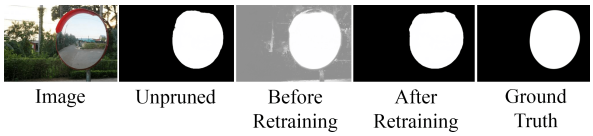


Figure 6: Visual Example of Performance of Pruned Model Before and After Retraining

5.3 Model Component Analysis

To demonstrate the contribution of our proposed edge extraction and prediction module, we conducted ablation experiments on our unpruned model (Tables 7 and 8). On MSD and PMD, incorporating our module outperformed not including any edge semantics-related module and utilizing PMDNet's original edge detection and fusion module. On DLSU-OMRS, using PMDNet's original module resulted in the highest performance, albeit only by 0.0001 F_β and 0.0114 MAE points. Visual examples are given in Figure 7.

To investigate the effects of our choice of loss functions, we also measured the performance of our best-

performing model if simple BCE and IoU loss functions were used to supervise the training of the boundary and final mirror maps, respectively. As seen in Tables 9 and 10, our proposed loss function resulted in the best F_β and MAE scores on MSD and the highest F_β on PMD. A visual example is also provided in Figure 8.

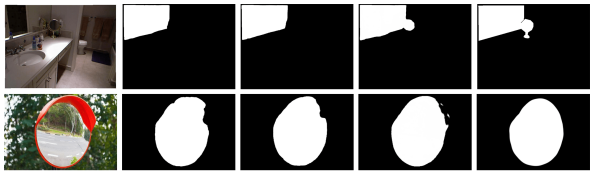


Figure 7: Visual Example of Performance of Ablated Models. The use of our edge extraction and prediction module helps in capturing boundaries of small objects that may otherwise be missed (first row). However, in certain cases, it may also result in the inclusion of noise in the predicted mask (second row).

	MSD	PMD	DLSU-OMRS
RCCL	0.8052	0.7957	0.8300
RCCL + EDF	0.8224	0.8001	**0.8389**
Ours	**0.8483**	**0.8117**	0.8388

Table 7: F_β After Ablation. *RCCL* and *EDF* refer to PMDNet's relational contextual contrasted local module and edge detection and fusion module. Our model modifies the EDF module (Section 4.1).

	MSD	PMD	DLSU-OMRS
RCCL	0.0957	0.0332	0.0956
RCCL + EDF	0.0949	0.0335	**0.0918**
Ours	**0.0800**	**0.0313**	0.1032

Table 8: MAE After Ablation

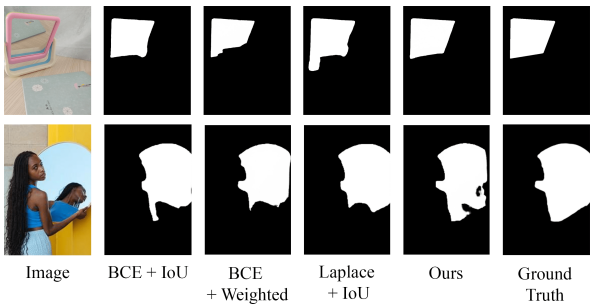


Figure 8: Visual Example of Performance Under Different Loss Functions. Using our compound loss function resulted in the most accurate mirror map in the image in the first row. Although its use in the image in the second row increased sensitivity to boundaries proximate to the reflection's chest area, the overall contour of the ground-truth mask was better captured.

Loss	MSD	PMD	DLSU-OMRS
BCE + IoU	0.8352	0.8038	0.8314
BCE + Weighted	0.8163	0.8073	0.8470
Laplace + IoU	0.8148	0.7989	**0.8553**
Ours	**0.8483**	**0.8117**	0.8388

Table 9: F_β Under Different Loss Functions. *Ours* refers to our use of a Laplacian-based loss function for the boundary map and an additive loss function combining weighted IoU and BCE loss for the final mirror map (Section 4.3).

Loss	MSD	PMD	DLSU-OMRS
BCE + IoU	0.0949	0.0320	0.0969
BCE + Weighted	0.0967	0.0319	0.0995
Laplace + IoU	0.0950	**0.0302**	**0.0881**
Ours	**0.0800**	0.0313	0.1032

Table 10: MAE Under Different Loss Functions

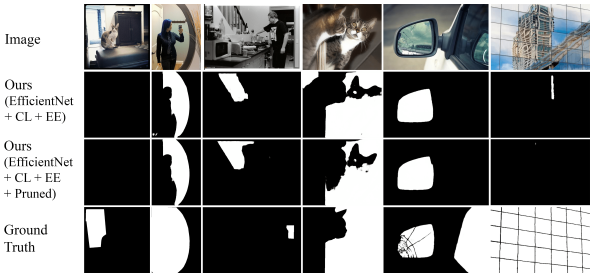


Figure 9: Failure Cases. *CL* and *EE* indicate that the model uses our proposed *compound loss* and *edge extraction and prediction module*, respectively. Some failure cases, such as the fourth image, may be confusing even for human observers. Moreover, fine details such as cracks (second to last image) are generally not preserved, although the mirror's overall contour is correctly captured.

5.4 Failure Cases

Figure 8 shows the limitations of our model. Since our model exploits contextual discontinuities and similarities, it has some difficulty handling cases where the contextual features inside and outside the mirror appear continuous (first image) or where the available contextual features are inadequate due to the mirror occupying the entire image (last image).

Sharp discontinuities within the mirror (second image) may also result in the reflection being treated as part of the predicted mask's boundary. Some transparent glass objects may be falsely flagged as mirrors, whereas small mirrors in the background (third image) and heavily tinted reflective surfaces (fifth image) may be challenging to recognize.

6 CONCLUSION

In this study, we propose DLSU-OMRS, a dataset of 454 images of outdoor mirrors and reflective surfaces, which are not well represented in existing mirror datasets. We also modified the architecture of PMDNet and extensively tested different feature extraction backbones and edge-related modules to guide the segmentation. Our best-performing model uses EfficientNetV2-Medium as its backbone and employs an edge detection module consisting of parallel convolutional layers and a lightweight convolutional block attention module to capture both low-level and high-level edge semantics.

Our model performs competitively with the state-of-the-art PMDNet, registering F_β scores of 0.8483, 0.8117, and 0.8388 on MSD, PMD, and our proposed dataset, respectively. Compressing this model by pruning via geometric median resulted in F_β scores of 0.8498, 0.7902, and 0.8456, respectively, maintaining competitive performance but with 78.20× fewer FLOPS and 238.16× fewer parameters.

Future directions include addressing the discussed limitations of our work and extending our approach to further realize the applicability of mirror detection and segmentation models to resource-constrained devices, such as those for autonomous navigation (e.g., drones).

7 ACKNOWLEDGMENT

We thank Mr. Gregory G. Cu and Mr. Fritz Kevin S. Flores for providing us with access to the University's computing resources. We also thank Dr. Macario O. Cordel, II and Dr. Ann Franchesca B. Laguna for their feedback on the initial manuscript.

8 REFERENCES

[Ach09] Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. Frequency-tuned salient region detection. 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1597-1604, 2009.

[Alo19] Alonso, I., Yuval, M., Eyal, G., Treibitz, T., and Murillo, A.C. CoralSeg: Learning coral segmentation from sparse annotations. Journal of Field Robotics, 36, 8, pp. 1456-1477, 2019.

[And18] Anderson, P. et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3674-3683, 2018.

[Ber18] Berman, M., Triki, A.R., and Blaschko, M. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[Can86] Canny, J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8, 6, pp. 679-698, 1986.

[Cha17] Chang, A. et al. Matterport3D: Learning from RGB-D Data in indoor environments. 2017 International Conference on 3D Vision (3DV), pp. 667-676, 2017.

[Cha22] Chahal, E.S., Patel, A., Gupta, A., Purwar, A., and Dhanalekshmi, G. Unet based Xception model for prostate cancer segmentation from MRI images. Multimedia Tools and Applications, 81, 26, pp. 37333-37349, 2022.

[Cho17] Chollet, F. Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800-1807, 2017.

[Den09] Deng, J. et al. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[Den18] Deng, R., Shen, C., Liu, S., Wang, H., and Liu, X. Learning to predict crisp boundaries. Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI, pp. 570-586, 2018.

[Gua22] Guan, H., Lin, J., and Lau, R.W.H. Learning semantic associations for mirror detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5931-5940, 2022.

[Hao22] Hao, Z., Wang, Z., Bai, D., and Tong, X. Surface defect segmentation algorithm of steel plate based on geometric median filter pruning. Frontiers in Bioengineering and Biotechnolology, 10, pp. 945248, 2022.

[He16] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[He17] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988, 2017.

[He19] He, Y., Liu, P., Wang, Z., Hu, Z., and Yang, Y. Filter pruning via geometric median for deep convolutional neural networks acceleration. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4335-4344, 2019.

[How19] Howard, A. et al. Searching for MobileNetV3. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314-1324, 2019.

[Hua17] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017.

[Kra11] Krähenbühl, P., and Koltun, V. Efficient inference in fully connected CRFs with Gaussian edge potentials. Advances in Neural Information Processing Systems, 24, 2011.

[Lee19] Lee, Y., Hwang, J.W., Lee, S., Bae, Y., and Park, J. An energy and GPU-computation efficient backbone network for real-time object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 752-760, 2019.

[Lin20a] Lin, J., Wang, G., and Lau, R.H. Progressive mirror detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3694-3702, 2020.

[Lin20b] Lin, Z., Sun, J., Davis, A., and Snavely, N. Visual chirality. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12292-12300, 2020.

[Lin22] Lin, J. et al. Efficient heterogeneous video segmentation at the edge. Sixth Workshop on Computer Vision for AR/VR (CV4ARVR), 2022.

[Liu21] Liu, N., Zhang, N., and Wan, K., Shao, L., and Han, J. Visual saliency transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4722-4732, 2021.

[Mei21] Mei, H. et al. Depth-aware mirror segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3044-3053, 2021.

[Mei22] Mei, H. et al. Large-field contextual feature learning for glass detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, 01, pp. 1-17, 2022.

[Par21] Park, D., and Park, Y.H. Identifying reflected images from object detector in indoor environment utilizing depth information. IEEE Robotics and Automation Letters, 6, 2, pp. 635-642, 2021.

[Ren20] Renda, A., Frankle, J., and Carbin, M. Comparing rewinding and fine-tuning in neural network pruning. International Conference on Learning Representations, 2020.

[Sil12] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from RGBD images. European Conference on Computer Vision (ECCV), 2012.

[Tan19] Tan, M., and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 97, pp. 6105-6114, 2019.

[Tan22] Tan, X. et al. Mirror detection with the visual chirality cue. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-13, 2022.

[Tin16] Tin, S.K., Ye, J., Nezamabadi, M., and Chen, C. 3D reconstruction of mirror-type objects using efficient ray coding. 2016 IEEE International Conference on Computational Photography (ICCP), pp. 1-11, 2016.

[Wei20] Wei, J., Wang, S., and Huang, Q. F^3Net: Fusion, feedback and focus for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence, 34, pp.12321-12328, 2020.

[Whe18] Whelan, T. et al. Reconstructing scenes with mirror and glass surfaces. ACM Trans. Graph., 37, 4, 2018.

[Woo18] Woo, S., Park, J., Lee, J., Lee, J., and Kweon, I.S. CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19, 2018.

[Xie17] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987-5995, 2017.

[Yan19] Yang, X. et al. Where is my mirror? Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8809-8818, 2019.

[Zen17] Zendel, O., Honauer, K., Murschitz, M., Humenberger, M., and Domínguez, G. Analyzing computer vision data - the good, the bad and the ugly. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6670-6680, 2017.

[Zha17] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230-6239, 2017.

[Zha18] Zhang, Y., Ye, M., Manocha, D., and Yang, R. 3D reconstruction in the presence of glass and mirrors by acoustic and visual fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 8, pp. 1785-1798, 2018.

[Zha19] Zhao, T., and Wu, X. Pyramid feature attention network for saliency detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3080-3089, 2019.

[Zho17] Zhou, B. et al. Scene parsing through ADE20K dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.