Robust Shadow Detection by Exploring Effective Shadow Contexts

Xianyong Fang
Anhui University
Hefei, China
fangxianyong@ahu.edu.cn

Xiaohao He
Anhui University
Hefei, China
hexiaohaoahu@163.com

Linbo Wang
Anhui University
Hefei, China
wanglb@ahu.edu.cn

Jianbing Shen
University of Macau
Macau, China
shenjianbingcg@gmail.com

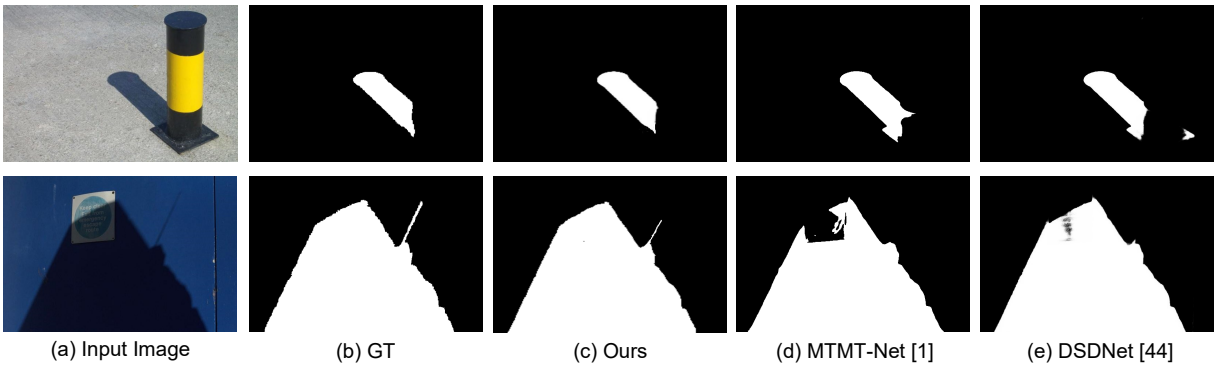


Figure 1: Example of shadow detections with the SBU dataset [34]. The top and bottom scenes can lead to fake shadow (the dark parts of the pillar) and foreground (the human shadow) respectively.

ABSTRACT

Effective contexts for separating shadows from non-shadow objects can appear in different scales due to different object sizes. This paper introduces a new module, Effective-Context Augmentation (ECA), to utilize these contexts for robust shadow detection with deep structures. Taking regular deep features as global references, ECA enhances the discriminative features from the parallelly computed fine-scale features and, therefore, obtains robust features embedded with effective object contexts by boosting them. We further propose a novel encoder-decoder style of shadow detection method where ECA acts as the main building block of the encoder to extract strong feature representations and the guidance to the classification process of the decoder. Moreover, the networks are optimized with only one loss, which is easy to train and does not have the instability caused by extra losses superimposed on the intermediate features among existing popular studies. Experimental results show that the proposed method can effectively eliminate fake detections. Especially, our method outperforms state-of-the-arts methods and improves over 13.97% and 34.67% on the challenging SBU and UCF datasets respectively in balance error rate.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475199

CCS CONCEPTS

• **Computing methodologies** → **Object detection**.

KEYWORDS

Shadow detection, deep learning, encoder-decoder

ACM Reference Format:
Xianyong Fang, Xiaohao He, Linbo Wang, and Jianbing Shen. 2021. Robust Shadow Detection by Exploring Effective Shadow Contexts. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475199

1 INTRODUCTION

Shadow is the light effect due to surface occlusion, which exists almost everywhere in our daily lives. It can be hard or soft, depending on the number of light sources. Accurately detecting shadows is important for computing illumination [20, 26], layout [17], camera calibration [16], object tracking [23], etc.

Recently, deep learning based approaches demonstrate better performances [12, 18, 24, 34] than traditional physical [3, 5] or handcrafted methods [33, 41, 46]. They can train optimally deep

classifiers supported by big sample data. Various methods have been proposed based on convolutional neural networks (CNN) by direct application [9, 18, 22, 28, 34], extending with more contexts [12, 36, 46] or adversarial learning with generative adversarial networks (GAN) [6] for small data training [21, 25] or higher discrimination [35]. Some researchers [1, 11, 12, 36, 44, 46] prefer to more constraints through additional loss functions from layers or networks.

However, current methods may still suffer from two difficulties (Figure 1): 1) Dark surfaces similar to the normal shadows and 2) light shaded shadows similar to the backgrounds. The former can lead to fake shadows which should be real surfaces, while the latter can lead to fake backgrounds which should be shadows. We can conclude that the shadow detection cannot be stable without correcting these fakes.

Humans can easily recognize shadows. How can we do that? Our common sense shows that the surroundings or contexts of objects of interest provide important cues for judgment. Looking at the scenes shown in Figure 1, we not only just check the shadow or foreground surface themselves independently, but also their surroundings to validate their shadow properties. For example, the dark surface won't be taken as shadow because it is on a pillar standing on the floor. Therefore, we can conclude that appropriate surroundings or contexts are important for recognition.

Existing deep neural network based studies do not emphasize the importance of appropriate contexts pertaining to varying-sized objects, even though they have already tried to capture rich contexts. When coming to shadow detection, they either overlook the importance of those appropriate ones or eliminate them during the convolution processes. Therefore, it is necessary to figure out a novel method to explore object related scales effectively so that right contexts can be utilized for shadow localization.

To this end, we propose a new module, Effective-Context Augmentation (ECA), which adopts multiple parallel convolutions with different kernels to augment effective object contexts in proper scales for shadow detection. Taking the regular deep features as the global guidance and fusing them with the discriminative features from the fine scales by the convolutions, ECA can boost the effective object contexts embedded in the final output feature and thus help the object detection process.

The robust object contexts enhanced by ECA features can also assist the generation process from the strong representation for a robust segmentation, i.e., ECA features can act as strong cues to separate the shadow and non-shadow regions. Consequently, we propose an encoder-decoder based neural network method to fully explore the merits of ECA. It adopts ECA in the encoder to obtain discriminative contexts for the image representation and also takes the ECA features to guide the decoding process for an efficient classification. Our method can integrate more effective scales for shadow contexts than the existing networks and thus eliminate fakes significantly. Unlike state-of-the-arts studies [1, 11, 44, 46], it does not take additional constraints from the intermediate features. Such a one-loss design is easy to train. In addition, this design reduces the possible affections incurred by the inaccurate intermediate features and thus improves stability.

At last, note that the parallel design with multiple convolutions adopted by ECA are also appeared in the Inception module [30] and its variants [2, 15, 29, 31]. However, there are significant differences between ECA and these modules. They aims at discover redundant scales by sparse structures, so they stack those convolution and pooling operations repeatedly for more scales. Their direct concatenation of the convolved features cannot augment the object detection related effective contexts. ECA, on the other hand, aims at discovering the discriminative object contexts. Therefore, it only computes those convolutions once to initialize fine scales but, instead, further pool the convolved features for better discriminative features and fuse them with the global backbone deep features to boost the effective object contexts.

In summary, our contributions are mainly twofold. On one hand, an efficient context discovery module, ECA, is proposed. It takes different convolutions simultaneously to obtain initial features in fine scales and then enhances their discriminative features with the regular backbone features as global references. Output features with the boosted effective contexts for object detection can be obtained. On the other hand, an end-to-end deep framework is introduced. It follows the encoder-decoder idea to fulfill an easy-training shadow detection without fakes, where ECA is adopted into both parts to obtain an robust feature abstraction and classifier with effective shadow contexts. Designed with one-loss, it is easy to train without the distraction of the inaccurate intermediate features. Experiment results demonstrate the state-of-the-arts shadow detection performance, where our method outperforms existing ones with over 13.97% and 34.67% decreases on the challenging SBU and UCF [45] datasets in balance error rate (BER) [32] respectively.

In the following, after introducing related studies (Section 2), the details of our method (Section 3), including ECA, will be discussed. The experimental results are presented in Section 4 with the whole paper concluded at last in Section 5.

2 RELATED WORK

Related studies can be classified as traditional or deep methods according to whether they adopt the deep network based ideas.

Among the traditional methods, the early models mainly builds on the physical models using illumination-invariant priors [3–5]. Later on, handcreafted features are adopted based on the information from edge [13], color [3], texture [45], etc. Perhaps the representative methods are those based on region classification [7, 8, 33, 41]. These methods need specifically designed artificial features which are often not discriminative enough and thus difficult to apply in practice.

Deep learning based methods are popular in recent years, thanks to their high performances from the deep feature extraction. CNN can extract the effective multi-scale features for robust shadow detections [18, 22, 28, 34]. For example, Vicente *et al.* [34] adopted a patch-based CNN with image-level prior and image patches to detect shadows. Some of them [18, 28] do not take the end-to-end design. Those CNN based framework may not capture rich shadow characteristics.

Some authors [12, 36, 46] extended CNN by exploring more contextual cues. For example, Zhu [46] proposed the bidirectional pyramid network which combines the deeply global and shallowly local information together to obtain rich contextual features; Wang *et al.* [36] obtained global and local information by stacked multiple

parallel fusion branches. Our methods adopts ECA to find effective contexts through in-layer local convolutions and global references from the backbone networks.

Some authors [21, 25, 35] utilized generative adversarial networks (GAN) [6] to work around the limited training data [21, 25] or improve the discrimination power [35]. For example, Le *et al.* adopted two U-Nets [27] based GAN with the illumination priors to augment the dataset for more samples. More complicated adversarial model is proposed by Chen *et al.* [1], where a multi-task model learns shadow edge detection, shadow region detection, and shadow count detection simultaneously and the adversarial training is applied to the student-teacher networks.

Another idea is proposed by Zheng *et al.* [44] who took fake detection regions as the distraction areas and fused several existing results with ground truths to obtain labels for robust discrimination. This method requires to compute new ground truths for the distraction regions with several existing models.

Recently, Wang *et al.* [38] introduced an interesting work for detecting the shadow associated foreground instance. They presented a new dataset and an evaluation metric. However their focus is on the overall instance but not the shadow itself. Similar instance oriented work is also proposed by Wang *et al.* [37]. Inoue and Yamasaki [14] proposed a novel large-scale synthetic shadow/shadow-free/matte image triplets dataset and demonstrated improved performance on it.

Existing deep learning based shadow detection methods use the regular features from the layer-by-layer convolution, which can be unstable, considering the vast scale losses during the convolution. Some of them [1, 36, 44, 46] even rely on several constraints by computing more loss functions and thus may require additional computational resources due to inaccurate intermediate features. Hu *et al.* [11] additionally introduced the detail enhancement module (DEM) for complex shadows. Our model, however, includes ECA to boost discriminative multi-scale features by referring to the global regular features from the backbone networks. Therefore, it can augment effective contexts for object detection in comparison with the existing methods. In addition, our model only uses one loss and thus is easy to train with less computational load and also improves the stability.

The parallel design with multiple convolutions adopted by ECA have previously considered by the Inception module [30] and its several variants [2, 15, 29, 31]. They targeted at redundant scales in sparse structures and thus stacked those convolution and pooling operations repeatedly with direct concatenation for the output. This structure cannot discovery effective context for object discrimination which is exactly the target of ECA. Therefore, ECA only need compute the convolutions once, but additionally pools the convolved fine features for discriminative scales and further fuses them with the global deep features to boost effective object contexts.

3 OUR PROPOSED METHOD

The new method aims at discovering effective object contexts for a robust shadow detection. In the literature, contexts in different scales have already been considered where regular layer-by-layer convolution can provide layer-wise abstraction. However, this type of convolution does not weight the effective scales for the varying-sized objects, *i.e.*, those scales can be either overlooked as normal ones or even lost after the long deep learning process. As shown in Row *A* of Figure 2, the shadow feature of the normal object O_1 can always be found during the convolution process and thus it can be segmented successfully as shown on the right of this row. However, the dark surface of O_2 can always be taken as shadow during the abstraction process because its shadow context may not be properly considered in the intermediate layers. Similar observation can also be seen for the light shaded S_3 whose shadow feature gradually disappears without proper reception fields. The regular multi-scale features from the layer-by-layer convolution is not enough for a robust detection and effective object contexts are necessary as important cues for a robust detection.

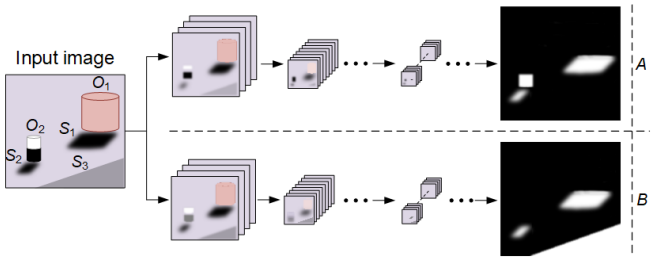


Figure 2: The principle of incorporating effective contexts. Object O_1 is the normal object with a shadow S_1; Object O_2 whose shadow is S_2 has a dark lower surface similar to shadow; Shadow S_3 is light shaded. Row *A* shows the general detection process in existing studies while Row *B* illustrates the detection process with effective contexts incorporated.

But how to obtain the effective contexts for shadow detection? It is observed that the human takes the surrounding distributions as references to judge where the dark areas are objects or shadows. For our deep network, regular deep features encode such global distributions and thus can be taken to weight the importance of different object features. And consequently, a new module, ECA, is proposed, which takes the regular deep features as global cues to boost the effect objects contexts.

In particular, ECA convolves the input feature with several different scales in parallel for features with fine contexts and then fuses their discriminative features as representatives with the regular deep features so that the significant contextual information can be globally augmented. Through incorporating ECA in each layer of the deep structure, more effective contexts can be gradually obtained in the subsequent layers of wider reception fields.

As demonstrated in Row *B* of Figure 2, the novel layer-wise features can dig discriminative contexts through the layer-by-layer convolution and, then, the dark surface of O_2 is gradually filtered out, while the light shaded S_3 are kept salient. Such strong features can ensure the success of the final detection as shown on the right side of Row *B*.

Discriminative contexts can also guide the final classification. Therefore, an encoder-decoder based framework is proposed so that the appropriate scales are integrated into the decoding process at multiple levels.

3.1 The general pipeline

The proposed framework takes the U-Net [27] like encoder-decoder structure (Figure 3). The global contexts for the correspondingly discriminative contexts are generated with the regular features from ResNet-101 [40] in four scales.

Accordingly, the propose method works as follows to obtain an end-to-end shadow detection for an input image. First the encoder part processes through layer-by-layer coding to obtain its condensed representation. Then the decoder is applied to map the representation scale by scale to generate its final shadow distribution. In each layer of the encoder, ECA discovers discriminative contexts through multi-scale convolutions, feature pooling and final fusion with the regular deep feature in corresponding scale as ECA feature of this layer. This feature is then input to the next ECA for the next layer to continue the abstraction process for discriminative contexts. When decoding, the ECA feature is also fused again with the feature map of the same convolution layer so that effective-contexts guided generation can be obtained in the decoder. At last, a 1×1 convolution layer is applied and thresholded by Sigmoid to obtain the final prediction.

We now discuss the details of ECA.

3.2 ECA

ECA consists of two main parts: Fine-scale-feature preparation and effective-context boosting. The first step aims at collecting features in fine scales so that richer contexts can be discovered, while the second step can enhance the effective contexts with the discriminative features by the global reference of the regular deep features. The output of ECA is then a feature with robust contexts for object detection.

First comes the fine-scale-feature preparation in ECA. It convolves the input feature with 1×1, 3×3 and 5×5 convolutions simultaneously. Consequently, there will be triple of the original channels and, therefore, a preliminary fusion is then required to reduce parameters and avoid overfitting for capturing important information. Here, a 3×3 convolution is applied to them. Assume the convolution operation $f^m_{(P,Q)}(\boldsymbol{V})$ for the input feature $\boldsymbol{V}$ by the mth $P \times Q$ kernel, $\boldsymbol{w}^m_{(P,Q)}$, as

$$f^m_{(P,Q)}(\boldsymbol{V}) = ReLU(\boldsymbol{w}^m_{(P,Q)} \otimes \boldsymbol{V} + \boldsymbol{b}^m_{(P,Q)})), \tag{1}$$

where $\boldsymbol{b}^m_{(P,Q)}$ is the bias for the mth kernel and $ReLU$ is the rectified linear activation function. The preliminary feature F^m_{pre} from the preliminary fusion can be formulated as

$$F^m_{pre}(\boldsymbol{V}) = f^m_{(3,3)}(C((f^m_{(1,1)}(\boldsymbol{V}), f^m_{(3,3)}(\boldsymbol{V})), f^m_{(5,5)}(\boldsymbol{V}))), \tag{2}$$

where C denotes the concatenation.

Then, the effective-context boosting is applied. Here the preliminary feature is first max pooled to obtain more discriminative features in different scales for further augmentation of effective contexts. Then the regular deep feature from the corresponding layer is input and fused with pooled features for more discriminative scales on the current layer. Such a concatenated feature may contain noise after many convolutions and fusions and, therefore, it is refined again by a 3×3 filter kernel as the ECA feature. This refinement also reduce the feature dimension and avoid overfitting. Assume the regular feature being F^m_{reg}. The ECA feature $F^m_{ECA}(\boldsymbol{V})$ for $\boldsymbol{V}$ computed by these operations can be formulated as

$$F^m_{ECA}(\boldsymbol{V}) = f^m_{(3,3)}(C(\mathcal{P}(F^m_{pre}(\boldsymbol{V})), F^m_{reg})), \tag{3}$$

where $\mathcal{P}$ is the 2×2 maxpooling.

This ECA feature is then input to the next ECA for further abstraction with next regular scales, so that robust contexts in different scales can be encoded at last.

On decoding, the ECA feature can act as guidance to the inference process for a robust classifier. Here the skip connection similar to U-Net is applied. Details on our connection method is explained in the experimental part (Section 4).

3.3 Training and testing

The proposed method is implemented with Pytorch. ResNet-101 is pre-trained by ImageNet to supply the regular features.

3.3.1 Loss design. Intermediate features are sometimes inaccurate and, therefore, taking them as loss constraints may degrade the performance. In addition, those intermediate losses may complicate the convergence process. Therefore, to simplify the training process and reduce the effect of intermediate features, we prefer to computing the loss once with the final output. In particular, we adopt the idea of binary cross-entropy and take a weighted cross-entropy to balance the contributions from positive and negative samples. It will incur heavy computational burdens if computing the weights in each batch. Therefore, these weights are computed from the samples in advance. This function can be formally written as

$$L = -\sum_i (\lambda y^{lab}_i \log y^{pre}_i + (1-\lambda)(1-y^{lab}_i) \log (1-y^{pre}_i)), \tag{4}$$

where: y^{pre}_i and y^{lab}_i represent the model prediction and the ground truth of the shadow class of the i-th pixel, respectively; and λ is the weight which is set to 0.7 by our experiments.

3.3.2 Training and testing parameters. Training is done with a single GeForce GTX 1080 Ti. SBU, ISTD [35] and CUHK [11] datasets are taken as the training sets. All samples are resized as 256×256. Adamax is taken as the optimizer and the learning rate is set to 0.0005.

Testing takes the same resize process as training. The results are then upsampled by bilinear interpolation to restore their original resolutions as final detection results. Note that we don't postprocess the results by CRF [19] as some other methods do [1, 12, 44, 46].

4 EXPERIMENTAL RESULTS

Our main experimental comparisons are undertaken with SBU, UCF, ISTD and CUHK datasets. The numbers of training images are 4089, 1330 and 7350 while the numbers of testing images are 638, 540 and 2100 for SUB, ISTD and CUHK respectively. UCF contains only 221 images and, therefore, is used for testing.

The performance of ECA is experimented with different configurations. Popular methods, such as modules with stacked parallel convolutions (Inception-v4 [29] and Xception [2]) and deep feature oriented structures (ResNet-101 [40]) are also taken for performance comparison. These methods are for rich contexts, while ECA aims

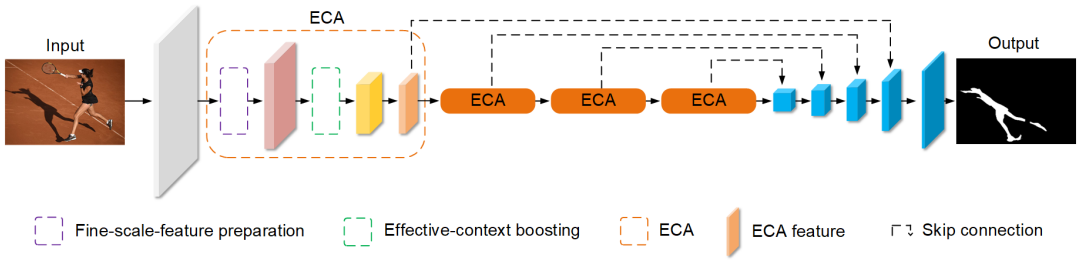


Figure 3: The architecture of the propose method. It fulfills an end-to-end shadow detection in the encoder-decoder style, where ECA is incorporated into the encoder as the main building blocks to gradually extract layer-wise features with effective detection contexts.

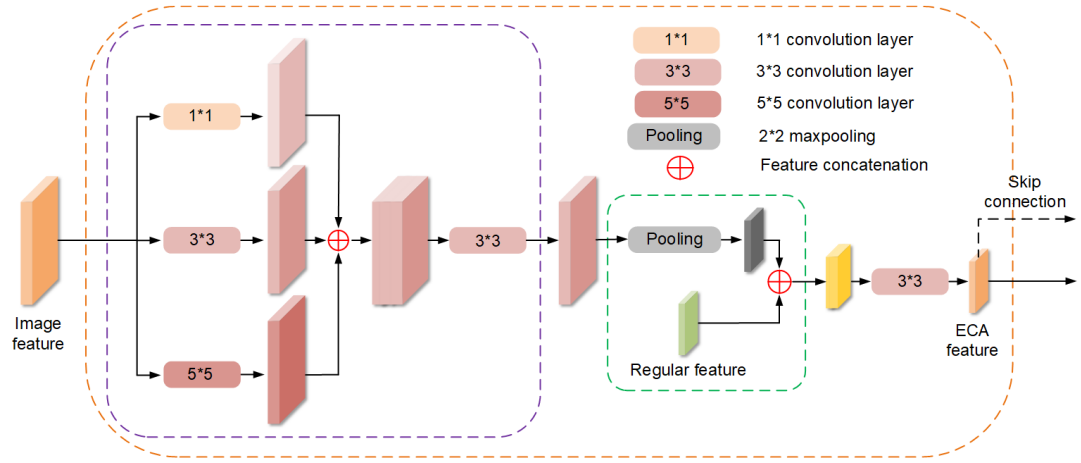


Figure 4: The structure of ECA.

at effective object contexts. Note that in our experiments the experiments on Inception-v4 and Xception are fulfilled by replacing ECA in our method with them directly.

Several state-of-the-art shadow detection methods are taken for the performance comparison, including FSDNet [11], MTMT-Net [1], DSDNet [44], BDRAR [46], DC-DSPF [39], ADNet [21], DSC [12], ST-CGAN [35], scGAN [25] , stacked-CNN [34] and patched-CNN [9]. Three saliency detection models, i.e., SRM [36], Amulet [42] and EGNet [43] are also included. For fair comparison, we take the quantitative results from the authors directly. However, for the qualitative results, not all methods provide the complete testing results for all data. Therefore, we simply copy their results directly if they provided. Among the left methods, we implement ADNet for visual comparisons because there is no open code for it; and we also train ST-CGAN with the provided codes for testing.

4.1 Qualitative Results

We first show the context enhancement performance of ECA with extracted 16×16 and 32×32 features (Figure 5). Here, two types of features from ECA, ECA-*before fusion* and ECA-*after fusion* are adopted to show its incremental efficacy. The former represents the discriminative features pooled from the multi-scale features with the latter for the final ECA output. The discriminative features from ECA can apparently capture the shadow areas more significantly than Inception-v4, Xception and ResNet-101, while the shadow responses in the final ECA features are generally most strong and apparent among all corresponding features, thanks to the discriminative-context augmentation from ECA. The features of Inception-v4, Xception and ResNet-101 are full of rich contexts but not sensitive to the shadow, which are also in line with their rich scale discovery initiatives.

Figure 6 shows the detection results among existing methods and ours. Our method is good at the light shaded shadows and resistant to the dark surfaces and obtains better performances than other methods. Incorporating ECA as the main building block for the robust features, our method shows its strong abstraction and robust generation abilities.

The merits of ECA and our shadow detection method are also tested with challenging images as shown in Figure 7. They represent the scenes with a lot of dark surfaces (*e.g.*, Brick), mixed shadows of different types (*e.g.*, Bottle) and extremely light shaded shadows in comparison with the backgrounds (*e.g.*, Stage). However, our method still achieves the best accuracies among all methods.

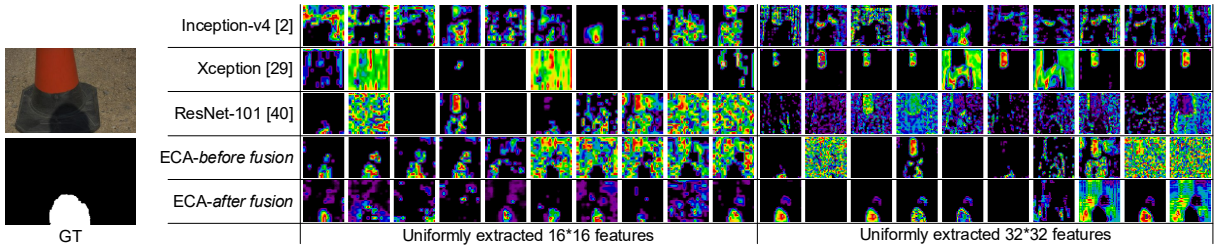


Figure 5: Context enhancement comparison among ECA and other modules by abstracted features.

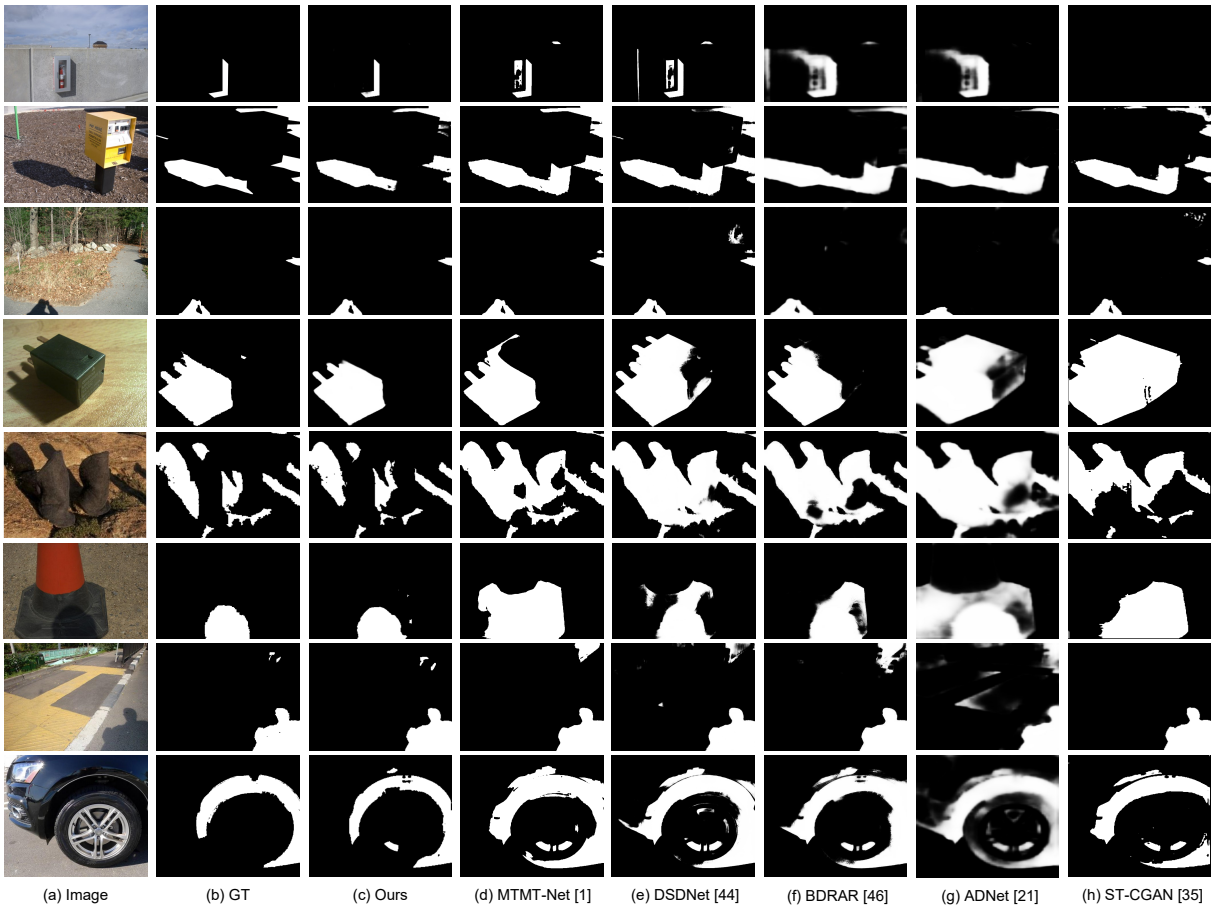


Figure 6: Shadow detection results of our method in comparison with existing ones.

4.2 Quantitative Results

The popular metric, BER, is adopted to evaluate the ability in obtaining balanced results,

$$BER = 1 - \frac{1}{2}\left(\frac{T_P}{T_P + F_N} + \frac{T_N}{T_N + F_P}\right), \tag{5}$$

where T_P, T_N, F_N and F_P are the numbers of true positives, true negatives, false negatives and false positives, respectively. The lower BER is, the better the performance is. Note that $T_P + F_N$ and $T_N + F_P$ represent the numbers of shadow and non-shadow pixels respectively.

Table 1 shows the statistical comparison results. Our method is good at all datasets. Especially, for SBU and UCF respectively, it obtains 13.97% and 34.67% lower in BER than the state-of-the-arts method, MTMT-Net.

4.3 Cross comparison

The generalization ability is also tested. For fair comparison, our method is compared with ST-CGAN, scGAN and stacked-CNN whose results are provided by the authors of ST-CGAN. The test is performed in the same way as them, where testing images from three datasets (SBU, ISTD and UCF) are evaluated with the model

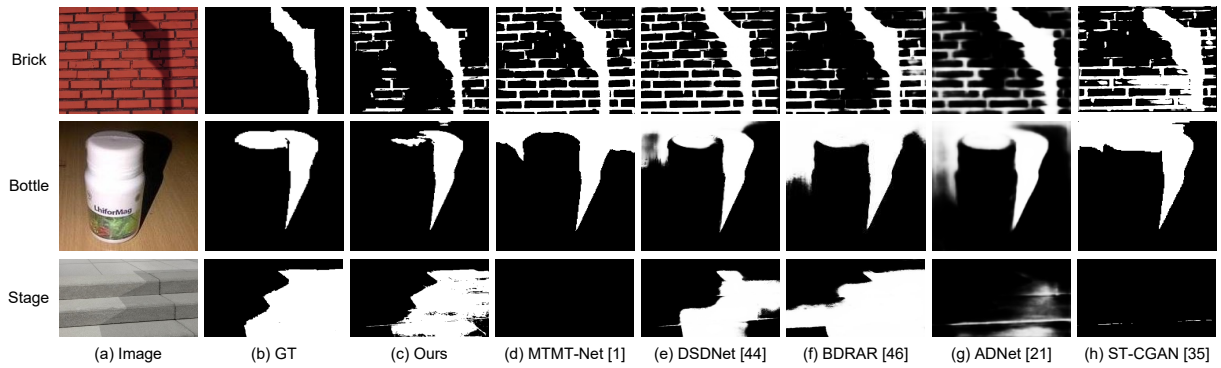


Figure 7: More detection results among existing methods and ours for challenging images.

Table 1: Quantitative comparisons among state-of-the-arts methods and ours in BER.

	SBU	ISTD	UCF	CUHK
Our method	**2.71**	**1.57**	**4.88**	**8.05**
FSDNet [11]	-	-	-	8.65
MTMT-Net [1]	3.15	1.72	7.47	-
DSDNet [44]	3.45	2.17	7.59	8.27
BDRAR [46]	3.64	2.69	7.81	9.18
DC-DSPF [39]	4.90	-	7.90	-
ADNet [21]	5.37	3.26	9.25	12.43
DSC [12]	5.59	3.42	10.54	8.65
ST-CGAN [35]	8.14	3.85	11.23	-
scGAN [25]	9.10	4.70	11.50	-
stacked-CNN [34]	11.00	8.6	13.00	-
patched-CNN [9]	11.56	-	-	-
SRM [36]	6.51	7.92	12.51	-
Amulet [42]	15.13	-	15.17	-
EGNet [43]	4.49	1.85	9.20	-

trained from SBU or ISTD respectively. CUHK is not included because there are no available results for ST-CGAN, scGAN and stacked-CNN. Table 2 shows that our model can reach higher generalization than ST-CGAN, scGAN and stacked-CNN.

Table 2: Cross comparison results between ST-CGAN and our method in BER.

	Training with SBU			Training with ISTD		
Testing dataset	SBU	ISTD	UCF	SBU	ISTD	UCF
Ours	**2.71**	**3.79**	**4.88**	**5.37**	**1.57**	**8.25**
ST-CGAN [35]	8.14	7.35	11.23	11.34	3.85	16.18
scGAN [25]	9.10	8.98	11.50	13.26	4.70	16.41
stacked-CNN [34]	11.00	10.45	13.00	15.94	8.60	18.67

4.4 Ablation studies

Several ablated versions of ECA are first taken to evaluate its effectiveness for shadow detection:

- w/o F^m_{pre}: ECA using 3×3 convolution kernel only without F^m_{pre}.
- Skip connection w/ F^m_{pre}: The full ECA but the skip connection starting from F^m_{pre}.
- Skip connection w/ F^m_{ECA}: The full ECA.

Table 3 shows the experimental results in BER, which validates that the efficacy of ECA feature for efficient shadow detection.

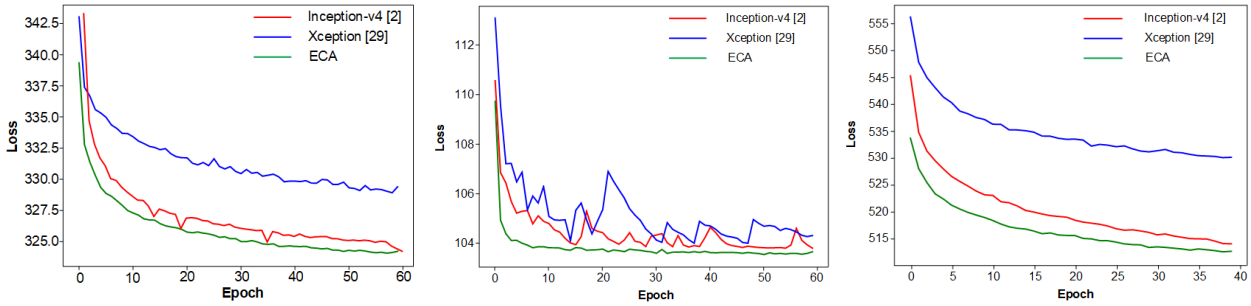
Table 3: Results of the ablation experiment for the effectiveness of ECA in BER.

	SBU	ISTD	UCF	CUHK
w/o F^m_{pre}	4.21	2.88	6.33	10.26
Skip connection w/ F^m_{pre}	3.99	1.79	5.46	9.31
Skip connection w/ F^m_{ECA}	**2.71**	**1.57**	**4.88**	**8.05**

Table 4: Results of the ablation experiment for different feature abstraction methods in BER.

	SBU	ISTD	UCF	CUHK
ECA(Our method)	**2.71**	**1.57**	**4.88**	**8.05**
ECA-*Simplified*	5.97	4.45	7.48	9.18
Inception-v4 [2]	2.91	1.69	6.99	8.91
Xception [29]	3.04	1.76	7.46	9.32
ResNet-101 [40]	4.91	3.01	8.24	9.98
U-Net (baseline) [27]	7.59	3.34	10.42	12.36

The ablation study with different feature abstraction modules and methods is also performed (Table 4). Here U-Net [27] is taken as a baseline, while pure multi-resolution features from ECA modules without the regular features (ECA-*Simplified*) are also compared. ECA achieves the best performances among all methods. This study shows that our method will not work without the regular features to provide general object information (e.g., position, silhouettes) as global guidance to augment object discrimination with effective contexts after fusing with the fine-scale features from the multiple convolutions. It also shows that just adopting some pretrained backbone networks (e.g., ResNet) may not be enough for low-level



(a) SBU (b) ISTD (c) CUHK

Figure 8: The convergence speeds of different modules on SBU, ISTD and CUHK during training. Note: UCF is not considered because it is too small to train stably.

tasks. They can obtain the general object features, but be difficult to discover effective contexts without enough object details. Therefore, they may not be robust enough for the low-level applications.

Further ablation study on ECA can be explored by the convergence speeds during training (Figure 8). The data from the top three modulesin previous ablation study (Table 4) are collected for comparisons, which are based on the results from the first 40 epochs. ECA can reach the fastest convergence speeds with the most stable loss decreasing performance for all datasets among all methods.

4.5 Extension to shadow removal

The proposed end-to-end framework can be also taken as the general backbone network. ResNet features act as global guidance when fusing with the fine-scale discriminative features and thus help obtaining the effective contexts for robust object abstraction. Such a strong-discrimination network can be applied to various applications (such as classification, regression, etc.) for better performances. Here we show such an example by extending the propose framework to shadow removal for the images from ISTD (Figure 9). In this case, the output layer of our method is convolved back to have three channels for restoration. The loss is defined as the L_1 loss [10]. Empowered by the strong context-discovery ability of ECA, our method shows potential applications to various areas.

Figure 9: Test results on applying our method to shadow removal.

5 CONCLUSION

Deep network can easily overlook or even discard contexts not in the specified scales and thus may miss some important cues for an effective shadow judgment. This paper introduce an effective-object-context augmentation module, ECA, which can fuse regular deep features with discriminative features from the simultaneous multi-scale convolutions and thus boost the appropriate object contexts for effective object detection. Taking the ECA feature to the next layer and applying the same ECA feature exploration iteratively will finally obtain a robust deep feature with strong object contexts. The ECA feature can also be taken to guide the generation of shadow distribution. Therefore, a novel end-to-end shadow detection network is introduced, which integrates ECA into both encoder and decoder to enhance feature abstraction and guide the classification. Designed with only one loss, our method is easy to train without the instability brought by the additional losses with the inaccurate intermediate features. Experimental results show our method can eliminate fakes and achieve better performances than existing methods. They also demonstrate that the proposed framework can be potentially applied to various areas as a novel backbone due to the strong discriminative power of ECA.

The proposed method can achieve better detections than other methods for very complicated scenes with several types of shadows (*e.g.*, the Bottle in Figure 7) or extremely fake objects (*e.g.*, the Brick and Stage in Figure 7). But it still has space to improve. The deep network may generate incorrect predictions and thus bias toward them. Overcoming this difficulty is our future work.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of Anhui Province, China (Grant No. 2108085MF210).

REFERENCES

[1] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. 2020. A Multi-Task Mean Teacher for Semi-Supervised Shadow Detection. In *CVPR*. 5610–5619. https://doi.org/10.1109/cvpr42600.2020.00565

[2] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*. 1251–1258.

[3] G.D. Finlayson, S.D. Hordley, Cheng Lu, and M.S. Drew. 2006. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 1 (2006), 59–68. https://doi.org/10.1109/tpami.2006.18

[4] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. 2004. Intrinsic Images by Entropy Minimization. In *Lecture Notes in Computer Science*. 582–595. https://doi.org/10.1007/978-3-540-24672-5_46

[5] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. 2009. Entropy Minimization for Shadow Removal. *International Journal of Computer Vision* 85, 1 (2009), 35–57. https://doi.org/10.1007/s11263-009-0243-z

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[7] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2011. Single-image shadow detection and removal using paired regions. In *CVPR*. 2033–2040. https://doi.org/10.1109/cvpr.2011.5995725

[8] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2013. Paired Regions for Shadow Detection and Removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2956–2967. https://doi.org/10.1109/tpami.2012.214

[9] Sepideh Hosseinzadeh, Moein Shakeri, and Hong Zhang. 2018. Fast Shadow Detection from a Single Image Using a Patched Convolutional Neural Network. In *IROS*. 3124–3129. https://doi.org/10.1109/iros.2018.8594050

[10] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. 2020. Direction-Aware Spatial Context Features for Shadow Detection and Removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 11 (2020), 2795–2808. https://doi.org/10.1109/tpami.2019.2919616

[11] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. 2021. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing* 30 (2021), 1925–1934.

[12] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. 2018. Direction-Aware Spatial Context Features for Shadow Detection. In *CVPR*. 2795–2808. https://doi.org/10.1109/cvpr.2018.00778

[13] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. 2011. What characterizes a shadow boundary under the sun and sky?. In *ICCV*. 898–905. https://doi.org/10.1109/iccv.2011.6126331

[14] Naoto Inoue and Toshihiko Yamasaki. 2020. Learning from Synthetic Shadows for Shadow Detection and Removal. *IEEE Transactions on Circuits and Systems for Video Technology* (2020), 1–1.

[15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*. 448–456.

[16] Imran N. Junejo and Hassan Foroosh. 2008. Estimating Geo-temporal Location of Stationary Cameras Using Shadow Trajectories. In *ECCV*. 318–331. https://doi.org/10.1007/978-3-540-88682-2_25

[17] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. 2011. Rendering synthetic objects into legacy photographs. In *SIGGRAPH Asia*. 157:1–157:12. https://doi.org/10.1145/2024156.2024191

[18] Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. 2014. Automatic Feature Learning for Robust Shadow Detection. In *CVPR*. 1939–1946. https://doi.org/10.1109/cvpr.2014.249

[19] Philipp Krhenbühl and Vladlen Koltun. 2011. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NIPS*. 109–117.

[20] Jean-Francois Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. 2009. Estimating natural illumination from a single outdoor image. In *ICCV*. 183–190. https://doi.org/10.1109/iccv.2009.5459163

[21] Hieu Le, Tomas F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. 2018. A+D Net: Training a Shadow Detector with Adversarial Shadow Attenuation. In *ECCV*. 680–696. https://doi.org/10.1007/978-3-030-01216-8_41

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 640–651. https://doi.org/10.1109/cvpr.2015.7298965

[23] I. Mikic, P.C. Cosman, G.T. Kogut, and M.M. Trivedi. 2000. Moving shadow and object detection in traffic scenes. In *ICPR*. 321–324. https://doi.org/10.1109/icpr.2000.905341

[24] Vu Nguyen, Tomas F. Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. 2017. Shadow Detection with Conditional Generative Adversarial Networks. In *ICCV*. 4510–4518. https://doi.org/10.1109/iccv.2017.483

[25] Vu Nguyen, Tomas F. Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. 2017. Shadow Detection with Conditional Generative Adversarial Networks. In *ICCV*. 4510–4518. https://doi.org/10.1109/iccv.2017.483

[26] Alexandros Panagopoulos, Dimitris Samaras, and Nikos Paragios. 2009. Robust shadow and illumination estimation using a mixture model. In *CVPR*. https://doi.org/10.1109/cvpr.2009.5206665

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*. 234–241.

[28] Li Shen, Teck Wee Chua, and Karianto Leman. 2015. Shadow optimization from structured deep edge detection. In *CVPR*. https://doi.org/10.1109/cvpr.2015.7298818

[29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*. 4278–4284.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9. https://doi.org/10.1109/cvpr.2015.7298594

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[32] Tomas F. Yago Vicente, Minh Hoai, and Dimitris Samaras. 2015. Leave-One-Out Kernel Optimization for Shadow Detection. In *ICCV*. 3388–3396. https://doi.org/10.1109/iccv.2015.387

[33] Tomas F. Yago Vicente, Minh Hoai, and Dimitris Samaras. 2018. Leave-One-Out Kernel Optimization for Shadow Detection and Removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 3 (2018), 682–695. https://doi.org/10.1109/tpami.2017.2691703

[34] Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-Scale Training of Shadow Detectors with Noisily-Annotated Shadow Examples. In *ECCV*. 816–832. https://doi.org/10.1007/978-3-319-46466-4_49

[35] Jifeng Wang, Xiang Li, and Jian Yang. 2018. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. In *CVPR*. 1788–1797. https://doi.org/10.1109/cvpr.2018.00192

[36] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. 2017. A Stagewise Refinement Model for Detecting Salient Objects in Images. In *ICCV*. 4039–4048. https://doi.org/10.1109/iccv.2017.433

[37] Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, and Pheng-Ann Heng. 2021. Single-Stage Instance Shadow Detection with Bidirectional Relation Learning. In *CVPR*.

[38] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. 2020. Instance Shadow Detection. In *CVPR*. https://doi.org/10.1109/cvpr42600.2020.00195

[39] Yupei Wang, Xin Zhao, Yin Li, Xuecai Hu, and Kaiqi Huang. 2018. Densely Cascaded Shadow Detection Network via Deeply Supervised Parallel Fusion. In *IJCAI*. 1007–1013. https://doi.org/10.24963/ijcai.2018/140

[40] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*. 5987–5995. https://doi.org/10.1109/cvpr.2017.634

[41] Xingsheng Yuan, Marc Ebner, and Zhengzhi Wang. 2015. Single-image shadow detection and removal using local colour constancy computation. *IET Image Processing* 9, 2 (2015), 118–126. https://doi.org/10.1049/iet-ipr.2014.0242

[42] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. 2017. Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection. In *ICCV*. 202–211. https://doi.org/10.1109/iccv.2017.31

[43] Jiaxing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. 2019. EGNet: Edge Guidance Network for Salient Object Detection. In *ICCV*. 8778–8787. https://doi.org/10.1109/iccv.2019.00887

[44] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. 2019. Distraction-Aware Shadow Detection. In *CVPR*. 5167–5176. https://doi.org/10.1109/cvpr.2019.00531

[45] Jiejie Zhu, Kegan G. G. Samuel, Syed Z. Masood, and Marshall F. Tappen. 2010. Learning to recognize shadows in monochromatic natural images. In *CVPR*. 223–230. https://doi.org/10.1109/cvpr.2010.5540209

[46] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. 2018. Bidirectional Feature Pyramid Network with Recurrent Attention Residual Modules for Shadow Detection. In *ECCV*. 122–137. https://doi.org/10.1007/978-3-030-01231-1_8