Mitigating Intensity Bias in Shadow Detection via Feature Decomposition and Reweighting

Lei Zhu[1] Ke Xu[2†] Zhanghan Ke[1] Rynson W.H. Lau[1†]

[1]City University of Hong Kong [2]Shanghai Jiao Tong University

lzhu68-c@my.cityu.edu.hk, kkangwing@gmail.com, zhanghake2-c@my.cityu.edu.hk, Rynson.Lau@cityu.edu.hk

Abstract

Although CNNs have achieved remarkable progress on the shadow detection task, they tend to make mistakes in dark non-shadow regions and relatively bright shadow regions. They are also susceptible to brightness change. These two phenomenons reveal that deep shadow detectors heavily depend on the intensity cue, which we refer to as intensity bias. *In this paper, we propose a novel feature decomposition and reweighting scheme to mitigate this intensity bias, in which multi-level integrated features are decomposed into intensity-variant and intensity-invariant components through self-supervision. By reweighting these two types of features, our method can reallocate the attention to the corresponding latent semantics and achieves balanced exploitation of them. Extensive experiments on three popular datasets show that the proposed method outperforms state-of-the-art shadow detectors.*

1. Introduction

Shadows may appear when lights cannot directly reach an object surface. While they provide cues on light source directions and scene illuminations, which facilitate scene understanding [20, 22], they may adversely affect the performance of computer vision tasks [7, 28]. Hence, shadow detection is crucial.

Earlier, many works were proposed to detect shadows using hand-crafted features. However, these methods are unreliable and may fail in complex scenes. Recently, deep learning based shadow detection methods have shown superior performance over traditional methods by a large margin. Being trained in an end-to-end manner, deep shadow detectors can automatically learn discriminative features for detection, without the efforts to specify which cues to use and how they should be represented. However, there is a cost to this convenience. There are signs that existing deep shadow detectors heavily rely on the intensity cue. While they may mis-recognize relatively bright shadow regions as non-shadows (Fig. 1 row 1) or dark non-shadow regions as shadows (Fig. 1 row 3), with a small shift in brightness (which should not change the image semantics), the detection results may change significantly (Fig. 1 rows 2 and 4).

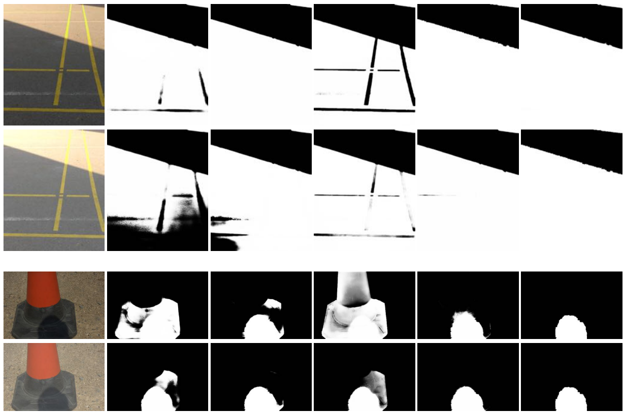

(a) Input Image (b) DSC [16] (c) DSD [50] (d) MTMT [5] (e) Ours (f) GT

Figure 1. Intensity bias in shadow detection. Rows 1 and 3 show two original images, while rows 2 and 4 show the two images with 20% increase in intensity. Existing methods [16, 50, 5] heavily rely on the intensity cue, and suffer from two problems. On the one hand, they mis-recognize a relatively brighter region inside the shadow as non-shadow (*e.g.*, yellow lanes in row 1), and dark non-shadow region as shadow (*e.g.*, the traffic cone in row 3). On the other hand, their predictions change significantly due to the brightness change (rows 2 and 4). Our method mitigates this intensity bias and produces more consistent and accurate results.

Although low intensity is a strong indication of shadows, other cues such as object-shadow correspondences, shadow edges, and region connectivity may also contribute to the shadow detection task [33]. However, deep models tend to be attracted by some dominant cues while leaving the less dominant ones underexplored. For example, Geirhos *et al.* [12] show that ImageNet-trained classifiers typically have a bias towards textures, and increasing the attention to shapes helps improve classification accuracy and robustness. Choi *et al.* [6] also show that reducing scene bias improves the generalization of action recognition networks.

To mitigate such a bias towards the intensity cue in shadow detection, the most straightforward solution is to

[†] Ke Xu and Rynson Lau are joint corresponding authors. Rynson Lau leads this project.

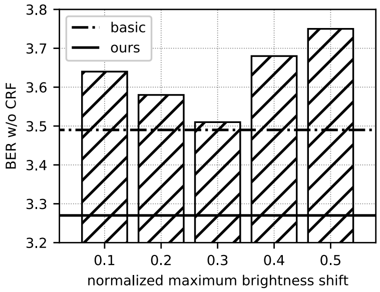


Figure 2. Augmenting the training data with a brightness shift of different extents (vertical bars) increases the balanced error rate (BER, lower is better), compared to the result without using data augmentation (dash horizontal line). Our method mitigates the intensity bias and produces a much better performance. All results are from a FPN-like network [26] trained on the SBU dataset [40].

apply data augmentation, such as random brightness shift, to implicitly impose the intensity-invariant constraint on the network. Unfortunately, such a simple strategy does not work and may even degrade the detection accuracy in practice, as shown in Fig. 2. This is due to distributional discrepancy between the true data and the augmented data [5]. Alternatively, existing works have tried to explicitly introduce or reinforce other cues. For example, Hu *et al.* [16] propose a module to model the contrast information in a direction-aware manner. Chen *et al.* [5] propose to incorporate shadow edges and shadow count in the detection. However, introducing these specific cues cannot address this intensity bias problem well, as demonstrated in Fig. 1.

Instead, we propose in this paper a novel feature decomposition and reweighting scheme to combat the undue attention to intensity. Specifically, we first decompose deep shadow features into intensity-variant (*i.e.*, responding to intensity changes) and intensity-invariant (*i.e.*, without responding to intensity changes) components, so that the network can mine the two types of features individually, and then re-integrate them with appropriate weights.

The challenge of this approach is how to decompose highly coupled features into intensity-variant and intensity-invariant components. To guide the learning of such decomposition, we construct two novel self-supervised tasks. While one aims to minimize the difference between the intensity-invariant features extracted from the input image and its brightness-shifted counterpart, the other is to predict the brightness shift from the intensity-variant features (Sec. 3.1). To reallocate the attention to the decomposed features, during the training stage, we gradually shift the learning focus from the intensity-variant features to the intensity-invariant features via cumulative learning [51]. We then search for an optimal weight on the validation set and fix it for the inference stage (Sec. 3.2). In summary, our main contributions are three-fold:

- We propose a novel feature decomposition and reweighting scheme to mitigate the intensity bias in shadow detection, which allows our shadow detector to reallocate its attention between the intensity-variant and intensity-invariant features.
- We propose a novel self-supervised approach, which is tailored for shadow detection, to guide the decomposition of deep features.
- Extensive experiments on three public datasets demonstrate that the proposed method outperforms state-of-the-art shadow detection methods.

2. Related Works

Shadow detection. To detect shadows in a single image, earlier works propose physical models [11, 10, 38] or machine learning classifiers based on handcrafted features [23, 53, 14]. Typically, they exploit one or more heuristic cues, such as chromacity [11, 10, 23, 14, 39], edge [11, 23, 53, 17], intensity [38, 14, 53, 17, 39], and texture [53, 14, 39]. However, these methods cannot handle real-world complex scenes well, when the assumptions (*e.g.*, uniform illumination) made in these physical models are violated, or the hand-crafted features used in the traditional machine learning classifiers fail to represent the shadow patterns.

Recently, deep learning based methods have shown great success in shadow detection. Nguyen *et al.* [29] first propose a tailored conditional GAN for this task. Hu *et al.* [16] propose to detect shadows by learning global contextual features in a direction-aware manner. Le *et al.* [24] propose to jointly learn a shadow detector with another network for generating augmented training data using adversarial training. Zhu *et al.* [55] propose to fuse multi-level features recursively and bidirectionally. Wang *et al.* [41] propose a stacked conditional GAN to jointly learn shadow detection and removal. Zheng *et al.* [50] propose to learn distraction-aware features for shadow detection by learning from the false predictions of other deep shadow detectors. Most recently, Chen *et al.* [5] introduce a teacher-student framework [37] to exploit additional unannotated shadow images. They also explicitly detect shadow boundaries to improve the detection accuracy.

While compelling results are obtained by deep shadow detectors, we note that they tend to predict a brighter region within a shadow as non-shadow and a dark region as shadow. In addition, their predictions may change significantly as we adjust the brightness of the input image. Such phenomenons indicate that they rely too much on the intensity cue to make their predictions. This motivates our work in attempting to balance the impact of intensity-variant and intensity-invariant features.

Self-supervised learning. Towards task-agnostic image representation learning, self-supervised learning has received remarkable attention in the representation learning community recently. As its name implies, in self-supervised learning, the supervision signals are derived not from human annotations, but from the input images themselves. One line of self-supervised learning relies on elaborate pretext tasks such as predicting relative patches [9], solving jigsaw puzzles [31], colorization [47], and predicting the degree of image rotation [13, 4, 25]. Without explicitly constructing artificial labels, another line of works adopt the contrastive learning strategy [44, 2, 3], in which the image representation is obtained by contrasting positive and negative pairs in the feature embedding space.

Unlike those methods designed for task-agnostic visual representation, we apply self-supervised learning for a task-specific objective: decomposing the discriminative shadow features into intensity-variant and intensity-invariant components. In addition, instead of constructing a single self-superivised task, a pair of self-supervised tasks, including a novel pretext task and the contrastive learning approach, are jointly exploited to learn such a decomposition.

3. Proposed Method

Existing deep shadow detectors place undue importance to the intensity cue. To mitigate such a bias, our key idea is to readjust the network's focus on the dominant intensity cue and excavate other less dominant cues. However, as deep features encode all cues together in a coupled manner, it is not easy to reallocate the attention to some specific ones. Hence, we propose a feature decomposition and reweighting (FDR) scheme to achieve such controllability.

Fig. 3 shows the workflow of the proposed FDR scheme, with both training and inference stages.

3.1. Self-supervised Feature Decomposition

We introduce a pair of contradictory self-supervised tasks to guide the learning of intensity-variant and intensity-invariant features in the training stage. They are separately applied to bilateral branches to encourage such a mutually complementary decomposition.

Formally, given a training image $\mathbf{I}$ as input, we first randomly shift its brightness to produce a counterpart $\mathbf{I}'$:

$$\mathbf{I}' = \mathbf{I} + \gamma, \tag{1}$$

where γ is the shift amount. It is a random variable uniformly sampled from the range of $[-\Delta, \Delta]$. We then feed both $\mathbf{I}$ and $\mathbf{I}'$ to the Feature Extraction Subnetwork followed by the FDR module. The bilateral projection branches of the FDR modules output four intermediate feature maps: $\mathbf{F}_i$ (intensity-invariant features of $\mathbf{I}$), $\mathbf{F}_v$ (intensity-variant features of $\mathbf{I}$), $\mathbf{F}'_i$ (intensity-invariant features of $\mathbf{I}'$), and $\mathbf{F}'_v$ (intensity-variant features of $\mathbf{I}'$).

Since we do not expect the intensity-invariant features to change under such a brightness shift, we introduce the first self-supervised task here to force these two intensity-invariant features (*i.e.*, $\mathbf{F}'_i$ and $\mathbf{F}_i$) to be consistent, as:

$$\mathcal{L}_i = \Phi_{MAE}(\mathbf{F}_i, \mathbf{F}'_i), \tag{2}$$

where $\Phi_{MAE}(\cdot,\cdot)$ is a mean-absolute-error loss function. In addition, we also expect that the intensity related information is encoded in the intensity-variant features. Hence, we formulate the second self-supervised task as learning to predict the perturbation amount γ from $\mathbf{F}'_v$. Specifically, we attach an auxiliary regression head $\phi(\cdot;\theta)$ parameterized by global average pooling (GAP), followed by a fully connected (FC) layer. It takes $\mathbf{F}'_v$ as input to predict γ, and the corresponding pretext loss is defined as:

$$\mathcal{L}_v = \Phi_{MAE}(\phi(\mathbf{F}'_v;\theta), \gamma). \tag{3}$$

Obviously, predicting the brightness shift from $\mathbf{I}'$ without referring to the original image $\mathbf{I}$ is ill-posed and challenging. However, it is worth noting that an ill-posed pretext task is widely used in self-supervised representation learning, such as predicting image rotation [13, 4, 25] or flip [27] *without reference*. On the one hand, for representation learning, it is crucial to avoid the network learning to solve the pretext task by exploiting low-level visual features [32], which can easily happen when we provide reference images. On the other hand, as a *pretext* task, what we concern about is the representation induced by it, instead of how well the network can do for itself. In our case, predicting the brightness shift without reference images forces the network to encode the intensity prior based on how a given image should look like at the average exposure level. Our experiments (Sec. 5.5.1) show that providing the reference image $\mathbf{I}$ (*i.e.*, using a well-posed pretext task) would deteriorate the final detection performance.

The two contradictory losses $\mathcal{L}_i$ and $\mathcal{L}_v$ guide the network to decompose coupled features into intensity-invariant and intensity-variant features, allowing sufficient excavation of both and further re-evaluation of their individual contributions to the final prediction.

3.2. Feature Reweighting via Cumulative Learning

As discussed above, since intensity is still a dominant cue in shadow detection, we should include both intensity-variant features $\mathbf{F}_v$ and intensity-invariant features $\mathbf{F}_i$ into the shadow detection task. What we need is to balance the impact of the two types of features. Since the whole model is trained in an end-to-end manner, we may formulate the feature fusion step with a summation, as in the feature pyramid network [26]. In addition, to re-weight the contributions of these features, we introduce a trade-off parameter μ to formulate feature reweighting as:

$$\mathbf{F}_r = \mu\mathbf{F}_v + (1-\mu)\mathbf{F}_i, \tag{4}$$

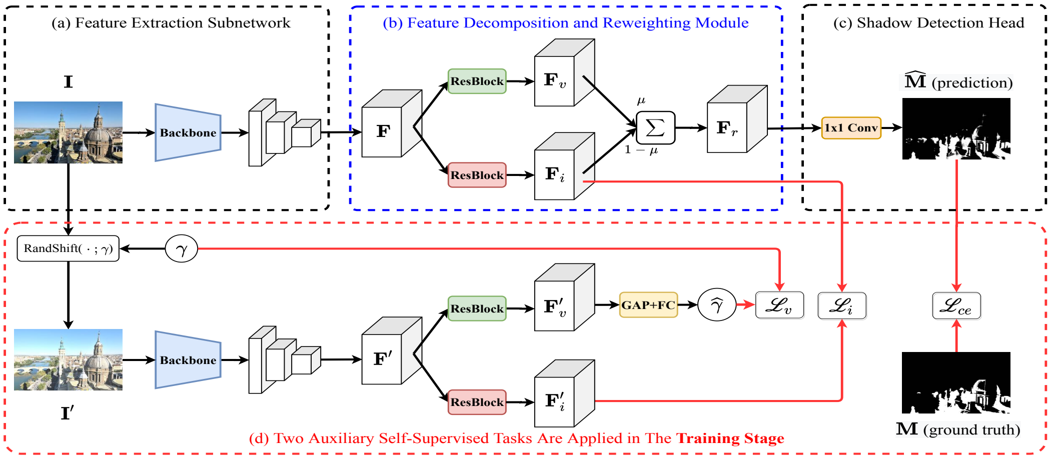


Figure 3. Training and inference pipeline of the proposed method. Our network incorporates three modules: (a) Feature Extraction Subnetwork for extracting multi-level integrated features $\mathbf{F}$; (b) Feature Decomposition and Reweighting Module for decomposing $\mathbf{F}$ into intensity-variant features $\mathbf{F}_v$ and intensity-invariant features $\mathbf{F}_i$, which are further re-combined by a weighted summation to produce reweighted features $\mathbf{F}_r$; and (c) Shadow Detection Head for predicting the shadow mask $\hat{\mathbf{M}}$. (d) In the training stage, we construct two auxiliary self-supervised tasks to guide the learning of feature decomposition: jointly optimizing two auxiliary self-supervision losses $\mathcal{L}_v$ and $\mathcal{L}_i$ with the shadow detection loss $\mathcal{L}_{ce}$. Note that operation nodes with the same color share their parameters.

where $\mathbf{F}_r$ denotes the output shadow features.

A straightforward choice to determine μ is to learn it from data, by setting it as a differentiable parameter [18] or predicting it with a network branch [19]. However, these two strategies do not work in our case, as the intensity bias comes from the data. In fact, through experiments (more details in Sec. 5.5.2), we observe that μ would continue to grow to near 1 as training proceeds while the detection performance shows no improvement. Instead, in the training stage, we adopt cumulative learning [51] to gradually shift the focus of the network from intensity-variant features to intensity-invariant features. Given the current training epoch T and total training epoch T_{max}, μ is:

$$\mu = 1 - (\frac{T}{T_{max}})^{\beta}, \tag{5}$$

where β is a hyper-parameter to control the descending pace of μ over the training stage.A larger β will induce a smoother focus transition from intensity-variant features to intensity variant features at the beginning of training.

Once the training is finished, we need to determine a proper μ for inference. Since μ is a scalar between $[0, 1]$, its value can be obtained through a grid search on the validation set. In practice, we set the search step to 0.1.

The cumulative learning strategy is similar to Dropout [35]. While in Dropout, some neurons are randomly dropped in a *hard* manner, in cumulative learning, we progressively drop intensity-variant features in a *soft* manner. The choice of μ is then to obtain the suitable weights for the two types of features. This implicitly creates an ensemble of the two detectors to exploit both.

Table 1. EfficientNet-B3 [36] has a similar classification performance to ResNext101 [45] on ImageNet [8], but contains significantly fewer parameters and requires less computation.

backbone	#Params	#FLOPS	Top-1 Acc.	Top-5 Acc.
EfficientNet-B3	**12M**	**1.8**	81.1	95.5
ResNext101	84M	32	80.9	95.6

4. Training and Test Strategies

Choice of backbone. We first explain our choice of the backbone network for extracting the multi-level integrated features (MLIF). Recent state-of-the-art shadow detectors [55, 50, 5] rely on a heavy backbone (*e.g.*, ResNext101 [45]) to extract the backbone features. However, this is not suitable for us, as our method requires extra forward and backward pass on brightness shifted input in training stage. To maintain a stable training with a reasonable batch size, we choose the light-weight EfficientNet-B3 [36] as our backbone. Nonetheless, as shown in Table 1, EfficientNet-B3 [36] has a similar classification performance to ResNext101 [45] on ImageNet [8]. Such a replacement should not affect our shadow detection performance and analysis.

With EfficientNet-B3 [36] as the backbone, we extract multi-level features from every two consecutive blocks, producing 13 groups of feature maps in total. We use bilinear upsampling to keep their spatial resolutions to remain half of that of the input, and use 1×1 convolution to reduce their channels into 16. These features are then concatenated and fused into 32-channel features via 1×1 convolution, producing the multi-layer integrated features (MLIF) for further feature decomposition and reweighting.

Loss function. Considering the imbalanced numbers of shadow and non-shadow pixels in natural scenes, we adopt the balanced binary cross entropy as shadow detection loss:

$$\mathcal{L}_{ce}(\mathbf{M}, \hat{\mathbf{M}}) = -\sum_i [\frac{N_n}{N} M_i \log \hat{M}_i + \frac{N_p}{N}(1 - M_i) \log(1 - \hat{M}_i)], \tag{6}$$

where i is the index of the spatial location. $\mathbf{M}$ is the ground truth shadow map. $\hat{\mathbf{M}}$ is the predicted shadow mask. N_p, N_n and N are the number of shadow and non-shadow pixels, and the total number of pixels in the image, respectively.

Together with the two proposed loss terms $\mathcal{L}_i$ and $\mathcal{L}_v$ for feature decomposition, the final loss function $\mathcal{L}_{total}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_i \mathcal{L}_i + \lambda_v \mathcal{L}_v, \tag{7}$$

where λ_i and λ_v are two balancing parameters, which are empirically set to 1 and 0.1, respectively.

Training details. Following recent state-of-the-art shadow detectors [50, 5, 55, 16], we initialize the backbone with weights pretrained on ImageNet [8]. Other newly introduced trainable parameters are randomly initialized. We optimize the whole network for 10 epochs using the Adam optimizer, with an initial learning rate of $5e-4$, which is adjusted by the exponential decay strategy (decay rate = 0.7). In each training iteration, the input images are resized to a resolution of 400 × 400 and fed into the network with a mini-batch size of 6. We apply random horizontal flipping for data augmentation. Δ (for determining the maximum amount of intensity shift for feature decomposition) is set to 0.3, and β (for controlling the cumulative learning pace) is set to 2. The training is run on a single RTX2080Ti GPU. As the datasets (SBU [40] and ISTD [41]) used in our training do not provide validation split, as suggested in [1], we randomly hold out 10% of the data in the training set for validation. For fair comparison to existing works, once the optimal μ is determined, we put them back and retrain the model with the whole training set. Training takes 2 hours on SBU [40] and 1 hour on ISTD [41].

Inference. Following recent shadow detection works [16, 55, 24, 50, 5], we use the fully connected CRF [21] to refine our predictions (with a threshold value of 0.5) to obtain the final shadow mask. The inference of images at a resolution of 400 × 400 runs at 161 frames per second.

5. Experiments

5.1. Evaluation Datasets and Metric

Evaluation datasets. We conduct our experiments on three public datasets, *i.e.*, SBU [40], UCF [54] and ISTD [41], to evaluate our shadow detector. The UCF dataset contains 135 training images and 110 test images. The SBU dataset contains 4,089 training images and 638 test images. The ISTD dataset contains 1,330 training images and 540 test images. We follow previous shadow detection methods to train on the SBU training set, and test on both SBU and UCF test sets. For the evaluation on the ISTD test set, we train our model on its training set.

Evaluation metric. For quantitative performance evaluation, we use the popular metric, balanced error rate (BER):

$$BER = (1 - \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP})) \times 100, \tag{8}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives.

Methods for comparison. We first compare our method with 11 state-of-the-art shadow detectors, including MTMT [5], DSD [50], DC-DSPF [43], ADNet [24], DSC [16], BDRAR [55], ST-CGAN [41], patched-CNN [15], scGAN [30], stacked-CNN [40], and Unary-Pairwise [14]. All of them are deep-learning based methods, except Unary-Pairwise [14], which is based on handcrafted features.

As shadow detection is a kind of pixel-level classification problem, it is related to saliency object detection (SOD) and semantic segmentation. For a comprehensive study, we also compare our method with four state-of-the-art SOD methods, EGNet [49], ITSD [52], SRM [42], and Amulet [46], and one semantic segmentation method, PSPNet [48]. All these methods are deep learning based. They are trained and tested in the same way as the deep shadow detectors.

5.2. Quantitative Comparison

Table 2 shows the quantitative results on the three benchmark datasets. We can see that our method achieves the best BER scores over all state-of-the-art methods, on the three datasets. Compared to the second best-performing method, MTMT-Net [5], our method reduces the BER scores by 3.49%, 2.14% and 9.9% on SBU [40], UCF [54], and ISTD [41], respectively. Note that MTMT-Net is a semi-supervised method that exploits both **full** labeled shadow images as well as extra unlabeled shadow images. Our proposed method does not require additional training data. This demonstrates the effectiveness of our proposed decomposition and reweighting scheme on mining both intensity-variant and intensity-invariant features.

5.3. Qualitative Comparison

We further compare our method with the most recent shadow detectors qualitatively, as shown in Fig. 4. We can see that our method has clear visual advantages over existing shadow detection methods on challenging scenes. When shadows are cast on dark objects (*e.g.*, first three rows) or regions with drastically varying colors (*e.g.*, last three

Table 2. Quantitative comparison of our method with the state-of-the-art methods on three shadow detection benchmark datasets. For each dataset, we list the error rates for shadow region and non-shadow region as well as the balanced error rate (BER). The best results are marked in **bold**. (*) MTMT is trained with extra unlabelled data; (**) DSD is trained with extra supervision from other models.

		SBU [40]			UCF [54]			ISTD [41]		
methods	year	BER↓	Shadow↓	Non Shad.↓	BER↓	Shadow↓	Non Shad.↓	BER↓	Shadow↓	Non Shad.↓
FDRNet(Ours)	-	**3.04**	**2.91**	3.18	**7.28**	8.31	6.26	**1.55**	1.22	1.88
MTMT* [5]	2020	3.15	3.73	2.57	7.47	10.31	4.63	1.72	1.36	2.08
DSD** [50]	2019	3.45	3.33	3.58	7.59	9.74	5.44	2.17	1.36	2.98
DC-DSPF [43]	2019	4.90	4.70	5.10	7.90	6.50	9.30	-	-	-
ADNet [24]	2018	5.37	4.45	6.30	9.25	8.37	10.14	-	-	-
DSC [16]	2018	5.59	9.76	1.42	10.54	18.08	**3.00**	3.42	3.85	3.00
BDRAR [55]	2018	3.64	3.40	3.89	7.81	9.69	5.44	2.69	**0.50**	4.87
ST-CGAN [41]	2018	8.14	3.75	12.53	11.23	**4.94**	17.52	3.85	2.14	5.55
patched-CNN [15]	2018	11.56	15.60	7.52	-	-	-	-	-	-
scGAN [30]	2017	9.10	8.39	9.69	11.50	7.74	15.30	4.70	3.22	6.18
stacked-CNN [40]	2016	11.00	8.84	12.76	13.00	8.84	12.76	8.60	7.69	9.23
Unary-Pariwise [14]	2011	25.03	36.26	13.80	-	-	-	-	-	-
ITSD [52]	2020	5.00	8.65	**1.36**	10.16	17.13	3.19	2.73	2.05	3.40
EGNet [49]	2019	4.49	5.23	3.75	9.20	11.28	7.12	1.85	1.75	1.95
SRM [42]	2017	6.57	10.52	2.50	12.51	21.41	3.60	7.92	13.97	**1.86**
Amulet [46]	2017	15.13	-	-	15.17	-	-	-	-	-
PSPNet [48]	2017	8.57	-	-	11.75	-	-	4.26	4.51	4.02

rows), existing methods fail to differentiate such differences well, by either over-segmenting or under-segmenting the shadow regions. In contrast, our method detects the shadows more accurately with fine structures and details. This again demonstrates the effectiveness of our proposed decomposition and reweighting scheme in mitigating the intensity bias in shadow detection.

5.4. Feature Visualisation

In Fig. 5, we give an example of feature visualization using GradCAM [34]. In this example, we show the original image and its two brightness shifted versions (one brighter and one darker). We can see that: (1) while the activation map of the intensity-invariant features (column 3) is uniformly distributed inside the shadow region, that of the intensity-variant features (column 4) is highly correlated to pixel intensity; and (2) under brightness shift, the activation of intensity-invariant features remains stable, but that of the intensity-variant features changes accordingly. These visualizations show that our method can successfully decompose these two types of features. More visualization results are provided in supplemental materials.

5.5. Ablation Study

We first perform an ablation study on the SBU dataset, to verify the design choices of our network, including the components for feature decomposition and different cumulative learning strategies. The average BER scores with CRF refinement are reported. We then analyze the model's sensitivity to μ in the test phase, to shed some light on the importance of feature reweighting.

5.5.1 Components for Feature Decomposition

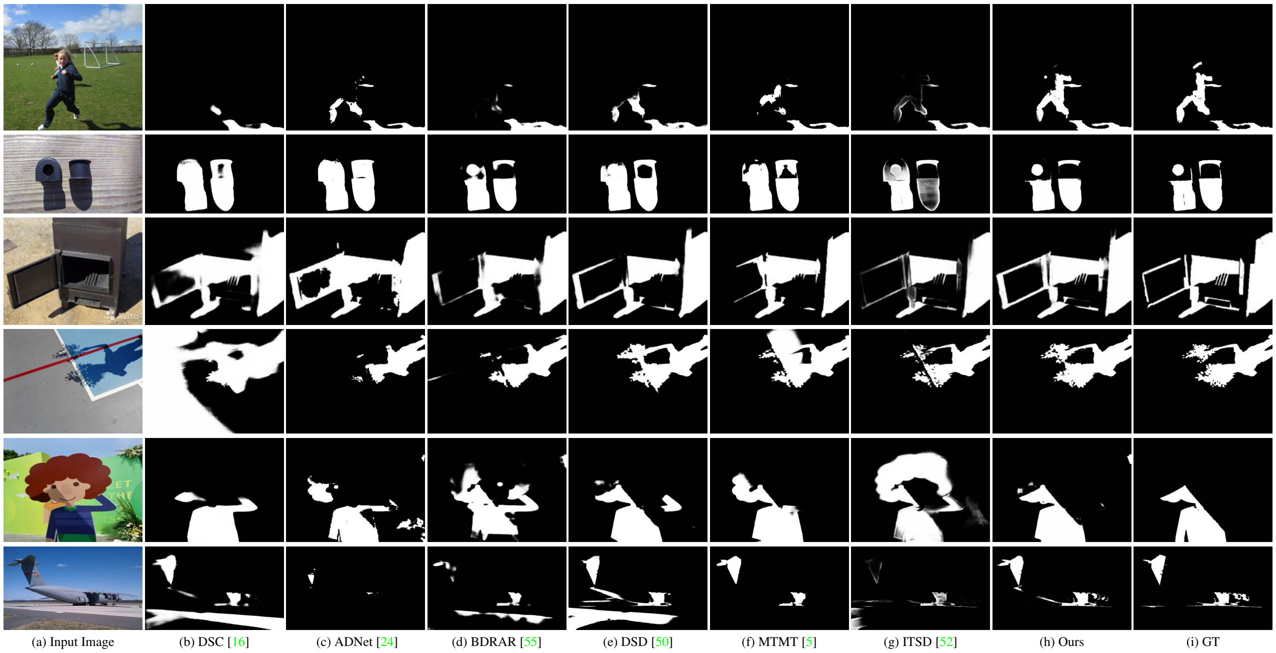
To verify the effectiveness of the components used in the proposed feature decomposition, we compare the full model (Ours) with the following configurations:

Table 3. Ablation study on the components of the feature decomposition. We show BER scores on SBU.

	BB	$\mathcal{L}_i$	$\mathcal{L}_v$	BER ↓
basic	×	×	×	3.32
basic+BB	✓	×	×	3.32
basic+BB+$\mathcal{L}_i$	✓	✓	×	3.24
basic+BB+$\mathcal{L}_v$	✓	×	✓	3.36
well-posed $\mathcal{L}_v$	✓	✓	✓	3.21
Ours	✓	✓	✓	**3.04**

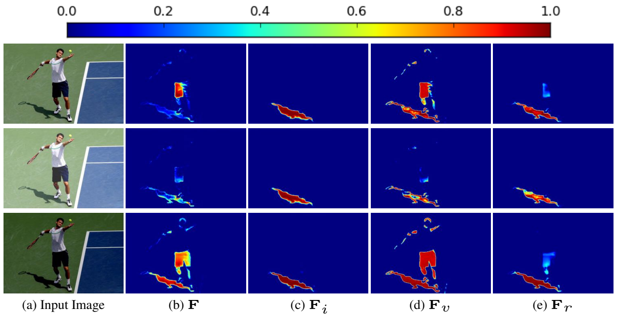
- Basic: we remove the bilateral branches after multi-level integrated feature extraction, and the two corresponding loss functions (*i.e.*, $\mathcal{L}_i$ and $\mathcal{L}_v$) used for feature decomposition. This forms our baseline.
- Basic+BB: we add the bilateral branches after the extraction of multi-level integrated features, without the two decomposition losses (*i.e.* $\mathcal{L}_i$ and $\mathcal{L}_v$).
- Basic+BB+$\mathcal{L}_i$: we remove the $\mathcal{L}_v$ loss for guiding intensity-variant projection, from our full model.
- Basic+BB+$\mathcal{L}_v$: we remove the $\mathcal{L}_i$ loss for guiding intensity-invariant projection, from our full model.
- Well-posed $\mathcal{L}_v$: in our full model, we replace $\mathcal{L}_v$ loss (Eq. 3) with $\mathcal{L}_v = \Phi_{MAE}(\phi(\mathbf{F}'_v - \mathbf{F}_v; \theta), \gamma)$. It yields a deterministic pretext task where the brightness shift is predicted with reference to both the brightness shifted and the original image.

As shown in Table 3, with only the bilateral branches added, the detection accuracy shows no improvement. This means that a deeper architecture does not help improve detection accuracy. In addition, by adding only $\mathcal{L}_i$, the performance is improved as it moderately suppressed the ex-



(a) Input Image (b) DSC [16] (c) ADNet [24] (d) BDRAR [55] (e) DSD [50] (f) MTMT [5] (g) ITSD [52] (h) Ours (i) GT

Figure 4. Qualitative comparison of the proposed method with the most recent state-of-the-art methods.



(a) Input Image (b) $\mathbf{F}$ (c) $\mathbf{F}_i$ (d) $\mathbf{F}_v$ (e) $\mathbf{F}_r$

Figure 5. Feature map visualization using GradCAM [34]. We show the original image (top row), its brighter counterpart (second row) and its darker counterpart (third row). From left to right: (a) input image, (b) $\mathbf{F}$: features before decomposition, (c) $\mathbf{F}_i$: intensity-invariant features, (d) $\mathbf{F}_v$: intensity-variant features, (d) $\mathbf{F}_r$: recombined features.

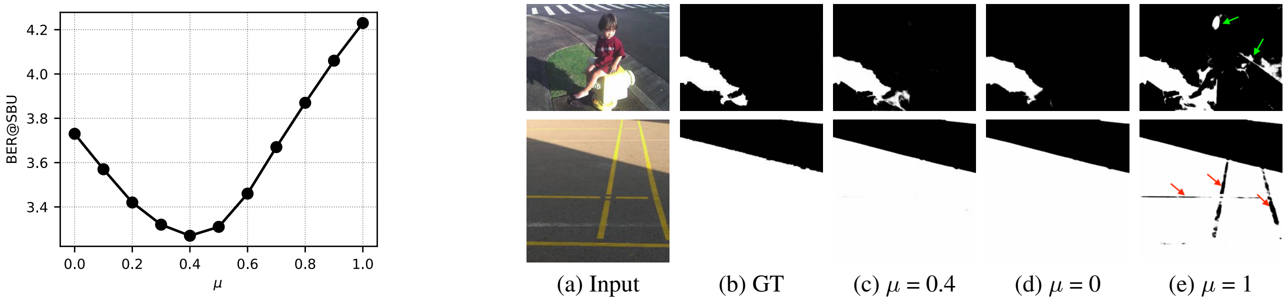
cessive attention to intensity. In contrast, adding only $\mathcal{L}_v$ further emphasizes on intensity, and the result gets worse.

Overall, we can see that our improvement mainly comes from the use of paired self-supervision losses (both $\mathcal{L}_v$ and $\mathcal{L}_i$), instead of from just one of them. Intuitively, a joint usage of them is crucial for feature decomposition: it guarantees that the two feature maps $\mathbf{F}_v$ and $\mathbf{F}_i$ projected by the bilateral branches encode the expected information, as these two tasks are contrary and complementary (*i.e.*, while one forces a part of the network to encode intensity related information, the other one imposes intensity-invariant constraint on the other part of the network). In contrast, if only one of them is used, as the bilateral branches stem from the same backbone, the constraint will affect both of the projected features, making it difficult to decompose them well.

Finally, the last second row shows that using well-posed $\mathcal{L}_v$, *i.e.* predicting brightness shift with reference to both the brightness shifted and the original image, produces a performance similar to the case where only $\mathcal{L}_i$ is added. This is because such a pretext task becomes too simple, as the network can even solve it by just trivially learning an identity mapping from input to the intensity-variant features. As a result, the intensity-variant features may fail to encode high-level semantics for shadow detection.

5.5.2 Different Cumulative Learning Strategies

We observe from our experiment in Fig. 7 that if we just leave the weight parameter μ learnable (*i.e.*, setting it differentiable [18]) or using an extra branch to predict it [19], μ will continuously increase to $\sim$1 before the training ends. This verifies our observation that deep shadow detectors tends to be biased to the intensity-variant features. Intuitively, this is because CNNs are biased towards local features, and intensity is exactly a local feature. Besides, intensity is indeed the most evident cue for shadow detection. To avoid this, the cumulative learning strategy is introduced to gradually shift the learning focus of the network from the intensity-variant features to the intensity-invariant features. Table 4 compares different strategies for adjusting μ during the training stage. It is worth noting that decay strategies (*i.e.*, linear decay and parabolic decay) perform better than increment strategies (*i.e.*, linear increment and parabolic increment). This suggests that we should gradually shift the attention from intensity-variant features to intensity-invariant ones. In addition, the parabolic decay strategy gives better performance than the linear decay one.



(a) Input (b) GT (c) $\mu = 0.4$ (d) $\mu = 0$ (e) $\mu = 1$

Figure 6. Effect of varying μ in the test phase. Left: the BER score with respect to μ on SBU testset. Right: two visual examples. Setting $\mu = 0$ blocks intensity-variant features, $\mu = 1$ blocks intensity-invariant features, and $\mu = 0.4$ provides best trade-off of using the two.

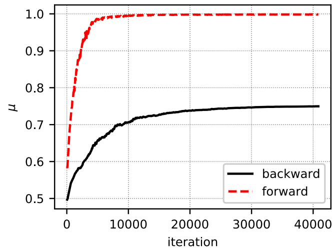


Figure 7. A learnable μ continuously increases to a $\sim$1, suggesting that the network is biased to the intensity-variant features. **Forward**: we add an extra branch consisting of global average pooling and a fully connected layer to predict μ from multi-layer integrated features. **Backward**: we set μ as a differentiable parameter and let it be optimized together with the network parameters.

Table 4. Ablation study of different cumulative strategies.

Strategy	μ	BER ↓
Constant	0.5	3.45
Backward	-	3.47
Forward	-	3.42
Linear increment	$\frac{T}{T_{max}}$	3.63
Parabolic increment	$(\frac{T}{T_{max}})^2$	3.73
Linear decay	$1 - \frac{T}{T_{max}}$	3.26
Parabolic decay (Ours)	$1 - (\frac{T}{T_{max}})^2$	**3.04**

This implies that a smooth transition at the beginning of training is important, as it allows the detector to thoroughly exploit intensity-variant features first.

5.5.3 Test-phase Sensitivity to μ

While we have determined an optimal value for μ by grid search on the validation set (Sec. 3.2), we alter it here in the test stage to see how the shadow detection performance may respond to the change of μ. In Fig. 6(left), the BER score with respect to the changing μ on SBU testset shows that there is an optimal value of μ, which provides the best trade-off between intensity-variant and intensity-invariant features: putting more or less emphasis on the intensity-variant features will degrade the detection performance.

Fig. 6(right) shows two visual examples to illustrate the effect of changing μ. Note that setting $\mu = 1$ means that only the intensity-variant features are used for shadow prediction, where the results are indeed susceptible to pixel intensity (*e.g.*, the regions pointed to by arrows in Fig. 6(e)). In contrast, if we set $\mu = 0$ (Fig. 6(d)), which means that only the intensity-invariant features are used, there is only a small proportion of shadow pixels being mis-classified as non-shadow pixels. This justifies that the intensity-invariant cues provide discriminative information for shadow detection. However, the best results are achieved when μ is set to neither extreme. Overall, these observations verify the importance of our idea to reallocate the attention between intensity-variant and intensity-invariant features.

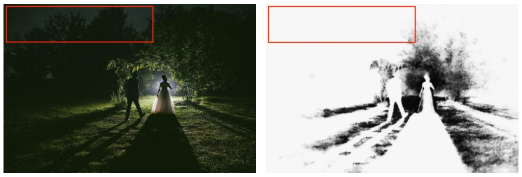


Figure 8. Failure case. It may fail on night-time image as the training set contains only day-time images.

6. Conclusion

This paper has presented a novel feature decomposition and reweighting scheme to mitigate the bias of deep shadow detectors to the intensity cue. The key idea is to decompose multi-level integrated features into intensity-variant and intensity-invariant components, and then reallocate the attention between them. We have introduced two auxiliary self-supervised tasks to guide the learning of this task-specific feature decomposition. Experimental results on three datasets show that our model achieves favorable performances, compared to the state-of-the-art methods.

Although our deep shadow detector achieves a balanced exploitation of intensity-variant and intensity-invariant features, which helps mitigate the intensity bias, in inference time it still relies on the prior learned from the training set. Hence, it may fail if the illumination of the image strongly deviates from that of the training set, as shown in Fig. 8. As a future work, we plan to explore domain adaptation for shadow detection.

References

[1] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *ICLR*, 2019. 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020. 3

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv:2006.10029*, 2020. 3

[4] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *CVPR*, 2019. 3

[5] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7

[6] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 1

[7] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *TPAMI*, 25(10):1337–1342, 2003. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4, 5

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 3

[10] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *IJCV*, 2009. 2

[11] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *TPAMI*, 2005. 2

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 1

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[14] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, 2011. 2, 5, 6

[15] Sepideh Hosseinzadeh, Moein Shakeri, and Hong Zhang. Fast shadow detection from a single image using a patched convolutional neural network. In *IROS*, 2018. 5, 6

[16] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, 2018. 1, 2, 5, 6, 7

[17] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *ICCV*, 2011. 2

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4, 7

[19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019. 4, 7

[20] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM TOG*, 30(6):1–12, 2011. 1

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 5

[22] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, pages 183–190, 2009. 1

[23] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, 2010. 2

[24] Hieu Le, Tomas F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+d net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, 2018. 2, 5, 6, 7

[25] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Rethinking data augmentation: Self-supervision and self-distillation. *arXiv:1910.05872*, 2019. 3

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 3

[27] Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *CVPR*, pages 12295–12303, 2020. 3

[28] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *TPAMI*, 26(8):1079–1087, 2004. 1

[29] Vu Nguyen, Tomas F. Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017. 2

[30] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017. 5, 6

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[32] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM TOG*, 2019. 3

[33] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern Recognition*, 2012. 1

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 6, 7

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 4

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[38] Jiandong Tian, Xiaojun Qi, Liangqiong Qu, and Yandong Tang. New spectrum ratio properties and features for shadow detection. *Pattern Recognition*, 2016. 2

[39] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *TPAMI*, 2017. 2

[40] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, 2016. 2, 5, 6

[41] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018. 2, 5, 6

[42] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 5, 6

[43] Yupei Wang, Xin Zhao, Yin Li, Xuecai Hu, and Kaiqi Huang. Densely cascaded shadow detection network via deeply supervised parallel fusion. In *IJCAI*, 2018. 5, 6

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[45] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 4

[46] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 5, 6

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 3

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5, 6

[49] 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*, 2019. 5, 6

[50] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *CVPR*, 2019. 1, 2, 4, 5, 6, 7

[51] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 2, 4

[52] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, 2020. 5, 6, 7

[53] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2010. 2

[54] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2010. 5, 6

[55] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018. 2, 4, 5, 6, 7