Triple-cooperative Video Shadow Detection

Zhihao Chen[1*], Liang Wan[1*], Lei Zhu[2†], Jia Shen[1], Huazhu Fu[3], Wennan Liu[4], Jing Qin[5]
[1] College of Intelligence and Computing, Tianjin University
[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge
[3] Inception Institute of Artificial Intelligence, UAE
[4] Academy of Medical Engineering and Translational Medicine, Tianjin University
[5] The Hong Kong Polytechnic University

Abstract

Shadow detection in a single image has received significant research interests in recent years. However, much fewer works have been explored in shadow detection over dynamic scenes. The bottleneck is the lack of a well-established dataset with high-quality annotations for video shadow detection. In this work, we collect a new video shadow detection dataset (ViSha), which contains 120 *videos with* 11,685 *frames, covering 60 object categories, varying lengths, and different motion/lighting conditions. All the frames are annotated with a high-quality pixel-level shadow mask. To the best of our knowledge, this is the first learning-oriented dataset for video shadow detection. Furthermore, we develop a new baseline model, named triple-cooperative video shadow detection network (TVSD-Net). It utilizes triple parallel networks in a cooperative manner to learn discriminative representations at intra-video and inter-video levels. Within the network, a dual gated co-attention module is proposed to constrain features from neighboring frames in the same video, while an auxiliary similarity loss is introduced to mine semantic information between different videos. Finally, we conduct a comprehensive study on ViSha, evaluating 12 state-of-the-art models (including single image shadow detectors, video object segmentation, and saliency detection methods). Experiments demonstrate that our model outperforms SOTA competitors.*

1. Introduction

As a common phenomenon in our daily life, shadows in natural images provide hints for extracting scene geometry [41, 24], light direction [27], and camera location and its parameters [23]. Shadows can also benefit diverse image understanding tasks, e.g., image segmentation [10], object detection [8], and object tracking [37]. The last decade has witnessed a growing interest in image shadow detection. Many methods have been developed by examining color and illumination priors [14, 13], by developing data-driven approaches with hand-crafted features [21, 28, 54], or by learning discriminative features from a convolutional neural network (CNN) [25, 43, 39, 20, 29, 56, 19, 53].

However, in striking contrast with the flourishing development of image shadow detection, much fewer works have been explored in shadow detection over dynamic scenes. On the other hand, we also notice that video processing has become an urgent topic in recent years, and a lot of methods were proposed for video salient object detection [30, 47, 11] and video object segmentation [35, 40]. What makes video shadow detection lag far behind these video processing tasks? Compared with shadow detection of a single image, video shadow detection (VSD) needs to utilize temporal information to identify shadow pixels of each video frame. Although there exist multiple datasets for image shadow detection, video salient object detection, and video object segmentation, such standard widespread benchmark (with a sufficient number of video clips, covering diverse content) is missing for video shadow detection. What's more, CNN-based methods have not been exploited for this problem due to the lack of such a dataset.

In this work, **we first collect a new video shadow detection (ViSha) dataset.** It contains 120 videos with 11,685 image frames and 390 seconds duration, covering shadows of 7 object classes and 60 object categories, various motion/lighting conditions, and different instance numbers. All the video frames are carefully annotated with a high-quality pixel-level shadow mask. To the best of our knowledge, this is the first learning-oriented dataset for video shadow detection, which could facilitate the community to explore further in this field. Second, **we develop a new baseline model, a triple-cooperative video shadow detection network (TVSD-Net), for this task.** Instead of just exploiting temporal information within one video clip

*Zhihao Chen and Liang Wan are the joint first authors of this work.
†Lei Zhu (lz437@cam.ac.uk) is the corresponding author of this work.

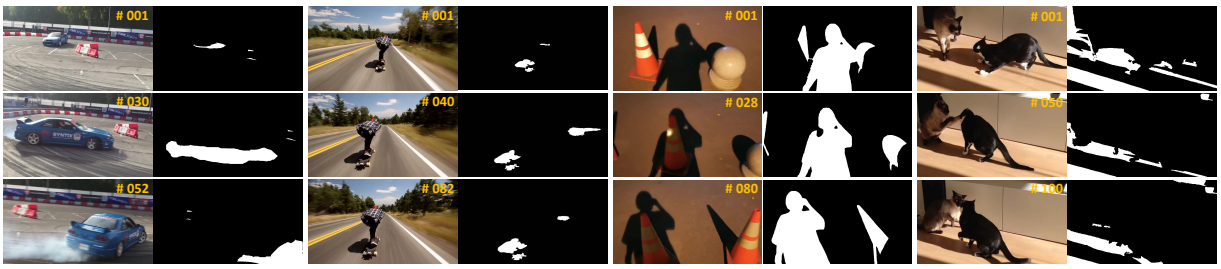

Figure 1: The examples of proposed Video Shadow Detection (***ViSha***) dataset, with pixel-level shadow annotations.

as most current video object detection networks did, we propose to learn at both intra-video and inter-video levels then model their correlation features. Our TVSD-Net utilizes triple parallel networks in a cooperative manner. To be specific, we take two neighboring frames from the same video and one image frame from another video as inputs. Then a dual gated co-attention (DGC) module is devised to learn a global intra-video correlation on the two frames of the same video, and a triple-cooperative (T) module encodes the inter-video property, which promotes the similarity between the same-video frames and suppresses the similarities between different-video frames. Finally, **we present a comprehensive evaluation of 12 state-of-the-art models on our ViSha dataset**, making it the most complete VSD benchmark. Results show that our model significantly outperforms existing methods, including single image shadow detectors [4, 56, 53], single saliency detectors [18, 9], semantic segmentation method [34, 52], video object segmentation [35], and video saliency detection methods [32, 42]. In summary, our work forms the first learning-oriented VSD benchmark, thereby providing a new view to video object detection from a shadow perspective. Our dataset and code have been released at https://github.com/eraserNut/ViSha.

2. Related Works

Single-image Shadow detection. Existing shadow detection works mainly focus on detect shadow pixels from a single input image. Deep learning-based methods have achieved dominated results. Please refer to [4, 48, 17] for a review on traditional shadow detectors based on hand-crafted features. The first shadow detection CNN [25] identified shadow pixels by building a seven-layer CNN to extract deep features from superpixels, and then employed a conditional random field (CRF) to further smooth shadow detection results. Vicente et al. [43] trained a patch-based CNN with an image-level shadow prior. Nguyen [39] introduced a generative adversarial network with a conditional generator to generate a shadow mask. Hu et al. [20] learned spatial context features in a direction-aware manner, while Zhu et al. [56] utilized two series of recurrent attention residual (RAR) modules to aggregate context information at different CNN layers for shadow detection. Zheng et al. [53] learned distraction-aware features to explicitly predict false positives and false negatives for robust shadow detection. Rather than relying on only annotated shadow data, Chen et al. [4] explored the complementary information of shadow region detection, shadow boundary detection, and shadow count detection and embedded the multi-task learning into a semi-supervised learning framework to fuse unlabeled data for helping shadow detection.

Video shadow detection aims to detect the shadow regions from each frame of a video. Existing video shadow detection methods almost relied on hand-crafted features and were developed one decade ago. For instance, Nadimi et al. [38] leveraged a spatio-temporal albedo test and dichromatic reflection model. Jr et al. [22] detected the moving shadow in videos by employing improved background subtraction techniques. Benedek et al. [2] combined the color and microstructural features to detect the shadow in surveillance Videos. Note that these methods work well on high-quality scenarios (*e.g.*, stable lighting, single shadow, moving objects) due to the limited generalization capability of these hand-craft features. In addition, no large-scale datasets are publicly available for fairly evaluating different video shadow detection methods. In order to exploit the capability of CNN-based methods for VSD tasks, it is desirable to collect a large-scale VSD dataset and develop a CNN model to provide a complement evaluation.

Video object segmentation automatically detects primary foreground objects from their background in all frames of a video. It can be roughly categorized into unsupervised video object segmentation (UVOS) and semi-supervised video object segmentation (SVOS) (please refer to [35, 40] for a detailed review). Compared with the counterpart for image object segmentation, VOS exploits the temporal information across frames. Lu et al. [35] formulated a co-attention siamese network (COSNet) to model UVOS from a global perspective via a co-attention mechanism. Oh et al. [40] leveraged memory networks and learned to read relevant information from all past frames with object masks for resolving SVOS. However, current CNN-based VOS mainly learned appearance or motion representations in intra-video, while ignoring the valuable discriminative

Table 1: Statistics of the proposed ViSha dataset. See Section 3.3 for details.

ViSha	Shadow Motion		Camera Motion		# Shadow Instances				Shadow Type		Scene Type			
	Stable	Moving	Stable	Moving	1	2	3	≥4	Hard	Soft	Day	Night	Indoor	Outdoor
# Training	10	40	12	38	8	4	6	22	36	12	33	17	18	32
# Testing	10	60	20	50	17	15	9	29	52	18	53	17	17	53
# Total	20	100	32	88	25	19	15	51	88	30	86	34	35	85

Table 2: Video sources of our ViSha dataset.

Source	OTB [49]	VOT [26]	LaSOT [12]	TC-128 [33]	NfS [15]	Self
# Videos	11	7	18	16	9	59

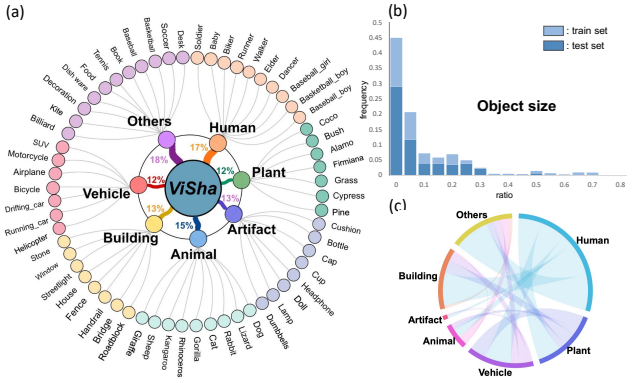

Figure 2: Statistics of the proposed ViSha. (a) Shadow categories. (b) Ratio distribution of the shadows. (c) Mutual dependencies among shadow categories in (a).

inter-video representations across different videos.

Video saliency detection identifies most distinctive objects for each video frame [5, 46, 1, 7, 6, 45]. Recently, many CNN-based video saliency detection methods achieved dominant results. The first CNN attempt was a fully convolutional network proposed by Wang et al. [47]. Le et al. [30] adopted 3D filters to combine spatial and temporal information in a spatio-temporal CRF framework, while Li et al. [31] presented optical flow guided recurrent neural network. Song et al. [42] passed concatenated spatial features at multiple scales into an extended deeper bidirectional ConvLSTM to obtain spatio-temporal information. Fan et al. [11] collected a video saliency detection dataset and developed a saliency-shift-aware convLSTM module to extract both spatial and temporal information. Similar to the task of video object detection, video saliency detection methods usually focus on extracting spatial-temporal features from multiple frames of a video, which also ignores the intra-video discriminative property of salient objects.

3. ViSha: Video Shadow Detection Dataset

We introduce **ViSha**, a new dataset for video shadow detection. Our dataset includes 120 videos with diverse content, varying length, and object-level annotations. Some example video frames can be found in Figure 1. We will show details of ViSha from the following key aspects.

3.1. Data Collection

In order to provide a solid basis for video shadow detection, we think that the dataset should (1) cover diverse realistic-scenes, and (2) contain sufficient challenging cases. As shown in Table 2, more than half videos are from 5 widely-used video tracking benchmarks (*i.e.*, OTB [49], VOT [26], LaSOT [12], TC-128 [33], and NfS [15]). Note that these video tracking datasets are not originally designed for shadow detection, and hence there are limited videos with shadows, which are all included in our dataset. The remaining 59 videos are self-captured with different hand-held cameras, over different scenes, at varying times. We then manually trim the videos to make sure that each frame has at least one shadow area, and remove dark-screen transitions. The frame rate is adjusted to 30 fps for all video sequences. For instance, the videos from NfS [15] have a high-speed frame rate of 240, for which we make sampling at every 8 frames. Eventually, our video shadow detection dataset (ViSha) contains 120 video sequences, with a totally of 11,685 frames and 390 seconds duration. The longest video contains 103 frames and the shortest contains 11 frames.

3.2. Dataset Annotation and Split

For each video frame, we provide pixel-accurate, manually created segmentation in the form of a binary mask. In realistic scenarios, shadows can be distorted, ambiguous , and hard to identify (see examples in Figure 1). Eight human annotators are pretrained and instructed to carefully annotate all the shadows by tracing shadow boundaries. Then, two viewers are assigned to inspect and validate the labeled shadows. In the annotation process, we notice that two cases deserve special attention. First, the soft shadow is usually subjected to unclear boundaries. Considering the temporal consistency between adjacent frames, we demand that the labeling of soft shadows should be consistent across frames. Second, the back-light parts of objects often appear in dark colors, yet they do not form shadows, for which we treat them as non-shadow areas.

To provide guidelines for future works, we randomly

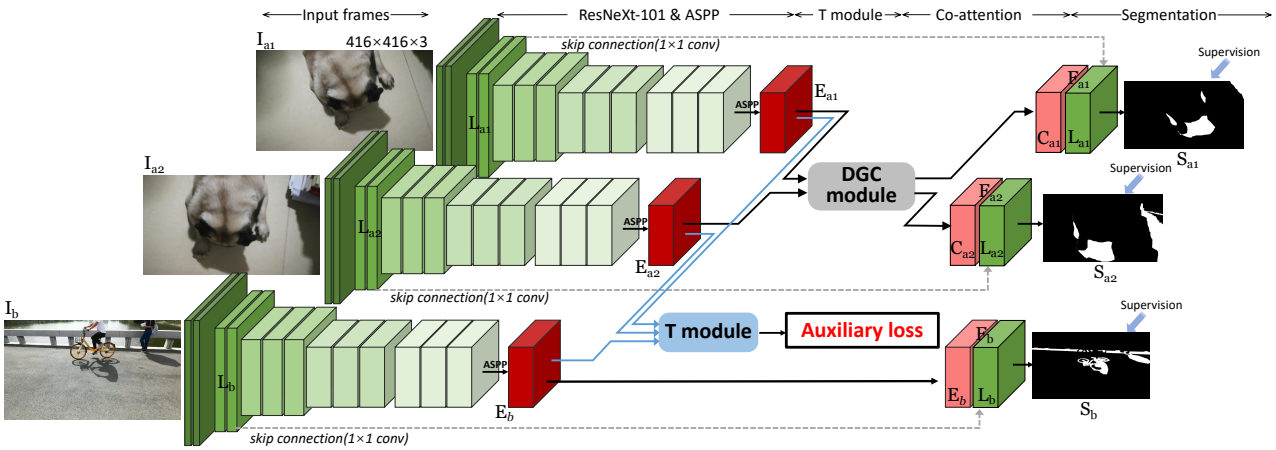

Figure 3: The schematic illustration of our proposed TVSD-Net. DGC module denotes dual-gated co-attention module and T module denotes triple-cooperative module. See Section 4.1 for details.

split the dataset into training and testing sets with a ratio of 5:7. Table 1 shows the statistics for the dataset. We can see that both the training set and testing set have sufficient diversity. It is also worth noting that we allocate more video sequences for testing sets because small testing sets may lead to model over-fitting.

3.3. Dataset Features and Statistics

Sufficient Shadow Diversity. The diversity of shadow refers to the diversity of shadow sources, i.e. objects which cast the corresponding shadows. In ViSha, the shadow source is composed of seven main categories: Human, Plant, Vehicle, Animal, Artifact, Building, and Others. Figure 2 (a)&(c) show these categories (7 main classes with 60 sub-classes) and their mutual dependencies, respectively. Figure 2 (b) shows the shadow ratio distribution in ViSha. The shadow ratio is defined as the proportion of the number of shadow pixels to that of the entire image. Besides, as shown in Table 1, there are 95 videos having more than one shadow instance, making this dataset challenging for the video shadow detection task.

Motion of Camera and Objects. As a video dataset, ViSha contains ample motion diversity for objects and cameras (summarized in Table 1). In the viewpoint of shadow motion, 20 videos have the object shadows stay relatively static to the background, while the other 100 videos witness shadows suffer from fast-moving or distortion. Similarly, from the aspect of camera motion, there are 32 videos in which the camera is fixed and the shadow changes in the scene are relatively stable (*e.g.*, surveillance cameras), while in the rest 88 videos, the shadows exhibit drastic changes and/or motion blur lead by camera shaking or movement.

Various Lighting Conditions. Different lighting conditions can lead to hard shadow, which has an obvious boundary, or soft shadow, which has a rather blurry or unclear boundary. Hard shadow is usually produced when the scene contains only one single strong light source (*e.g.*, sunshine). Soft shadow is usually caused under lighting condition with multiple light sources. 88 videos in Visha contain the hard shadows while 30 videos contain the soft shadows.

Richness of the Scene. As we all know, data-driven models are subjected to the domain shift. For example, if all videos in the training set of ViSha are taken in the daytime, the trained models can hardly handle the shadows in night scenes. The same phenomenon also applies to the cases of indoor and outdoor scenes. In order to avoid such a problem, we build ViSha with 86 daytime videos and 34 night videos. Furthermore, ViSha contains 35 indoor videos and 85 ourdoor videos. More examples of ViSha can be found in the supplementary material.

4. Proposed Method

4.1. Overview of Our Network

Figure 3 shows the schematic illustration of our triple-cooperative video shadow detection network (TVSD-Net). The intuition behind our network is to leverage discriminative feature information at both intra-video and inter-video levels. That is, for neighboring frames from the same video, their features shall be similar; while frames from different videos will have features to be distinguishable.

Our TVSD-Net takes three shadow images as inputs. The first two images (denoted as $\{\mathbf{I}_{a1}, \mathbf{I}_{a2}\}$) are from the same video, while the third image $\mathbf{I}_b$ is randomly selected from another video. We devise three branches to pass each input image into a feature embedding module to extract three high-level semantic features, which are denoted as $\{\mathbf{E}_{a1}, \mathbf{E}_{a2}, \mathbf{E}_b\} \in \mathbb{R}^{26\times26\times256}$. The feature embedding module consists of a feature extraction backbone (ResNeXt-

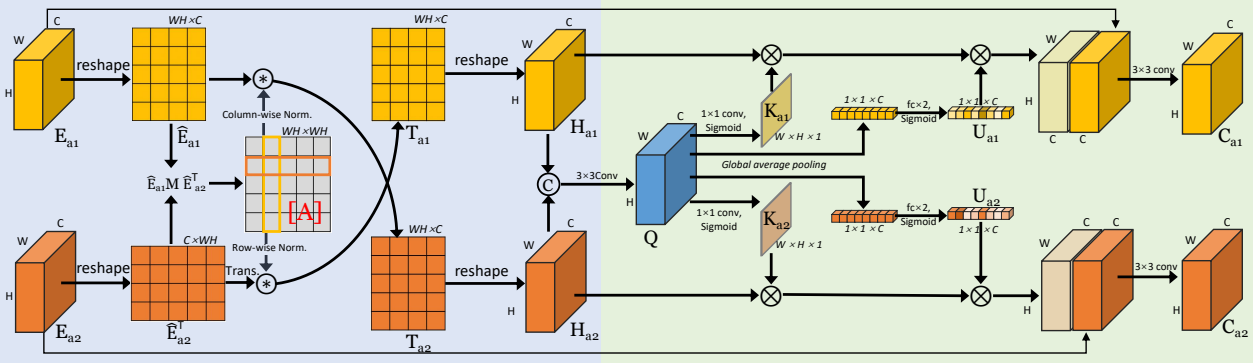

Figure 4: The schematic illustration of Dual Gated Co-attention (DGC) Module. Blue and green parts represent the collaboration attention (co-attention) module and dual-gated module, respectively. See Section 4.2 for details.

101) and an atrous spatial pyramid pooling (ASPP) module, which has a 1×1 point-wise convolution, three 3×3 convolutions with dilation rates of 12, 24, and 36 respectively, and a global average pooling layer. We empirically replace the last CNN layers of ResNeXt-101 with the dilation convolution (dilation rate of 2) and set the first convolutional stride to 1 to balance the spatial resolution of features and GPU memory size.

To learn global intra-video features, we devise a co-attention mechanism to emphasize the coherent information in $\mathbf{E}_{a1}, \mathbf{E}_{a2}$ from the same video (DGC module; see Figure 4). The refined features are denoted as $\mathbf{C}_{a1}$ and $\mathbf{C}_{a2}$, respectively. Note that deep CNN layers are able to capture highly semantic features tending to describe global attributes of shadow regions, while shallow CNN layers are responsible for extracting subtly fine features to represent delicate structures. We concatenate the refined high-level feature $\mathbf{C}_{a1}$ with a low-level feature map $\mathbf{L}_{a1}$ from the feature extraction backbone via a ship connection in the first network branch and then apply 3×3 and 1×1 convolutional layers on the concatenated features to generate a shadow detection result $\mathbf{S}_{a1}$. Similarly, the second branch concatenates $\mathbf{C}_{a2}$ with a low-level feature map $\mathbf{L}_{a2}$ of the feature extraction backbone to generate another shadow detection result S_{a2}.

In the third branch, without any co-attention module, we directly concatenate high-level features $\mathbf{E}_b$ with a low-level feature map $\mathbf{L}_b$, and predict one more shadow detection result $\mathbf{S}_b$. What's more, we devise a triple-cooperative module (T module; see Section 4.3) to learn inter-video features in helping shadow detection. The auxiliary loss adopted in T module makes $\mathbf{E}_{a1}$ and $\mathbf{E}_{a2}$ from two frames of the same video similar, while $\mathbf{E}_b$ from another video should be dissimilar to them.

Loss Function. To better handle the scale variance of shadows, we fuse the binary cross entropy (BCE) loss function with a lovász-hinge loss [3] function to compute the shadow detection loss ($\mathcal{L}_{seg}$) of all three inputs I_{a_1}, I_{a_2}, and I_b:

$$\mathcal{L}_{seg} = \mathcal{L}_{a1} + \mathcal{L}_{a2} + \mathcal{L}_b, \tag{1}$$

where

$$\begin{aligned}
\mathcal{L}_{a1} &= \Phi_{BCE}(\mathbf{S}_{a1}, \mathbf{G}_{a1}) + \Phi_{Hinge}(\mathbf{S}_{a1}, \mathbf{G}_{a1}) \,, \\
\mathcal{L}_{a2} &= \Phi_{BCE}(\mathbf{S}_{a2}, \mathbf{G}_{a2}) + \Phi_{Hinge}(\mathbf{S}_{a2}, \mathbf{G}_{a2}) \,, \\
\mathcal{L}_b &= \Phi_{BCE}(\mathbf{S}_b, \mathbf{G}_b) + \Phi_{Hinge}(\mathbf{S}_b, \mathbf{G}_b) \,.
\end{aligned} \tag{2}$$

Here, $\Phi_{BCE}(\cdot)$ and $\Phi_{Hinge}(\cdot)$ denote the BCE loss and the lovász-hinge loss, respectively; $\mathbf{S}_{a1}$/$\mathbf{S}_{a2}$/$\mathbf{S}_b$ and $\mathbf{G}_{a1}$/$\mathbf{G}_{a2}$/$\mathbf{G}_b$ are the predicted shadow detection map and the corresponding ground truth of $\mathbf{I}_{a1}$/$\mathbf{I}_{a2}$/$\mathbf{I}_b$;

Finally, we use a combination of the shadow detection segmentation loss $\mathcal{L}_{seg}$ and the devised auxiliary task loss $\mathcal{L}_{aux}$ (described in Section 4.3) to train our whole network. The total loss of our network is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \beta\mathcal{L}_{aux}, \tag{3}$$

where β is to control the weight of auxiliary loss, and we empirically set $\beta = 10$ in our experiments.

4.2. Dual Gated Co-attention Module

The dual gated co-attention module explicitly encodes intra-video correlations between a pair of frames in a video, via a co-attention mechanism and a dual-gated mechanism. This enables TVSD-Net to focus on frequently coherent regions, thus further helping to discover the shadow regions and produce reasonable VSD results.

Co-attention Mechanism. The blue region of Figure 4 shows the collaboration-attention (co-attention) module, which takes two features $\mathbf{E}_{a1} \in \mathbb{R}^{W\times H\times C}$, $\mathbf{E}_{a2}$ as the inputs to compute their correlations. Inspired by [35], we first reshape $\mathbf{E}_{a1}$ to be a new feature map $\hat{\mathbf{E}}_{a1} \in \mathbb{R}^{WH\times C}$ and reshape $\mathbf{E}_{a2}$ to be $\hat{\mathbf{E}}_{a2} \in \mathbb{R}^{WH\times C}$, and compute an affinity matrix $\mathbf{A} \in \mathbb{R}^{WH\times WH}$:

$$\mathbf{A} = \hat{\mathbf{E}}_{a1}\mathbf{M}\hat{\mathbf{E}}_{a2}^{\top}, \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^{C\times C}$ is a weight matrix. Intuitively, each element of $\mathbf{A}$ represents the similarity between each column of $\mathbf{E}_{a1}$ and each row of $\mathbf{E}_{a2}$.

From $\mathbf{A}$, we employ a $\texttt{Softmax}$ function to column-wisely and row-wisely normalize $\mathbf{A}$ respectively, and multiply the resultant normalization features with $\mathbf{A}$ to compute two co-attention enhanced features $\mathbf{T}_{a1}$ and $\mathbf{T}_{a2}$:

$$\begin{aligned}\mathbf{T}_{a1} &= \texttt{Softmax}(\mathbf{A}) * \hat{\mathbf{E}}_{a2} \in \mathbb{R}^{C\times WH}\ , \\ \mathbf{T}_{a2} &= \texttt{Softmax}(\mathbf{A}^{\top}) * \hat{\mathbf{E}}_{a1} \in \mathbb{R}^{C\times WH}\ ,\end{aligned} \tag{5}$$

Then, we reshape $\mathbf{T}_{a1}$ to be $\mathbf{H}_{a1} \in \mathbb{R}^{C\times W\times H}$ and reshape $\mathbf{T}_{a2}$ to be $\mathbf{H}_{a2} \in \mathbb{R}^{C\times W\times H}$. By this way, we intuitively transform the $a1$ features ($\mathbf{E}_{a1}$) to a fake $a2$ features ($\mathbf{H}_{a2}$). Compared to the original $\mathbf{E}_{a2}$, the fake one ($\mathbf{H}_{a2}$) encodes more temporal information.

Dual-gated Mechanism. Since there may exist potential appearance variations (*e.g.*, occlusion, out-of-view) between two neighboring frames, using co-attention module enhances the coherent features, yet may also introduce some noises from adjacent frames. Hence, it is better to weight co-attention enhanced features from two input frames, instead of treating the learned co-attention information equally. To achieve this goal, we propose a dual-gated mechanism to obtain co-attention confidences.

Unlike the self-gated mechanism [35], we learn the co-attention confidences by leveraging $\mathbf{H}_{a1}$ and $\mathbf{H}_{a2}$ together. Our dual-gated mechanism consists of a spatial gated operation and a channel gated operation.

Specifically, we fuse $\mathbf{H}_{a1}$ and $\mathbf{H}_{a2}$ by applying a 3×3 convolution on the concatenation of $\mathbf{H}_{a1}$ and $\mathbf{H}_{a2}$ to compute a fused feature map $\mathbf{Q}$:

$$\mathbf{Q} = \texttt{Conv}(\texttt{Concat}(\mathbf{H}_{a1}, \mathbf{H}_{a2}))\ . \tag{6}$$

Then, two spatial gated maps ($\{\mathbf{K}_{a1}, \mathbf{K}_{a2}\} \in \mathbb{R}^{W\times H\times 1}$) are computed by utilizing a $\texttt{Sigmoid}$ function and a 1×1 convolution on $\mathbf{Q}$:

$$\begin{aligned}\mathbf{K}_{a1} &= \texttt{Sigmoid}(\texttt{Conv}(\mathbf{Q}))\ , \\ \mathbf{K}_{a2} &= \texttt{Sigmoid}(\texttt{Conv}(\mathbf{Q}))\ .\end{aligned} \tag{7}$$

Moreover, we generate two channel-wise gated maps $\mathbf{U}_{a1}$ and $\mathbf{U}_{a2}$:

$$\begin{aligned}\mathbf{U}_{a1} &= \texttt{Sigmoid}(\texttt{fc}(\texttt{GAP}(\mathbf{Q})))\ , \\ \mathbf{U}_{a2} &= \texttt{Sigmoid}(\texttt{fc}(\texttt{GAP}(\mathbf{Q})))\ .\end{aligned} \tag{8}$$

Once obtaining spatial and channel gated maps, we multiply the spatial gated map with the co-attention enhanced features $\{\mathbf{H}_{a1}, \mathbf{H}_{a2}\}$, and then multiply the resultant features with the channel gate map to produce gated features $\{\mathbf{D}_{a1}, \mathbf{D}_{a2}\}$. We then apply a 3×3 convolution on the concatenation of $\mathbf{D}_{a1}$/$\mathbf{D}_{a2}$ and $\mathbf{E}_{a1}$ ($\mathbf{E}_{a2}$) to produce output features of the dual gated co-attention module, i.e., $\mathbf{C}_{a1}$ and $\mathbf{C}_{a2}$. The definitions of $\mathbf{C}_{a1}$ and $\mathbf{C}_{a2}$ are given by:

$$\begin{aligned}\mathbf{C}_{a1} &= \texttt{Conv}(\texttt{Cancat}(\mathbf{E}_{a1}, \mathbf{H}_{a1}\otimes\mathbf{K}_{a1}\otimes\mathbf{U}_{a1}))\ , \\ \mathbf{C}_{a2} &= \texttt{Conv}(\texttt{Cancat}(\mathbf{E}_{a2}, \mathbf{H}_{a2}\otimes\mathbf{K}_{a2}\otimes\mathbf{U}_{a2}))\ ,\end{aligned} \tag{9}$$

where $\otimes$ denotes element-wise product.

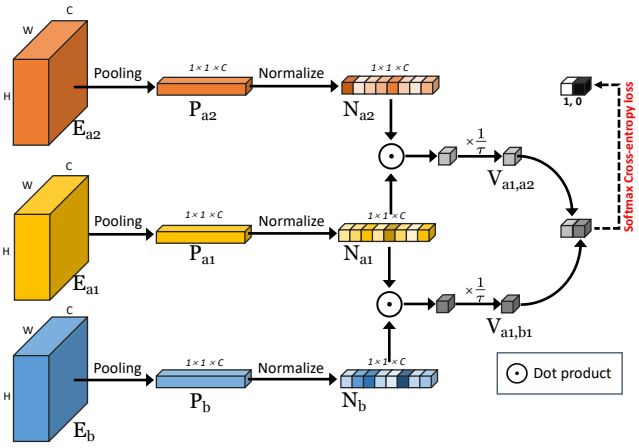

Figure 5: The schematic illustration of our triple-cooperative (T) module; See Section 4.3 for details.

4.3. Triple-cooperative Module

Given high-level features $\{\mathbf{E}_{a1}, \mathbf{E}_{a2}, \mathbf{E}_b\}$ of the three input images $\{\mathbf{I}_{a1}, \mathbf{I}_{a2}, \mathbf{I}_b\}$, we devise a triple-cooperative module (T-module) to make features from the same video similar and features from different videos dissimilar. Figure 5 shows the schematic illustration of T-module, which computes an auxiliary loss based on $\{\mathbf{E}_{a1}, \mathbf{E}_{a2}, \mathbf{E}_b\}$. Intuitively, we expect the similarity between two frames from the same video to be close to 1, while the similarity between two frames from different videos approaches to 0.

To be specific, we apply three global average pooling operations on $\mathbf{E}_{a1}$, $\mathbf{E}_{a2}$, and $\mathbf{E}_b$ to obtain three features $\{\mathbf{P}_{a1}, \mathbf{P}_{a2}, \mathbf{P}_b\} \in \mathbb{R}^{1\times 1\times 256}$, which are then normalized as $\{\mathbf{N}_{a1}, \mathbf{N}_{a2}, \mathbf{N}_b\}$:

$$\begin{aligned}\mathbf{N}_{a1} &= \frac{\mathbf{P}_{a1}}{\texttt{max}(\|\mathbf{P}_{a1}\|_2, \epsilon)}\ ,\ \mathbf{N}_{a2} = \frac{\mathbf{P}_{a2}}{\texttt{max}(\|\mathbf{P}_{a2}\|_2, \epsilon)}\ , \\ \mathbf{N}_b &= \frac{\mathbf{P}_b}{\texttt{max}(\|\mathbf{P}_b\|_2, \epsilon)}\ ,\end{aligned} \tag{10}$$

where ϵ is a small positive number and it is set as $\epsilon = 1e^{-12}$ to avoid division by zero.

After that, we compute the similarity $V_{a1,a2}$ of $\mathbf{N}_{a1}$ and $\mathbf{N}_{a2}$ from the same video by computing the dot produce of $\mathbf{N}_{a1}$ and $\mathbf{N}_{a2}$, and also the similarity $V_{a1,b}$ of $\mathbf{N}_{a1}$ and $\mathbf{N}_b$ from different videos, given by:

$$V_{a1,a2} = \mathbf{N}_{a1}\cdot\mathbf{N}_{a2},\ \text{and},\ V_{a1,b} = \mathbf{N}_{a1}\cdot\mathbf{N}_b. \tag{11}$$

Then, we multiple two similarities with a temperature $\frac{1}{\tau}$ [16] and concatenated them to form a two-element vector. The rest is to compare whether this two-element vector is close to the target distribution of (1, 0). Here, after applying a softmax function on the two-element vector for normalization, we compute a cross-entropy loss between the two vectors as the auxiliary loss of our T module.

In summary, the definition of our auxiliary loss $\mathcal{L}_{aux}$ is given by:

$$\mathcal{L}_{aux} = -\log \frac{\exp(V_{a1,a2}/\tau)}{\exp(V_{a1,a2}/\tau) + \exp(V_{a1,b}/\tau)}, \quad (12)$$

where τ=0.7 is a temperature constant to control degree of two similarities. The sum is over one positive and one negative samples. It is clear that the auxiliary loss makes the similarity $V_{a1,a2}$ from same video to be 1 while making the similarity of $V_{a1,b}$ from different videos to be 0. Hence, $\mathcal{L}_{aux}$ tends to have a small score when $\mathbf{N}_{a1}$ is similar to $\mathbf{N}_{a2}$ from the same video, and dissimilar to $\mathbf{N}_b$ from a different video.

4.4. Implementation Details

We implement our TVSD-Net using PyTorch. Adam optimizer is employed to train the network with mixed precision training [36] on a NVIDIA GTX 2080Ti. We initialize the feature extraction backbone via a pre-trained ResNeXt-101 [50] on ImageNet while other layers are trained from scratch. The weight decay, batch size, epoch number are set as 0.0005, 5, and 12, respectively. We set the initial learning rate as 0.0005 for scratch layers and 0.00005 for pretrained layers, and then use the cosine decay with a warm-up period to adjust the learning rate. TVSD-Net requires about 0.06s to process an image of 416×416.

In the testing phase, we take the shadow detection result $\mathbf{S}_{a1}$ in the first branch as the output of the TVSD-Net. Given an input video, to obtain the shadow detection result of each frame (we call it target frame), we follow [35] to empirically select the subsequent five frames of the target frame, and then pass the target frame as $\mathbf{I}_{a1}$, and each of fives frames as $\mathbf{I}_{a2}$ to the TVSD-Net. By doing so, we obtain five segmentation results and then average the five results as the final shadow detection result of the target frame.

5. Experiments

5.1. Experimental Settings

Evaluation Metrics. We adopt four common evaluation metrics to quantitatively compare video shadow detection methods. They are Mean Absolute Error (**MAE**) [18, 57] and F-measure ($\mathbf{F}_\beta$) [18, 55], Intersection over Union (**IoU**) [51], and Balance Error Rate (**BER**) [20]. In general, a better video shadow detection method shall have smaller BER and MAE scores, and larger $\mathbf{F}_\beta$ and IoU scores.

Comparative Methods. Since there is no CNN-based method for video shadow detection, we make comparison against 12 state-of-the-art methods for relevant tasks, including FPN [34], PSPNet [52], DSS [18], R^{3}Net [9], BDRAR [56], DSD [53], MTMT [4], PDBM [42], COSNet [35], MGA [32], FEELVOS [44] and STM [40].

Table 3: Comparing our network (TVSD-Net) against the state-of-the-art methods.

Method	Year	MAE ↓	F_β ↑	IoU ↑	BER↓
BDRAR	2018	0.050	0.695	0.484	21.29
DSD	2019	0.044	0.702	0.518	19.88
MTMT	2020	0.043	0.729	0.517	20.28
FPN	2017	0.044	0.707	0.512	19.49
PSPNet	2017	0.051	0.642	0.476	19.75
DSS	2017	0.045	0.696	0.502	19.77
R^{3}Net	2018	0.043	0.710	0.502	20.40
PDBM	2018	0.066	0.623	0.466	19.73
COSNet	2019	0.040	0.705	0.514	20.50
MGA	2019	0.067	0.601	0.399	25.77
FEELVOS	2019	0.043	0.710	0.512	19.76
STM	2019	0.068	0.597	0.408	25.69
TVSD-Net (Ours)	-	**0.033**	**0.757**	**0.567**	**17.70**

Table 4: Ablation analysis. Here, "Co-att" denotes the original Co-attention module in [35]; "DGM" denotes our proposed Dual-Gated Mechanism; "T-module" denotes the Triple-cooperative module.

Network	Co-att	DGM	T-m	MAE ↓	F_β ↑	IoU ↑	BER↓
basic	×	×	×	0.042	0.743	0.538	19.17
basic+Co-att	✓	×	×	0.041	0.739	0.545	18.53
basic+T-module	×	×	✓	0.039	0.739	0.549	18.57
ours-w/o-T-module	✓	✓	×	0.038	0.744	0.551	18.72
ours-w/o-DGM	✓	×	✓	0.038	0.756	0.540	19.55
our method	×	✓	✓	**0.033**	**0.757**	**0.567**	**17.70**

Among them, FPN and PSPNet are developed for single-image semantic segmentation. DSS and R^{3}Net are dedicated for single-image saliency detection, while BDRAR, DSD, and MTMT are utilized for single-image shadow detection. Lastly, PDBM, COSNet, MGA, FEELVOS and STM are for video saliency detection and object object segmentation. We use their public codes, and re-train these methods on our training set for a fair comparison.

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

Table 3 shows the performances on our ViSha dataset, where COSNet has the best MAE score of 0.040; MTMT has the best F_β score of 0.729; DSD has the best IoU score of 0.518; PDBM has the best BER score of 19.73. Compared to the best-performing existing methods, our method obtains improvements with large margins, with MAE improvement of 17.50%, F_β improvement of 3.84%, IoU improvement of 9.46%, and BER improvement of 10.29% on our dataset, respectively. That shows the superiority of our method on video shadow detection.

Figure 6 visually compares the video shadow detection maps produced by our method and the state-of-the-arts. From the results, we can see that our TVSD-Net (3rd column of Figure 6) can more accurately identify shadow pixels than compared methods. It effectively locates different shadows under various backgrounds, and successfully dis-

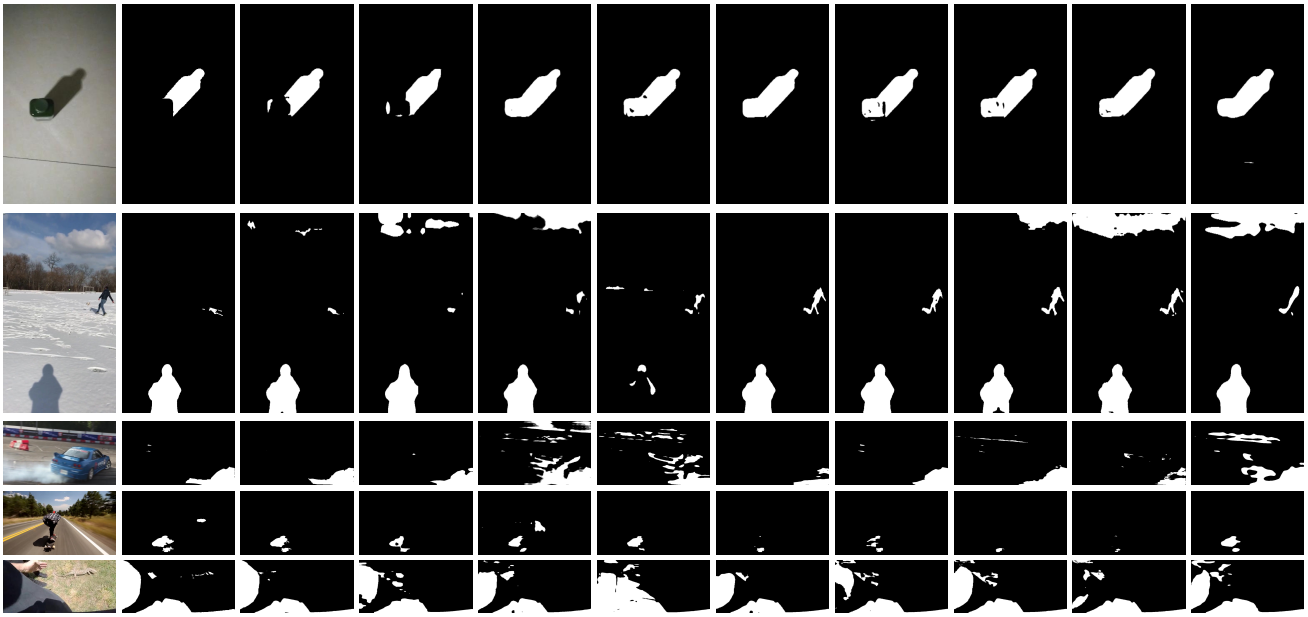

Figure 6: Visual comparison of shadow maps produced by our method and other methods.

criminates true shadows from those non-shadow dark regions. For example, in the 1st row, most compared methods regard both the bottle and its shadow as the shadow regions, while our TVSD-Net can discriminate them successfully. A similar situation is also reflected in the 2nd row, our TVSD-Net can better capture the distinction between the soccer player and his shadow.

5.3. Ablation Study

We perform ablation study experiments to verify the performance of Dual-gated co-attention module and Triple-cooperative module in TVSD-Net. Here, we consider four baseline networks. The first baseline network (denoted as "basic") is constructed by removing the co-att, DGM, and T module from the TVSD-Net. The second (denoted as "basic+co-att") is to add the original co-attention module [35] into "basic". The third baseline (denoted as "basic+T-module") is to add our T-module into "basic". The fourth baseline (denoted as "ours-w/o-T-module") removes T-module from our network. The fifth baseline (denoted as "ours-w/o-DGM") removes DGM from our network.

Table 4 summarizes the BER values of our network and seven baseline networks on the ViSha dataset. From the results, we have the following observations: (i) "basic+co-att" have the superior performance of four evaluation metrics over "basic", which means that learning intra-video coherent information can provide helpful information for video shadow detection. (ii) "basic+T-module" has smaller BER and MAE scores and larger F_β and IoU scores than "basic", demonstrating that the additional auxiliary loss from the T module incurs detection improvement. (iii) "ours-w/o-T-module" can more accurately detect shadow pixels than "basic+co-att" due to its better results of F_β, IoU, BER, and MAE. It indicates dual gated module helps to increase the co-attention confidences than the original co-attention module [35]. (iv) our TVSD-Net has better metric results than "ours-w/o-T-module" and "ours-w/o-DGM", showing that combining the two modules achieves a higher video shadow detection accuracy.

6. Conclusion

This paper presents a novel network for video shadow detection. One of our key contributions is to first collect a learning-oriented video shadow detection (ViSha) dataset, which contains 120 videos with 11,685 frames covering various objects and scenes, with pixel-level shadow annotations. The second contribution is the development of a novel network for video shadow detection, by learning intra-video and inter-video discriminative properties of shadows. Experimental results on the collected dataset demonstrated that our method consistently outperforms 12 state-of-the-art methods by a large margin. To the best of our knowledge, this work is the first annotated dataset for video shadow detection, and our ViSha dataset can facilitate further research in video shadow detection.

Acknowledgments: The work is supported by the National Natural Science Foundation of China (Grant No. 61902275, 61572354).

References

[1] Ç. Aytekin, H. Possegger, T. Mauthner, S. Kiranyaz, H. Bischof, and M. Gabbouj. Spatiotemporal saliency estimation by spectral foreground detection. *IEEE Transactions on Multimedia*, 20(1):82–95, 2017. 3

[2] C. Benedek and T. Sziranyi. Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. *IEEE Transactions on Image Processing*, 17(4):608–621, 2008. 2

[3] M. Berman, Amal R.T., and Matthew B B. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018. 5

[4] Z. Chen, L. Zhu, L. Wan, S. Wang, W. Feng, and P.-A. Heng. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*, pages 5611–5620, 2020. 2, 7, 8

[5] M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, and S.M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014. 3

[6] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10):2941–2959, 2019. 3

[7] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE Transactions on Image Processing*, 28(10):4819–4831, 2019. 3

[8] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003. 1

[9] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690. AAAI Press, 2018. 2, 7, 8

[10] A. Ecins, C. Fermuller, and Y. Aloimonos. Shadow free segmentation in still images using local density measure. In *ICCP*, pages 1–8, 2014. 1

[11] D.-P. Fan, W. Wang, M. Cheng, and J. Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 1, 3

[12] H. Fan, H. Ling, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, and C. Liao. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, pages 5374–5383, 2019. 3

[13] G.D. Finlayson, M.S. Drew, and C. Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 1

[14] G.D. Finlayson, S.D. Hordley, C. Lu, and M. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. 1

[15] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, pages 1134–1143, 2017. 3

[16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[17] S. Hosseinzadeh, M. Shakeri, and H. Zhang. Fast shadow detection from a single image using a patched convolutional neural network. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3124–3129, 2018. 2

[18] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019. 2, 7

[19] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2795–2808, 2020. 1

[20] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, pages 7454–7462, 2018. 1, 2, 7

[21] X. Huang, G. Hua, J. Tumblin, and L. Williams. What characterizes a shadow boundary under the sun and sky? In *ICCV*, pages 898–905, 2011. 1

[22] J. C. S. Jacques, C. R. Jung, and S. R. Musse. Background subtraction and shadow detection in grayscale video sequences. In *Brazilian Symposium on Computer Graphics & Image Processing*, 2005. 2

[23] I.N. Junejo and H. Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. In *ECCV*, pages 318–331, 2008. 1

[24] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics*, 30(6):157:1–157:12, 2011. 1

[25] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Automatic feature learning for robust shadow detection. In *CVPR*, pages 1939–1946, 2014. 1, 2

[26] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016. 3

[27] J.-F. Lalonde, A.A. Efros, and S.G. Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, pages 183–190, 2009. 1

[28] J.-F. Lalonde, A.A. Efros, and S.G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, pages 322–335, 2010. 1

[29] H. Le, Y. Vicente, F Tomas, V. Nguyen, M. Hoai, and D. Samaras. A+ D Net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, pages 662–678, 2018. 1

[30] T.-N. Le and A. Sugimoto. Deeply supervised 3D recurrent fcn for salient object detection in videos. In *BMVC*, pages 1–13, 2017. 1, 3

[31] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, pages 3243–3252, 2018. 3

[32] H. Li, G. Chen, G. Li, and Y. Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 2, 7, 8

[33] P. Liang, E. Blasch, and H. Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing*, 24(12):5630–5644, 2015. 3

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 2, 7

[35] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, pages 3623–3632, 2019. 1, 2, 5, 6, 7, 8

[36] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al. Mixed precision training. 2018. 7

[37] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1079–1087, 2004. 1

[38] S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1079–1087, 2004. 2

[39] V. Nguyen, T.F.Y. Vicente, M. Zhao, M. Hoai, and D. Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, pages 4510–4518, 2017. 1, 2

[40] S.W. Oh, J.-Y. Lee, N. Xu, and S.J. Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 1, 2, 7

[41] T. Okabe, I. Sato, and Y. Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *ICCV*, pages 1693–1700, 2009. 1

[42] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018. 2, 3, 7, 8

[43] T.F.Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pages 816–832, 2016. 1, 2

[44] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, June 2019. 7

[45] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[46] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, 2015. 3

[47] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017. 1, 3

[48] Y. Wang, X. Zhao, Y. Li, X. Hu, K. Huang, et al. Densely cascaded shadow detection network via deeply supervised parallel fusion. In *IJCAI*, pages 1007–1013, 2019. 2

[49] Y. Wu, J. Lim, and M. H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 3

[50] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 7

[51] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for real-time semantic segmentation on high-resolution images. In *ECCV*, pages 405–420, 2018. 7

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2, 7, 8

[53] Q. Zheng, X. Qiao, Y. Cao, and R.W. Lau. Distraction-aware shadow detection. In *CVPR*, pages 5167–5176, 2019. 1, 2, 7, 8

[54] J. Zhu, K.G. Samuel, S.Z. Masood, and M.F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, pages 223–230, 2010. 1

[55] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng. Aggregating attentional dilated features for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3358–3371, 2019. 7

[56] L. Zhu, Z. Deng, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pages 121–136, 2018. 1, 2, 7, 8

[57] L. Zhu, X. Hu, C.-W. Fu, J. Qin, and P.-A. Heng. Saliency-aware texture smoothing. *IEEE Transactions on Visualization and Computer Graphics*, 26(7):2471–2484, 2018. 7