Learning Shadow Correspondence for Video Shadow Detection

Xinpeng Ding[1[0000−0001−7653−1199]], Jingwen Yang[1[0000−0003−2420−0130]], Xiaowei Hu[2[0000−0002−5708−7018]], and Xiaomeng Li[1,3[0000−0003−1105−8083]⋆]

[1] The Hong Kong University of Science and Technology
{xdingaf,jyangbv}@connect.ust.hk
[2] Shanghai AI Laboratory
huxiaowei@pjlab.org.cn
[3] The Hong Kong University of Science and Technology Shenzhen Research Institute
eexmli@ust.hk

Abstract. Video shadow detection aims to generate consistent shadow predictions among video frames. However, the current approaches suffer from inconsistent shadow predictions across frames, especially when the illumination and background textures change in a video. We make an observation that the inconsistent predictions are caused by the shadow feature inconsistency, *i.e.*, the features of the same shadow regions show dissimilar proprieties among the nearby frames. In this paper, we present a novel **S**hadow-**C**onsistent **Cor**respondence method (**SC-Cor**) to enhance pixel-wise similarity of the specific shadow regions across frames for video shadow detection. Our proposed SC-Cor has three main advantages. Firstly, without requiring the dense pixel-to-pixel correspondence labels, SC-Cor can learn the pixel-wise correspondence across frames in a weakly-supervised manner. Secondly, SC-Cor considers intra-shadow separability, which is robust to the variant textures and illuminations in videos. Finally, SC-Cor is a plug-and-play module that can be easily integrated into existing shadow detectors with no extra computational cost. We further design a new evaluation metric to evaluate the temporal stability of the video shadow detection results. Experimental results show that SC-Cor outperforms the prior state-of-the-art method, by 6.51% on IoU and 3.35% on the newly introduced temporal stability metric.

Keywords: Shadow detection, video understanding, and correspondence learning.

1 Introduction

Shadows in natural images or videos present different colors and brightness. Known where the shadow is, we can infer light source directions [26, 36], scene geometry [22, 35, 21], and camera locations or parameters [21]. Therefore, shadow detection has attracted a lot of attention and achieved remarkable progress.

⋆ Corresponding Authors

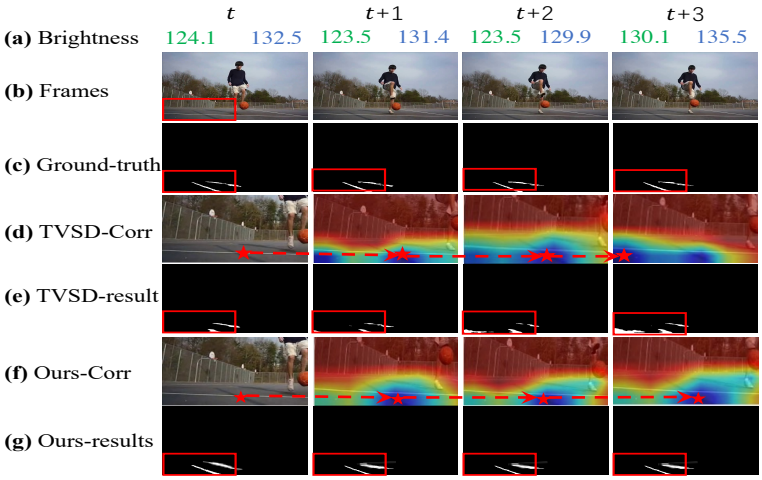

Fig. 1: Comparison of correspondences and results of TVSD [7] and Ours. We compute the brightness of four selected frames. Green and blue values indicate the non-shadow and shadow regions respectively. Given a query shadow region in the t-th frame, *i.e.*, the orange pentagram, we find the most similar features in nearby frames for the query, and regard the found features as its correspondences. "TVSD-Corr" and "Ours-Corr" indicate the correspondences found by TVSD [7] and our method respectively. "TVSD-result" and "Our-result" refer to the results predicted by TVSD and our method. It is clear that the found correspondences in the $(t+3)$-th frame are in non-shadow regions (dark areas). This shadow features inconsistency, *i.e.*, features of the same shadow region may be dissimilar across frames, would result inconsistent prediction (red boxes). Our method can address the shadow feature inconsistency, and generate the contiguous results.

However, most of the recent methods [55, 23, 45, 33, 8, 16, 56] detect shadows from single images while shadow detection over dynamic scenes, *i.e.*, in videos, is less explored.

To explore the powerful representation capability of deep learning for video shadow detection (VSD), Chen *et al.* [7] collect a large-scale video shadow detection (ViSha) dataset covering various scenarios. Then, a global contrastive objective is applied on the frame-level, which enhances the similarity between frames in the same video and push away the representations of frames from different videos. However, video shadow detection is a fine-grained pixel-level detection task, this frame-level semantic constraint may ignore shadow details, i.e., the same shadow regions across frames show dissimilar, resulting in inconsistent predictions; see red boxes in Fig. 1 (d). Video data has the inherent property of *temporal consistency*, where the nearby frames are expected to contain similar shadow regions. Hence, we aim to explore both frame-level accuracy and temporal-level consistency for video shadow detection. In this paper, we make a critical observation that this inconsistent prediction is caused by the shadow

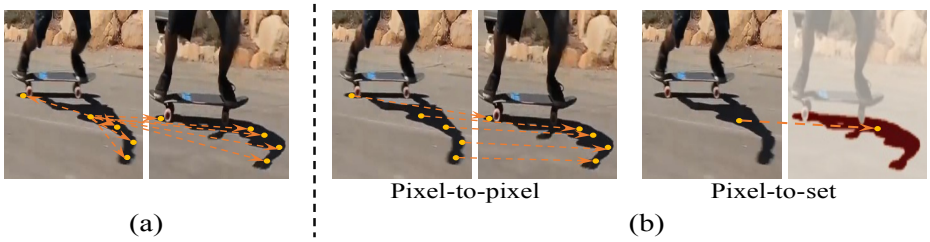

Fig. 2: (a)Supervised contrastive learning pulls close all pixels in shadow regions, which is too strict to generate complete shadow detection; see Fig. 6 and Table 2 for details. (b) We aims to leverage correspondence learning to consider intra-shadow separability. Unlike existing correspondence learning [31, 20] that require *pixel-to-pixel* labels among video frames, for each pixel in the shadow, we only know its corresponding pixel is within a shadow region in another frame, denoted as *pixel-to-set* correspondence learning.

feature inconsistency, *i.e.*, the features of the same shadow regions show dissimilar proprieties among the nearby frames. For example, due to the illumination change (see Fig. 1 (a)), the extracted features in a specific shadow region may show higher similarity with dark non-shadow regions in the nearby frames; see orange pentagram in the $(t+3)$-th frame in Fig. 1 (c).

To address this problem, we aim to enhance temporal pixel-wise similarity for the specific shadow regions across frames, thus improving the detection accuracy and consistency in shadow videos; see Figs. 1 (e) and (f). The supervised contrastive learning aims to increase intra-class compactness and inter-class separability, and has been used for image classification [24] or semantic segmentation [54, 51]. An intuitive way is to use the supervised contrastive learning to pull close pixels in shadow regions across frames, and push away shadowed pixels and non-shadowed pixels. However, as objects move and illumination changes, the same shadow region may appear on backgrounds with different textures across frames. Simply adopting the supervised contrastive learning for video, *i.e.*, pulling close all pixels in shadows, leads to incomplete shadow regions; see Fig. 6 and Table 2 for the comparison results.

Hence, in this paper, we leverage the correspondence learning to learn a more fine-grained pixel-wise similarity. *i.e.*, only encouraging a pixel to be similar with its corresponding pixels in nearby frames. However, unlike the existing correspondence learning [31, 20, 11] that requires *pixel-to-pixel* correspondence labels across video frames, we do not need the pixel-wise correspondence labels. To this end, we present a novel **S**hadow-**C**onsistent **Cor**respondence method, namely **SC-Cor**, to learn the dense shadow correspondence in a *pixel-to-set* way, based on a key prior knowledge that a corresponding *(pixel)* of shadow is within a shadow region *(set)* in another frame; see Fig. 2 (a). Different from the supervised contrastive learning [54, 24, 51], our proposed SC-Cor keeps the pixel most similar to the anchor in the shadow region and considers inter-shadow

separability, which is robust to the variant textures and illuminations in videos. Note that our SC-Cor is a plug-and-play module and is only used in the training process. Therefore, SC-Cor can be easily applied to any deep-learning-based video/image shadow detection method without additional computational cost in testing.

Finally, existing metrics only evaluate the performance of VSD in frame-level, *e.g.*, frame-level Balance Error Rate (BER), and ignores the temporal consistency of shadow predictions. To this end, we introduce a new evaluation metric, temporal stability (TS), which computes the intersection over union score between the adjacent frames, thus helping to evaluate the temporal consistency of shadow predictions in videos. Below, we summarize the major contributions of this work:

- We present a novel and plug-and-play shadow-consistent correspondence (SC-Cor) method for video shadow detection. Compared with the existing *pixel-to-pixel* learning, our proposed SC-Cor is learned in a *pixel-to-set* way, without requiring pixel-wise correspondence labels.
- To fairly evaluate the temporal consistency of different shadow detection approaches, we introduce a new evaluation metric, which evaluates the flow-warped IoU between the adjacent video frames.
- We evaluate our SC-Cor on the benchmark dataset for video shadow detection and the experimental results show that our method clearly outperforms various state-of-the-art approaches in terms of both frame-level and temporal-level evaluation metrics.

2 Related Work

Image shadow detection. Early traditional methods are based on the hand-crafted shadow features, *e.g.*, intensity, chromaticity, physical properties, geometry, and textures [37]. Recently, deep-learning-based methods become the mainstream algorithms for shadow detection [23, 40, 45, 33, 18, 16, 27, 56, 55, 9]. Khan *et al.* [23] build the first method based on deep neural network, which is a seven-layer CNN that learns from super-pixel level features and object boundaries. Hu *et al.* [17] present a fast shadow detection network by designing a detail enhancement module to refine shadow details. In the most recent work, Zhu *et al.* [57] design a feature decomposition and re-weighting scheme, which leverages intensity-variant and intensity-invariant features via self-supervision to mitigate the susceptibility of the intensity cue. Except the general shadow detection, Wang *et al.* [48, 47] detecte the shadow regions associated with the objects simultaneously.

Video shadow detection. Early traditional video shadow detection (VSD) methods adopt the hand-crafted spectral and spatial features [1, 19, 32] to detect the shadow regions. To exploit the capability of deep-learning-based methods on this task, Chen *et al.* [7] collect the first large-scale VSD dataset ViSha. To detect the shadows in videos, they design a deep-learning-based method that contains

a dual gated co-attention module and an auxiliary similarity loss to mine frame-level consistency information between different videos. Hu *et al.* [15] capture the temporal consistency by an optical-flow-based warping module to align and combine features between video frames. However, due to lack of the temporal pixel-level relation, these methods would suffer from shadow feature inconsistency and generate temporal-inconsistent results. Unlike existing methods, this paper presents a novel solution to learn pixel-wise consistency by formulating the dense shadow correspondence objective. Our method is flexible and can be easily integrated into many existing methods designed for both single-image and video shadow detection methods.

Correspondence learning. Finding correspondences between pairs of images is a fundamental task in computer vision [43, 39, 3, 42, 31, 20]. However, these methods require pixel-level correspondence labels and can hardly be obtained in videos. Hence, numerous works aim to learn temporal correspondence in the unsupervised way [49, 50, 52]. These methods perform unsupervised correspondence learning on videos and show obvious improvement on the obvious foreground objects. However, shadows are usually less obvious than the foreground, and may show different appearances and deformation due to illumination and texture changes. Our SC-Cor can address the above problems in a weakly supervised way, which is proved by experiments (see Fig. 6 and Table 2). In this paper, different from all of these methods, we aim to learn pixel-wise similarity in a *pixel-to-set* way.

Contrastive learning. Contrastive learning pulls close an anchor and a positive sample, and pushes the anchor away from many negative samples, which has show great success in self-supervised learning [6, 4, 13, 5, 11, 12, 10]. Recently, the supervised contrastive learning aims to increase intra-class compactness and inter-class separability to improve image classification [24, 28] or semantic segmentation [54, 51]. However, as objects move and illumination changes, the same shadow region may appear on backgrounds with different textures across frames. The supervised contrastive learning, *i.e.*, simply pulling close all pixels in shadow regions is too strict, resulting in generating incomplete shadow regions; see Fig. 6 and Table 2 for details. Differently, our proposed SC-Cor aims to keep the pixel most similar to the anchor in the shadow regions, which considers inter-class separability due to the varying shadows in the videos.

3 Methodology

Fig. 3 (a) shows the training process of the overall SC-Cor framework, which can generate temporal-consistent and accurate shadow detection results. Formally, we denote a video sequence and the corresponding ground-truth (GT) masks as $\{V_t\}_{t=1}^{T}$ and $\{\mathbf{Y}_t\}_{t=1}^{T}$, respectively, where T is the frame number of this video sequence. Given two video frames, which are denoted as V_t and $V_{t+\delta}$ and δ is the time interval, we feed them into two branches of the framework; see Fig. 3 (a). Each branch contains a feature extractor, which is used to capture the spatial features, *i.e.*, $\mathbf{F}_t$ and $\mathbf{F}_{t+\delta}$ of the input frames. Note that the weights in these two

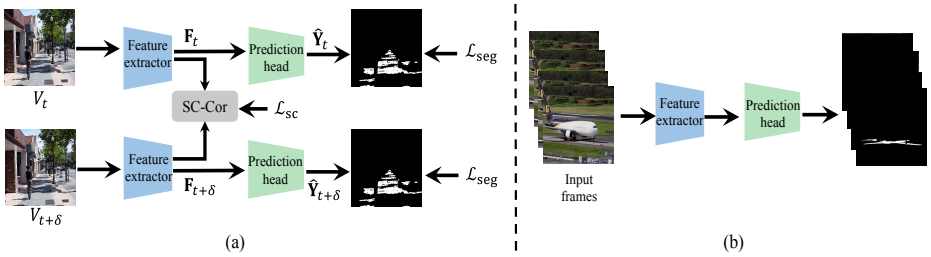

Fig. 3: (a) Illustration of the training process integrated with our shadow-consistent correspondence (SC-Cor) learning objective. Given two frames from one video, besides using the the segmentation loss $\mathcal{L}_{seg}$ to supervise their frame-level predictions individually, we also enhance their temporal consistency by SC-Cor, described in Section 3.1. (b) Illustration of the inference phase. The proposed SC-Cor is only applied during training. We can improve the temporal consistency as well as the frame-level accuracy without any extra parameters or computation cost during inference.

feature extractors are shared. Then, we adopt shadow-consistent correspondence (Fig. 4) to extract the temporal information of $\mathbf{F}_t$ and $\mathbf{F}_{t+\delta}$; see details in Sec. 3.1 and 3.2. Next, we send $\mathbf{F}_t$ and $\mathbf{F}_{t+\delta}$ into the shared prediction head to obtain the shadow detection results $\hat{\mathbf{Y}}_t$ and $\hat{\mathbf{Y}}_{t+\delta}$, which are supervised by the ground-truth masks $\mathbf{Y}_t$ and $\mathbf{Y}_{t+\delta}$. Note that the SC-Cor module is flexible and is only used in the training stage without any extra parameters introduced in the test stage, as shown in Fig. 3 (b). Therefore, it can serve as a plug-and-play component and can be used in many single-image or video shadow detection methods.

3.1 Shadow-Consistent Correspondence

To explore the temporal consistency for VSD, we aim to learn shadow correspondence to capture the pixel-wise relations between shadows across frames in the video, which acts as a regularizer to optimize the framework. As discussed in Sec. 1, instead of the dense *pixel-to-pixel* labels [39, 3], in this paper, we only obtain the *pixel-to-set* labels. To learn dense shadow correspondence, we introduce a novel shadow consistent correspondence method. The proposed shadow-consistent correspondence contains three modules: (a) a shadow guidance module, (b) a cross-frame correspondence module, and (c) a consistency regularization, as shown in Fig. 4.

(a) Shadow guidance module. The shadow guidance module aims to obtain a feature map that only contains feature vector on the shadow regions. Let $\mathbf{F}_t \in \mathbb{R}^{H\times W\times D}$ be the feature map of the frame V_t, where D, H and W denote the dimension, height, and width of the feature map, respectively. Here, we define the ground-truth shadow mask as $\mathbf{Y}_t \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{Y}_t = \{0, 1\}$, where $\mathbf{Y}_t(h, w) = 0$ indicates that the position (h, w) in $\mathbf{Y}_t$ is in the non-shadow regions and $\mathbf{Y}_t(h, w) = 1$ represents the position (h, w) is in the shadow regions. Then,

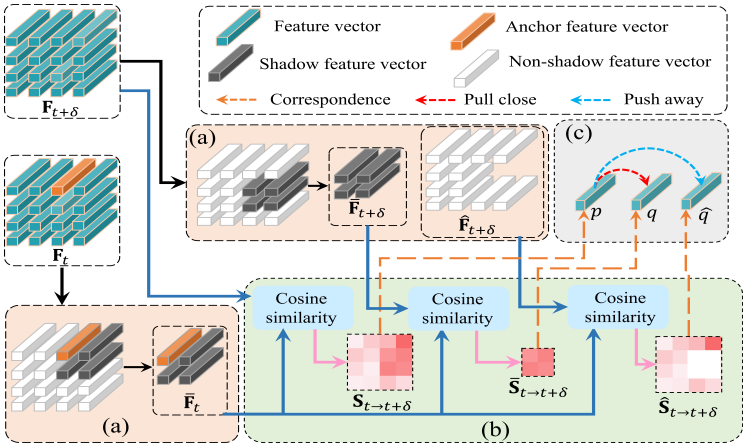

Fig. 4: Procedure of shadow-consistent correspondence from frame t to $t+\delta$. The proposed method consists of three modules: (a) the shadow guidance module (Sec. 3.1 (a)), (b) the cross-frame correspondence module (Sec. 3.1 (b)), and (c) the consistency regularization (Sec. 3.1 (c)).

we can obtain the set of shadow indexes $\mathcal{O}$:

$$\mathcal{O} = \{(h,w) \mid \mathbf{Y}_t(h,w) = 1\} . \tag{1}$$

$$\overline{\mathbf{F}}_t = \{\mathbf{F}_t(h,w) \mid (h,w) \in \mathcal{O}\}, \tag{2}$$

where $\overline{\mathbf{F}}_t \in \mathbb{R}^{N_t \times D}$ and N_t indicates the number of shadow feature vectors on $\mathbf{F}_t$, *i.e.*, $|\mathcal{O}| = N_t$. We define the operation of the shadow guidance module as $\overline{\mathbf{F}}_t = \mathrm{SG}(\mathbf{F}_t)$.

(b) Cross-frame correspondence module. For each shadow feature vector of a frame, cross-frame correspondence module aims to find its correspondence feature vector, *i.e.*, most relevant one, from another frame. Formally, we define the features of two frames from a video as $\mathbf{F}_t$ and $\mathbf{F}_{t+\delta}$, where $\delta > 0$ is the time interval. In the following, we will illustrate how to find the correspondence from $\mathbf{F}_t$ to $\mathbf{F}_{t+\delta}$, as well as in the other way around, *i.e.*, from $\mathbf{F}_{t+\delta}$ to $\mathbf{F}_t$.

To find the most correlated feature vector from $\mathbf{F}_{t+\delta}$, we first obtain the shadow feature map of $\mathbf{F}_t$ by $\overline{\mathbf{F}}_t = \mathrm{SG}(\mathbf{F}_t)$. Then, we measure the similarity between $\overline{\mathbf{F}}_t$ and $\mathbf{F}_{t+\delta}$ by the cosine similarity:

$$\mathbf{S}_{t\to t+\delta} = \frac{\overline{\mathbf{F}}_t \cdot \mathbf{F}_{t+\delta}}{\|\overline{\mathbf{F}}_t\| \|\mathbf{F}_{t+\delta}\|} , \tag{3}$$

where $\mathbf{S}_{t\to t+\delta} \in \mathbb{R}^{N_t \times L}$ and $L = H \times W$. Here, we take the n-th ($n \in N_t$) feature vector on $\overline{\mathbf{F}}_t$ as the anchor vector, and compute its most relevant feature vector on $\mathbf{F}_{t+\delta}$ based on the similarity map $\mathbf{S}_{t\to t+\delta}$:

$$p = \max_m \mathbf{S}_{t\to t+\delta}(n,m), m \in [1, L] , \tag{4}$$

where p is the index of the corresponding location on $\mathbf{F}_{t+\delta}$. Note that we perform the same operation to find the corresponding p for each feature vector on $\overline{\mathbf{F}}_t$.

(c) Consistency regularization. Here, we perform the consistency regularization to enforce the found correspondence inside the ground-truth sets. Specifically, we pull the anchor feature vector close to the shadow ground-truth on the second frame and push it away from the non-shadow regions on the second frame. More specifically, besides measuring the similarity between $\overline{\mathbf{F}}_t$ and $\mathbf{F}_{t+\delta}$, we additionally compute the similarity between $\overline{\mathbf{F}}_t$ and $\overline{\mathbf{F}}_{t+\delta}$ through Eq. 3, which can be defined as $\overline{\mathbf{S}}_{t\to t+\delta} \in \mathbb{R}^{N_t \times N_{t+\delta}}$, where $N_{t+\delta}$ is the number of shadow feature vectors on $\mathbf{F}_{t+\delta}$. Then, for the anchor n, we find its correspondence on $\overline{\mathbf{F}}_{t+\delta}$ based on $\overline{\mathbf{S}}_{t\to t+\delta}$ in the same way as Eq. 4, and we denote the found corresponding location as q. Note that p is the corresponding location found from the whole feature map while q is the corresponding location only found from the shadow set (indicated by the ground-truth mask). To pull p and q together, we minimize the discrepancy in two feature similarities by Eq. 5.

$$\mathcal{L}_{\text{shadow}}^{t\to t+\delta} \;=\; \frac{1}{N_t}\sum_{n=1}^{N_t}\left(\mathbf{S}_{t\to t+\delta}(n,p) - \overline{\mathbf{S}}_{t\to t+\delta}(n,q)\right)^2 \;. \tag{5}$$

To push the anchor p away from the non-shadow regions, we first compute the set of non-shadow indexes $\hat{\mathcal{O}}$:

$$\hat{\mathcal{O}} \;=\; \{(h,w) \mid \mathbf{Y}_{t+\delta}(h,w) = 0\} \;. \tag{6}$$

Based on $\hat{\mathcal{O}}$, we obtain the non-shadow feature $\hat{\mathbf{F}}_{t+\delta}$. Then, we compute the similarity between $\overline{\mathbf{F}}_t$ and $\hat{\mathbf{F}}_{t+\delta}$ in the same way as Eq. 3 to obtain $\hat{\mathbf{S}}_{t\to t+\delta} \in \mathbb{R}^{N_t \times M_{t+\delta}}$, where $M_{t+\delta}$ is the number of non-shadow features on $\mathbf{F}_{t+\delta}$, *i.e.*, $|\hat{\mathcal{O}}| = M_{t+\delta}$. For the anchor n, we find its correspondence on $\hat{\mathbf{F}}_{t+\delta}$ based on $\hat{\mathbf{S}}_{t\to t+\delta}$ in the same way as Eq. 4, which is denoted as $\hat{q}$. To push away p and $\hat{q}$, we maximize the margin in two feature similarities by the following loss function:

$$\begin{aligned}&\mathcal{L}_{\text{n-shadow}}^{t\to t+\delta} = \\ &\frac{1}{N_t}\sum_{n=1}^{N_t}\max\left(0, \beta - |\mathbf{S}_{t\to t+\delta}(n,p) - \hat{\mathbf{S}}_{t\to t+\delta}(n,\hat{q})|\right) \;,\end{aligned} \tag{7}$$

where β controls the margin between $\mathbf{S}_{t\to t+\delta}(n,p)$ and $\hat{\mathbf{S}}_{t\to t+\delta}(n,\hat{q})$. In the same way, we can obtain the consistency regularization in the other way around, *i.e.*, from $\mathbf{F}_{t+\delta}$ to $\mathbf{F}_t$, and we define the loss functions as $\mathcal{L}_{\text{shadow}}^{t+\delta\to t}$ and $\mathcal{L}_{\text{n-shadow}}^{t+\delta\to t}$. Finally, shadow-consistent correspondence learning can be formulated as follows:

$$\mathcal{L}_{\text{sc}} \;=\; \mathcal{L}_{\text{shadow}}^{t\to t+\delta} + \mathcal{L}_{\text{n-shadow}}^{t\to t+\delta} + \mathcal{L}_{\text{shadow}}^{t+\delta\to t} + \mathcal{L}_{\text{n-shadow}}^{t+\delta\to t} \;. \tag{8}$$

3.2 Brightness-invariant Correspondence

Due to the changes of brightness in a video, the shadow and non-shadow regions in different frames may present similar appearance. To learn brightness-invariant correspondence, we randomly shift the brightness of one frame and

learn the brightness-invariant shadow consistency between the shifted frame and another frame in the same video [57]. Formally, for two frames in a video, *i.e.*, V_t and $V_{t+\delta}$, we randomly shift the intensity of $V_{t+\delta}$ to produce the shifted frame $V'_{t+\delta} = V_{t+\delta} + \gamma$, where $\gamma \in [-\Delta, \Delta]$ is a randomly generated shift parameter and Δ is a hyper-parameter to control the shift range. Next, we use the shadow-consistent learning to learn the cross-frame correspondence between V_t and $V'_{t+\delta}$, as introduced in Sec. 3.1.

3.3 Overall Objective

The overall objective of our framework is defined as:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{sc}} \;, \tag{9}$$

where λ is a hyper-parameter to control the trade-off between these two losses and $\mathcal{L}_{\text{seg}}$ is the segmentation loss to supervise the pixel-wise prediction. The segmentation loss is different for different shadow detection works. For example, [55, 57, 7] adopt the binary cross entropy (BCE) loss as the segmentation loss:

$$\mathcal{L}_{\text{seg}} \;=\; -\frac{1}{T}\sum_{t=1}^{T} \mathbf{Y}_t \cdot \log\left(\hat{\mathbf{Y}}_t\right) + (1-\mathbf{Y}_t) \cdot \log\left(1-\hat{\mathbf{Y}}_t\right), \tag{10}$$

where $\mathbf{Y}_t$ and $\hat{\mathbf{Y}}_t$ are ground-truth and predicted shadow masks, respectively. TVSD-Net [7] uses BCE combined with a lovász-hinge loss [2]. In the experiments, to highlight the effectiveness of the proposed shadow-consistent correspondence, the segmentation loss keeps consistent with existing papers [55, 7].

4 Experimental Results

4.1 Evaluation Metrics and Datasets.

Temporal stability. Compared with the previous works that only evaluates the performance on each single image (frame-level), in this paper, we introduce a new evaluation metric to evaluate the temporal stability across the video frames, motivated by [25, 44]. In detail, different from [25, 44] that compute the optical flow between RGB frames, we calculate the optical flow between the ground-truth labels of two adjacent frames, *i.e.*, $\mathbf{Y}_t$ and $\mathbf{Y}_{t+1}$ through ARFlow [30], since the motions of shadows are hard to be captured on the RGB frames. For instance, the optical flows generated by RGB are focus on objects, which can not capture shadows since the motions of shadows are hard to be captured on the RGB frames; see supplementary materials for more details. Then, assume $I_{t\to t+1}$ as the optical flow between $\mathbf{Y}_t$ and $\mathbf{Y}_{t+1}$, and we define the reconstructed result that warps $\hat{\mathbf{Y}}_{t+1}$ by the optical flow $I_{t\to t+1}$ as $\mathbb{Y}_t$. Next, we measure the temporal stability of VSD for a video based on the flow warping IoU between the adjacent frames as:

$$\text{TS} \;=\; \frac{1}{T-1}\sum_{t=1}^{T-1} IoU(\hat{\mathbf{Y}}_t, \mathbb{Y}_t) \;. \tag{11}$$

Table 1: **Comparison with the state-of-the-art methods.** "↓" indicates the lower the scores, the better the results, while "↑" indicates the higher the scores, the better the results. "AVG" is the average score of IoU and TS, which presents the frame-level and temporal-level IoUs. "ISD" and "VSD" stand for the single-image shadow detection and video shadow detection, respectively. "SOD" stands for salient object detection. "VOS" stands for video object segmentation. "S-BER" and "N-BER" stand for BER of shadow regions and non-shadow regions, respectively.

		Frame-level						*Temporal-level*	
Task	Method	MAE ↓	F_β ↑	BER ↓	S-BER ↓	N-BER ↓	IoU [%] ↑	TS [%] ↑	AVG ↑
Scene Parsing	FPN [29]	0.044	0.707	19.49	36.59	2.40	51.28	74.27	62.78
	PSPNet [53]	0.052	0.642	19.75	36.44	3.07	47.65	76.63	62.14
SOD	DSS [14]	0.045	0.697	19.78	36.96	2.59	50.28	75.02	62.65
VOS	PDBM [41]	0.066	0.623	19.74	34.32	5.16	46.65	80.00	63.33
	FEELVOS [46]	0.043	0.710	19.76	37.27	2.26	51.20	74.89	63.05
	STM [34]	0.064	0.639	23.77	43.88	3.65	44.69	75.30	60.00
ISD	BDRAR [56]	0.050	0.695	21.30	40.28	2.31	48.39	72.63	60.51
	MTMT [8]	0.043	0.729	20.29	38.71	1.86	51.69	74.44	63.07
	FSD [17]	0.057	0.671	20.57	38.06	3.06	48.56	74.88	61.72
	DSD [55]	0.044	0.702	19.89	37.89	1.88	51.89	74.68	63.29
	DSD + ours	0.039	0.730	15.15	27.78	2.52	58.40	78.03	68.22
	-	**+0.05**	**+0.028**	**+4.65**	**+10.11**	-0.64	**+6.51**	**+3.35**	**+4.93**
VSD	Hu *et al.* [15]	0.078	0.683	17.03	30.13	3.93	51.03	83.67	67.35
	TVSD [7]	0.033	0.757	17.70	33.97	1.45	56.57	78.25	67.41
	TVSD + ours	0.042	0.762	13.61	24.31	2.91	61.50	81.44	71.47
	-	-0.09	**+0.005**	**+4.09**	**+9.66**	-0.46	**+4.93**	**+3.19**	**+4.06**

Frame-level accuracy. Except using the proposed evaluation metric to measure temporal stability, we follow the previous works [7, 55, 57] and adopt four common evaluation metrics that have been widely used in image/video shadow detection to evaluate the detection accuracy in frame-level. Specifically, they are Mean Absolute Error (MAE) [7], F-measure (F_β) [7, 17], Intersection over Union (IoU) [7], and Balance Error Rate (BER) [18, 57].

Evaluation dataset. We conduct our experiments on the ViSha dataset [7] to evaluate the performance. ViSha consists of 11,685 image frames and 390s duration, which is adjusted to 30 fps for all video sequences. This dataset is split into 50 videos for training and 70 videos for testing.

4.2 Implementation Details

Since our framework is a plug-and-play module that can be used in any shadow detectors, we insert our framework into two state-of-the-art methods on single-image shadow detection and video shadow detection, *i.e.*, DSD [55] and TVSD [7], for evaluation. During the training process of our shadow-consistent correspondence module, λ in Eq. 9 and β in Eq. 7 are set to 10 and 0.5, respectively; please refer to Sec. 4.4 for the analysis of λ. Δ is used to control the range of the brightness shift and it is set to 0.3 following [57]. Note, shifting brightness of frames will change the distribution of the images, resulting in degrading the detection

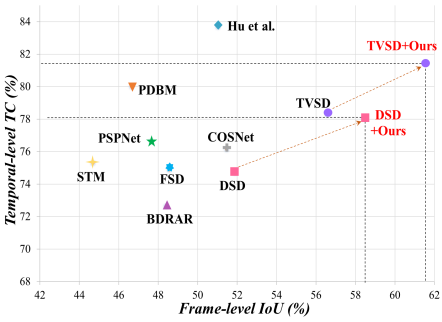

Fig. 5: **Trade-off between temporal and frame-level accuracy.**

Table 2: **Comparison of the supervised contrastive learning and ours**. Baseline is DSD [55]. "SCon" indicates the supervised contrastive learning. "Ratio" indicates the ratio of the found correspondence in the shadow regions. "AVG" refers to the average score of IoU and TS.

Method	Ratio ↑	AVG ↑
Baseline	36.71	63.29
Baseline + [50]	82.33	66.17
Baseline + SCon	85.30	65.13
Baseline + Ours	85.12	68.22

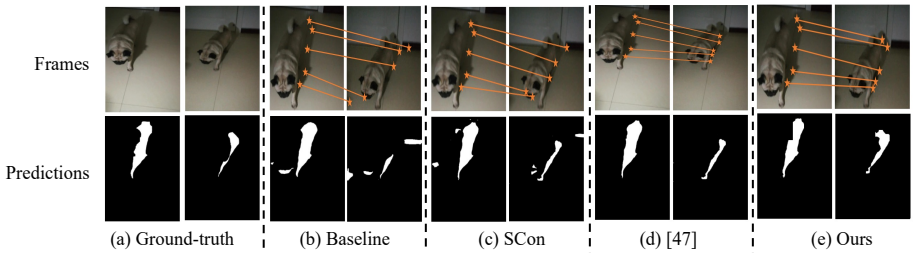

Fig. 6: Visualization of the correspondence found by different models. Note that Baseline is DSD [55]. We sample five pixels in one frame and find their correspondence in the other one.

performance [57]. Hence, we only shift the brightness of the frames after 2,000 training iterations and freeze the batch normalization [38]; see supplementary materials for more details.

4.3 Comparison with the State-of-the-art Methods

We conduct the experiments on ViSha [7] to compare with the state-of-the-art methods designed for scene parsing, salient object detection, video object segmentation, single-image shadow detection, and video shadow detection; please see the compared methods in Table 1. We obtain the results of these methods by retraining them on the ViSha dataset for video shadow detection with the recommendation training parameters or by downloading their results directly from Internet.

Table 1 provides the comparison results, which clearly shows that our proposed method can largely improve the performance of both single-image and video shadow detection approaches, *i.e.*, DSD [55] and TVSD [7], in terms of both frame-level and temporal-level accuracy. DSD is designed for single-image

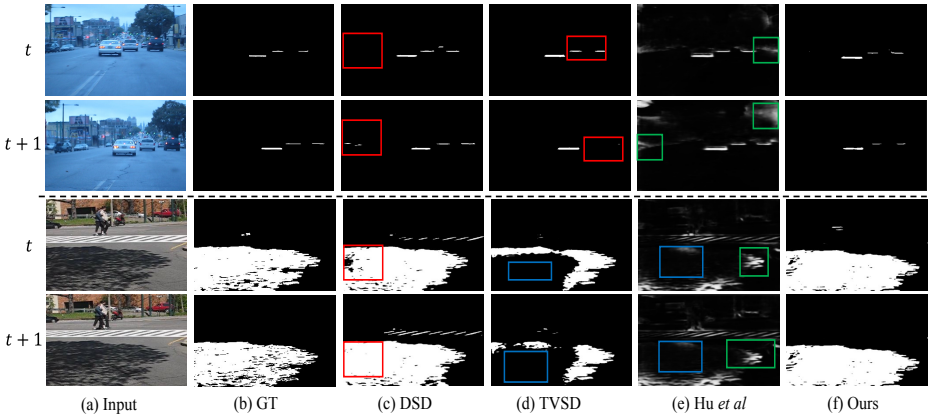

Fig. 7: Visual comparison of video shadow detection results produced by different methods. (a) is the input images and (b) is the ground-truth (GT) images. (c)-(f) are the results predicted by DSD [55], TVSD [7], Hu *et al.* [15], and our method, respectively. Our method takes the DSD as the basic network. Note that red boxes indicate the inconsistent predictions across video frames, blue boxes indicate the inaccurate static predictions, and green boxes show the blurry predictions.

shadow detection and our approach improves the performance a lot by further considering the dense correspondence among different video frames. Although TVSD is designed for video shadow detection and has explored the temporal consistency in videos, our method further explores the shadow-region correspondence and learns the brightness-invariant features. Besides, our method achieves the best trade-off on both temporal-level and frame-level accuracy, as shown in Fig. 5.

Fig. 7 illustrates the visual comparison of the shadow masks produced by DSD [55], TVSD [7], Hu *et al.* [15], and ours. From the results, we can see that our method provides more accurate and consistent shadow detection results across different videos frames than others. More examples and failure cases please refer to supplementary materials.

Comparison of the supervised contrastive learning, unsupervised correspondence learning and our SC-Cor. Specifically, for the sampled pixel of one frame, we find the most correlated pixel in the other frame and denote the found pixel as its correspondence. It is clear that the correspondences found by the baseline would be in dark non-shadow regions. Those found by the supervised contrastive learning would only focus on the shadow regions with similar textures. For unsupervised correspondence learning, we select [50], a recently SOTA, for comparison. It is clear that ours outperforms [47]. In order to further evaluate the effect of our method, we report the ratio of the found correspondence in ground truth masks and the average performance in Table 2. The results show that our method can largely improve the accuracy of found correspondence, *e.g.*,

Table 3: **Ablation on the effectiveness of SC-Cor and BS**. "SC-Cor" indicates the shadow-consistent correspondence and "BS" indicates the brightness shift operation.

Method	SC	BS	*Frame-level* MAE ↓	BER ↓	IoU [%] ↑	*Temporal-level* TS [%] ↑	AVG ↑
Baseline	✗	✗	0.044	19.89	51.89	74.68	63.29
Ours (SC-Cor)	✓	✗	0.040	15.67	56.89	77.09	67.00
Ours (BS)	✗	✓	0.043	16.82	53.65	74.92	64.29
Ours (Full)	✓	✓	**0.039**	**14.89**	**58.40**	**78.03**	**68.22**

Table 4: **Ablation Study on Shadow-consistent correspondence**.

(a) **Bidirectional correspondence**. "Bi-D" refers to the bidirectional correspondence, which has been defined by $t \to t+\delta$ and $t+\delta \to t$ in Eq. 8.

Bi-D	*Frame-level* BER ↓	IoU [%] ↑	*Temporal-level* TS [%] ↑	AVG ↑
✗	15.54	57.25	77.08	67.17
✓	**14.89**	**58.40**	**78.03**	**68.22**

(b) **Consistency regularization**. "Shadow" and "N-Shadow" indicate the regularization method described in Eq. 5 and Eq. 7.

	Frame-level BER ↓	IoU [%] ↑	*Temporal-level* TS [%] ↑	AVG ↑
Shadow	15.67	57.18	76.89	67.04
N-shadow	15.79	57.09	76.72	66.91
Full	**14.89**	**58.40**	**78.03**	**68.22**

over 43.41% on DSD [55]. Although the ratio of correspondences found by supervised contrastive learning in ground-truth is high, the generated shadow mask may be incomplete; see Fig. 6 (c).

4.4 Ablation Study

We conduct ablation experiments to show how each module in our framework design contributes to video shadow detection. We regard DSD [55] as our baseline module in this section. All the detection results are reported on the testing set of the ViSha dataset [7].

Effectiveness of SC-Cor and BS. Table 3 reports the effectiveness of the shadow-consistent correspondence (SC-Cor) and the brightness shift (BS) operation. Training with SC-Cor, we can see a clear improvement in terms of both frame-level accuracy and temporal accuracy, *i.e.*, 4.22 on BER and 2.40% on TS. It is worth noting that only adopting with BS cannot obtain the clear improvement on the temporal stability, *i.e.*, 74.92% vs. 74.68%, due to the lack of exploring temporal information. By combining with both SC-Cor and BS, the model achieves the best performance.

Bidirectional correspondence and consistency regularization. Table 4a reports the results of bidirectional correspondence in Eq. 8 and shows the effectiveness of the designed bidirectional correspondence. Furthermore, we perform

Table 5: **Ablation Study on different frame setting**.

(a) **Multiple frames**. "Frame number" denotes the number of sampled frames.

Frame number	*Frame-level* BER ↓	IoU [%] ↑	*Temporal-level* TS [%] ↑	AVG ↑
2	15.91	58.06	77.96	68.01
3	15.86	58.28	78.46	68.37
4	**15.82**	**58.26**	**78.51**	**68.39**

(b) **Frame interval** δ.

δ	*Frame-level* BER ↓	IoU [%] ↑	*Temporal-level* TS [%] ↑	AVG ↑
1	**14.51**	**58.62**	76.94	67.78
3	14.73	58.55	77.23	67.89
5	14.89	58.40	78.03	**68.22**
7	15.64	57.12	**78.23**	67.68

the ablation study on the shadow and non-shadow consistency regularization in Table 4b, showing that the the combination of them achieves the best results.

Multiple frames and Frame interval. We integrate our SC-Cor with multiple pairs of frames in a video and analyze the effectiveness in Table 5a. We observe that training with more frames brings a slight improvement on both frame-level and temporal-level accuracy. Considering the training efficiency, we choose two pairs of frames. Furthermore, we study the frame sampling strategy and report the detection results in Table 5b. It is clear that the longer time interval achieves the higher temporal stability while the short one performs better in frame-level accuracy. For instances, $\delta = 1$ achieves the best BER, *i.e.*, 14.51, and the lowest TS, *i.e.*, 76.94%. On the contrary, $\delta = 7$ obtains the best TS performance 78.23%. In this paper, we set δ as five to balance the temporal-level accuracy and the frame-level accuracy.

5 Conclusion

In this paper, we present a novel and plug-and-play shadow-consistent correspondence (SC-Cor) method for video shadow detection (VSD). A shadow-consistent correspondence is formulated to enforce the network to learn temporal-consistent shadows. A brightness shifting operation is employed to further regularize the network to be brightness-invariant. Considering current metrics only evaluate the frame-level accuracy, we introduce a new temporal stability metric, namely TS, for VSD. Experimental results on the benchmark dataset prove that our SC-Cor outperforms various shadow detection methods.

6 Acknowledgement

This work was supported by a research grant from HKUST-BICI Exploratory Fund (HCIC-004).

References

1. Benedek, C., Szirányi, T.: Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. IEEE Transactions on Image Processing **17**(4), 608–621 (2008)
2. Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4413–4421 (2018)
3. Bian, J., Lin, W.Y., Matsushita, Y., Yeung, S.K., Nguyen, T.D., Cheng, M.M.: Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4181–4190 (2017)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
7. Chen, Z., Wan, L., Zhu, L., Shen, J., Fu, H., Liu, W., Qin, J.: Triple-cooperative video shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2715–2724 (2021)
8. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5611–5620 (2020)
9. Ding, B., Long, C., Zhang, L., Xiao, C.: Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
10. Ding, X., Liu, Z., Li, X.: Free lunch for surgical video understanding by distilling self-supervisions. arXiv preprint arXiv:2205.09292 (2022)
11. Ding, X., Wang, N., Zhang, S., Cheng, D., Li, X., Huang, Z., Tang, M., Gao, X.: Support-set based cross-supervision for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11573–11582 (2021)
12. Ding, X., Wang, N., Zhang, S., Huang, Z., Li, X., Tang, M., Liu, T., Gao, X.: Exploring language hierarchy for video grounding. IEEE Transactions on Image Processing (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
14. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(4), 815–828 (2019). https://doi.org/10.1109/TPAMI.2018.2815688
15. Hu, S., Le, H., Samaras, D.: Temporal feature warping for video shadow detection. arXiv preprint arXiv:2107.14287 (2021)
16. Hu, X., Fu, C.W., Zhu, L., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence **42**(11), 2795–2808 (2020)

17. Hu, X., Wang, T., Fu, C.W., Jiang, Y., Wang, Q., Heng, P.A.: Revisiting shadow detection: A new benchmark dataset for complex world. IEEE Transactions on Image Processing **30**, 1925–1934 (2021). https://doi.org/10.1109/TIP.2021.3049331
18. Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7454–7462 (2018)
19. Jacques, J.C.S., Jung, C.R., Musse, S.R.: Background subtraction and shadow detection in grayscale video sequences. In: XVIII Brazilian symposium on computer graphics and image processing (SIBGRAPI'05). pp. 189–196. IEEE (2005)
20. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: Cotr: Correspondence transformer for matching across images. arXiv preprint arXiv:2103.14167 (2021)
21. Junejo, I.N., Foroosh, H.: Estimating geo-temporal location of stationary cameras using shadow trajectories. In: European conference on computer vision. pp. 318–331. Springer (2008)
22. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. ACM Transactions on Graphics (TOG) **30**(6), 1–12 (2011)
23. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1939–1946. IEEE (2014)
24. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
25. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
26. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Estimating natural illumination from a single outdoor image. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 183–190. IEEE (2009)
27. Le, H., Vicente, T.F.Y., Nguyen, V., Hoai, M., Samaras, D.: A+d net: Training a shadow detector with adversarial shadow attenuation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
28. Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via cf labels and distillation. IEEE Transactions on Image Processing **30**, 2016–2028 (2021)
29. Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.106, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.106
30. Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Luo, D., Wang, C., Li, J., Huang, F.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6489–6498 (2020)
31. Melekhov, I., Tiulpin, A., Sattler, T., Pollefeys, M., Rahtu, E., Kannala, J.: Dgc-net: Dense geometric correspondence network. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1034–1042. IEEE (2019)
32. Nadimi, S., Bhanu, B.: Physical models for moving shadow and object detection in video. IEEE transactions on pattern analysis and machine intelligence **26**(8), 1079–1087 (2004)

33. Nguyen, V., Vicente, T.F.Y., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4510–4518 (2017)
34. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
35. Okabe, T., Sato, I., Sato, Y.: Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1693–1700. IEEE (2009)
36. Panagopoulos, A., Samaras, D., Paragios, N.: Robust shadow and illumination estimation using a mixture model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 651–658. IEEE (2009)
37. Sanin, A., Sanderson, C., Lovell, B.C.: Shadow detection: A survey and comparative evaluation of recent methods. Pattern Recognition **45**(4), 1684–1695 (2012). https://doi.org/https://doi.org/10.1016/j.patcog.2011.10.001, https://www.sciencedirect.com/science/article/pii/S0031320311004043
38. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? In: Proceedings of the 32nd international conference on neural information processing systems. pp. 2488–2498 (2018)
39. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
40. Shen, L., Chua, T.W., Leman, K.: Shadow optimization from structured deep edge detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2067–2074 (2015). https://doi.org/10.1109/CVPR.2015.7298818
41. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
42. Truong, P., Danelljan, M., Timofte, R.: Glu-net: Global-local universal network for dense flow and correspondences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6258–6268 (2020)
43. Tyszkiewicz, M.J., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. arXiv preprint arXiv:2006.13566 (2020)
44. Varghese, S., Bayzidi, Y., Bar, A., Kapoor, N., Lahiri, S., Schneider, J.D., Schmidt, N.M., Schlicht, P., Huger, F., Fingscheidt, T.: Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 336–337 (2020)
45. Vicente, T.F.Y., Hou, L., Yu, C.P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: European Conference on Computer Vision. pp. 816–832. Springer (2016)
46. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.: FEELVOS: fast end-to-end embedding learning for video object segmentation. CoRR **abs/1902.09513** (2019), http://arxiv.org/abs/1902.09513
47. Wang, T., Hu, X., Fu, C.W., Heng, P.A.: Single-stage instance shadow detection with bidirectional relation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–11 (2021)
48. Wang, T., Hu, X., Wang, Q., Heng, P.A., Fu, C.W.: Instance shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1880–1889 (2020)

49. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
50. Xu, J., Wang, X.: Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. arXiv preprint arXiv:2103.17263 (2021)
51. Zhang, F., Torr, P., Ranftl, R., Richter, S.: Looking beyond single images for contrastive semantic segmentation learning. Advances in Neural Information Processing Systems **34** (2021)
52. Zhang, Q., Xiao, T., Efros, A.A., Pinto, L., Wang, X.: Learning cross-domain correspondence for control with dynamics cycle-consistency. arXiv preprint arXiv:2012.09811 (2020)
53. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
54. Zhao, X., Vemulapalli, R., Mansfield, P.A., Gong, B., Green, B., Shapira, L., Wu, Y.: Contrastive learning for label efficient semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10623–10633 (2021)
55. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5167–5176 (2019)
56. Zhu, L., Deng, Z., Hu, X., Fu, C.W., Xu, X., Qin, J., Heng, P.A.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 121–136 (2018)
57. Zhu, L., Xu, K., Ke, Z., Lau, R.W.: Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4702–4711 (2021)