IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 July 2023, accepted 7 August 2023, date of publication 15 August 2023, date of current version 22 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305576

RESEARCH ARTICLE

UnShadowNet: Illumination Critic Guided Contrastive Learning for Shadow Removal

SUBHRAJYOTI DASGUPTA[1,2], ARINDAM DAS[3,4], SENTHIL YOGAMANI[5], SUDIP DAS[3], CIARÁN EISING[4,6], (Senior Member, IEEE), ANDREI BURSUC[7], AND UJJWAL BHATTACHARYA[8], (Senior Member, IEEE)

[1]Mila—Quebec AI Institute, Montreal, QC H2S 3H1, Canada
[2]Département D'Informatique et de Recherche Opérationnelle, Université de Montréal, Montreal, QC H3T 1J4, Canada
[3]Department of Driving Software and Systems, Valeo India, Chennai 600130, India
[4]Department of Electronic and Computer Engineering, University of Limerick, Limerick, V94 T9PX Ireland
[5]Valeo Visions Systems, Galway, H54 Y276 Ireland
[6]Lero (the Science Foundation Ireland Research Centre for Software), Tierney Building, University of Limerick, Limerick, V94 NYD3 Ireland
[7]Valeo.ai, 75017 Paris, France
[8]Computer Vision and Pattern Recognition (CVPR) Unit, Indian Statistical Institute, Kolkata 700108, India

Corresponding author: Ciarán Eising (ciaran.eising@ul.ie)

ABSTRACT Shadows are frequently encountered natural phenomena that significantly hinder the performance of computer vision perception systems in practical settings, e.g., autonomous driving. A solution to this would be to eliminate shadow regions from the images before the processing of the perception system. Yet, training such a solution requires pairs of aligned shadowed and non-shadowed images which are difficult to obtain. We introduce a novel weakly supervised shadow removal framework *UnShadowNet* trained using contrastive learning. It is composed of a *DeShadower* network responsible for the removal of the extracted shadow under the guidance of an *Illumination* network which is trained adversarially by the illumination critic and a *Refinement* network to further remove artefacts. We show that *UnShadowNet* can be easily extended to a fully-supervised set-up to exploit the ground-truth when available. *UnShadowNet* outperforms existing state-of-the-art approaches on three publicly available shadow datasets (ISTD, adjusted ISTD, SRD) in both the weakly and fully supervised setups.

INDEX TERMS Shadow removal, weakly-supervised learning, contrastive learning.

I. INTRODUCTION

Shadows are a common phenomenon that exists in most natural scenes. It occurs due to inadequate illumination that makes part of the image darker than the other region of the same image. It causes a significant negative impact on the performance of various computer vision tasks such as object detection, semantic segmentation, and object tracking. Image editing [1] using shadow matting is one of the common ways to remove shadows. Shadow detection and correction can improve the efficiency of the machine learning model for a broad spectrum of vision-based problems such as image restoration [2], satellite image analysis [3], information recovery in urban high-resolution panchromatic satellite images [4], face recognition [5], and object detection [6]. In this work, we focus on natural images captured in a terrestrial setting, such as may be obtained by commercial devices and, particularly, automotive cameras.

Shadows are prevalent in almost all images in automotive scenes. The complex interaction of shadow segments with the objects of interest such as pedestrians, roads, lanes, vehicles, and riders makes the scene understanding challenging. Additionally, it does not have any distinct geometrical shape or size similar to soiling [7], [8]. Thus, they commonly lead to poor performance in road segmentation [9], [10], pedestrian pose estimation [11], [12], [13], segmentation [14], [15] and trajectory prediction [16]. Moving shadows can be incorrectly detected as a dynamic object in background subtraction [17], motion segmentation [18], depth estimation [19], [20] and SLAM algorithms [21], [22]. The difficulty of shadows is

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

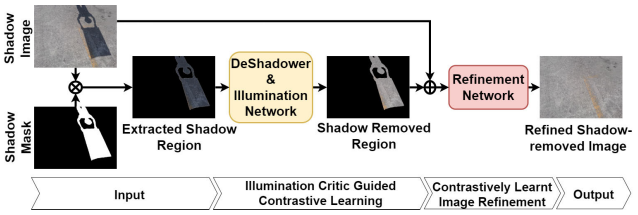


FIGURE 1. The proposed shadow removal framework. The shadow image and its shadow mask are subjected to pixel-wise product operation ⊗ to obtain the shadow extracted which is fed as input to the DeShadower ($\mathcal{D}$) and Illumination network ($\mathcal{I}$) simultaneously. $\mathcal{D}$ learns contrastively from $\mathcal{I}$ and the resultant shadow-removed region is embedded via ⊕ in the input image before feeding it to the Refinement network which produces the final Shadow-free image. The end-to-end network is trained in a weakly supervised manner.

further exacerbated in strong sun glare scenes where the dynamic range is very high across shadow and glare regions [23]. These issues lead to an incomplete or partial understanding of 360° surrounding region of the vehicle and bring major safety concerns for the passengers and Vulnerable Road Users (VRU) while performing automated driving [24]. Alternate sensor technologies like thermal camera [25], [26] are resistant to shadow issues and can be used to augment cameras.

In recent times, convolutional neural networks (CNNs) based approaches have significantly surpassed classical computer vision-based shadow removal techniques [27], [28], [29], [30], [31], [32]. The majority of the recent deep learning-based shadow removal approaches are fully-supervised in nature. However, such an end-to-end training setup requires *paired data*, namely shadow images and their shadow-free versions of the same images. These paired data are used to train CNNs [33], [34], [35]. Practically, the paired data is difficult to obtain particularly when the vehicle is moving fast. Some of the challenges include highly controlled lighting sources, object interactions, occlusions, and static scenes. Data acquisition through such a controlled setting suffers from diversity and often reports color inconsistencies [31] between shadow and shadow-free reference of the same image. Additionally, it is very difficult to capture any High Dynamic Range (HDR) natural scene without any presence of shadow for a shadow-free reference sample.

Some of the recent studies [36], [37], [38], [39], [40], [41], [42], [43] address the above-mentioned challenges and solve the shadow removal problem using *unpaired* data. They studied the physical properties of shadows such as illumination, color, and texture extensively. Motivated by these recent works, we propose an end-to-end trained weakly-supervised architecture for shadow removal as illustrated in Figure 1. In brief, we pass the shadow region of an input image to the DeShadower network that is aided by the Illumination network to contrastively learn to ''remove'' shadow from the region by exploiting the illumination properties. It is followed by the Refinement network that helps to remove any artifacts and maintain the overall spatial consistency with the input image and finally generates a shadow-free image.

Summary of Contributions and Distinctively Novel Features of This Work:

1) We develop a novel weakly-supervised training scheme namely *UnShadowNet* using contrastive learning to build a shadow remover in unconstrained settings where the network can be trained even without any shadow-free samples.
2) We propose a contrastive loss-guided DeShadower network to remove the shadow effects and a refinement network for efficient blending of the artifacts from shadow removed area.
3) We achieved state-of-the-art results on three public datasets namely ISTD, adjusted ISTD, and SRD in both constrained and unconstrained setups.
4) We perform extensive ablation studies with different proposed network components, diverse augmentation techniques, shadow inpainting, and tuning of several hyper-parameters.

II. RELATED WORK

Removing shadows from images has received a significant thrust due to the availability of large-scale datasets. In this section, first, we briefly discuss the classical computer vision methods reported in the literature. Then we discuss the more recent deep learning-based approaches. Finally, we summarize the details of contrastive learning and its applications since it is a key component in our framework.

A. CLASSICAL APPROACHES

1) ILLUMINATION-BASED SHADOW REMOVAL

Initial work [2], [27], [44], [45] on removing shadows were primarily motivated by the illumination and color properties of shadow region. In one of the earliest research, Barrow et al. [46] proposed an image-based algorithm that decomposes the image into a few predefined intrinsic parts based on shape, texture, illumination, and shading. Later Guo et al. [28] reported the simplified version of the same intrinsic parts by establishing a relation between the shadow pixels and the shadow-free region using a linear system. Likewise, Shor et al. [47] designed a model based on the illumination properties of shadows that makes a hard association between shadow and shadow-free pixels. In another study, Finlayson et al. [48] proposed a model that generates illumination invariant image for shadow detection followed by removal. The main idea of this work is that the pixels with similar chromaticity tend to have similar albedo. Further, histogram equalization-based models performed quite well for shadow removal, where the color of the shadow-free area was transferred to the shadowed area as reported by Vicente et al. [49], [50].

2) SHADOW MATTING

Porter and Duff [51] introduced a matting-based technique that became effective while handling shadows that are less distinct and fuzzy around the edges. The matting method

was only helpful to some extent, as computing shadow matte from a single image is difficult. To overcome this problem, Chuang et al, [1] applied matting for shadow editing and then transferred the shadow regions to the different scenes. Later shadow matte was computed from a sequence of video frames captured using a static camera. Shadow matte was adopted by Guo et al. [28] and Zhang et al. [30] in their framework for shadow removal.

B. DEEP LEARNING-BASED APPROACHES

1) SHADOW REMOVAL USING PAIRED DATA

Deep neural networks have been able to learn the properties of a shadow region efficiently when the network is trained in a fully supervised manner. Such setup requires paired data which means the shadow and shadow-free versions of the same image are fed as input to the network. Qu et al. [33] proposed an end-to-end learning framework called Deshadownet for removing shadows where they extract multi-scale contextual information from different layers. This information containing density and color offset of the shadows finally helped to predict the shadow matte. The method ST-CGAN, a two-stage approach proposed by Wang et al. [31], presents an end-to-end network that jointly learns to detect and remove shadows. This framework was designed based on conditional GAN [52]. In SP+M-Net [32], physics-based priors were used as inductive bias. The networks were trained to obtain the shadow parameters and matte information to remove shadows. However, these parameters and matte details were pre-computed using the paired samples, and the same were regressed in the network. Further, Hu et al. [35] designed a shadow detection and removal technique by analyzing the contextual information in image space in a direction-aware manner. These features were then aggregated and fed into an RNN model. In ARGAN [34], an attentive recurrent generative adversarial network was reported. The generator contained multiple steps where shadow regions were progressively detected. A negative residual-based encoder was employed to recover the shadow-free area and then a discriminator was set up to classify the final output as real or fake. In another recent framework, RIS-GAN [53] used adversarial learning shadow removal was performed using three distinct discriminators negative residual images. Subsequently, shadow-removed images and the inverse illumination maps were jointly validated.

2) SHADOW REMOVAL USING UNPAIRED DATA

Mask-ShadowGAN [36] is the first deep learning-based method that learns to remove shadows from unpaired training samples. Their approach was conceptualized on CycleGAN [54] where a mapping was learned from a source (shadow area) to a target (shadow-free area) domain. Le and Samaras [37] presented a learning strategy that crops the shadow area from an input image to learn the physical properties of shadow in an unpaired setting. CANet [38] handles the shadow removal problem in two stages. First, contextual information was extracted from the non-shadow area and then transferred the same to the shadow region in the feature space. Finally, an encoder-decoder setup was used to fine-tune the final results. LG-ShadowNet [39] explored the lightness and color properties of shadow images and put them through multiplicative connections in a deep neural network using unpaired data. Cun et al. [40] handled the issues of color inconsistency and artifacts at the boundaries of the shadow-removed area using a Dual Hierarchically Aggregation Network (DHAN) and a Shadow Matting Generative Adversarial Network (SMGAN). Weakly-supervised method G2R-ShadowNet [41] designed three sub-networks dedicated to shadow generation, shadow removal, and image refinement. Fu et al. [42] modeled the shadow removal problem from a different perspective, which is auto-exposure fusion. They proposed shadow-aware FusionNet and boundary-aware RefineNet to obtain the final shadow-removed image. Further in [43] a weakly-supervised approach was proposed that can be trained even without any shadow-free samples.

3) MISCELLANEOUS

In video sequences, cast shadows are often misinterpreted as moving objects. It was highlighted in [55] and considered as *insignificant* shadows. These cast shadows were removed in [56] by conditional random field. Liu et al. [57] investigated the cast shadows in detail by proposing a Gaussian Mixture Model at the pixel level in HSV color space followed by a pre-classifier and finally using Markov Random Fields for shadow removal. Patch-based illumination-invariant features such as binary patterns of local color constancy (BPLCC) and light-based gradient matching (LGM) were introduced in [58]. These features were used to create two dictionaries each for objects and shadows respectively. Each patch was assigned to an independent class in each iteration based on the distance from the reference dictionary. A feature fusion-based approach was followed in [59] where Spatio-Temporal Kernel Density Estimation (ST-KDE) based model was proposed for background modeling and Local Binary Pattern (LBP) features of this model were fused with the Gabor features probabilistically. Apart from shadow removal, shadow detection is also a well-studied area, some of the recent works include [6], [60], [61]. Inoue et al. [62] highlighted the problem of preparing a large-scale shadow dataset. They proposed a pipeline to synthetically generate shadow/shadow-free/matte image triplets.

C. CONTRASTIVE LEARNING

Learning the underlying representations by contrasting the positive and the negative pairs have been studied earlier in the community [63], [64]. This line of thought has inspired several works that attempt to learn visual representations without human supervision. While one family of works uses the concept of a memory bank to store the class representations [65], [66], [67], another set of works develops on the idea of maximization of mutual information [68], [69], [70]. Recently, Park et al. [71] presented an approach for

unsupervised image-to-image translation by maximizing the mutual information between the two domains using contrastive learning. In our work, we adopt the problem of shadow removal to solve it without using shadow-free ground truth samples with the help of contrastive learning.

III. PROPOSED METHOD

In this work, we define the problem of shadow removal as the translation of images from the shadow domain $\mathcal{S} \subset \mathbb{R}^{H\times W\times C}$ to shadow-free domain $\mathcal{F} \subset \mathbb{R}^{H\times W\times C}$ by utilizing only the shadow image and its mask and alleviating the use of its shadow-free counterpart. The proposed architecture *UnShadowNet* is illustrated in Figure 2. We briefly summarize the high-level characteristics here and discuss each part in more detail in the following subsections. In this section, we present the overall architecture of our proposed end-to-end shadow removal network, namely *UnShadowNet*. The architecture can be divided into *three* parts: DeShadower Network ($\mathcal{D}$), Illumination Network ($\mathcal{I}$) and Refinement Network ($\mathcal{R}$). These three networks are jointly trained in a weakly-supervised manner. Let us consider a shadow image $S \in \mathcal{S}$ and its corresponding shadow mask S_M. We obtain the shadow region S_s by cropping the masked area from S_M in the shadow image S. The DeShadower Network learns to remove the shadow from the region using a contrastive learning setup. It is aided by the Illumination Network which generates bright samples for $\mathcal{D}$ to learn from. The Refinement Network finally combines the shadow-free region S_f with the real image and refines it to form the shadow-free image $\hat{S}$.

A. DeShadower NETWORK ($\mathcal{D}$)

The DeShadower Network is designed as an encoder-decoder-based architecture that generates a shadow-removed region (S_r) from the shadow region (S_s). The shadow-removed regions generated by this network S_r should *associate* more with the bright samples and *dissociate* itself from the shadow samples. We employ a contrastive learning approach to help the DeShadower network achieve this and learn to generate shadow-free regions. In a contrastive learning framework, a *"query"* maximizes the mutual information with a *"positive"* sample in contrast to other samples that are referred to as *"negatives"*. In this work, we use a "noise contrastive estimation" framework [68] to maximize the mutual information between S_f and the bright sample B. We treat the bright samples generated by the Illumination Network as the "positive" and the shadow regions as the "negatives" in this contrastive learning setup. Thus, the objective function for maximizing (and minimizing) the mutual information can be formulated with the InfoNCELoss [68], a criterion derived from both statistics [68], [72] and metric learning [63], [64], [73]. Its formulation bears similarities with the cross-entropy loss:

$$\ell(x, x^+, x^-) = -log\left[\frac{exp(x\cdot x^+/\tau)}{exp(x\cdot x^+/\tau)+\sum_{i=1}^{N} exp(x\cdot x_i^-/\tau)}\right] \quad (1)$$

where x, x^+, x^- are the *query*, *positive* and *negatives* respectively. τ is the temperature parameter that controls the sharpness of the similarity distribution. We set it to the default value from prior work [65], [66]: τ=0.07.

The feature stack in the encoder of the DeShadower Network, represented as $\mathcal{D}_{enc}$, already contains latent information about the input shadow region S_s. From $\mathcal{D}_{enc}$, L layers are selected, and following practices from prior works [70], we pass these features through a projection head, an MLP (M_l) with two hidden layers. Subsequently, we obtain features:

$$s_l = M_l(\mathcal{D}^l_{enc}(S_s)); \quad l \in \{1, 2, \ldots, L\} \quad (2)$$

where $\mathcal{D}^l_{enc}$ is the l-th chosen layer in $\mathcal{D}^l_{enc}$. Similarly the output or the 'unshadowed' region S_f and the bright region B are encoded respectively as:

$$f_l = M_l(\mathcal{D}^l_{enc}(S_f)); \; b_l = M_l(\mathcal{D}^l_{enc}(B)) \quad (3)$$

We adjust the InfoNCE loss [68] into a layer-wise NCE loss:

$$\mathcal{L}_{NCE}(f_l, b_l, s_l) = \mathbb{E}_{S_f\sim\mathcal{F}, S_s\sim\mathcal{S}, B\sim\mathcal{B}} \quad \ell(f_l, b_l, s_l) \quad (4)$$

The generator should not change the contents of an image when there is no need to. In other words, given a shadow-free sample as input, it is expected to generate the same output without any change. To enforce such a regularization, we employ an identity loss [54], [74]. It is formulated using an $L1$ loss as:

$$\mathcal{L}_{iden} = \mathbb{E}_{S_f\sim\mathcal{F}} ||\mathcal{D}(S_f), S_f||_1 \quad (5)$$

Additionally, as described further in the following sections, the Illumination Critic $\mathcal{I}_\mathcal{C}$ is trained on real non-shadow samples and augmented bright samples. Therefore, we can additionally use the cues provided by the Illumination Critic to distil its knowledge of illumination to the DeShadower Network. This is achieved by computing the loss:

$$\mathcal{L}_{critic} = [1 - \mathcal{I}_\mathcal{C}(\mathcal{D}(S_s))]^2 \quad (6)$$

B. ILLUMINATION NETWORK ($\mathcal{I}$)

Shadow regions have a lower level of illumination compared to their surroundings. The exact illumination level can vary according to scene lighting conditions as illustrated in Fig. 3. To show that a real shadow image and an image with a region where brightness is reduced are similar even semantically, we designed a small experimental setup. We fine-tune a ResNet [75] with samples containing real shadows and no shadows for a Shadow/Non-shadow classification task and then test the images where we reduce the brightness in the shadow region. In the majority of the cases, the network classifies it to be a 'Shadow' image.

Using this heuristic, the Illumination Network ($\mathcal{I}$) is designed as a Generative Adversarial Network [76] to serve as a complementary augmentation setup to generate synthetic images where the illumination level is increased in a shadow region. The shadow region S_s is passed through the generator $\mathcal{I}_\mathcal{G}$ to produce brighter samples B of the shadow region.

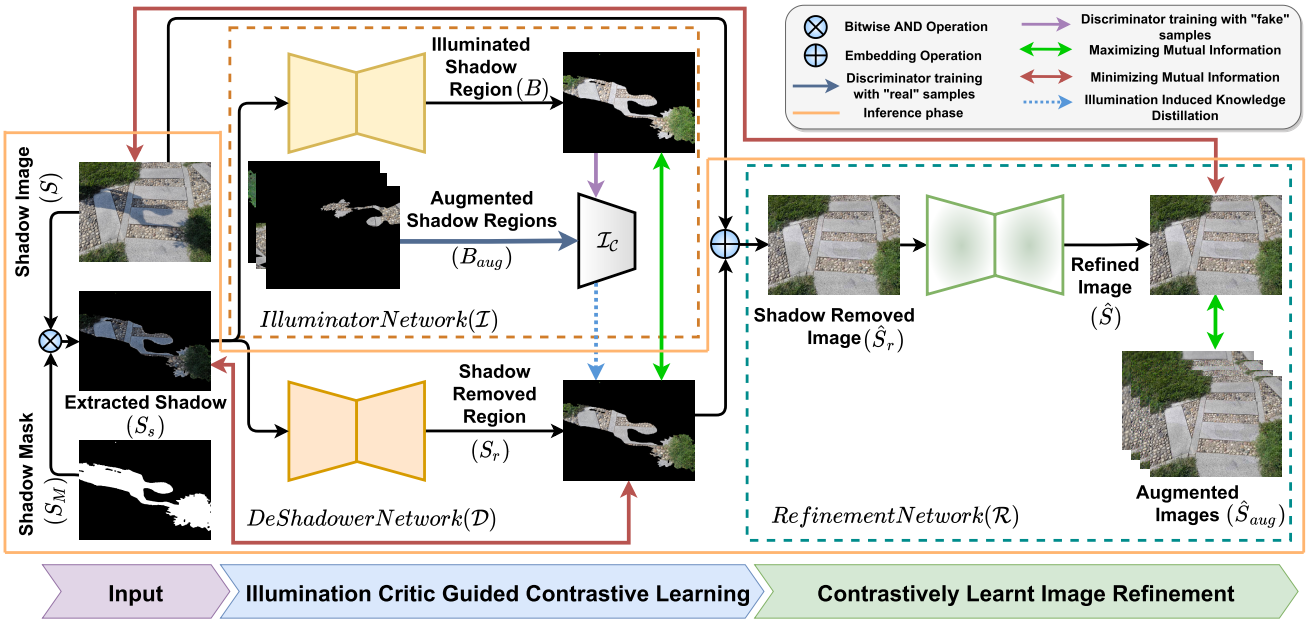


FIGURE 2. ***UnShadowNet*** **is the proposed end-to-end weakly-supervised shadow removal architecture. It has three main sub-networks: DeShadower Network ($\mathcal{D}$), Illumination Network ($\mathcal{I}$) and Refinement Network ($\mathcal{R}$). The pixelwise product operation $\otimes$ between shadow image ($\mathcal{S}$) and its shadow mask ($\mathcal{S_M}$) extracts the shadow region ($\mathcal{S_s}$), which is then fed to $\mathcal{D}$ and $\mathcal{I}$ simultaneously. The generator of the adversarially trained Illumination network generates an illuminated version ($\mathcal{B}$) of $\mathcal{S_s}$ which is subjected to validation by a discriminator, called Illumination Critic ($\mathcal{I_c}$) trained on augmented shadow-free regions ($\mathcal{B_{aug}}$). DeShadower is trained to produce shadow-removed region ($\mathcal{S_r}$) of $\mathcal{S_s}$. To create a more realistic illumination region $\mathcal{S_r}$, a contrastive approach is employed between $\mathcal{B}$ and $\mathcal{S_r}$. Finally, shadow-removed image ($\hat{\mathcal{S}}_r$) is obtained by applying embedding operation $\oplus$ to become input to the Refinement network. $\mathcal{R}$ is trained to efficiently blend the areas between shadow-removed and non-shadow regions so that it is robust to noise, blur, etc. Here contrastive learning approach was followed where** ***positive*** **samples ($\hat{\mathcal{S}}_{aug}$) were generated as per the method in [70].**

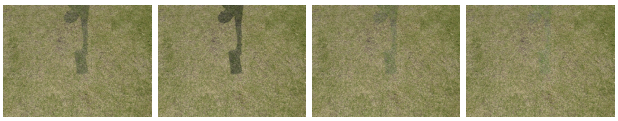


FIGURE 3. Illustration of different illumination control levels of shadow region.

The illumination critic ($\mathcal{I_C}$) learns to classify these samples generated by $\mathcal{I_G}$ as 'fake'. The motivation of this discriminator is detailed in the following section. The generator $\mathcal{I_G}$ and the discriminator $\mathcal{I_C}$ thus learns from the adversarial loss as:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{S_s\sim\mathcal{S}}\left[(1-\mathcal{I_C}(\mathcal{I_G}(S_s)))^2\right] \\ &+ \mathbb{E}_{S_s\sim\mathcal{S}}\left[\mathcal{I_C}(\mathcal{I_G}(S_s))^2\right] \\ &+ \mathbb{E}_{B'\sim\mathcal{B}}\left[(1-\mathcal{I_D}(B'))^2\right] \end{aligned} \tag{7}$$

We observe that the more optimal samples the Illumination Network generates, the better it aids $\mathcal{D}$ to create more realistic shadow-removed samples. Therefore, to improve $\mathcal{I}$ to create well-illuminated samples we employ the illumination loss as an $L1$ loss between the $\mathcal{I_G}$ generated bright sample B and the shadow-removed sample S_f as:

$$\mathcal{L}_{illum} = \frac{1}{N}\sum_{i=0}^{N}||S_f - B||_1 \tag{8}$$

The adversarial loss with the help of the discriminator and the illumination loss together play a role in generating well-illuminated samples, which in turn helps $\mathcal{D}$ to create better shadow-removed samples. In this regard, both $\mathcal{D}$ and $\mathcal{I}$ complement each other for the task. The Illumination Network supervises $\mathcal{D}$ to generate shadow-removed regions and likewise, $\mathcal{D}$ encourages $\mathcal{I}$ to create well-illuminated samples by learning from it. The choice of using $\mathcal{I}$ is experimentally justified in the ablation study section, as it helps to generate better results rather than relying solely on a pre-determined illumination level increase.

C. ILLUMINATION CRITIC ($\mathcal{I_C}$)

The role of the Illumination Critic ($\mathcal{I_C}$) is two-fold. Firstly, in the Illumination Network which generates well-illuminated variations of the shadow region S_s, the $\mathcal{I_C}$ is designed as a discriminator to the $\mathcal{I_G}$. The knowledge $\mathcal{I_C}$ learns from representations of shadow-free regions allows it to encourage $\mathcal{I_G}$ to create well-illuminated variations of the shadow region S_s which is later used as *positive* pair to contrastively train $\mathcal{D}$.

Additionally, the DeShadower Network utilizes the knowledge of the $\mathcal{I_C}$ to create realistic shadow-removed regions from the S_s. Having learned the representations of shadow-free regions and augmented samples with varying illumination, $\mathcal{I_C}$ can influence $\mathcal{D}$ to ''remove' shadows from

shadow regions using the $\mathcal{L}_{critic}$ in Eqn. 6. This two-fold characteristic of $\mathcal{I}_{\mathcal{C}}$ facilitates the complementary nature of $\mathcal{D}$ and $\mathcal{I}$ where they mutually improve each other.

To train $\mathcal{I}_{\mathcal{C}}$, we crop randomly masked non-shadow areas from S as well as other samples in the dataset similar to [41]. Additionally, $\mathcal{I}_{\mathcal{C}}$ is trained by augmented samples where each shadow region S_s is converted to 3 different samples by varying the illumination levels. The illumination levels are increased by a factor $\mu - 5, \mu, \mu + 5$ where μ is fixed empirically as presented in Table 3. It is trained using the same adversarial loss as the Illumination Network.

D. REFINEMENT NETWORK ($\mathcal{R}$)

After obtaining the shadow-removed region S_r, it is embedded with the original shadow image S. The embedding operation can be defined as:

$$\hat{S}_r = S - S * S_M + S_r * S_M \tag{9}$$

Following the embedding operation, there remain additional artefacts around the inpainted area due to improper blending. The Refinement Network $\mathcal{R}$ is designed to get rid of such artefacts by making use of the global context in the image. The absence of explicit ground truths in this setting motivated us to design a contrastive setup to train $\mathcal{R}$. To generate the *positive* samples, we follow [70] to augment the generated shadow-removed image ($\hat{S}_r$) by using random cropping of non-shadow regions. It is followed by additional transformations like resizing the cropped region back to the original size, random cutout, Gaussian blur, and Gaussian noise, represented as $\hat{S}_{aug}$. The objective is to maximize the information between the *query* image and the *positive* image pairs and reduce the same with the *negative* ones. In this phase, we reuse the existing encoder of $\mathcal{R}$ represented as $\mathcal{R}_{enc}$ as a feature extractor. We extract the layer-wise features of the *query* F_l, *positive* F_l^+ and *negative* F_l^- images and pass them through an MLP with two-hidden layers, similar to $\mathcal{D}$. Thus, we obtain the feature representations of F_l, F_l^+ and F_l^- respectively as follows:

$$\begin{aligned} F_l &= \hat{M}_l(\mathcal{R}^l_{enc}(\hat{S})); F_l^+ = \hat{M}_l(\mathcal{R}^l_{enc}(\hat{S}_{aug})); \\ F_l^- &= \hat{M}_l(\mathcal{R}^l_{enc}(S)) \end{aligned} \tag{10}$$

Therefore the objective function for the contrastive learning setup can be represented as:

$$\mathcal{L}_{NCE}(F_l, F_l^+, F_l^-) = \mathbb{E}_{\hat{S}\sim\mathcal{F}, S\sim\mathcal{S}} \quad \ell(F_l, F_l^+, F_l^-) \tag{11}$$

Additionally, we find that following [77] and [78], using a "layer-selective" perceptual loss along with the contrastive loss helps to preserve the integrity of the overall spatial details present in the input and output images. It is computed based on the features extracted by `relu_5_1` and `relu_5_3` of a VGG-16 [79] feature extractor as:

$$\mathcal{L}_{ref} = \frac{1}{2}\sum_{i=0}^{2} ||VGG_i(\hat{S}) - VGG_i(S)||_2^2 \tag{12}$$

E. SUPERVISED SETUP

Paired data is difficult to obtain for large-scale real-world datasets, however, it can be collected for a controlled smaller dataset. Here we demonstrate that *UnShadowNet* can be easily extended to exploit when paired shadow-free ground-truths (G) are available. Since the optimal level of illumination in the regions are available from G itself, we remove $\mathcal{I}$ in the fully-supervised setup and use different augmented versions of the G directly. Additionally, we make use of different losses that help to generate more realistic shadow-free images. To avoid loss of details in terms of content [80], we employ the pixel-wise *L1*-norm:

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=0}^{N} ||\hat{S}_i - G_i|| \tag{13}$$

Color plays an important role in preserving the realism of the generated image and maintaining consistency with the real image. To this end, we follow a recent study in the literature [81] to formulate the color loss as:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{P} \angle(\hat{S}_i, G_i) \tag{14}$$

where $\angle(,)$ computes an angle between two colors regarding the RGB color as a 3D vector [81], and P represents the number of pixel-pairs.

In addition, style plays an important role in an image that corresponds to the texture information [82]. We follow [83] to define a Gram matrix as the inner product between the vectorised feature maps i and j in layer l:

$$\gamma^l_{i,j} = \sum_k V^l_{i,k} \cdot V^l_{j,k} \tag{15}$$

The Gram matrix is the style for the feature set extracted by the l-th layer of VGG-16 net for an input image. Subsequently, the style loss can be defined as:

$$\mathcal{L}_s = \frac{1}{N_l}\sum_{i=0}^{N_l} ||\hat{S}_i - \gamma_i||^2 \tag{16}$$

where S_i and γ_i are the gram matrices for the generated shadow-free image and ground truth image respectively using VGG-16.

Therefore, the complete supervised loss can be formulated as a weighted sum ($\mathcal{L}_{sup}$) of the pixel ($\mathcal{L}_p$), color ($\mathcal{L}_c$) and style ($\mathcal{L}_s$) losses:

$$\mathcal{L}_{sup} = \lambda_1 \cdot \mathcal{L}_p + \lambda_2 \cdot \mathcal{L}_c + \lambda_3 \cdot \mathcal{L}_s \tag{17}$$

where λ_1, λ_2 and λ_3 are the weights corresponding to the pixel, color, and style losses respectively and are set empirically to 1.0, 1.0 and 1.0×10^4 following [81], [83] and [84] respectively in our experiments.

TABLE 1. Ablation study of the various components of UnShadowNet in both weakly-supervised and fully-supervised setup on ISTD [31] dataset using RMSE, PSNR and SSIM metrics.

Methods	Shadow Region			Non-Shadow Region			All		
	$RMSE\downarrow$	$PSNR\uparrow$	$SSIM\uparrow$	$RMSE\downarrow$	$PSNR\uparrow$	$SSIM\uparrow$	$RMSE\downarrow$	$PSNR\uparrow$	$SSIM\uparrow$
Weakly-Supervised Framework									
D-Net	11.8	32.46	0.978	4.6	34.85	0.972	6.1	30.26	0.947
D+I-Net	9.2	33.68	0.981	3.2	35.03	0.974	4.7	30.28	0.949
D+R-Net	9.9	33.43	0.980	3.4	35.47	0.974	5.0	30.17	0.950
D+I+R-Net	8.9	34.01	0.982	2.9	35.48	0.976	4.4	30.41	0.951
UnShadowNet	**8.3**	**34.47**	**0.983**	**2.9**	**35.51**	**0.977**	**3.8**	**30.63**	**0.951**
Fully-Supervised Framework									
D-Net	7.7	35.57	0.984	4.6	35.24	0.972	5.2	31.24	0.952
D+R-Net	6.3	36.13	0.989	2.8	36.03	0.978	3.8	31.76	0.958
UnShadowNet *Sup.*	**5.9**	**36.19**	**0.989**	**2.7**	**36.44**	**0.978**	**3.3**	**31.98**	**0.959**

TABLE 2. Ablation study of UnShadowNet using different training strategies on adjusted ISTD dataset.

Curriculum Learning	Shadow Inpainting	Data Augmentation	Shadow Region			Non-Shadow Region			All		
			RMSE ↓	PSNR ↑	SSIM ↑	RMSE ↓	PSNR ↑	SSIM ↑	RMSE ↓	PSNR ↑	SSIM ↑
Weakly-Supervised Framework											
✓	✗	✗	9.08	33.88	0.983	3.65	35.34	0.977	3.99	30.08	0.949
✓	✓	✗	8.65	34.12	0.984	3.11	35.47	0.977	3.87	30.41	0.951
✓	✗	✓	9.01	33.91	0.983	3.54	35.36	0.977	3.96	30.15	0.950
✗	✓	✗	8.97	34.01	0.983	3.42	35.40	0.977	3.96	30.23	0.950
✗	✗	✓	9.23	33.59	0.981	3.74	35.22	0.976	4.08	29.88	0.947
✗	✓	✓	8.89	34.03	0.983	3.38	35.43	0.977	3.94	30.32	0.950
✓	✓	✓	**8.31**	**34.47**	**0.984**	**2.92**	**35.51**	**0.977**	**3.80**	**30.63**	**0.951**
Fully-Supervised Framework											
✓	✗	✗	6.90	35.75	0.986	3.03	36.20	0.977	3.76	31.08	0.958
✓	✓	✗	6.23	36.12	0.989	2.82	36.38	0.978	3.58	31.17	0.959
✓	✗	✓	6.86	35.92	0.986	2.99	36.12	0.977	3.74	31.01	0.958
✗	✓	✗	6.81	36.07	0.987	3.01	36.21	0.977	3.67	31.27	0.959
✗	✗	✓	7.13	34.82	0.982	3.28	36.02	0.976	3.93	31.46	0.956
✗	✓	✓	6.47	36.09	0.987	2.96	36.26	0.978	3.63	31.85	0.959
✓	✓	✓	**5.92**	**36.20**	**0.989**	**2.71**	**36.44**	**0.978**	**3.33**	**31.98**	**0.959**

IV. EXPERIMENTATION DETAILS

A. DATASET AND EVALUATION METRICS

1) DATASETS

In this work, we train and evaluate our proposed method on three publicly available datasets discussed below.

2) ISTD

ISTD [31] contains image triplets: a shadow image, a shadow mask, and a shadow-free image captured at different lighting conditions that make the dataset significantly diverse. A total of 1, 870 image triplets were generated from 135 scenes for the training set, whereas the testing set contains 540 triplets obtained from 45 scenes.

3) ISTD+

The samples of ISTD [31] dataset were found to have color inconsistency issues between the shadow and shadow-free images as mentioned in the original work [31]. The reason was that shadow and shadow-free image pairs were collected at different times of the day which led to the effect of different lighting appearance in the images. This color irregularity issue was fixed by Le et al. [32] and an adjusted ISTD (ISTD+) dataset was published.

4) SRD

There are total 408 pairs of shadow and shadow-free images in SRD [33] dataset without the shadow-mask. For the training and evaluation of our both constrained and unconstrained setup, we use the shadow masks publicly provided by Cun et al. [40].

5) EVALUATION METRICS

For all the experiments conducted in this work, we use Root Mean-Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) respectively as metrics to evaluate and compare the proposed approach with other state-of-the-art methods. Following the prior-art [28], [31], [32], [33], [36], [37], [39], we compute the RMSE on the recovered shadow-free area, non-shadow area and the entire image in LAB color space. In addition to RMSE, we also compute PSNR and SSIM scores in RGB color space. RMSE is interpreted as better when it is lower, while PSNR and SSIM are better when they are higher.

B. IMPLEMENTATION DETAILS

The configuration of the generator is adopted from the DenseUNet architecture [84]. Unlike the conventional UNet architecture [85], it uses skip connections to facilitate better

FIGURE 4. Qualitative results of progressive addition of various components in UnShadowNet. DeShadower network (D-Net) alone is capable to remove shadow but it fails to match up the illumination level of the shadow-removed area with the shadow-free region. The DeShadower network accurately handles the illumination level when it is trained contrastively with the Illumination network (D+I Net). There remain some visible artifacts due to improper blending that is taken care of by the Refinement network.

information sharing among the symmetric layers. For the discriminator, we employ the architecture of the PatchGAN [52] discriminator that penalizes generated image structure at the scale of patches instead of at the image level. We develop and train all our models using the PyTorch framework. The proposals are trained using Momentum Optimizer with 1×10^{-4} as the base learning rate for the first 75 epochs, then we apply linear decay for the rest of the epochs. We train the whole model for a total of 200 epochs. Momentum was set to 0.9. All the models were trained on a system comprising one NVIDIA GeForce GTX 2080Ti GPU and the batch size was set to 1 for all experiments. In the testing phase, shadow-removed outputs are re-sized to 256×256 to compare with the ground truth images, as followed in [37] and [43]. We used the shadow detector by Ding et al. [34] to extract the shadow masks during the testing phase.

C. ABLATION STUDY

We considered the adjusted ISTD [31] dataset to perform our ablation studies due to its large volume and common usage in most of the recent shadow removal literature. We design an extensive range of experiments on this dataset in both *weakly-supervised* and *fully-supervised* settings to evaluate the efficacy of the proposed several network components of *UnShadowNet* and find out the best configuration of our model.

1) NETWORK COMPONENTS:

DeShadower network ($\mathcal{D}$) is the basic unit that acts as the overall shadow remover in the proposal. In the weakly-supervised setup, first, we experiment with only $\mathcal{D}$ for shadow removal (D-Net). We then add the Illumination network ($\mathcal{I}$) to include diverse illumination levels on the non-shadow regions in the image. We couple $\mathcal{I}$ with $\mathcal{D}$ in contrastive learning setup (D+I-Net). After shadow removal, the shadow-free region needs refinement for efficient blending with the non-shadow area. Hence we add a Refinement network ($\mathcal{R}$) with $\mathcal{D}$ where L1 loss guides to preserve the structural details (D+R-Net). Next, we consider illumination-guided contrastive learned refinement (D+R-Net) network where we add $\mathcal{I}$ and that becomes D+I+R-Net. Further improvement is achieved when we add contrastive loss in $\mathcal{R}$ which completes the UnShadowNet framework. In the fully-supervised setup, as described earlier, $\mathcal{I}$ is not used. As a result, we present the study of D-Net, D+R-Net, and UnShadowNet respectively.

Table 1 summarizes the ablation study of various proposed network components. Improvement in accuracy is observed due to the addition of $\mathcal{I}$ in contrastive learning setup. $\mathcal{R}$ adds further significant benefit when L1 loss is replaced with contrastive loss. The improvements of the proposed components are consistent in both self-supervised and fully-supervised learning as reported in the same table. All further experiments are performed based on the configuration marked as UnShadowNet.

2) CURRICULUM LEARNING

Curriculum Learning [86] is a type of learning strategy that allows one to feed easy examples to the neural network first and then gradually increase the complexity of the data. This helps to achieve stable convergence of the global optimum. As per Table 2, it is observed that the curriculum learning technique provides considerable improvement when applied along with shadow inpainting and data augmentation.

3) SHADOW INPAINTING

Appearance of shadows is a natural phenomenon and yet it is not an easy task to define the strong properties of shadow. This is because it does not have distinguishable shape, size, texture, etc. Hence it becomes important to augment the available shadow samples extensively so that they can be effectively learned by the network.

In this work, we estimate the mean intensity values of the existing shadow region of an image (I_P). Then we randomly select a shadow mask (S_M) from the existing set of shadow

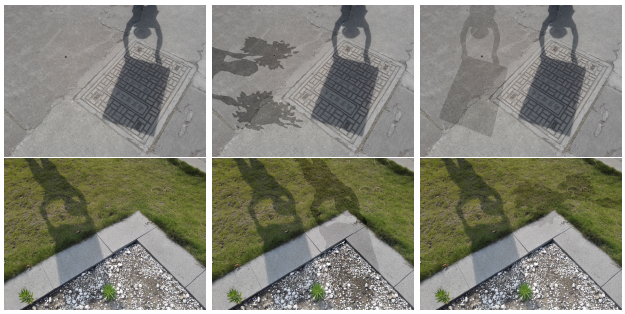


FIGURE 5. Augmented samples with inpainted shadow regions.

TABLE 3. Ablation study on illuminance factor (μ) of the proposed UnShadowNet in both weakly-supervised and fully-supervised setup on ISTD [31] dataset using RMSE, PSNR, and SSIM metrics.

	Shadow Region			All		
μ	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
Weakly-Supervised Framework						
5	24.2	27.57	0.959	19.1	24.76	0.936
25	17.1	30.13	0.974	12.4	27.30	0.942
50	10.3	33.71	0.982	6.2	30.01	0.951
75	**8.3**	**34.47**	**0.983**	**3.8**	**30.63**	**0.951**
100	9.8	31.56	0.980	5.6	28.52	0.946
Fully-Supervised Framework						
-15	13.2	31.57	0.979	9.7	29.26	0.949
0	5.9	36.19	0.989	3.3	31.98	0.959
15	6.8	35.53	0.988	4.1	31.07	0.959
30	11.6	33.87	0.983	6.3	30.56	0.952

samples. The mask (S_M) is inpainted on the shadow-free region of the image (I_P). The pixels that belong to the S_M in I_P will have brightness adjusted as the earlier computed mean. We do not apply the same mean every time, in order to generate diverse shadow regions, the estimated mean value is adjusted by $\pm 5\%$. The main motivations of this inpainting are two-fold: 1) It is difficult to learn complex shadows when it interacts with diverse light sources and other objects in the scene. The inpainted shadows are standalone and will provide an easier reference sample to another shadow segment in I_P. 2) It also increases the robustness of the network towards shadow removal by inpainting shadows with more diverse variations. Figure 5 shows the proposed shadow inpainting with random shadow masks and different shadow intensities. Table 2 indicates the significant benefits of inpainting complementing the standard data augmentation.

4) DATA AUGMENTATION

Data augmentation is an essential constituent to regularize any deep neural network-based model. We make use of some of the standard augmentation techniques such as image flipping with a probability of 0.3, random scaling of images in the range 0.8 to 1.2, adding Gaussian noise, blur effect, and enhancing contrast.

Table 2 sums up the role of curriculum learning, shadow inpainting, and data augmentation individually and the various combinations. This ablation study is performed on both weakly-supervised and fully-supervised setups indicating that both these training strategies are beneficial to learn shadow removal tasks.

5) ILLUMINANCE FACTOR (μ)

The DeShadower Network maximizes the information with "bright" synthetic augmentations generated by the Illumination Network. The effectiveness of the Illumination Network is verified from the results in Fig. 4. To train the Illumination Network, we sample shadow regions from the dataset and vary their brightness by $\mu - 5$, μ, $\mu + 5$. The different values experimented for the Illuminance factor (μ) are presented in Table 3. We find that setting the value of μ at 50 gives the most optimal results in shadow removal performance. For the fully-supervised setup, since the ground-truth images are available, the optimal level of brightness is obtained from those samples itself, consequently, $\mu = 0$ gives the best performance.

D. QUANTITATIVE STUDY

We evaluate our proposals and compare quantitatively with the state-of-the-art shadow removal techniques on ISTD [31], Adjusted ISTD [32], and SRD [33] benchmark datasets.

1) ISTD

Table 4 compares the proposed method with the state-of-the-art shadow removal approaches using RMSE, PSNR, and SSIM metrics for shadow, shadow-free, and all regions. We achieve state-of-the-art results and the improvement with respect to all metrics for shadow area in both training setups, namely weakly-supervised (UnshadowNet) and fully-supervised (UnshadowNet *Sup.*), are quite significant. There are a few other fully-supervised shadow removal methods evaluated on ISTD [31] dataset, which we compared with our proposed fully-supervised setup. In this setup as well, as per Table 5, our proposed method outperforms other state-of-the-art approaches.

2) ISTD+

Table 6 shows the performance of our proposed shadow remover on the adjusted ISTD [32] dataset using RMSE metric. The comparison of our method in a fully-supervised setup with other techniques trained in the same fashion demonstrates the robustness of our framework as it shows incremental improvement over the most recent state-of-the-art methods. In addition, we have performed experiments using a weakly-supervised setup where the metrics are comparable and only slightly behind the fully-supervised model.

3) SRD

We report and compare our shadow removal results in both the constrained and unconstrained setups with existing fully-supervised methods on SRD [33] using RMSE metric. Table 7 indicates that our proposal trained in a fully-supervised fashion obtains the lowest RMSE in all regions and outperforms the most recent state-of-the-art methods [42], [53].

TABLE 4. Quantitative comparison of two variants of *UnShadowNet* with other state-of-the-art shadow removal methods using RMSE, PSNR and SSIM metrics. Methods marked with '*' were evaluated on the adjusted ISTD [31] dataset. Scores of the other methods are computed on the ISTD dataset and obtained from their respective publications.

		Shadow Region			Non-Shadow Region			All		
Methods	TrainingData	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑
Yang *et al.* [87]	–	23.2	21.57	0.878	14.2	22.25	0.782	15.9	20.26	0.706
Gong and Cosker [88]	–	13.0	30.53	0.972	2.6	36.63	0.982	4.3	28.96	0.943
Guo *et al.* [28]	Non.Shd.+Shd (Paired)	20.1	26.89	0.960	3.1	35.48	0.975	6.1	25.51	0.924
ST-CGAN [31]	Non.Shd.+Shd (Paired)	12.0	31.70	0.979	7.9	26.39	0.956	8.6	24.75	0.927
SP+M-Net* [32]	Non.Shd.+Shd (Paired)	8.1	35.08	0.984	2.8	36.38	0.979	3.6	31.89	0.953
G2R-ShadowNet *Sup.** [41]	Non.Shd.+Shd (Paired)	7.9	36.12	0.988	2.9	35.21	0.977	3.6	31.93	0.957
UnShadowNet *Sup.**	**Non.Shd.+Shd (Paired)**	**5.9**	**36.19**	**0.989**	**2.7**	**36.44**	**0.978**	**3.3**	**31.98**	**0.959**
Mask-ShadowGAN* [36]	Shd.Free(Unpaired)	10.8	32.19	0.984	3.8	33.44	0.974	4.8	28.81	0.946
LG-ShadowNet* [39]	Shd.Free(Unpaired)	9.9	32.44	0.982	3.4	33.68	0.971	4.4	29.20	0.945
Le *et al.** [37]	Shd.Mask	10.4	33.09	0.983	2.9	35.26	0.977	4.0	30.12	0.950
G2R-ShadowNet* [41]	Shd.Mask	8.9	33.58	0.979	2.9	35.52	0.976	3.9	30.52	0.944
UnShadowNet	**Shd.Mask**	**8.3**	**34.47**	**0.984**	**2.9**	**35.51**	**0.977**	**3.8**	**30.63**	**0.951**

FIGURE 6. Qualitative comparison of our proposed method with other state-of-the-art shadow removal methods that use shadow mask and shadow image as input, on four challenging samples from ISTD [31] dataset.

E. QUALITATIVE STUDY

Figure 6 shows qualitative results of the proposed model trained in weakly-supervised format on a total of three challenging samples from the ISTD [31] dataset. We also visually compare with two existing and most recently published weakly-supervised shadow removal methods by Le et al. [37] and G2R-ShadowNet [41] respectively. It is clearly observed that UnShadowNet is accurate while removing shadows in complex backgrounds. In addition to the unconstrained setup, Figure 7 shows the results of our UnshadowNet *Sup.* model on ISTD [31] dataset. It is to be noted that the visual results are not shown on the adjusted ISTD [32] dataset because the test samples are the same as in the ISTD dataset, the only difference is in the color of the ground truth. In addition, we consider SRD [33] dataset and this is the first work where visual results are presented on the samples from the same dataset. Figure 8 and 9 demonstrate the results of UnShadowNet in weakly-supervised and fully-supervised setup.

F. RUNTIME ANALYSIS

We compare the runtime performance of our model with recent other contemporary architectures. For this purpose, the available code bases were used to estimate the run-time. During inference, LG-ShadowNet [39] takes 0.874 seconds, G2R-ShadowNet [41] takes 0.805 seconds and UnShadowNet takes 0.822 seconds.

G. EVALUATION OF GENERALIZATION IN AN UNCONSTRAINED AUTOMOTIVE DATASET

Automotive object detection and segmentation datasets do not provide shadow labels; thus, it is impossible to quantitatively evaluate these datasets extensively. We sampled a

FIGURE 7. Qualitative results on the ISTD [31] dataset using *fully-supervised* UnShadowNet setup.

FIGURE 8. Qualitative results on the SRD [33] dataset using *weakly-supervised* UnShadowNet setup.

FIGURE 9. Qualitative results on the SRD [33] dataset using *fully-supervised* UnShadowNet setup.

few shadow scenes from the challenging IDD dataset [94] which contains varied lighting condition scenes on Indian roads. It was impossible to train our model as shadow masks were unavailable. Thus we used this dataset to evaluate the robustness and generalization of our pre-trained model on novel scenes. The qualitative results are illustrated in Figure 10. Although the performance of the proposed shadow removal framework is either comparable to the state-of-the-art or superior, it is still not robust to be used in real-world autonomous driving systems. We feel that more extensive datasets have to be built for shadows to perform more detailed studies and we hope this work encourages the creation of these datasets or annotations of shadows in existing datasets.

TABLE 5. Comparative study of fully and weakly-supervised UnShadowNet with other fully supervised state-of-the-art shadow removal methods on ISTD [31] dataset using RMSE metric. The (*) marked method was trained using unpaired data.

Methods	$Shadow\downarrow$	$Non-Shadow\downarrow$	$All\downarrow$
Guo *et al.* [89]	18.95	7.46	9.30
Zhang *et al.* [30]	9.77	7.12	8.16
Iizuka *et al.* [90]	13.46	7.67	8.82
Wang *et al.* [91]	10.63	6.73	7.37
DeshadowNet [33]	12.76	7.19	7.83
MaskShadow-GAN* [36]	12.67	6.68	7.41
ST-CGAN [31]	10.31	6.92	7.46
Cun *et al.* [40]	11.4	7.2	7.9
AngularGAN [92]	9.78	7.67	8.16
RIS-GAN [53]	8.99	6.33	6.95
CANet [38]	8.86	6.07	6.15
Hu *et al.* [35]	7.6	3.2	3.9
Fu *et al.* [42]	7.77	5.56	5.92
UnShadowNet *Sup.*	**7.01**	**4.58**	**5.17**
UnShadowNet	**9.18**	**5.16**	**6.08**

TABLE 6. Comparative study of fully and weakly supervised UnShadowNet with other state-of-the-art shadow removal methods on adjusted ISTD [32] dataset using RMSE metric. The (*) marked method was trained using unpaired data.

Methods	$Shadow\downarrow$	$Non-Shadow\downarrow$	$All\downarrow$
Guo *et al.* [89]	22.0	3.1	6.1
Gong *et al.* [93]	13.3	–	–
ST-CGAN [31]	13.4	7.7	8.7
DeshadowNet [33]	15.9	6.0	7.6
MaskShadow-GAN* [36]	12.4	4.0	5.3
SP+M-Net [43]	7.9	3.1	3.9
Fu *et al.* [42]	6.5	3.8	4.2
SP+M+I-Net [43]	6.0	3.1	3.6
UnShadowNet *Sup.*	**5.9**	**2.7**	**3.3**
UnShadowNet	**8.3**	**2.9**	**3.8**

TABLE 7. Comparative study of fully and weakly-supervised UnShadowNet with other fully-supervised state-of-the-art shadow removal methods on SRD [33] dataset using RMSE metric. No other prior art was found to remove shadows in a weakly supervised fashion on the same dataset.

Methods	$Shadow\downarrow$	$Non-Shadow\downarrow$	$All\downarrow$
Guo *et al.* [89]	31.06	6.47	12.60
Zhang *et al.* [30]	9.50	6.90	7.24
Iizuka *et al.* [90]	19.56	8.17	16.33
Wang *et al.* [91]	17.33	7.79	12.58
DeshadowNet [33]	17.96	6.53	8.47
ST-CGAN [31]	18.64	6.37	8.23
Hu *et al.* [35]	11.31	6.72	7.83
AngularGAN [92]	17.63	7.83	15.97
Cun *et al.* [40]	8.94	4.80	5.67
Fu *et al.* [42]	8.56	5.75	6.51
RIS-GAN [53]	8.22	6.05	6.78
CANet [38]	7.82	5.88	5.98
UnShadowNet *Sup.*	**7.78**	**5.31**	**5.74**
UnShadowNet	**8.92**	**5.96**	**6.61**

V. LIMITATIONS AND FUTURE DIRECTIONS

As presented in our experiments, UnShadowNet outperforms the existing state-of-the-art in several standard shadow removal datasets. However, there are certain areas that can

FIGURE 10. Qualitative results (*bottom*) on a few input samples (*top*) from IDD dataset [94]. *UnShadowNet* trained on ISTD [31] dataset enables to remove shadow reasonably in automotive scenes.

be improved. Our model relies upon an external shadow detector [34] which may not always accurately predict the shadow regions. This may cause resultant areas where the shadow is not removed. In future work, we intend to build a single-stage architecture to incorporate both shadow detection and removal. Since shadows are physical phenomena, another interesting direction would be to exploit the inherent physical properties of illumination that result in shadows.

Moreover, in our research, we observed that the focus is mainly on datasets that have images of a narrow field-of-view and lacks complex situations that may arise in real-life automotive scenes. For future works, we think it will be important to develop a suitable dataset that comprises such challenging scenarios as in real-world automotive settings.

The proposed method is not optimized for run-time and we still obtained a reasonable inference time of 0.822 seconds. With optimization techniques like pruning and multi-task learning, real-time performance can potentially be achieved.

VI. CONCLUSION

In this work, we have developed a novel end-to-end framework consisting of a deep learning architecture for image shadow removal in unconstrained settings. The proposed model can be trained with full or weak supervision. We achieve state-of-the-art results in all the major shadow removal datasets. Although weak supervision has slightly lesser performance, it eliminates the need for shadowless ground truth which is difficult to obtain. To enable the weakly supervised training, we have introduced a novel illumination network which is composed of a generative model used to brighten the shadow region and a discriminator trained using shadow-free patches of the image. It acts as a guide (called illumination critic) for producing illuminated samples by the generator. DeShadower, another component of the proposed framework is trained in a contrastive way with the help of illuminated samples which are generated by the preceding part of the network. Finally, we propose a refinement network that is trained in a contrastive way and is used for fine-tuning the shadow-removed image obtained as an output of the DeShadower. We perform ablation studies to show that the three components of our proposed framework, namely the illuminator, Deshadower, and refinement network work effectively together. To evaluate the generalization capacity of the proposed approach, we tested a few novel samples of shadow-affected images from a generic automotive dataset and obtained promising results of shadow removal. Shadow removal continues to be a challenging problem in dynamic automotive scenes and we hope this work encourages further dataset creation and research in this area.

ACKNOWLEDGMENT

The authors would like to thank Valeo for encouraging advanced research, and also would like to thank Tuan-Hung Vu (valeo.ai, France), Saikat Roy (DKFZ, Germany), and Aniruddha Saha (University of Maryland, Baltimore County) for providing a detailed review prior to submission.

(Subhrajyoti Dasgupta, Arindam Das, and Senthil Yogamani are co-first authors.)

REFERENCES

[1] Y.-Y. Chuang, D. B. Goldman, B. Curless, D. H. Salesin, and R. Szeliski, "Shadow matting and compositing," in *Proc. ACM SIGGRAPH Papers*, Jul. 2003, pp. 1–5.

[2] M. S. Drew, G. D. Finlayson, and S. D. Hordley, "Recovery of chromaticity image free from shadows via illumination invariance," in *Proc. ICCV Workshop Color Photometric Methods Comput. Vis.*, 2003, pp. 1–8.

[3] P. M. Dare, "Shadow analysis in high-resolution satellite imagery of urban areas," *Photogrammetric Eng. Remote Sens.*, vol. 71, no. 2, pp. 169–177, Feb. 2005.

[4] N. Su, Y. Zhang, S. Tian, Y. Yan, and X. Miao, "Shadow detection and removal for occluded object information recovery in urban high-resolution panchromatic satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2568–2582, Jun. 2016.

[5] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, "Improving shadow suppression for illumination robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 611–624, Mar. 2019.

[6] H. Le, T. F. Y. Vicente, V. Nguyen, M. Hoai, and D. Samaras, "A+D net: Training a shadow detector with adversarial shadow attenuation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–17.

[7] A. Das, "SoildNet: Soiling degradation detection in autonomous driving," in *Proc. NeurIPS Mach. Learn. Auto. Driving Workshop*, 2019, pp. 1–9.

[8] A. Das, P. Křížek, G. Sistu, F. Bürger, S. Madasamy, M. Uricár, V. R. Kumar, and S. Yogamani, "TiledSoilingNet: Tile-level soiling detection on automotive surround-view cameras using coverage metric," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[9] A. Dahal, E. Golab, R. Garlapati, V. R. Kumar, and S. Yogamani, "RoadEdgeNet: Road edge detection system using surround view camera images," *Electron. Imag.*, vol. 33, no. 17, p. 210, Jan. 2021.

[10] S. Chennupati, G. Sistu, S. Yogamani, and S. Rawashdeh, "AuxNet: Auxiliary tasks enhanced semantic segmentation for automated driving," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 1–15.

[11] A. Das, S. Das, G. Sistu, J. Horgan, U. Bhattacharya, E. Jones, M. Glavin, and C. Eising, "Deep multi-task networks for occluded pedestrian pose estimation," 2022, *arXiv:2206.07510*.

[12] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "ClueNet: A deep framework for occluded pedestrian pose estimation," in *Proc. BMVC*, 2019, p. 245.

[13] S. Das, P. S. R. Kishore, and U. Bhattacharya, "An end-to-end framework for pose estimation of occluded pedestrians," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1446–1450.

[14] R. Hazem, E. Mohamed, V. R. K. Sistu, S. Ganesh, C. Eising, A. El-Sallab, and S. Yogamani, "FisheyeYOLO: Object detection on fisheye cameras for autonomous driving," in *Proc. NeurIPS Workshop Mach. Learn. Auto. Driving*, 2020, pp. 1–5.

[15] A. Das, S. Kandan, S. Yogamani, and P. Křížek, "Design of real-time semantic segmentation decoder for automated driving," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 1–15.

[16] S. Sharma, G. Sistu, L. Yahiaoui, A. Das, M. Halton, and C. Eising, "Navigating uncertainty: The role of short-term trajectory prediction in autonomous vehicle safety," 2023, *arXiv:2307.05288*.

[17] B. R. Kiran, A. Das, and S. Yogamani, "Rejection-cascade of Gaussians: Real-time adaptive background subtraction framework," in *Proc. Nat. Conf. Comput. Vis., Pattern Recognit., Image Process., Graph.*, 2019, pp. 1–10.

[18] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "MODNet: Motion and appearance based moving object detection network for autonomous driving," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2859–2864.

[19] V. R. Kumar, M. Klingner, S. Yogamani, S. Milz, T. Fingscheidt, and P. Mäder, "SynDistNet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 61–71.

[20] V. Ravi Kumar, S. Yogamani, S. Milz, and P. Mäder, "FisheyeDistanceNet++: Self-supervised fisheye distance estimation with self-attention, robust loss function and camera view generalization," *Electron. Imag.*, vol. 33, no. 17, p. 181, Jan. 2021.

[21] L. Gallagher, V. Ravi Kumar, S. Yogamani, and J. B. McDonald, "A hybrid sparse-dense monocular SLAM system for autonomous driving," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Aug. 2021, pp. 1–8.

[22] A. Dahal, V. Ravi Kumar, S. Yogamani, and C. Eising, "An online learning system for wireless charging alignment using surround-view fisheye cameras," 2021, *arXiv:2105.12763*.

[23] L. Yahiaoui, M. Uřičář, A. Das, and S. Yogamani, "Let the sunshine in: Sun glare detection on automotive surround-view cameras," *Electron. Imag.*, vol. 32, no. 16, p. 80, Jan. 2020.

[24] C. Eising, J. Horgan, and S. Yogamani, "Near-field perception for low-speed vehicle automation using surround-view fisheye cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 13976–13993, Sep. 2022.

[25] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15940–15950, Sep. 2022.

[26] A. Das, S. Das, G. Sistu, J. Horgan, U. Bhattacharya, E. Jones, M. Glavin, and C. Eising, "Revisiting modality imbalance in multimodal pedestrian detection," 2023, *arXiv:2302.12589*.

[27] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 59–68, Jan. 2006.

[28] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2956–2967, Dec. 2013.

[29] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 431–446, Mar. 2016.

[30] L. Zhang, Q. Zhang, and C. Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4623–4636, Nov. 2015.

[31] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1788–1797.

[32] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8577–8586.

[33] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau, "DeshadowNet: A multi-context embedding deep network for shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2308–2316.

[34] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10212–10221.

[35] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2795–2808, Nov. 2020.

[36] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-ShadowGAN: Learning to remove shadows from unpaired data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2472–2481.

[37] H. Le and D. Samaras, "From shadow segmentation to shadow removal," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. pp 264–281.

[38] Z. Chen, C. Long, L. Zhang, and C. Xiao, "CANet: A context-aware network for shadow removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4723–4732.

[39] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 30, pp. 1853–1865, 2021.

[40] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1–8.

[41] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4925–4934.

[42] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, "Auto-exposure fusion for single-image shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10566–10575.

[43] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9088–9101, Dec. 2022.

[44] G. D. Finlayson and M. S. Drew, "4-sensor camera calibration for image representation invariant to shading, shadows, lighting, and specularities," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 473–480.

[45] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2002, pp. 1–14.

[46] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, "Recovering intrinsic scene characteristics," *Comput. Vis. Syst.*, vol. 2, no. 2, pp. 3–26, Apr. 1978.

[47] Y. Shor and D. Lischinski, "The shadow meets the mask: Pyramid-based shadow removal," *Comput. Graph. Forum*, vol. 27, no. 2, pp. 577–586, Apr. 2008.

[48] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 35–57, Oct. 2009.

[49] T. F. Y. Vicente and D. Samaras, "Single image shadow removal via neighbor-based region relighting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 309–320.

[50] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 682–695, Mar. 2018.

[51] T. Porter and T. Duff, "Compositing digital images," in *Proc. 11th Annu. Conf. Comput. Graph. Interact. Techn.*, Jan. 1984, pp. 253–259.

[52] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[53] L. Zhang, C. Long, X. Zhang, and C. Xiao, "RIS-GAN: Explore residual and illumination with generative adversarial networks for shadow removal," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1–8.

[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[55] D. Xu, J. Liu, X. Li, Z. Liu, and X. Tang, "Insignificant shadow detection for video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 1058–1064, Aug. 2005.

[56] Y. Wang, "Real-time moving vehicle detection with cast shadow removal in video based on conditional random field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 437–441, Mar. 2009.

[57] Z. Liu, K. Huang, and T. Tan, "Cast shadow removal in a hierarchical manner using MRF," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 1, pp. 56–66, Jan. 2012.

[58] M. Russell, J. J. Zou, G. Fang, and W. Cai, "Feature-based image patch classification for moving shadow detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2652–2666, Sep. 2019.

[59] S. Sahoo and P. K. Nanda, "Adaptive feature fusion and spatio-temporal background modeling in KDE framework for object detection and shadow removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1103–1118, Mar. 2022.

[60] T. Wang, X. Hu, Q. Wang, P.-A. Heng, and C.-W. Fu, "Instance shadow detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1877–1886.

[61] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5610–5619.

[62] N. Inoue and T. Yamasaki, "Learning from synthetic shadows for shadow detection and removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4187–4197, Nov. 2021.

[63] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 1–8.

[64] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[65] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[66] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[67] M. Laskin, A. Srinivas, and P. Abbeel, "CURL: Contrastive unsupervised representations for reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5639–5650.

[68] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[69] P. Bachman, R. Devon Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, *arXiv:1906.00910*.

[70] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–11.

[71] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 1–29.

[72] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. AISTATS*, 2010, pp. 1–8.

[73] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 1–9.

[74] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[76] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 1–14.

[77] K. Xu, L. Wen, G. Li, H. Qi, L. Bo, and Q. Huang, "Learning self-supervised space-time CNN for fast video style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 2501–2512, 2021.

[78] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 1–18.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[80] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[81] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6842–6850.

[82] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 1–9.

[83] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[84] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8202–8211.

[85] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 1–8.

[86] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 1–8.

[87] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4361–4368, Oct. 2012.

[88] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014, pp. 1–11.

[89] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 2033–2040.

[90] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[91] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[92] O. Sidorov, "Conditional GANs for multi-illuminant color constancy: Revolution or yet another approach?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1748–1758.

[93] H. Gong and D. Cosker, "Interactive removal and ground truth for difficult shadow scenes," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 33, no. 9, p. 1798, 2016.

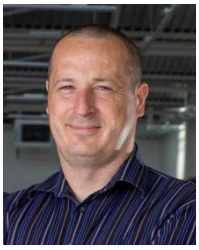
[94] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1743–1751.

SUBHRAJYOTI DASGUPTA received the bachelor's degree from the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, under the supervision of Prof. Ujjwal Bhattacharya. He is currently pursuing the dual master's degree with Mila and Université de Montréal. Later, he was a researcher. He was a Deep Learning Project Trainee with the Bhabha Atomic Research Center, Mumbai. His research interests include computer vision, such as multi-modal learning, scene understanding, generative models, and also its applications in inter-disciplinary domains, such as autonomous driving and climate research.

ARINDAM DAS is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, University of Limerick, Ireland. He is also an AI Software Architect with the Department of Driving Software and Systems (DSW), Valeo, India, where he was also an Expert in AI. He is responsible to design AI algorithms to support various features for autonomous driving systems. He has almost ten years of industry experience in computer vision, deep learning, and document analysis. He has authored 17 peer-reviewed publications and 48 patents. His current research interests include weakly-supervised learning, domain adaptation, image restoration, and multimodal learning.

SENTHIL YOGAMANI is currently an Artificial Intelligence Architect and holds a director-level technical leader position with Valeo, Ireland. He leads the research and design of AI algorithms for various modules of autonomous driving systems. He has over 16 years of experience in computer vision and machine learning, including 14 years of experience in industrial automotive systems. He is the author of more than 125 publications with 5200 citations and more than 100 filed patents. He serves on the editorial board of various leading IEEE automotive conferences, including ITSC and IV and the advisory board of various industry consortia, including Khronos, Cognitive Vehicles, and IS Auto. He was a recipient of the Best Associate Editor Award at ITSC 2015 and the Best Paper Award at ITST 2012.

SUDIP DAS received the B.Tech. degree in computer science and engineering from the West Bengal University of Technology (WBUT), India, in 2017. He is currently a Senior Research Engineer with the Department of Driving Software and Systems (DSW), Valeo, India. He was in a research position with the Indian Statistical Institute, Computer Vision and Pattern Recognition Unit, Kolkata. He is also passionate to work on the various problems of autonomous driving. His research interests include the relevant problems of unsupervised leaning, curriculum learning, transfer learning, domain adaptation, computer vision, and deep learning, with the goal of detecting, segmenting, and pose estimating of objects in images or videos.



CIARÁN EISING (Senior Member, IEEE) received the B.E. degree in electronic and computer engineering and the Ph.D. degree from the National University of Ireland, Galway, in 2003 and 2010, respectively. From 2009 to 2020, he was a Computer Vision Team Lead and an Architect with Valeo Vision Systems, where he was also a Senior Expert. In 2016, he was an Adjunct Lecturer with the National University of Ireland. In 2020, he joined the University of Limerick as a Lecturer in artificial intelligence and computer vision.

ANDREI BURSUC received the Ph.D. degree from Mines ParisTech, in 2012. He is currently a Research Scientist with Valeo.ai, Paris, France. He was a Postdoctoral Researcher with Inria Rennes and Inria, Paris. In 2016, he moved to industry to pursue research on autonomous systems. His current research interests include computer vision and deep learning, in particular annotation-efficient learning and predictive uncertainty quantification. He regularly serves as a reviewer for major computer vision and machine learning conferences and journals. He is teaching undergraduate courses with Ecole Normale Supérieure and Ecole Polytechnique.

UJJWAL BHATTACHARYA (Senior Member, IEEE) received the M.Sc. and M.Phil. degrees in pure mathematics from Calcutta University. He is currently a member of the Faculty of the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, where he joined as a Junior Research Fellow, in 1991. In the past, he collaborated with a few industries and research laboratories in India and abroad. In 1995, he received the Young Scientist Award from the Indian Science Congress Association. Also, he received a few best paper awards from various groups. He is a Life Member of IUPRAI, the Indian Unit of the IAPR. He has served as a program committee member for various reputed international conferences and workshops. Also, he was a co-guest editor of a few special issues of international journals. His current research interests include machine learning, computer vision, image processing, document processing, and handwriting recognition.