Learning to See in the Dark with Events

Song Zhang*[1,3], Yu Zhang*†[3,4], Zhe Jiang*[3], Dongqing Zou[3], Jimmy Ren[3], and Bin Zhou†[1,2]

[1] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] SenseTime Research, Beijing, China [4] Tsinghua University, Beijing, China
{zhangs,zhoubin}@buaa.edu.cn
{zhangyu1,jiangzhe,zoudongqing,rensijie}@sensetime.com

Abstract. Imaging in the dark environment is important for many real-world applications like video surveillance. Recently, the development of *Event Cameras* raises promising directions in solving this task thanks to its High Dynamic Range (HDR) and low requirement of computational sources. However, such cameras record sparse, asynchronous intensity changes of the scene (called *events*), instead of canonical images. In this paper, we propose learning to see in the dark by translating HDR events in low light to canonical sharp images as if captured in day light. Since it is extremely challenging to collect paired event-image training data, a novel unsupervised domain adaptation network is proposed that explicitly separates domain-invariant features (*e.g.* scene structures) from the domain-specific ones (*e.g.* detailed textures) to ease representation learning. A detail enhancing branch is proposed to reconstruct day light-specific features from the domain-invariant representations in a residual manner, regularized by a ranking loss. To evaluate the proposed approach, a novel large-scale dataset is captured with a DAVIS240C camera with both day/low light events and intensity images. Experiments on this dataset show that the proposed domain adaptation approach achieves superior performance than various state-of-the-art architectures.

Keywords: Domain Adaptation · Event Camera · Image Reconstruction · Low Light Imaging

1 Introduction

Event cameras [23, 5], a kind of bio-inspired vision sensors that mimic the human eye in receiving the visual information, have gained more and more attention in computer vision community. Different from traditional cameras capturing intensity frames at a fixed rate, event cameras transmit the changes of scene intensities asynchronously, which are called events. Endowed with microsecond

*Equal contribution.
†Corresponding authors: Bin Zhou (zhoubin@buaa.edu.cn) and Yu Zhang (zhangyulb@gmail.com).

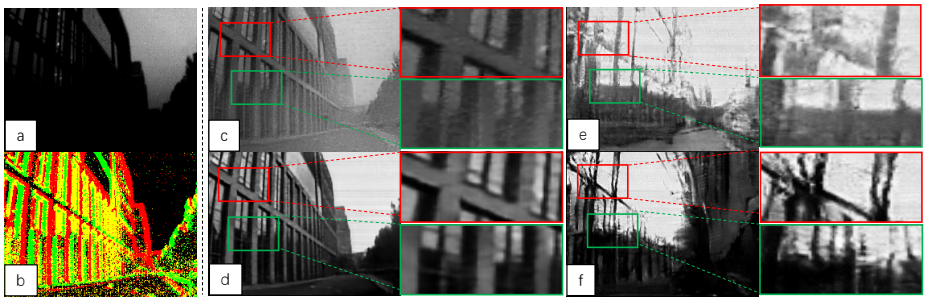

Fig. 1: Motivation of the proposed approach. (a) In low light the conventional camera fails to capture the scene clearly. (b) Event cameras, on the other side, can perceive the scene due to its high dynamic range, though with noisy measurements. (c) In low light, even the state-of-the-art event-to-image translation approach [31] fails reconstructing a clean image. (d) Despite the large domain gap, the proposed domain adaptation approach learns to reconstruct high-quality day-light image by observing low-light events. (e)(f) Domain translation results of strong baselines [39] and [29], respectively.

temporal resolution, low power assumption and a high dynamic range (e.g., 140 dB compared to 60 dB of most consumer-level standard cameras), event cameras have promising applications in various scenes with broad illumination range.

Reconstructing canonical images from asynchronous events has been explored in various research [33, 27, 21, 31]. Early works, which were primarily inspired by the physical formation model of events, are prone to real-world imperfectness and often generate noisy reconstructions. Recent deep reconstruction model [31] has demonstrated impressive performance through being trained on large amounts of simulated image/event pairs. Despite their successes, few works pay sufficient attention to event-based imaging in low-light environment. Like conventional cameras, events in low light have their own distributions. The sparity and noisiness of low-light events render their distribution rather different with that in normal light. As a result, even the state-of-the-art event-to-image translation approach [31] fails generating clean reconstruction due to the domain gap, as shown in Fig. 1 (c). Besides, collecting large datasets with events and reference sharp images in low light is hardly practical.

Motivated by recent advances of deep domain adaptation methods [11, 24, 17, 16, 29] and the HDR property of event cameras, a novel domain adaptation method is proposed to tackle the problem of generating clear intensity images from events in the dark. Specifically, features are extracted from the low-light event domain and transferred to the day-light one, in which the reference sharp images are much easier to collect. To this end, previous domain adaptation methods usually project features from different domains into a common feature space, which can be supervised with available rich labels [16, 29, 9, 35]. In this manner, discriminative cross-domain attributes are preserved while the distractive ones are neglected. While this is fine for high-level downstream tasks such as classification and segmentation, it may not be the best choice for the proposed task. In

fact, the domain-specifc "distractors", such as the high frequency scene details and textures, largely affects the reconstruction quality and should be complementary to the domain-shared attributes (*e.g.* scene structrues). Therefore, we propose distangled representation learning: we decompose the day-light domain features into a domain-shared part and a domain-specific part. The domain-shared part is expected to encode scene-level information such as structures, intensity distributions, *etc.*, which could also be perceived from the low-light events. The domain-specific part contains day-light exclusive patterns, *e.g.* high-frequency details, which may not be easily recovered from low-light events. A dedicated residual detail enhancing network is incorporated to recover plausible day-light details from the common representations in a generative manner. We show that such decomposition can be regularized with a simple ranking loss. To evaluate the proposed approach, we capture a large dataset with real-world low-light/day-light events and images, with a DAVIS240C camera [5]. Experiments on this dataset show that the proposed architecture outperforms various state-of-the-art domain adaptation and event-based image reconstruction approaches.

Our contributions are: 1) we propose a novel domain-adaptation approach for event-based intensity image reconstruction in low light. Our approach achieves the best reconstruction quality among various state-of-the-art architectures; 2) we propose to learn decomposed representations to smooth the adaptation process, which can be regularized by a simple ranking loss; 3) a large dataset with low-light/day-light events and images of various real-world scenes is compiled, and will be made publicly available to facilitate future research.

2 Related Work

Intensity image reconstruction from events. Dynamic Vision Sensors (DVS) [23, 5] are a kind of biology inspired sensors which sends signals when the scene exhibits illumination changes. An "on" signal is generated when the pixel brightness goes up to a certain degree, or an "off' signal otherwise. The signals, often referred as *events*, can be denoted with a tuple (u, v, t, p), where u, v are the pixel coordinates of the event, t is the timestamp, and $p \in \{-1, +1\}$ is the polarity, *i.e*, representing "on" and "off" states.

Early attempts on reconstructing the intensity image from events are inspired by their physical imaging principles. Generally, the spatiotemporal formation of events are formulated with various forms of pixel tracking models. For example, Kim *et al.* [20] estimate scene movements and gradients, and integrate them in a particle filter to generate final intensity reconstructions. Barua *et al.* [2] employs patch-based sparse dictionary learing on events to address image reconstruction. Bardow *et al.* [1] integrates optical flow estimation with intensity reconstruction into a cost function with various spatiotemporal regularizations. Scheerlinck *et al.* [33] updates each pixel location with incoming events via a complementary ordinary differential equation filter.

Recent research directs to learning deep event-to-image translation models from data [21, 31]. Generally, these approaches achieve much improved recon-

structions due to deep networks. However, a challenge is how to collect clean training pairs of events and reference images. Wang *et al.* [21] captures such data with off-the-shelf hybrid event camera simultaneously recording events and images with time calibration. Rebecq *et al.* applies a carefully designed event simulator [30] to synthesize event data from existing large-scale image datasets. However, in low-light scenario, while the former fails due to the limited dynamic range of image sensor, the latter may not well model real noise distributions.

Unsupervised domain adaptation. To overcome the above difficulty, we treat low-light image reconstruction from events as an unsupervised domain adaptation task. Unsupervised domain adaptation has been widely adopted for image classification [24, 34, 36, 4, 11, 13], segmentation [16, 9] and synthesis [3, 29, 15]. In the line of pixel-level adaptation networks, Hoffman *et al.* [15] proposes an early attempt that obtains domain-agnostic features. Ghifary *et al.* [13] and Murez *et al.* [29] develop further constraints (*e.g.* decodable shared features, cycle-consistent mapping originating from [39]), to further improve the robustness of adaptation. Based on these constraints, further extensions are proposed [35, 9] by exploring structural cues, tailored for segmentation. However, restricting the inference on domain-shared features may neglect domain-specific high-frequency information, which is not problematic or even desired for high-level tasks such as classification, but may sacrifice the sharpness and realistics for image generation. Bousmalis *et al.* [4] propose to explicitly decompose the features into domain-shared and domain-specific ones. However, their purpose is to obtain more discriminative and clean domain-shared features for classification, instead of leveraging domain-specific high-frequency details to improve reconstruction.

Deep image enhancement. Recently, impressive results were shown for training deep image enhancing models on enhancing bayer raw [7, 8] and HDR [19, 25, 12, 38, 6] images. However, collecting the reference aligned images for low-light events are extremely difficult. A widely adopted alternative is unpaired training, via modality translation paradigm [17, 39, 10, 18]. Nevertheless, the proposed task involves modality translation as a task (*i.e.* event-to-image translation), in two separate domains (*i.e.* low light and day light). In addition, the domain difference in the proposed task involves various factors like color, noise, scene textures and details. This introduces further challenges as opposed with existing success, for which domain difference often lies in color distribution [10, 18]. In fact, we find that existing domain translation approaches cannot well address the proposed task, as illustrated in Fig. 1 and our experimental section.

3 The Proposed Approach

Our approach assumes that a dataset with events and their reference intensity images are given in day light environment, while events and images are time-calibrated. Such data could be easily collected with mordern off-the-shelf hybrid cameras with calibrated sensors, *e.g.* DAVIS240C [5]. In addition, a set of captures of events in the target low-light environment are also required, but this

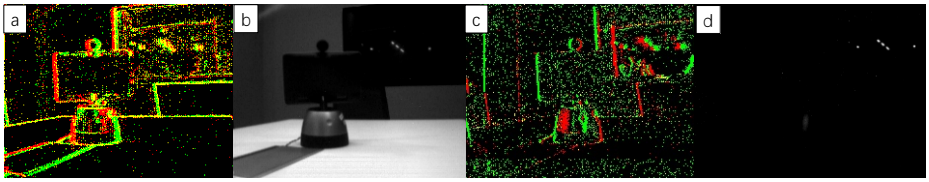

Fig. 2: Events and intensity images of the same scene in day-light and low-light environment, captured with the DAVIS240C camera in a lab room with controlled lighting. The first two columns show events (a) and corresponding intensity image (b) in day-light, respectively. The last two columns (c) (d) illustrate them in the dark.

time the reference images are allowed absent. The scene coverage in day light and low light do not need to be aligned. On this dataset, our objective is to train an event-to-image translation model that generalizes to the target domain. We start with introducing the high-level design of the proposed architecture, then goes into the technical details of each component.

3.1 Learning Architecture

As described previously, event cameras capture changes of scene intensities on each pixel location. Therefore, the distributions of events are closely related to lighting conditions, like standard frame-based cameras. In Fig. 2 we shown sample events and the corresponding image captures of the same scene. We can observe that 1) the low-light and day-light events distributions differ in a number of aspects including noise, event density and the sharpness of details. The events captured in low light are obviously much noisier and less sharper; 2) however, due to the high dynamic range of event camera, the coarse scene structures are still preserved in different lightings. Therefore, an ideal domain adaptation approach should be capable of 1) extracting the shared scene-level features in two domains; 2) "hallucinating" plausible day light-specific details that are missing in the low-light domain. While the former has been verified as a common rule for domain adaptation methods [29], the latter and how it can be integrated into the whole pipeline for image reconstruction, is still less explored.

Keeping this in mind, we propose a novel training architecture for domain-adapted low-light event-to-image translation, as summarized in Fig. 3. It consists of a day light-specific private encoder $\mathbf{E}_p$, a domain-shared encoder $\mathbf{E}_c$, a shared decoder $\mathbf{R}$ for image reconstruction, a domain discriminator $\mathbf{D}$, and a detail enhancing branch $\mathbf{T}_e$ which will be explained shortly after.

Given a source (day-light) domain event representation $\mathbf{x}^s$, we concatenate it with a spatial noise channel $\mathbf{z}$, and feed them into the source-private encoder $\mathbf{E}_p$, obtaining domain-private features $\mathbf{x}_p^f = \mathbf{E}_p\left(\mathbf{x}^s, \mathbf{z}; \boldsymbol{\theta}_p\right)$. Conditioning on both the real samples and a noise channel helps improve the generalization of the model, which is verified in previous works [16] and our experiments. Meanwhile, the source events $\mathbf{x}^s$ itself also go through the domain-shared encoder $\mathbf{E}_c$, yielding domain-public features $\mathbf{x}_c^f = \mathbf{E}_c\left(\mathbf{x}^s; \boldsymbol{\theta}_c\right)$. On the other side, the target domain

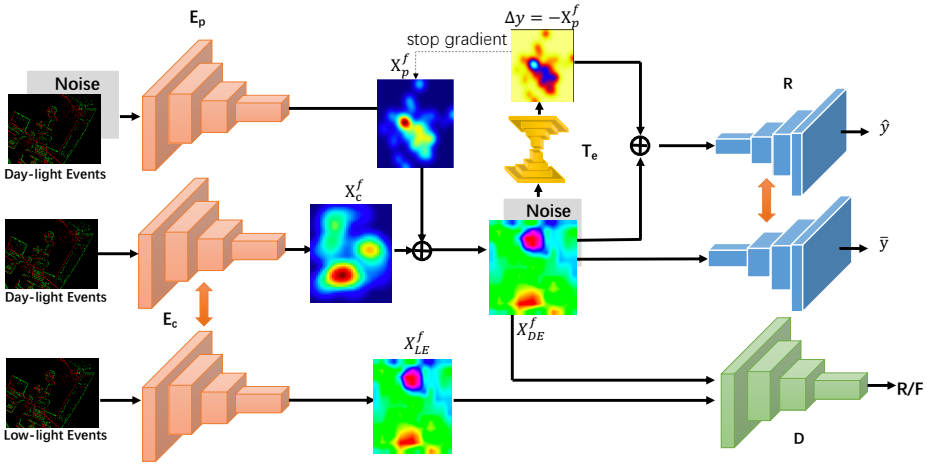

Fig. 3: The proposed framework. The day-light events and low-light events are fed into a shared encoder $\mathbf{E}_c$ to extract their representations $\mathbf{x}_c^f$ and $\mathbf{x}_{LE}^f$. Meanwhile, the day-light events are also fed into a private encoder $\mathbf{E}_p$ along with a noise channel, yielding source domain-specific residuals $\mathbf{x}_p^f$. By adding operation, the modulated day-light features $\mathbf{x}_{DE}^f = \mathbf{x}_c^f + \mathbf{x}_p^f$ and the low-light features $\mathbf{x}_{LE}^f$ lie in a domain-shared feature space, guaranteed by adversarial training with a discriminator $\mathbf{D}$. The detail enhancement branch $\mathbf{T}_e$ reconstructs day-light domain-specific residuals from the shared features. Finally, a shared decoder $\mathbf{R}$ reconstructs the intensity images using both the domain-specific and shared representations. "R/F" respresents Real or Fake logits.

(low-light) event representation $\mathbf{x}_{LE}^t$ is sent to the same shared encoder to get its encoding $\mathbf{x}_{LE}^f = \mathbf{E}_c\left(\mathbf{x}_{LE}^t, \boldsymbol{\theta}_c\right)$. Here, $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_c$ are trainable parameters.

We expect that the domain-private features $\mathbf{x}_p^f$ encode day-light exclusive patterns such as high-frequency details, which may not be easily captured by the shared features $\mathbf{x}_{LE}^t$ extracted from the low-light events. On the other hand, the domain-public features $\mathbf{x}_c^f$ from day-light events may still contain source-domain relevant attributes. We perform an "add" operation between $\mathbf{x}_c^f$ and $\mathbf{x}_p^f$, obtaining a modulated feature $\mathbf{x}_{DE}^f = \mathbf{x}_c^f + \mathbf{x}_p^f$, which lies in a shared feature space with $\mathbf{x}_{LE}^f$ as if extracted from the low-light domain. In this manner, $\mathbf{x}_p^f$ can be deemed as the "negative" complementary features of source domain. To guarantee the desired behavior, we feed $\mathbf{x}_{LE}^f$ and $\mathbf{x}_{DE}^f$ to a domain discriminator $\mathbf{D}\left(\cdot; \boldsymbol{\theta}_D\right)$, which serves to distinguishing between the transformed feature map $\mathbf{x}_{DE}^f$ from the day-light domain, and the features $\mathbf{x}_{LE}^f$ extracted from the real low-light domain.

The features $\mathbf{x}_{DE}^f$ projected into the domain-shared space can be then fed into the decoder $\mathbf{R}\left(\cdot; \boldsymbol{\theta}_R\right)$ to get the reconstruction result $\bar{\mathbf{y}}$, which can be directly supervised with source-domain labels. However, following the above analysis, the shared features may not preserve high-frequency source-domain relevant details, leading to suboptimal reconstruction. To address this limitation,

a detail enhancing branch $\mathbf{T}_e$ is proposed. It aims to inversely recovering the removed source domain-specific features $\mathbf{x}_p^f$ in conditional generation manner, taking as input the domain-shared features $\mathbf{x}_{DE}^f$ and the identical noise channel: $\Delta y = -\mathbf{x}_p^f \approx \mathbf{T}_e\left(\mathbf{x}_{DE}^f, \mathbf{z}; \boldsymbol{\theta}_t\right)$. The full day-light features are then constructed by adding Δy to $\mathbf{x}_{DE}^f$ in residual way. The recovered features are finally fed into the decoder $\mathbf{R}$ to get an improved reconstruction $\hat{\mathbf{y}}$.

To train the proposed architecture, we employ the following adversarial training objective:

$$\min_{\boldsymbol{\theta}_g, \boldsymbol{\theta}_r, \boldsymbol{\theta}_t} \max_{\boldsymbol{\theta}_d} \alpha \mathcal{L}_D(\mathbf{D}, \mathbf{G}) + \mathcal{L}_R(\mathbf{G}, \mathbf{T}_e, \mathbf{R}), \tag{1}$$

where α balances the interaction of different terms, and is set to 0.1. For the domain loss $\mathcal{L}_D(\mathbf{D}, \mathbf{G})$, we define it as a common logistic loss in training GANs:

$$\begin{aligned} \mathcal{L}_D(\mathbf{D}, \mathbf{G}) =& \mathcal{E}_{\mathbf{X}^t}[\log \mathbf{D}(\mathbf{x}^t; \boldsymbol{\theta}_d)] + \\ & \mathcal{E}_{\mathbf{X}^s, \mathbf{Z}}[\log(1 - \mathbf{D}(\mathbf{G}(\mathbf{x}^s, \mathbf{z}, \boldsymbol{\theta}_g); \boldsymbol{\theta}_\mathbf{d}))], \end{aligned} \tag{2}$$

where $\mathcal{E}$ denotes expectation over distribution.

Our reconstruction related loss $\mathcal{L}_R$ can be decomposed into three terms:

$$\mathcal{L}_R(\mathbf{G}, \mathbf{T}_e, \mathbf{R}) = L_p(\hat{\mathbf{y}}, \bar{\mathbf{y}}, \mathbf{y}^g) + \beta L_t(\Delta \mathbf{y}, \mathbf{x}_p^f) + \gamma L_r(\hat{\mathbf{y}}, \bar{\mathbf{y}}, \mathbf{y}^g), \tag{3}$$

where $\bar{\mathbf{y}}$ and $\hat{\mathbf{y}}$ are the reconstructions from shared and inversely recovered features (see Fig. 3), $\mathbf{y}^g$ is the groundtruth reconstruction available on the source day-light domain. The balancing weights β and γ are empirically set to 0.1.

The first term is the ℓ_1 photometric reconstruction loss:

$$L_p(\hat{\mathbf{y}}, \bar{\mathbf{y}}, \mathbf{y}^g) = \frac{1}{|\mathbb{I}^s|} \sum_{i=1}^{\mathbb{I}^s} \left(\|\hat{y}_i - y_i^g\|_1 + \|\bar{y}_i - y_i^g\|_1 \right), \tag{4}$$

in which $\mathbb{I}^s$ denotes the set of all the image pixels in source domain.

With the second term, we train the detail enhancing branch to regress the source domain-specific residuals:

$$L_t(\Delta \mathbf{y}, \mathbf{x}_p^f) = \frac{1}{|\mathbb{F}^s|} \sum_{i=1}^{\mathbb{F}^s} \| \Delta y_i + (\mathbf{x}_p^f)_i \|_1. \tag{5}$$

The set $\mathbb{F}_s$ denotes all the pixel locations on the immediate feature maps $\mathbf{x}_p^f$. Note that the target labels in (5) are not fixed but dynamically changed with network training. To avoid contaminating representation learning in early training stage, gradients from L_t to the private encoder $\mathbf{E}_p$ is blocked, as illustrated in Fig. 3. It effectively cancels the loops in information flows and gets rid of mode collapse.

Finally, we add further regularizations between $\bar{\mathbf{y}}$, the reconstructions from the shared-space features $\mathbf{x}_{DE}^f$, and $\hat{\mathbf{y}}$, reconstructions from the recovered domain-

specific features $\mathbf{x}_{DE}^{f} + \Delta\mathbf{y}$. We employ a ranking loss, preferring stronger performance for the latter:

$$L_r(\hat{\mathbf{y}}, \bar{\mathbf{y}}, \mathbf{y}^g) = \max\left(\frac{1}{|\mathbb{I}^s|}\sum_{i=1}^{\mathbb{I}^s}\|\hat{y}_i - y_i^g\|_1 - \frac{1}{|\mathbb{I}^s|}\sum_{i=1}^{\mathbb{I}^s}\|\bar{y}_i - y_i^g\|_1 + \epsilon, 0\right), \quad (6)$$

where $\epsilon > 0$ is a predefined margin. Eqn. (6) encourages the reconstruction $\hat{\mathbf{y}}$ to have a lower loss compared with that of $\bar{\mathbf{y}}$ so as to probe a better reconstruction. This is somewhat counter-intuitive, since that (6) can achieve the same minimizer by degrading $\bar{\mathbf{y}}$ instead of improving $\hat{\mathbf{y}}$. The rationale behind this regularization is that the domain loss L_D and the reconstruction loss L_R in (1) do not always have consistent objectives, especially in perceptual image synthesis tasks. Actually, as advocated by recent studies [35, 26], adversarial generation and domain adaptation without regularization tends to align the most discriminative (instead of many) modes of two distributions. To this end, it may underplay important distortion-critic features, which are often not the most discriminative ones between domains but crucial for reconstruction quality. It shares with a similar finding from GAN-based image superresolution [22, 37]. Including GAN objective may damages distortion-based metrics (though perceptual quality can be improved by training on carefully created, high-quality clean datasets). In our setting, the large distribution gap between day-light and low-light domains can readily guide the discriminator to focus on several main differences, *e.g.* noise and texture distributions, divating from reconstructing other details and damages reconstruction quality. Through the regularization (6), it would be not a great issue if this happens, as the detail enhancing branch can encode the lost less discriminative but important information to promote a better reconstruction. In the experimental section, we show that such regularization indeed improves the results in both quantitative and subjective comparisons.

In Fig. 4 we visualize the reconstructed images with and without the ranking-based regularization. Without the ranking loss, the reconstruction in Fig. 4 (a) comes from the shared feature representations. In this case, the inter-domain adversarial loss dominates the training process, which focuses on discriminative distributional differences but neglects tiny scene details. Adding the ranking loss guides the detail enhancing network to recover such details while leaving the discriminative adaptation unaffected. As the loss itself does not increases network capacity, it serves as *self-regularization* that smooths the training process by distangling representation learning for different network components.

3.2 Implementation Details

Event representation. Events are asynchronous signals that cannot be trivially processed by image-specialized networks like CNN. To overcome this issue, we employ the stacked event frame representations [21]. Specifically, we divide the events captured from a scene into equal-size temporal chunks. In each chunk, we sum up event polarities triggered in the time interval of this chunk at per

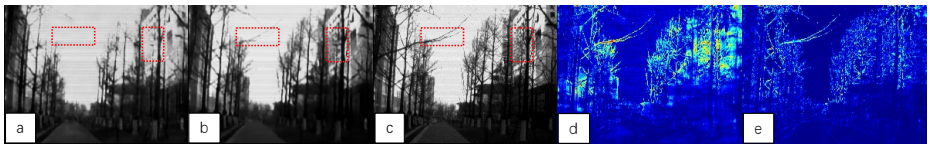

Fig. 4: Effectiveness of rank regularization. Reconstruction without and with rank regularization are visualized in (a) and (b), respectively. (c) The reference image. Reconstruction error maps of (a) and (b) are shown in (d) and (e).

pixel level. For each pixel $\mathbf{p}$, this writes to

$$\Phi(\mathbf{p}) = \sum_{\mathbf{e}=(e_{\mathbf{p}},e_t,e_l)\in\mathbb{E}} \mathbb{1}(e_{\mathbf{p}} = \mathbf{p} \wedge e_t \in [t, t+\tau]) \cdot e_l, \tag{7}$$

where $e_{\mathbf{p}}$ and e_t represent the pixel location and time stamp when the event is triggered, and $e_l \in \{-1, 1\}$ denotes the polarity of event. The indicator function $\mathbb{1}(\cdot)$ takes 1 if the input condition holds, or 0 otherwise. In our implementary, we employ 4 chunks, where the time length τ of each chunk spans across 1.25 miliseconds. In addition, we concatenate the time stamps of the events (normalized into $[0, 1]$) with the event frames, resulting into a 8-channel representation.

Noise channel. The noise channel is a 1-dimensional map with the same spatial resolution with that of event frames. Values of this noise map are randomly sampled from a spatial Gaussian distribution whose mean is zero and standard deviation is 1.

Layer configurations. The encoder and decoder forms the generator that aims to synthesizing plausible real-like images. Thus, they are implemented with residual networks. For encoder $\mathbf{E}_p$ and $\mathbf{E}_c$ in Fig. 3, the concatenated event frames and noise map go through a 7×7 convolution without striding, and two 3×3 convolutions with stride $= 2$. The feature dimensions are scaled up to 32, 64 and 128, sequentially. These features are then fed into 9 identical residual blocks, which has two 3×3 convolutions. For the decoder $\mathbf{R}$, it has 9 the same residual blocks, then upsample the features with two 3×3 deconvolution layers. The feature dimension reduces from 128 to 64 and 32 during upsampling. Finally a 7×7 convolution layer fuses the features to a single output map. The discriminator $\mathbf{D}$ consists of three 5×5 convolutions with stride $= 2$ and two 5×5 convolutions with stride $= 1$. The feature dimensions go up to 32, 64, 128 and 256, finally fused to a single output logit channel. Average pooling is taken to compute final score. All the convolution and deconvolution layers are interleaved with Instance Normalization and Leaky ReLU activation.

The detail enhancing branch $\mathbf{T}_e$ consumes encoded feature maps and the same noise channel as input of the private encoder $\mathbf{E}_p$. To make spatial dimensions consistent, the noise map individually goes through three 3×3 convolution layers with 32 output dimensions. The output are then concatenated with the input features, followed by a 3×3 processing convolution layer and 9 residual identity blocks. Again, instance normalization and ReLU (or leaky ReLU)

activation are used for interleaving the convolution layers. Please refer to our supplementary material for more detailed layer/parameter configurations.

Testing on low-light events. In testing phase, low-light event representations are passed through the shared encoder $\mathbf{E}_c$ to get $\mathbf{x}_{LE}^f$. A noise channel is sampled then combined with $\mathbf{x}_{LE}^f$, which are fed into the detail enhancing branch $\mathbf{T}_e$ to get the residual features $\Delta\mathbf{y}$. Finally, the shared decoder $\mathbf{R}$ consumes $\mathbf{x}_{LE}^f + \Delta\mathbf{y}$ to obtain the final reconstruction. Note that the private encoder $\mathbf{E}_p$ and the discriminator $\mathbf{D}$ are not involved in testing phase.

4 Experiments

4.1 Experimental Settings

Data collection. To the best of our knowledge, there does not exist a mature dataset for evaluating event-based low-light imaging. Thus, we introduce a large novel dataset to evaluate the proposed approach. The dataset is captured by a DAVIS240C camera [5], which has an Active Pixel Sensor (APS) to record intensity image sequences as well as an event sensor recording event streams. These two modalities are calibrated in spatial and temporal axis. The dataset consists of image scenes of urban environment at resolution 180×240, captured in both day-light and low-light conditions with various kinds of camera/scene movement. In summary, there are 8820 scenes captured in day time, and 8945 ones in the night. For each scene, the event stream spans across roughly 100ms. Note that in day light, the reference scenes and lightings are carefully chosen to avoid saturated pixels (though some still remains in high-light area), so that the images captured by APS sensors can serve as groundtruth. In the night, however, we do not control lighting and there is no high-quality reference images available. Across the experiments, this dataset is referred as DVS-Dark.

Baseline settings. To compare the proposed domain adaptation framework with state-of-the-art frameworks, we carefully choose and implement 4 representative baselines, as described as follows:

1) Pix2Pix [17, 21]. Pix2Pix is a successful conditional generative framework for image-to-image translation, while Wang *et al.* [21] extend it to event-to-image translation. We carefully follow Wang *et al.* [21] to set this baseline, and train it on the day-light event/image pairs. The trained network can be applied to low-light test. Thus, no adaptation is performed in this baseline.
2) CycleGAN [39]. The milestone unpaired domain translation framework is proposed by Zhu *et al.* [39]. We adapt this framework to translate between low-light events and day-light images. However we find the naive implementation hardly working in practice, potentially due to the large domain gap between low-light events and day-light images. Thus, we intead adopt a modification in semi-supervised setting: the paired events and images are used in supervised manner to train the forward/backward mapping networks, together with unpaired domain translation.

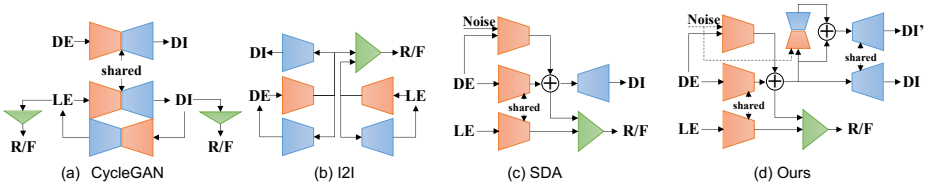

Fig. 5: Illustration of baseline architectures. The red, blue, and green graphics denote the encoder, decoder, and discriminator, respectively. Symbols "D" and "L" represents day-light and low-light, while "E" and "I" refer to events and intensity images.

3) I2I [29]. In the image-to-image translation architecture proposed by Murez *et al.* [29], the day-light and low-light events are projected into a shared feature space. A decoder then learns to reconstruct day-light images from the shared features, which are directly supervised by paired day-light events and images. Besides, two domain-specific decoders (weights are not shared) are involved, responsible for reconstructing the shared features to the original event representations.
4) SDA [16]. The structured domain adaptation framework proposed by Hong *et al.* [16] is originally for semantic segmentation, which we tailor to address image reconstruction. In this framework, the source samples are passed through a source-private encoder to learn source-specific residuals, and also through a shared encoder to get immediate shared features. The immediate shared features then cancels the residuals, yielding domain-shared features that lie in the same space with the features extracted from the target domain. Note that compared with our architecture, SDA does not address detail recovery, and the final reconstruction still comes from the shared feature space.

An illustration of baseline architectures except Pix2Pix is referred to Fig. 5. In addition, we also compare with state-of-the-art event-based image reconstruction approaches CIE [33] and RIRM [28] and E2V [31]. They directly reconstruct intensity images from events via physical or learning-based rules.

Training details. We train the network for 200 epochs, using a mini-batch of 4 images and 10^{-4} as initial learning rate. The learning rate linearly decays to zero from the 100th to the 200th epoch. As for CIE, RIRM and E2V, we directly adopts the author's open source code.

4.2 Comparisons with State-of-the-Art Models

Quantitative evaluation through simulated experiments. Since groundtruth is only available for day-light data, we perform simulated experiments by artificially translating a random half of day-light subset to low light to conduct quantitative analyss. In details, we apply a pixel-valued S-shape tone curve to adjust the day-light image brightness. This curve globally shrinks the pixel brightness by roughly 50%, and further shrinks high-light pixels by 60%. To account for event distribution gap between lightings, we discard the captured day-ight events

Table 1: PSNR and SSIM numbers on simulated data.

	RIRM	CIE	E2V	E2V$^+$	CGAN	Pix2Pix	I2I	SDA	Ours
PSNR	11.28	13.30	12.16	17.15	14.20	22.65	23.36	24.93	**26.03**
SSIM	0.29	0.45	0.33	0.63	0.27	0.68	0.69	0.75	**0.77**

Table 2: Inception scores(higher is better) and Frechet Inception Distance(lower is better) on real low-light data. Numbers in the parentheses denote standard deviation.

	CIE	RIRM	E2V	CGAN	Pix2Pix	I2I	SDA	Ours
IS	3.36 (±0.47)	2.28 (±0.37)	3.21 (±0.61)	2.89 (±0.15)	3.86 (±0.24)	2.44 (±0.15)	3.75 (±0.24)	**3.87** (±0.21)
FID	267.8	208.5	210.42	163.28	109.08	177.59	110.18	**104.21**

and regenerate them with the ESIM simulator [30] for day-light and artificially created low-light images. We use 20% data for testing, and the left for training.

In Table 1 we summarize PSNR and SSIM numbers of different approaches. We note that since RIRM, CIE and E2V all reconstruct intensity frames purely from events, they are not aware of dataset-specific color distributions thus their PSNRs and SSIMs are not meaningful. For fair comparison we propose the variant E2V$^+$, which is obtained by finetuning E2V on the day-light images of DVS-Dark, using a similar process as described in [31] to simulate events. However, it still falls behind our approach in training and testing phase. Among domain translation/adaptation approaches, the proposed approach achieves much improved performance by additionally considering detail recovery from the shared feature space. In contrast, conventional reconstruction-from-shared-space paradigm, as adopted by I2I and SDA, gets worse results.

Comparisons on real low-light data. Due to the lack of groundtruth reference images in low light, we measure the performance via two widely adopted perceptual measures, the Inception Scores (IS) [32] and the Frechet Inception Distance (FID) [14], as summarized in Table 2. The event-based reconstruction approaches CIE, RIRM and E2V do not get satisfactory perceptual score. Among domain adaptation approaches, the proposed approach still achieves the best performance in both metrics, with the best perceptual quality. In Fig. 6, we provide representative results generated by different approaches. The event-based reconstruction methods CIE, RIRM and E2V recover plenty of scene details, but tend to be noisy. The domain adaptation approaches effectively addresses the noise but may miss tiny details/structures of the scene. The proposed approach preserves scene details best while successfully suppressing the noise.

4.3 Performance Analysis

Analysing different components. We analyse the contributions of different network components on the simulated datasets, as summarized in Table 3.

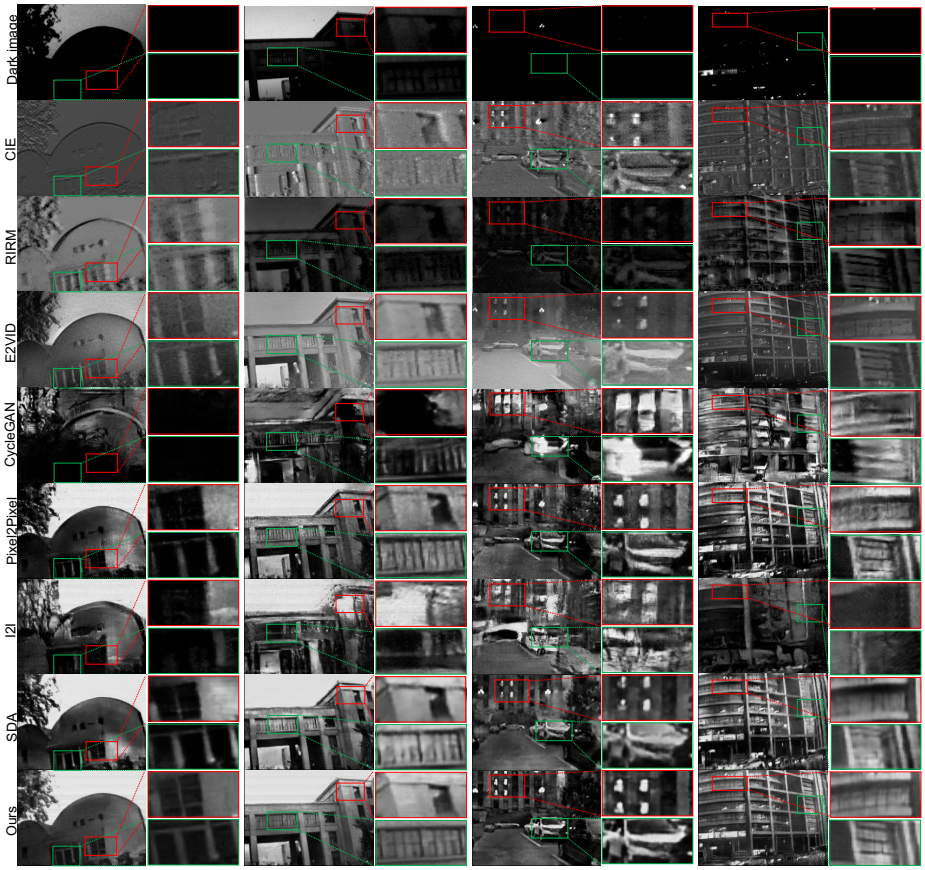

Fig. 6: Representative reconstruction results generated by different approaches from real low-light event data on the DVS-Dark dataset. More results are referred to our supplementary material. Best viewed in color with zoom.

noise.	detail.	rank.	PSNR	SSIM
✗	✗	✗	20.69	0.63
✓	✗	✗	21.24	0.66
✓	✓	✗	24.91	0.75
✓	✓	✓	26.03	0.77

Table 3: Ablation studies of different network components.

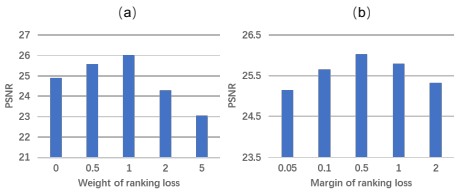

Fig. 7: PSNR as function of the weight (a) and margin (b) of ranking loss.

Here, "noise.", "detail." and "rank." denote the noise map augmentation, detail enhancing branch, and ranking mechanism, respectively. From the results, we observe that incorporating noise channel improve the results significantly as it

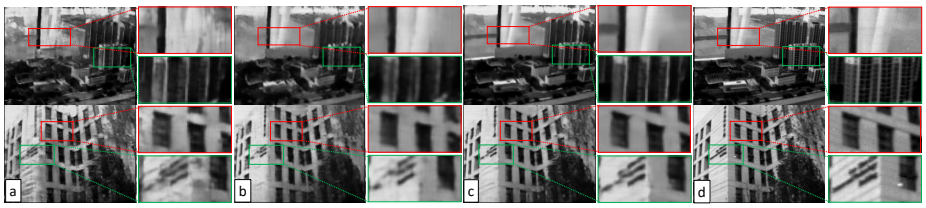

Fig. 8: Effectivenss of rank regularization. (a) w/o regularization; (b) w/ regularization, $\gamma = 5.0$; (c) w/ regularization, $\gamma = 1.0$; (d) reference image.

implicitly augments the data by introducing randomness into the inputs. The detail enhancing branch restores domain-specific details that are not well-modelled in shared feature learning, and leads to a substantial improvement. By further combining the rank regularization, the network is guided to learn both domain-discriminative and domain-exclusive features to promote better reconstruction.

Effectiveness of the rank regularization. The ranking loss (6) can effectively regularize the domain adaptation process of the proposed network. However, improper parameters may overwhelm the contribution of other loss terms, deteriorating the reconstruction quality in order to satisfy the ranking constraints. Fig. 7 shows that the optimal choices of the loss weight and margin on the DVS-Dark dataset are 1.0 and 0.5 respectively.

Besides recovering small details as analysed in Sect. 3.1, the rank regularization also leads to more smooth reconstruction. As shown in Fig. 8, without regularization, the discriminative adaptation would dominate the training process. The results are sharp, but with false textures. With over-regularization ($\gamma = 5.0$), false textures vanish but the results tend to be blurry. Proper strength of regularization ($\gamma = 1.0$) leads to sharp and clean results.

5 Conclusion

We present in this work a deep domain adaptation method for intensity image reconstruction from events captured in low light. The model explicitly learns the shared representation inter domains and domain-specific features via novel detail enhancing mechanism regularized by relative ranking. Our method outperforms related existing methods of image generation from events and unsupervised domain adaptation methods, in both quantitative and qualitative comparisons.

Acknowledgement. We thank the anonymous reviewers for their valuable comments. This work was supported in part by National Natural Science Foundation of China (U1736217 and 61932003), National Key R&D Program of China (2019YFF0302902), Pre-research Project of the Manned Space Flight (060601), and Beijing Posdoctorial Research Foundation (ZZ-2019-89).

References

1. Bardow, P., Davison, A., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 884–892 (2016)
2. Barua, S., Miyatani, Y., Veeraraghavan, A.: Direct face detection and video reconstruction from event cameras. In: IEEE winter conference on applications of computer vision (WACV). pp. 1–9 (2016)
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3722–3731 (2017)
4. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhans, D.: Domain separation networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 343–351 (2016)
5. Brandli, C., Berner, R., Yang, M., Liu, S., Delbruck, T.: A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014)
6. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing (TIP) **27**(4), 2049–2062 (2018)
7. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: IEEE International Conference on Computer Vision (ICCV). pp. 3185–3194 (2019)
8. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3291–3300 (2018)
9. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 33, pp. 865–872 (2019)
10. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) **17**(1), 2096–2030 (2016)
12. Gharbi, M., Chen, J., Barron, J., Hasinoff, S., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
13. Ghifary, M., Kleijn, B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision (ECCV). pp. 597–613 (2016)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NIPS). pp. 6626–6637 (2017)
15. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. ArXiv preprint. ArXiv:1612.02649 (2016)
16. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1335–1344 (2018)

17. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2017)
18. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. ArXiv preprint. ArXiv:1906.06972 (2019)
19. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics (TOG) **36**(4), 144–1 (2017)
20. Kim, H., Handa, A., Benosman, R., Ieng, S.H., Davison, A.: Simultaneous mosaicing and tracking with an event camera. British Machine Vision Conference (BMVC) **43**, 566–576 (2008)
21. L. Wang, S. M. Mostafavi, Y.S.H., Yoon, K.J.: Event-based High Dynamic Range Image and Very High Frame Rate Video Generation using Conditional Generative Adversarial Networks (2019)
22. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4681–4690 (2017)
23. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (2008)
24. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. International Conference on Machine Learning (ICML) (2015)
25. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)
26. Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1429–1437 (2019)
27. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. International Journal of Computer Vision (IJCV) **126**(12), 1381–1393 (2018)
28. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. International Journal of Computer Vision (IJCV) **126**(12), 1381–1393 (2018)
29. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4500–4509 (2018)
30. Rebecq, H., Gehrig, D., Scaramuzza, D.: ESIM: an open event camera simulator. Conf. on Robotics Learning (CoRL) (2018)
31. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
32. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (NIPS). pp. 2234–2242 (2016)
33. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: Asian Confence on Compututer Vision (ACCV) (2018)
34. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI Conference on Artificial Intelligence (AAAI) (2016)

35. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: IEEE International Conference on Computer Vision (ICCV). pp. 1456–1465 (2019)
36. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: IEEE International Conference on Computer Vision (ICCV). pp. 4068–4076 (2015)
37. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision (ECCV) (2018)
38. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: European Conference on Computer Vision (ECCV). pp. 117–132 (2018)
39. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International Conference on Computer Vision (ICCV) (2017)