Low Light Video Enhancement using Synthetic Data Produced with an Intermediate Domain Mapping

Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh

Huawei Noah's Ark Lab

{danaitri22, sean.j.moran}@gmail.com
{steven.mcdonagh, sarah.parisot, gregory.slabaugh}@huawei.com

Abstract. Advances in low-light video RAW-to-RGB translation are opening up the possibility of fast low-light imaging on commodity devices (*e.g.* smartphone cameras) without the need for a tripod. However, it is challenging to collect the required paired short-long exposure frames to learn a supervised mapping. Current approaches require a specialised rig or the use of *static* videos with no subject or object motion, resulting in datasets that are limited in size, diversity, and motion. We address the data collection bottleneck for low-light video RAW-to-RGB by proposing a data synthesis mechanism, dubbed *SIDGAN*, that can generate abundant dynamic video training pairs. SIDGAN maps videos found 'in the wild' (*e.g.* internet videos) into a low-light (short, long exposure) domain. By generating dynamic video data synthetically, we enable a recently proposed state-of-the-art RAW-to-RGB model to attain higher image quality (improved colour, reduced artifacts) and improved temporal consistency, compared to the same model trained with only static real video data.

1 Introduction

Low-light imaging (less than 5 lux) is a challenging scenario for camera image signal processor (ISP) pipelines due to the low photon count, low signal-to-noise ratio (SNR) and profound colour distortion [6]. The ISP is responsible for forming a high-quality RGB image with minimal noise, pleasing colors, sharp detail, and good contrast from the originally captured RAW data. Recently there has been growing research interest in end-to-end deep neural network architectures for modelling the entire ISP pipeline, both in well-lit [40] and low-light scenarios [6].

A major bottleneck in the learning of deep models for end-to-end RAW-to-RGB mapping is the availability of data. Existing models require a large amount of manually collected paired data (RAW sensor data and its corresponding RGB image) for training. However, collecting suitable amounts of paired data is often time consuming, error prone (*e.g.* misaligned pairs), and expensive. Chen *et al.* [5] resort to using a tripod to collect static videos for training. In [20] a

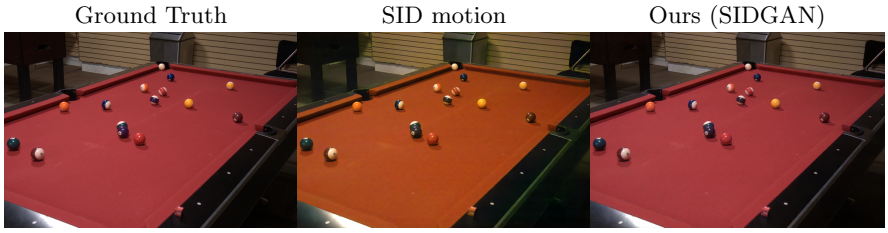

Fig. 1: Comparing the image quality with SID motion [20]. Training with synthetic data provides a more capable colour reproduction

novel optical system is designed to obtain dark and bright paired frames of the same scene simultaneously, but this rig is not publicly available, requires expertise to operate and only works if the scene is adequately illuminated. These challenges result in datasets that are limited in size, diversity in scene type, content and motion. This in turn typically produces models offering only limited colour reproduction and temporal consistency on real dynamic video (Figure 1 and Section 4). In this paper, we address the training data bottleneck for learning the RAW-to-RGB mapping, with a specific focus on low-light dynamic synthetic video data generation. Low-light video enhancement provides an ideal testbed for studying the potential of synthetic data as it is highly challenging to manually collect such data. In contrast to pre-existing work, we propose **S**eeing **I**n the **D**ark **GAN** (SIDGAN), a synthetic low-light video data generation pipeline leveraging Generative Adversarial Networks (GANs) [14].

GANs have proved to be a powerful modelling paradigm for learning complex high dimensional data manifolds for many types of real-world data such as natural images. The data distribution is modelled by framing learning as a competition between a generator network and a discriminator network. The generator attempts to produce samples from the desired data distribution that are as realistic as possible such that the discriminator network is fooled into classifying the synthetic samples as being real. The ensuing minimax game between the two networks can lead to a generator network that produces realistic samples from the data manifold. SIDGAN builds on the CycleGAN work of Zhu *et al.* [50] who demonstrate how to learn an unpaired mapping between two disparate domains (*e.g.* two sets of images with different styles). However, different to their work, we extend the mapping to three domains using a pair of CycleGANs while leveraging a weak supervisory signal in the form of an *intermediate* domain that has a paired data relationship with one of the remaining two domains. We argue that for an effective mapping between two domains that are very distant (*e.g.* internet videos and short exposure frames from a completely different sensor), that it is best to leverage an intermediate domain. Our approach is illustrated in Figure 2.

Our main contributions are three-fold:

- **Semi-supervised dual CycleGAN with intermediate domain mapping**: Mapping directly from internet videos (Figure 2, domain A) to short exposure (domain C) is difficult due to the large domain gap and lack of paired training examples. Instead, we bridge the gap using an intermediate long exposure domain (domain B) for which we have paired data (between domains B and C). This decomposes a difficult problem into two simpler problems, the latter with supervision.
- **Data abundance for RAW-to-RGB video**: The Dual CycleGAN allows synthesis of abundant video data in order to train high capacity models with, typically unavailable, dynamic and domain specific paired training data.
- **A practical strategy to combine synthetic and real data**: We propose an effective three-step training and fine-tuning scheme to address the remaining domain gap between synthetically generated and real video data. Combining our dynamic synthetic data with static real data yields a forward RAW-to-RGB video model with superior temporal consistency and colour reproduction compared to the same model trained with only real data.

2 Related Work

Low-light image and video quality enhancement topics are closely related to our contributions. In addition to these areas, we briefly review intermediate domain mappings, synthetic image generation and learning with unpaired data.

Low-light image enhancement. A large body of work exists on low-light image enhancement, spanning histogram equalization (HE) techniques [17, 3, 34] and approaches grounded in Retinex theory [25, 21, 22, 49]. Classical enhancement methods often make use of statistical techniques that typically rely on strong modeling assumptions, which may not hold true in real world scenes or scenarios. Deep learning techniques have also been readily applied to low-light image enhancement in recent years. The work of LLNet [28] employed an autoencoder towards low-light image denoising. Further convolutional works have used multiscale feature maps [45] and brightness transmission image priors [44] to enhance image contrast with strong qualitative results *c.f.* classical approaches.

Low-light video enhancement. The video enhancement problem is more recent and has received comparatively less attention. Analogous to static images, statistical Retinex theory has also been applied to video [27, 47]. Framing the problem in a joint-task setting was investigated by [23]; coupling low-light enhancement with denoising. Network based learning is also considered for video; Lv *et al.* [30] propose a multi-branch low-light enhancement network, applicable to both image and video domains. As earlier highlighted, learning-based mapping of (low-light) RAW-to-RGB work is highly relevant for our direction; Chen *et al.* [6, 5] learn this transformation considering both images and, latterly, video.

Capture of real video data in this problem setting is prohibitively expensive. However, as noted, systems have been proposed that can capture both bright and dark videos of identical scene content, providing training pairs for low-light video models. Jiang *et al.* [20] collected data and employed a standard CNN to learn

enhancement mappings for the transformation from low-light raw camera sensor data to bright RGB videos. Collected data was relatively small by deep learning standards (179 video pairs), illustrating the arduous burden of real-world video collection in scenarios that involve complex capture setups and custom hardware (here beam splitters and relay lenses). Uncommon specialist hardware, operator expertise requirements and support for (only) adequately illuminated scenes can be considered the main disadvantages of video enhancement work that depends exclusively on real-world data.

Intermediate domain mappings. The concept of harnessing intermediate domain bridges can be considered powerful and related strategies have been employed in a number of scenarios [13, 43, 26, 15, 8]. In addition to visual domains, evidence in support of the broader applicability of this family of strategies is also found in machine translation tasks [10], where intermediate domains enabled extension of bilingual systems to become multilingual. Relevant synthetic data work [7], that we draw from, leverages chains of image mappings ("indirect-paths") to gain a supervisory signal towards improving super-resolution. Combining intermediate domain mappings with synthetic data offers a promising direction for problem domains where acquisition of paired imagery, and therefore direct supervisory signal, is challenging.

Synthetic data augmentation. The use of synthetic data for model training and testing can be considered popular and datasets have been created for a multitude of image processing and computer vision problems [9, 32, 37, 11]. Early work performed successful scene text recognition with simplistic data generation [19] and, more recently, the benefits of combining synthetic data with Generative Adversarial Networks (GANs) [14] have been actively explored.

The work of [51] explores GAN data augmentation, generating artificial images using conditional Generative Adversarial Networks (cGANs). By conditioning on segmentation masks, realistic images were generated for their task (leaf segmentation), and related performance improved by $\sim$16% *c.f.* without synthetic augmentation. In [12] a semi-supervised adversarial framework is used to generate photorealistic facial images. By introducing pairwise adversarial supervision, two-way domain adaptation is constrained using only small paired real (and synthetic) images along with a large volume of unpaired data. Performance improves, due to the synthetic imagery, and consistently betters that of a face recognition network trained with Oxford VGG Face data. In [46] both paired and unpaired training data is utilised simultaneously in conjunction with generative models. Two generators and four discriminators are employed in a hybrid setting and qualitatively strong results, on multiple image-to-image translation tasks, are reported. Mixed and fully unsupervised approaches [29] begin to show great promise in faithfully generalising to real-world image distributions that are naturally sample-scarce or where data is otherwise hard to collect. These results motivate the use of inexpensive synthetic data for training GAN based tools. The need to collect large amounts of hand-annotated real-world data is avoided yet performance can surpass that of training with real-data exclusively.

In contrast to these successes, recent work [36] reports interesting findings when employing generative models (*e.g.* BigGAN), for data augmentation. Image classification error performance (Top-1 and Top-5) improves only marginally when additional synthetic data is added to an ImageNet training set. As GAN tools begin to be employed to aid downstream tasks, metrics that appropriately measure *downstream task performance* must be utilised *c.f.* solely evaluating synthetic sample image quality. Towards this our current work considers quantitative downstream performance evaluation, providing evidence towards the efficacy of our proposed data generation strategy (Section 4).

3 Learning the Low-Light Video RAW-to-RGB Mapping

Our objective is to learn short-to-long exposure mappings that provide accurate colour reproduction and temporal consistency. Given the lack of available real short, long exposure video pairs we propose a two-step approach that leverages *data synthesis*. The first step (Section 3.1), involves training a dual CycleGAN model for the purposes of data synthesis. The dual CycleGAN maps video frames to a domain characterised by short exposure images. A domain bridge (*e.g.* long exposure images) is used to regularise the mapping with available paired supervision. The trained CycleGAN permits videos 'from the wild' to be projected into the long and then short exposure domains, thereby generating the necessary paired supervision. Our second step, detailed in Section 3.2, utilises this synthetic data to train a forward model capable of mapping low-light video RAW to long exposure RGB. Finally, Section 3.3 provides details on synthetic data generation network architectures.

3.1 Synthetic Data Generation Using an Intermediate Domain

SIDGAN is modelled as a set of two CycleGANs in a dual configuration that learns the domain distributions for three domains; A, B and C. The model architecture is shown in Figure 2 and our domain B-C CycleGAN is shown in more detail in Figure 4. Domain A is characterised by a set of video frames defined by probability distribution p_A. The set of N videos from this domain $\{V_i\}_{i=1}^{N} \sim p_A$ are available for training. Similarly, we also consider M long exposure still images $\{L_i\}_{i=1}^{M} \sim p_B$ (domain B) and T short exposure still images $\{S_i\}_{i=1}^{T} \sim p_C$ (domain C). The remainder of this section details how the sample sets are leveraged in order to learn a mapping from domain A to domain C via bridge domain B.

The A-B CycleGAN learns, in a conditional GAN fashion, an *unpaired* mapping between domains A and B, transforming a set of RGB videos to a domain characterised by a set of RGB images (*i.e.* long exposure images) using generators G_{AB} and G_{BA}. This unsupervised CycleGAN does not require explicit sample pairings. Discriminator D_A attempts to distinguish generated video frames $\hat{V}{=}G_{BA}(L)$, $L \sim p_B$ from real video frames drawn from the input distribution

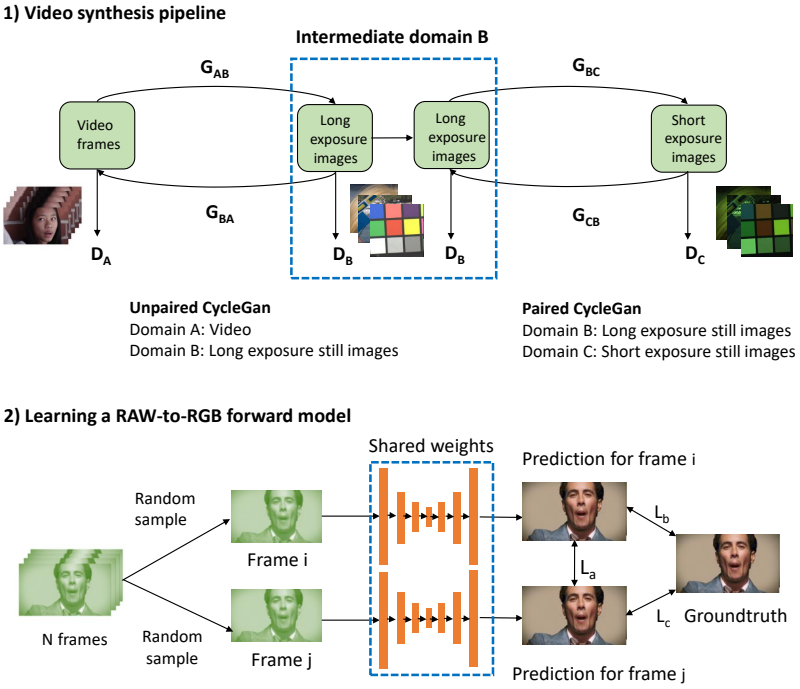

Fig. 2: **Step 1:** We use SIDGAN generators (G_{AB}, G_{BC}) to map Vimeo videos (domain A) into the long (domain B), then short (domain C) exposure domains, giving our synthetic training dataset. **Step 2:** The forward model can be very different from the generators of SIDGAN *e.g.* leveraging a mechanism for exploiting the temporal domain in the synthetic video data

$V \sim p_A$ in Domain A (Equation 1). Discriminator D_B tries to differentiate synthetic long exposure still images $\hat{L} = G_{AB}(V)$, $V \sim p_A$ from real long exposure images $L \sim p_B$, drawn from Domain B (Equation 2).

$$\mathcal{L}_{GAN}(G_{BA}, D_A) = \mathbb{E}_{L\sim p_B}[\log(1 - \log(D_A(G_{BA}(L))))] + \mathbb{E}_{V\sim p_A}[\log(D_A(V))] \quad (1)$$

$$\mathcal{L}_{GAN}(G_{AB}, D_B) = \mathbb{E}_{V\sim p_A}[\log(1 - \log(D_B(G_{AB}(V))))] + \mathbb{E}_{L\sim p_B}[\log(D_B(L))] \quad (2)$$

We regularise the mappings between domains such that G_{AB}, G_{BA} are approximate inverses of one another by employing a cycle consistency loss [50] (Equation 3):

$$\begin{aligned}\mathcal{L}_{cyc}(G_{AB}, G_{BA}) = \mathbb{E}_{L\sim p_B}\|[G_{AB}(G_{BA}(L)) - L]\|_1 + \\ + \mathbb{E}_{V\sim p_A}\|[G_{BA}(G_{AB}(V)) - V]\|_1.\end{aligned} \quad (3)$$

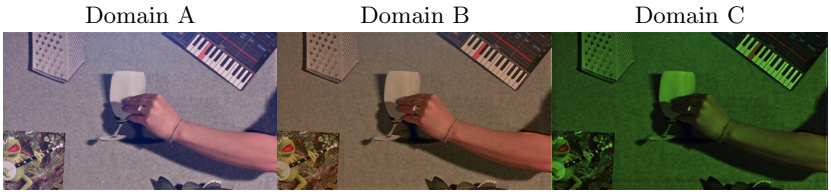

Fig. 3: Generating synthetic long and short exposure frame pairs in two steps: **Step 1**: Project videos from domain A to our sensor-specific long exposure domain B using generator G_{AB}. **Step 2:** Project the translated image from step 1 to the sensor-specific short exposure domain C using generator G_{BC}

Following [42, 50], we find it also important to add an identity loss $\mathcal{L}_{identity}$ in order to prevent colour inversion:

$$\mathcal{L}_{identity}(G_{AB}, G_{BA}) = \mathbb{E}_{V\sim p_A}\|[G_{BA}(V) - V]\|_1 + \mathbb{E}_{L\sim p_B}\|[G_{AB}(L) - L]\|_1. \quad (4)$$

Our final loss combines the introduced individual loss terms as a weighted combination, with individual components weighted by hyperparameters λ_1,λ_2 (Equation 5):

$$\begin{aligned}\mathcal{L}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{GAN}(G_{BA}, D_A) + \mathcal{L}_{GAN}(G_{AB}, D_B) \\ + \lambda_1\mathcal{L}_{cyc}(G_{AB}, G_{BA}) + \lambda_2\mathcal{L}_{identity}(G_{AB}, G_{BA}).\end{aligned} \quad (5)$$

In contrast to the domain A-B mapping, the B-C CycleGAN is *paired* (supervised) and employs generators G_{BC} and G_{CB}. This component of our dual CycleGAN model is responsible for mapping long exposure RGB images to short exposure counterparts. This domain mapping is paired as, in contrast to dynamic video, it is easier to collect short-long exposure pairs for still images by using a tripod and varying camera exposure time. SIDGAN leverages this supervision, using intermediate domain B, with the aim of enhancing the quality of the target task; mapping dynamic videos (domain A) to short exposure (domain C). The B-C CycleGAN component employs a loss (Equation 6), analogous to that of the domain A-B mapping, and additionally incorporates a $\mathcal{L}_{sup}$ term (Equation 7), harnessing the supervisory signal that is available.

$$\begin{aligned}\mathcal{L}(G_{BC}, G_{CB}, D_B, D_C) = \mathcal{L}_{GAN}(G_{CB}, D_B) + \mathcal{L}_{GAN}(G_{BC}, D_C) \\ + \lambda_1\mathcal{L}_{cyc}(G_{BC}, G_{CB}) + \lambda_2\mathcal{L}_{sup}(G_{BC}, G_{CB})\end{aligned} \quad (6)$$

Given a set of M short-long exposure pairs $\{(S_i, L_i)\}_{i=1}^{M}$, the supervised term $\mathcal{L}_{sup}(G_{BC}, G_{CB})$ is defined as:

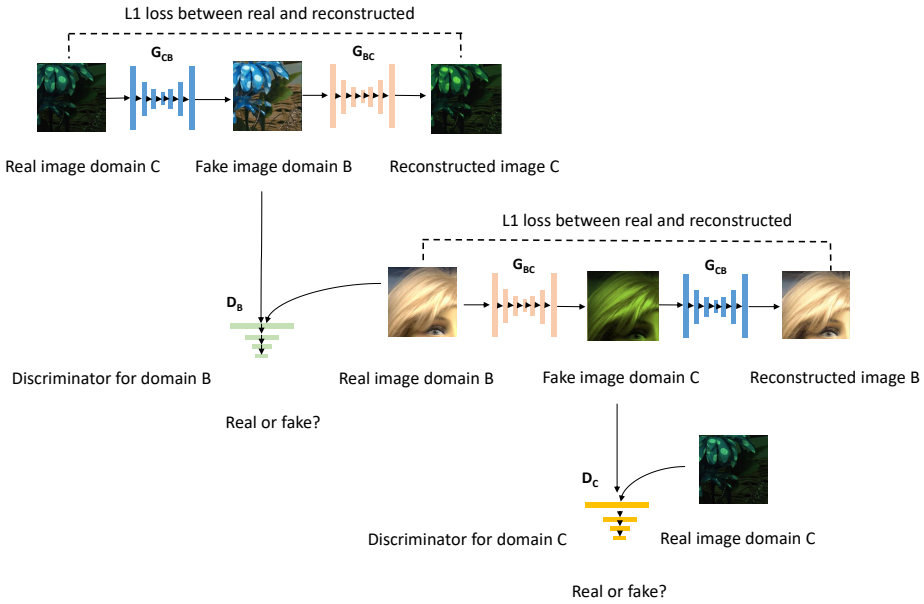

Fig. 4: Visualisation of our Domain B-C CycleGAN, a sub-component of the complete dual CycleGAN architecture found in Figure 2

$$\mathcal{L}_{sup}(G_{BC}, G_{CB}) = \mathbb{E}_{L\sim p_B}\|[G_{BC}(L) - S]\|_1 + \mathbb{E}_{S\sim p_C}\|[G_{CB}(S) - L]\|_1. \quad (7)$$

Our experimental evaluation (Section 4), demonstrates that a significant boost in translation quality is achieved when leveraging intermediate domain B to aid the weakly supervised CycleGAN mapping.

3.2 Training Low-Light RAW-to-RGB Forward Models

Our forward model training, fine-tuning schemes leverage a mixture of real and synthetic video data to learn a short-to-long exposure video mapping. We aim to extract an understanding of the correct colour and luminance distribution from real data (static video) while learning temporal consistency from the synthetic data (dynamic video). Our approach is shown in Figure 2. In the first step, synthetic data is generated by taking internet videos and passing them through generators G_{AB} and G_{BC}; using the process described previously (Section 3.1). In the second, step this synthetic video data is mixed with real data to train the forward model, adhering to the following three training and fine tuning steps:

1. **Training**: Train a forward model solely on real static video data
2. **Fine Tuning *a***: Fine tune solely on synthetic dynamic video data
3. **Fine Tuning *b***: Fine tune on real static video data

Our following experimental work (Section 4), employs a RAW-to-RGB forward model that follows the architecture of the SID motion model [5], reproduced in Figure 2. However, we note that our previously introduced synthetic data generation process is agnostic to specific model architectures. The model samples two frames from a static video and has three L_1 loss terms acting on the VGG features [41] of the two predicted frames ($\mathcal{L}_a$) and the two predicted frames and the groundtruth long exposure frame ($\mathcal{L}_b$, $\mathcal{L}_c$). As the training data is a static video there is no object and subject motion between frames, with noise being the only differentiator. We comment that while using our generator G_{CB} to model the short-to-long exposure mapping would also be possible, we instead leverage a temporal consistency term in the forward model to exploit the temporal dimension of the synthetic video generated using the dual CycleGAN (Section 3.1).

3.3 GAN Architectures for data generation

Generators G_{AB}, G_{BA}, G_{BC}, G_{CB} are modelled on the popular U-Net architecture [38]. In comparison to alternatives *e.g.* ResNet, we corroborate previous work [6] and find that the encoder-decoder architecture of the U-Net to be amenable to high-quality image translation. Nevertheless, we modify the components of the U-Net to further increase the quality of the produced images. Our final generators are comprised of 5 convolutional blocks with a stride of 1 followed by 2×2 max pooling layers. Upsampling is performed using a nearest neighbour bilinear interpolation followed by a 1×1 convolution, which we found important to reduce the prevalence of checkerboard (upsampling) artifacts.

The discriminators D_A, D_B, D_C are all PatchGAN discriminators [18] which attempt to penalize structure at the scale of patches by classifying them as real or fake. The discriminators ingest 192×192 patches which correspond to a receptive field that covers 75% of the input image. Finally, we note that CycleGANs in our dual CycleGAN setup are optimised independently. Joint training is theoretically possible but poses a more difficult optimisation problem and exhausted our available GPU memory in practice.

4 Experimental Results

4.1 Datasets and implementation details

We employed the Vimeo-90K dataset [48] to translate real-world videos into our low-light sensor specific domain. The dataset has 91,701 septuplet samples, each containing 7 video frames of resolution 448×255. For the sensor-specific long and short exposure domains (*i.e.* domains B and C), we use the Dark Raw Video (DRV) dataset [5], which contains 224 low-light raw video data and corresponding long-exposure images. Our intermediate domain B is represented by the long exposure DRV RGB images while for domain C we use the provided preprocessed DRV short exposure RAW video frames.

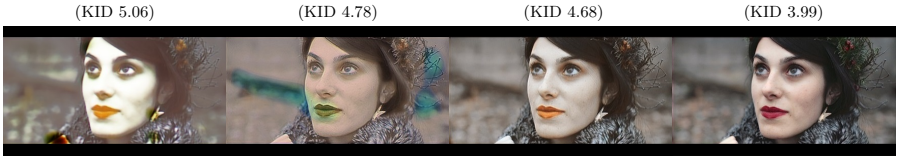

Fig. 5: Evolution of the KID distance during the unpaired training of CycleGAN$_{AB}$. The KID distance correlates well with visual quality and is used for model selection

Data is pre-processed using the pipeline of Chen *et al*. [5] which involves RAW-to-RGB conversion by averaging green pixels in each two-by-two block, black level subtraction, 2×2 binning, and global digital gain. Furthermore, noise is reduced using VBM4D [31] and pixel values are linearly scaled using exposure value (EV) difference. We resize the DRV long and short exposure RGB images such that resolution matches that of Vimeo-90K and normalize images in $[-1, 1]$. Experimentally we find that training on large patches is crucial in order to capture the global statistics and learn the correct white balance. For this reason, both CycleGAN$_{AB}$ and CycleGAN$_{BC}$ are trained using 256×256 crops corresponding to 50% of the resized DRV frames. We randomly select 400 Vimeo-90K videos and train our Dual CycleGAN, retaining the original train/val/test partitions of the DRV dataset. Finally, our forward RAW-to-RGB model is trained using the train partition of the DRV dataset and $9,366$ synthetic videos.

Models are implemented using Tensorflow and Keras [1, 2] and trained using an NVidia Tesla V100 GPU with 32GB memory. Our CycleGAN models are trained initially for 50 epochs with a learning rate 10^{-4} which then linearly decays for a further 20 epochs. Hyperparameters λ_1, λ_2 are found by empirical search and set to values 6.0, 6.0 in Equation 5, and values 10.0, 10.0 in Equation 6, respectively. The batch size is set to 1 and our forward model is trained using the training scheme described in Section 3.2 for a total of 1000 epochs. We employ a learning rate of 10^{-4} for the initial 500 epochs and reduce this to 10^{-5} for the latter half of training.

4.2 Synthetic data quality evaluation

We distinguish between the unpaired task that pertains to CycleGAN$_{AB}$ and the paired RGB-to-RAW mapping of CycleGAN$_{BC}$. Since CycleGAN$_{AB}$ is responsible for mapping videos from any source to our sensor-specific domain, no ground truth information is available for this task. In order to numerically evaluate generators G_{AB} and G_{BA} we adopt the following metrics: Fréchet Inception Distance (FID) [16] and Kernel Inception distance (KID) [4]. For CycleGAN$_{BC}$, we use the available ground truth; long and short exposure pairs of the test partition (DRV dataset) and evaluate performance using standard metrics; Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM).

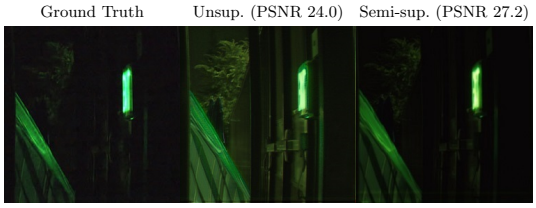

Fig. 6: Comparing CycleGAN with the proposed semi-supervised CycleGAN. Our semi-supervised variant shows better translation performance by exploiting the ground truth information in the optimization objective

We observe experimentally that the KID correlates better than FID with the visual quality of the generated samples (see Fig. 5) and we base final selection of models G_{BA} and G_{AB} solely on KID score. Our generator G_{BC}, responsible for mapping long exposure (domain B) to short exposure (domain C), achieves $27.28dB$ PSNR and 0.88 SSIM and G_{CB} performance is $25.28dB$ and 0.74, respectively. Quantitative results allude to the fact that long exposure captures more photons and images better represent scene colors and contrast. Intuitively the problem can be regarded as more ill-posed when mapping in the short to long direction. We also observe by ablation that training without the supervised term (Equation 7) resulted in significantly lower performance ($\sim 4dB$ less), providing evidence in support of our choice to decompose the data synthesis task into two separate learning problems and exploit the available paired data via the intermediate domain mapping. Figure 7 provides example predictions for generators G_{CB} and G_{BC}. We compare our trained RAW-to-RGB forward model against state-of-the-art approaches for low-light image and video processing. Following [20], we evaluate the performance on the static videos of the DRV dataset and examine both the image quality and the temporal stability of our method.

4.3 Output image and video quality evaluation

Image Quality: for consistent comparison with previous work [5], we compare the fifth frame of our output video with the respective long exposure ground truth image and evaluate the performance in terms of average PSNR and SSIM over the 49 DRV test videos. We compare performance with recent methods SID [6] and SID motion [5] as well as common baselines that combine performant denoising algorithms (VBM4D [31], KPN [33]) with traditional non-learning based enhancement tools (here using Rawpy[1]). Results are summarised in Table 1. Baselines are observed to perform poorly for this challenging task. Our forward model, trained purely on synthetic data, achieves a PSNR of $21.53dB$ and SSIM of 0.70. We attribute this fairly weak performance to a well understood domain shift between synthetic and real data [35, 39]. However, we observe

[1] https://pypi.org/project/rawpy/

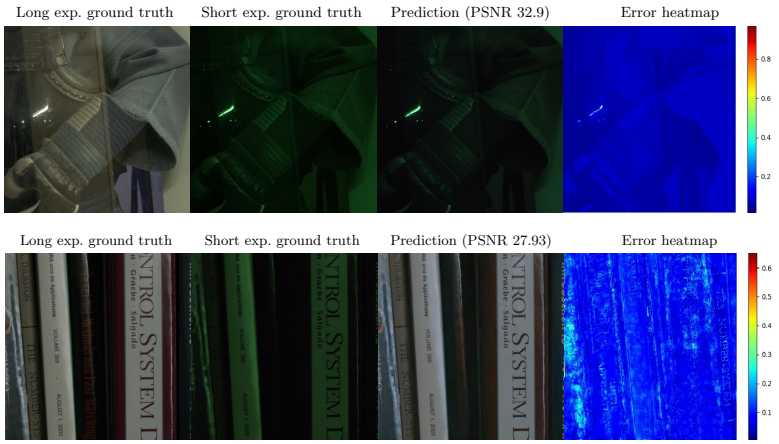

Fig. 7: First row: Mapping from domain B (long exposure) to domain C (short exposure) using generator G_{BC}. Second row: Mapping from domain C (short exposure) to domain B (long exposure) using generator G_{CB}

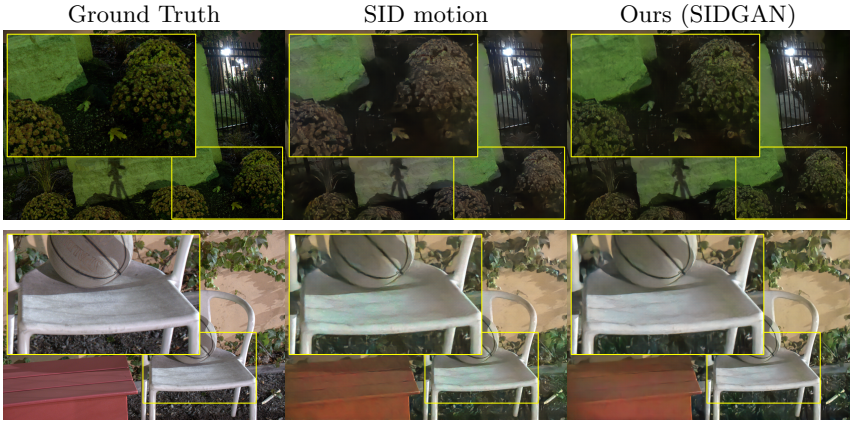

Fig. 8: Comparing the image quality with SID motion [5]. Note the improved colours in the marked regions

that training the model by adding a small fraction of real data (with a real : synthetic data ratio of $1:45$) successfully diminishes this domain gap yielding $28.94dB$ PSNR and 0.83 SSIM, constituting state-of-the-art performance on the DRV dataset.

Temporal Consistency: the DRV dataset contains static raw videos, thus temporal stability can be measured by computing temporal PSNR (TPSNR) and temporal SSIM (TSSIM) between pairs of consecutive frames, in similar fashion

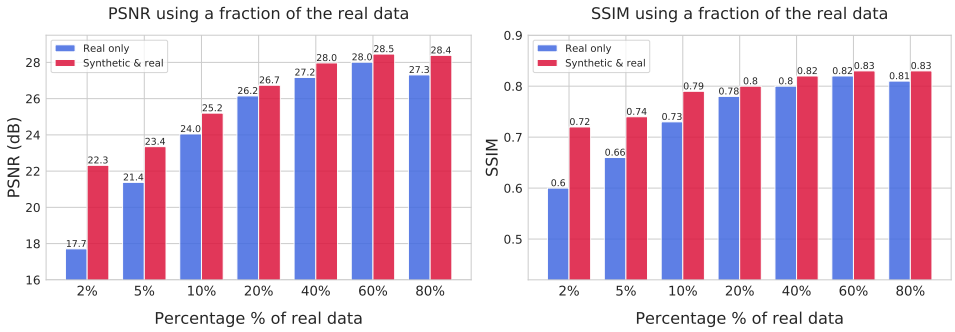

Fig. 9: The effect of real : synthetic training data ratios using portions of the DRV dataset. Left: PSNR, right: SSIM

Table 1: Output image quality on the DRV static dataset

Model	**PSNR↑**	**SSIM↑**
Input+Rawpy	12.94	0.165
VBM4D+Rawpy	14.77	0.315
KPN+Rawpy	18.81	0.540
SID w/o VBM4D	27.32	0.799
SID	27.69	0.803
SID Motion (real only)	28.26	0.815
SIDGAN (synthetic only)	21.53	0.704
SIDGAN (synthetic + real)	**28.94**	**0.830**

to [6]. Results are presented in Table 2. Our model offers competitive results when evaluated under these temporal metrics and we attribute strong performance to the extra information provided by our dynamic video synthetic data. We further evaluate dynamic video temporal stability by introducing synthetic training data, in a varying ratio with (scarce) real data. Average temporal warping error [24] is reported in Table 3. Largest improvements are observed when available real data is scarcest.

4.4 Real training data quantity and ratios

The addition of real image data was shown to help close synthetic training distribution domain gaps, resulting in quantitative improvements (Section 4.3). We further investigate the effect of adding real image data quantities in relation to synthetic data. Subsets of the DRV real dataset comprising 2%, 5%, 10%, 20% 40%, 60% and 80% are randomly sampled and model performance is evaluated when training solely on these real data subsets. We additionally train models on a set of 9,366 synthetic videos, generated by SIDGAN, and then fine-tune with the aforementioned real data subsets accordingly. All models are trained for

Table 2: Output video quality on the DRV static dataset

Model	**TPSNR↑**	**TSSIM↑**
SID [9] w/o VBM4D	33.72	0.939
SID	37.05	0.961
SID Motion (real only)	38.31	**0.974**
SIDGAN (synthetic + real)	**39.34**	0.966

Table 3: Dynamic video quality evaluation with varying real data ratios. Training the same model using increasing fractions of real data only (SID motion) and real data and synthetic data (SIDGAN).

Model	$E_{warp}\times 10^{-5}\downarrow$
SID Motion (2% real DRV data)	55.9
SID Motion (5% real DRV data)	54.3
SID Motion (20% real DRV data)	35.6
SID Motion (100% real DRV data)	29.3
SIDGAN (2% real DRV data + synthetic)	31.2
SIDGAN (5% real DRV data + synthetic)	32.7
SIDGAN (20% real DRV data + synthetic)	32.2
SIDGAN (100% real DRV data + synthetic)	**28.2**

1000 epochs using identical hyperparameters. PSNR and SSIM performance is reported in Figure 9. We observe that the addition of our synthetic data significantly boosts performance; increasing PSNR from 17.70 to 22.32, from 21.35 to 23.35 and from 24.04 to 25.19 for the cases of $2\%, 5\%$ and 10%, respectively. As the fraction of real data is increased, the gap in performance reduces indicating that the addition of our synthetic data again offers largest benefit in scenarios where real data is scarce.

5 Conclusions

We introduce **S**eeing **I**n the **D**ark **GAN** (SIDGAN), a data synthesis method addressing the training data bottleneck encountered when learning models for RAW-to-RGB problems. SIDGAN comprises two CycleGANs in order to leverage an intermediate domain mapping. Tasks that involve mapping between domains containing disparate appearance yet also lacking paired samples, can benefit from *intermediate domain* mappings that possess a paired data relationship with one of the original domains. We show that this strategy is capable of increasing the strength of the training signal and results in significant improvements for the investigated low-light RAW-to-RGB problem. Such tools may be widely applicable for domain mapping instances where data collection of directly paired samples between the domains of interest proves difficult or impossible.

References

1. Keras: Deep learning for humans. https://github.com/keras-team/keras
2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D.G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P.A., Vanhoucke, V., Vasudevan, V., Viégas, F.B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. CoRR **abs/1603.04467** (2016), http://arxiv.org/abs/1603.04467
3. Arici, T., Dikbas, S., Altunbasak, Y.: A histogram modification framework and its application for image contrast enhancement. IEEE Transactions on image processing **18**(9), 1921–1935 (2009)
4. Borji, A.: Pros and cons of GAN evaluation measures. CoRR **abs/1802.03446** (2018), http://arxiv.org/abs/1802.03446
5. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
6. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018)
7. Chen, S., Han, Z., Dai, E., Jia, X., Liu, Z., Xing, L., Zou, X., Xu, C., Liu, J., Tian, Q.: Unsupervised image super-resolution with an indirect supervised path. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
8. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X., Li, X.: Flowing on riemannian manifold: Domain adaptation by shifting covariance. IEEE transactions on cybernetics **44**(12), 2264–2273 (2014)
9. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
10. Escolano, C., Costa-jussà, M.R., Fonollosa, J.A.R.: Towards interlingua neural machine translation. CoRR **abs/1905.06831** (2019), http://arxiv.org/abs/1905.06831
11. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
12. Gecer, B., Bhattarai, B., Kittler, J., Kim, T.K.: Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 217–234 (2018)
13. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

15. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: 2011 international conference on computer vision. pp. 999–1006. IEEE (2011)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR **abs/1706.08500** (2017), http://arxiv.org/abs/1706.08500
17. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Transactions on Consumer Electronics **53**(4), 1752–1758 (2007)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016)
19. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. In: Workshop on Deep Learning, NIPS (2014)
20. Jiang, H., Zheng, Y.: Learning to see moving objects in the dark. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
21. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image processing **6**(7), 965–976 (1997)
22. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE transactions on image processing **6**(3), 451–462 (1997)
23. Kim, M., Park, D., Han, D.K., Ko, H.: A novel approach for denoising and enhancement of extremely low-light video. IEEE Transactions on Consumer Electronics **61**(1), 72–80 (2015)
24. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: European Conference on Computer Vision (2018)
25. Land, E.H.: The retinex theory of color vision. Scientific american **237**(6), 108–129 (1977)
26. Li, Y., Peng, X.: Learning domain adaptive features with unlabeled domain bridges. arXiv preprint arXiv:1912.05004 (2019)
27. Liu, H., Sun, X., Han, H., Cao, W.: Low-light video image enhancement based on multiscale retinex-like algorithm. In: 2016 Chinese Control and Decision Conference (CCDC). pp. 3712–3715. IEEE (2016)
28. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)
29. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. arXiv preprint arXiv:1909.09629 (2019)
30. Lv, F., Lu, F., Wu, J., Lim, C.: Mbllen: Low-light image/video enhancement using cnns. In: British Machine Vision Conference (BMVC) (2018)
31. Maggioni, M., Boracchi, G., Foi, A., Egiazarian, K.O.: Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. IEEE Transactions on Image Processing **21**, 3952–3966 (2012)
32. McDonagh, S., Klaudiny, M., Bradley, D., Beeler, T., Matthews, I., Mitchell, K.: Synthetic prior design for real-time face tracking. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 639–648. IEEE (2016)
33. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. CoRR **abs/1712.02327** (2017), http://arxiv.org/abs/1712.02327

34. Nakai, K., Hoshi, Y., Taguchi, A.: Color image contrast enhacement method based on differential intensity/saturation gray-levels histograms. In: 2013 International Symposium on Intelligent Signal Processing and Communication Systems. pp. 445–449. IEEE (2013)
35. Nowruzi, F.E., Kapoor, P., Kolhatkar, D., Hassanat, F.A., Laganiere, R., Rebut, J.: How much real data do we actually need: Analyzing object detection performance using synthetic and real data. arXiv preprint arXiv:1907.07061 (2019)
36. Ravuri, S., Vinyals, O.: Seeing is not necessarily believing: Limitations of biggans for data augmentation (2019)
37. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102–118. Springer (2016)
38. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015), http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a, (available on arXiv:1505.04597 [cs.CV])
39. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3752–3761 (2018)
40. Schwartz, E., Giryes, R., Bronstein, A.M.: DeepISP: Towards Learning an End-to-End Image Processing Pipeline. IEEE Transactions on Image Processing **28**(2), 912–923 (2019)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.1556
42. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=Sk2Im59ex
43. Tan, B., Zhang, Y., Pan, S.J., Yang, Q.: Distant domain transfer learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
44. Tao, L., Zhu, C., Song, J., Lu, T., Jia, H., Xie, X.: Low-light image enhancement using cnn and bright channel prior. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3215–3219. IEEE (2017)
45. Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., Xie, X.: Llcnn: A convolutional neural network for low-light image enhancement. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
46. Tripathy, S., Kannala, J., Rahtu, E.: Learning image-to-image translation using paired and unpaired training samples. In: Asian Conference on Computer Vision. pp. 51–66. Springer (2018)
47. Wang, D., Niu, X., Dou, Y.: A piecewise-based contrast enhancement framework for low lighting video. In: Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). pp. 235–240. IEEE (2014)
48. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. CoRR **abs/1711.09078** (2017), http://arxiv.org/abs/1711.09078
49. Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W.: A new low-light image enhancement algorithm using camera response model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 3015–3022 (2017)

50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
51. Zhu, Y., Aoun, M., Krijn, M., Vanschoren, J., Campus, H.T.: Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants. In: BMVC. p. 324 (2018)