Seeing Motion in the Dark Report

by

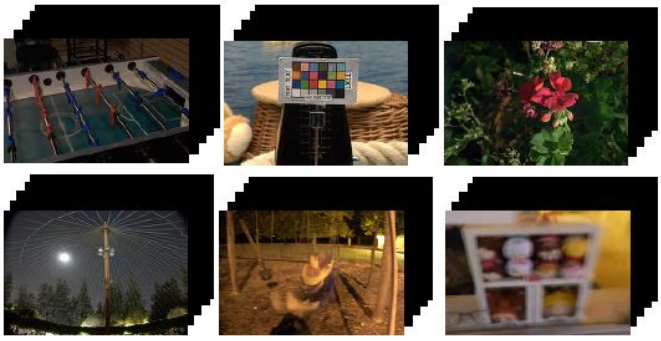
Sagi Shabtai

Shai Kimhi

18 March 2022

1. Introduction

Shooting video in low light is challenging because of low photon counting.
There are some physical solutions such as using a high ISO which can increase the brightness but also increases the amount of noise, or large aperture which is limited in consumer grade Cameras and mobile devices and the solution of flash changes the character of the scene and is problematic for video.
For images the method of long exposure times can work, but it is not suitable for dynamic video, as videos must be purchased at video rate.
Because of that, the solution on the paper refer to computational techniques for low light video processing.
In the article, the researchers deal with challenges beyond those presented by the individual low light images and try to receive good quality of output by deep processing of extremely weak light video and train a network with encourages temporary stability from the raw data to final sRGB output.
To study this problem, they shot 202 static raw videos with Sony RX100 VI camera, which can capture raw image sequences. Each set is equivalent to 5.5-second videos at 20 fps and has a long exposure reference image.



2. Related Work

The idea of creating a model that has the ability to generate visually compelling images from raw low-light images, in an **"end-to-end"** method instead of learning and improving specific sections of the image processing came from the " DeepISP: Towards Learning an End-to-End Image Processing Pipeline" article. In this article they presented a full end-to-end deep neural model of the camera image signal processing (ISP) pipeline called DeepISP. The model learns a mapping from the raw low-light mosaiced image to the final visually compelling image and encompasses low-level tasks such as demosaicing and denoising as well as higher-level tasks such as color correction and image adjustment.

A previous article by the same group of researchers, " Learning to See in the Dark", is heavily correlated to our article. There the idea was to train a deep network on a data set of raw short exposure and long exposure reference images (as ground truth) so that the web will study the image processing pipeline to maximize low-light imaging performance. However, these data sets contain static images scenes and burst and doesn't have a reference to data of video. In addition, in our paper they use the SID dataset and method from the " Learning to See in the Dark" for results comparison.

Another topic in the paper that has large amount of related work is **Single/Multiple-image denoising**. Most approaches are based on specific image priors such as smoothness, sparsity, low rank, or self-similarity. Learning-based methods further advanced performance in recent years, However, as demonstrated in the paper in Fig.5, approaching with those techniques to our problem, like temporal consistency, fail and cause visible artifacts.

In addition to single-image denoising, multiple-image denoising may achieve better results since more information is collected from the scene, for example Liu et al. and Hasinoff et al. propose a fast denoising method that produces a clean image from a burst of noisy images. However, video denoising is more challenging since every frame needs to be processed for the output video instead of only one resulting image in the method referring to burst denoising.

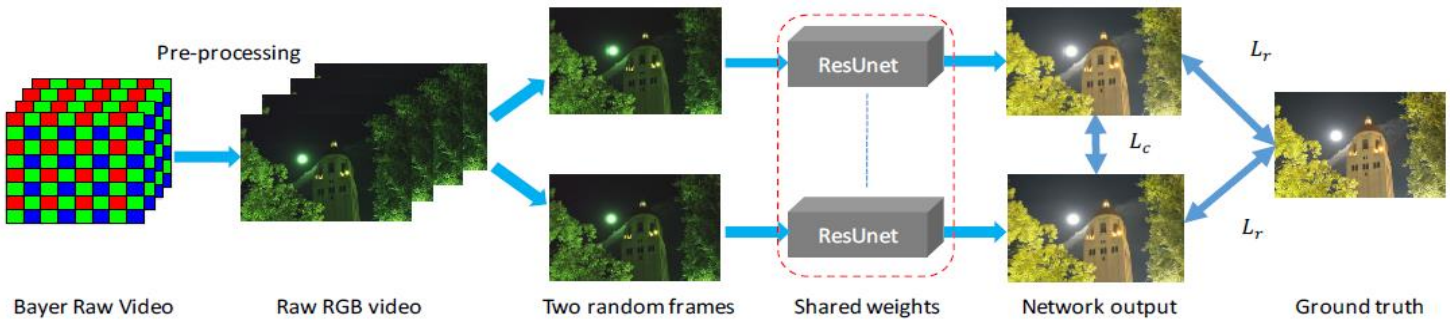
State-of-the-art video denoising methods according to the paper are both spatial and temporal. The spatial and temporal correlations should be utilized to reduce noise. The algorithms for video Spatial and temporal denoising in the paper include **VBM4D** and **KPN**, which rely on grouping similar patches and jointly filtering them to form the result.

3. Paper review

For the experiment the researchers collected a new Dark Raw Video (DRV) dataset. They used a Sony RX100 VI camera, which can capture raw image sequences at approximately 16 to 18 frames per second in continuous shooting mode, and the buffer can keep around 110 frames in total, which equivalent to 5.5-second videos at 20 fps. The resolution for the Bayer image is 3672 X 5496. The collection divided into two videos-sets, one for static videos with corresponding long-exposure ground truth, and the other contains dynamic videos without ground truth, under the assumption that the model trained on static videos can generalize to some extent to dynamic videos. The static videos set randomly divide them into approximately 64% for training, 12% for validation, and 24% for testing.

After collecting the data, they analyze the noise distribution guided by the work presented at "Burst denoising with kernel prediction networks". They show the distribution of the real data, compared to the distribution of the synthetic noise model corresponding to the related paper. The distributions are estimated via Parzen–Rosenblatt window estimator and the analysis shows that the real low-light data is severely biased, in part due to clipping and quantization. For example, there is a peak at zero because many sensor readings are too weak and are quantized to zero.

The researchers method of tackling the low light-video problem was to uses a deep network, trained by receiving raw RGB frames and comparing them to the reference ground-truth image.



The static/dynamic videos raw Bayer dataset was preprocessed once by Bayer to raw RGB conversion, black level subtraction, binning, global digital gain and noise reduction executed by the VBM4D algorithm which doesn't rely on training data and show good result in the "Non-local sparse models for image Restoration" paper.

The preprocessed raw RGB frames are fed to the "end-to-end" deep network by taking two frames of the same video sequence and send them through ResUnet structure with 16 residual blocks of Unet in a Siamese model of two parallel ResUnet structures with shared weights.

The Loss-function $\mathcal{L}$ defines by the 2 received frames, $\hat{Y}^1, \hat{Y}^2$,and the ground-truth image Y^*:

$$\mathcal{L}(\hat{Y}^1, \hat{Y}^2, Y^*) = \mathcal{L}_r + \mathcal{L}_c,$$

where $\mathcal{L}_r$ is referred to as the the recovery loss and $\mathcal{L}_c$ is called the self-consistency loss. They are defined as follows:

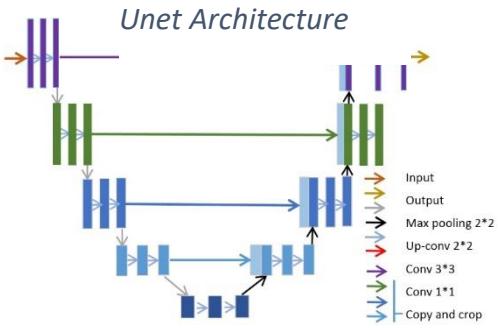
$$\mathcal{L}_r = \sum_l \frac{1}{N^l} \sum_{k=1,2} \|\Phi^l(Y^*) - \Phi^l(\hat{Y}^k)\|_1$$

$$\mathcal{L}_c = \sum_l \frac{\lambda}{N^l} \|\Phi^l(\hat{Y}^1) - \Phi^l(\hat{Y}^2)\|_1.$$

Here Φ^l represent the VGG features at the l-th layer and N^l is the number of such features, that used to increase the depth of the network using an architecture with very small (3×3) convolution filters, which prove to get improvement on image processing networks. Lambda refer as a regularization to the self-consistency loss and set to 0.05.

The initial learning rate was 10^{-4} and is reduced to 10^{-5} after 500 epochs for total 1000 epochs.

Notice that the train stage used only static videos dataset under the assumption that the trained



4. Experiment & Results

For the experiment several methods and models were tested, some of them more traditional and some are more advanced and include state-of-the-art models and processing algorithms such as SID.

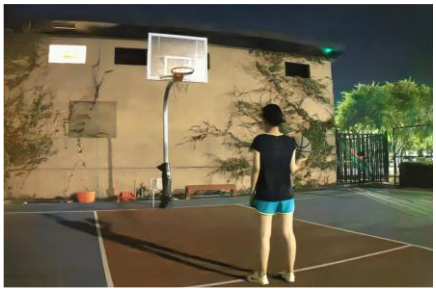
	PSNR (dB)	SSIM
Input+Rawpy	12.94	0.165
VBM4D [32]+Rawpy	14.77	0.315
KPN [35]+Rawpy	18.81	0.540
SID [9] w/o VBM4D	27.32	0.790
SID [9]	27.69	0.803
Ours	**28.26**	**0.815**
Ours w/o siamese	27.66	0.805
Ours w/o VBM4D [32]	27.62	0.803
Ours w/o both	27.26	0.793

For measurement they used PSNR and SSIM by comparing the 5th frame of the output video with the ground truth reference for each set of frames sequence, and then averaging the results.

As we shall see, the model presented in the paper surpasses the other methods, thus confirming the temporal preprocessing and the use of Siamese network can squeeze out better performance.

When comparing the result frames from each method they reveal that separating the processing stages on the low-light frames can lead to errors that are dragged and amplified throughout the processing. For example, the use of KPN denoising on the raw RGB leaves some residual noise that increased by the follows processing stages.

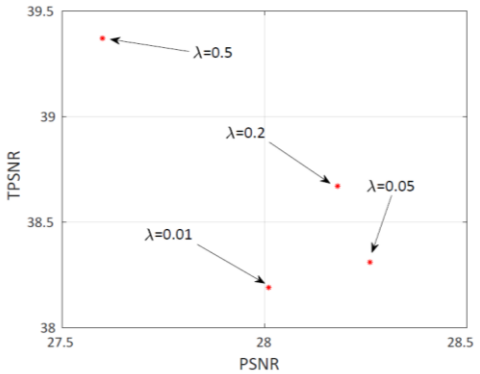
The article method is trained end-to-end to avoid such error accumulation.



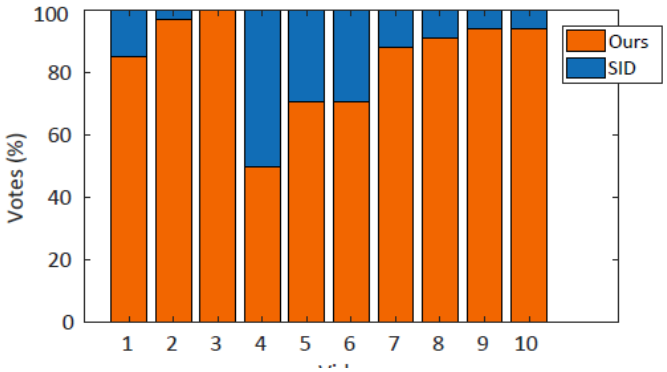
After that they tested the ability of their model to cope with video as a unit and not for specific frame (still static video). In this experiment they tested the model relative to the SID model and also tested the importance of the Siamese network architecture and the "end-to-end" theory on the outcome. In order to distinguish the temporal variants from single-image metrics, the measurement applied with temporal error on every pair of consecutive frames.
The results show that the new model has lower temporal errors than SID, and that the use of the Siamese network and "end-to-end" processing does indeed improve the model's abilities to overcome the low light exposure.

	TPSNR (dB)	TSSIM	TMAE ($\times 10^2$)
SID [9] w/o VBM4D	33.72	0.939	1.56
SID [9]	37.05	0.961	1.05
Ours	**38.31**	**0.974**	**0.89**
Ours w/o siamese	37.76	0.969	0.98
Ours w/o VBM4D	34.64	0.953	1.38
Ours w/o both	34.55	0.952	1.40

Additional test on the model was to change the value of λ from the loss function. The experiment found that a larger lambda leads to smaller temporary errors, but at the cost of lower spatial accuracy, which makes sense following the construction of the loss function as a combination of temporary and spatial error when a larger lambda amplifies the temporary error.



Finally, they wanted to see the generalization of the model on dynamic videos.
They first performed a perceptual evaluation that compares the results of the SID and the approach presented in the article to dynamic videos by asking 34 employees to indicate in which of the two videos is better quality blindly to the model.
The experiment was conducted on 10 randomly displayed video sequences as well, and as can be seen in the diagram below the employees showed a significant preference for the new model over the SID model, even over 90% in some cases.



One of the video sequences they used in the experiment was low light video of a birthday party(#10 video) from an iPhone X and the Sony RX100 VI camera. The iPhone video was captured using the auto mode. For the Sony video, they fixed the exposure time to 1/30 seconds while keeping the maximum aperture and ISO.
The raw image sequences for SID and our method were captured with ISO 2000 in continuous shooting mode.
Both SID and the new method were able to process the video such that the scene in the video has good visibility. However, the SID result suffers from both spatial and temporal artifacts, while our result is cleaner and more stable and 94% of the comparisons are in favor of the new method result.

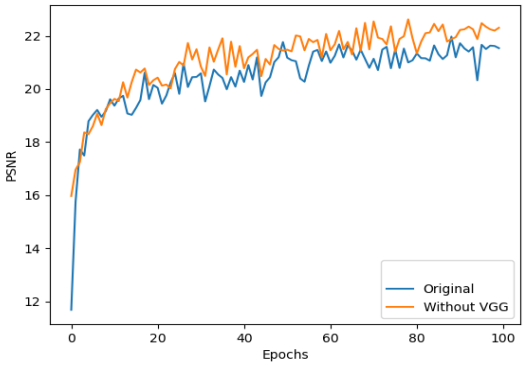

SID result


New method result

5. Ours additional work

In order to improve the performance and test the idea of the article a little more deeply, we changed the structure of the existing model in a number of ways and compared the results obtained in relation to the original model.

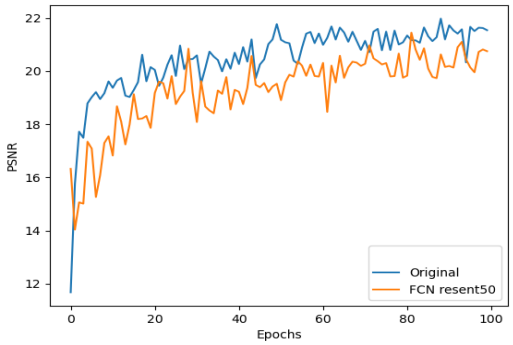
During the experiment we performed about 100 epochs per model.



First, we wanted to see the impact of using the VGG model to deepen the network as presented in the article, so we built a model that skips network output processing within the VGG architecture.

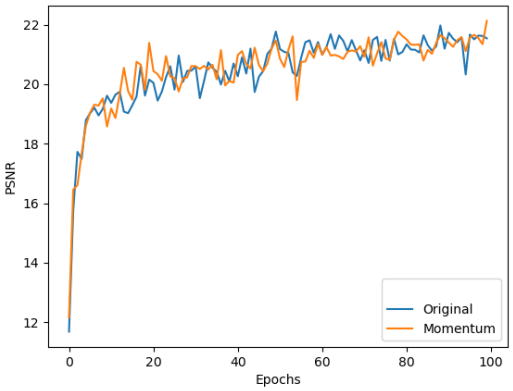
The PSNR comparison to the ground truth was better using MSE loss comparison instead of the paper's VGG fed model.

Next, we tested if using a different architecture then ResUnet pretrained on a different task as the Siamese network can squeeze out better performance. To do this we used Fully Convolutional Networks (FCN), FCN Resnet 50 pretrained on COCO train2017 segmentation dataset. In an article by Prof. Dr. Cem Gazioğlu it can be seen that a comparison between the two architectures (Unet and FCN) is not unequivocal and the superiority in performance varies depending on the processed information and the task of the model. In our case the better performance was by the Unet architecture from the original model and pretraining on segmentation dataset did not improve the performance as it has different properties than image enhancement and denoising.



After testing certain elements in the original model, we wanted to try to improve performance, so we added an aspect of momentum to the learning process as presented in the article "Momentum Contrast for Unsupervised Visual

Representation Learning". Momentum coefficient parameter initialized to be 0.999 because it assumes to be the optimal value by the article.



Input-Low Light Image

Original Method

Momentum

Without VGG

ResNet50 pretrained

GT

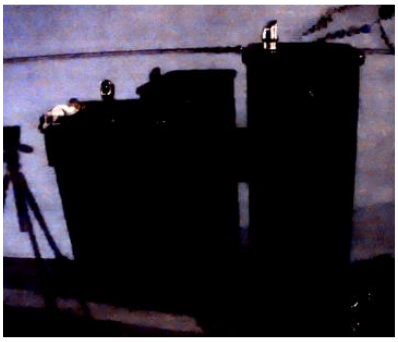
Like in the article we also measure the temporality of the results with the TPSNR measure in order to examine if the different models we tested can overcome the original model in terms of temporality like in the spatial measurement like the momentum model for example.

Model	*TPSNR*	*PSNR*
Original	31.2329	22.1291
Without VGG	29.4836	22.6250
FCN Resent50	30.5577	21.4462
Momentum	30.7537	21.97488
Momentum without VGG	28.9109	22.8015

As we can see in the table above the original model superior to all of our new models in terms of temporality, which is very important in processing videos, yet the momentum model achieves comparably good results in term of TPSNR and PSNR.

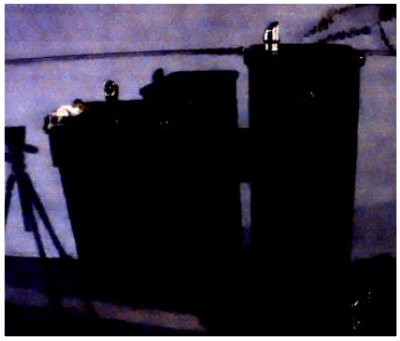
We can see from the images produces that the naïve loss approach achieves higher similarity to the Ground Truth yet produces a less smooth and more noisy images in comparison to the methods based on VGG models, the Segmentation pretrained model appear to produce more distortions which are probably caused by segmentation task ‘sacrificing’ information between similar and close pixels, which implies a reason that transfer learning from segmentation datasets don’t work in this domain.
The momentum method yields similar images to regular images with small differences when inspecting the images with slightly clearer and sharper details yet slightly more noise.

Input-Low Light Image

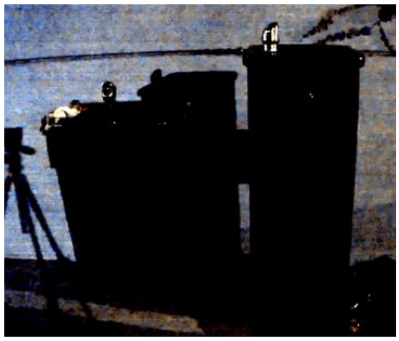


Original Method

FCN Resent50



Momentum



Without VGG

GT

Another experiment we wanted to perform was to try to discard the model presented in the article on another problem. Since the problem presented in the article is very specific and does not give much room for maneuver, we had a hard time finding a suitable idea, but finally we decided to try to see the model's dealing with the blur problem for burst images collections.

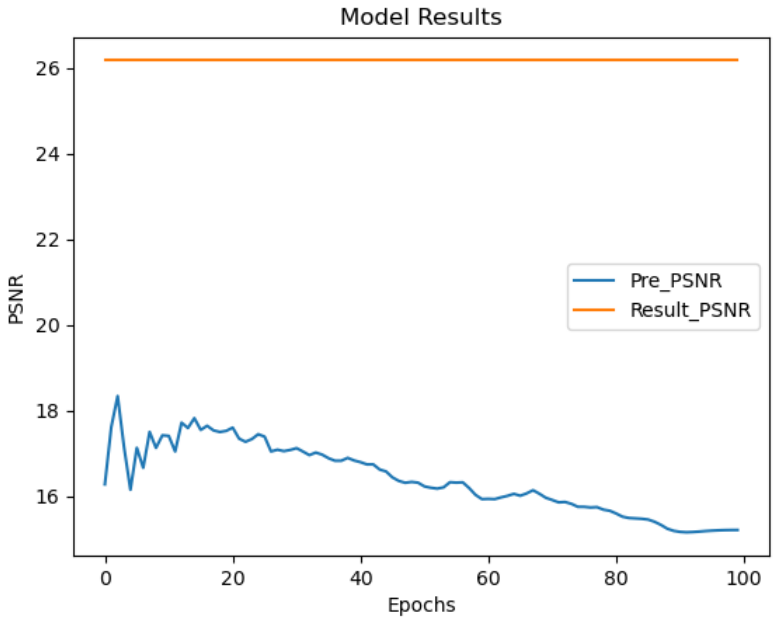
For this research we used Seungjun Nah's dataset, **GOPRO_Large dataset**, which contains blurry images created by generated and averaging varying number (7-13) of frames and the reference sharp image is the middle frame.

For training we used the train-sharp dataset by uniformity picking samples in 1/10 ratio. For each sample we randomly selected 2 frames from those that built the blurry image and fed them into the Siamese network of the model, here like the previous experiments we train the model along 100 epochs.

For each sample we randomly choice and for testing we used the test dataset by comparing the blur and sharp reference images. The average **Pre_PSNR** that we calculated was 26.17657.

we outputted the samples chosen for training and testing to CSV files:

Train_Samples.csv, Test_Pre_PSNR.csv



As can be seen in the graph above, the model failed to cope well with the task. It can be seen that the image that comes out of the model is more damaged than the one before processing and the longer you continue the more the average error increases.

The theory we think can explain this phenomenon is that our model is adapted to produce a temporal link on static images as training, so when given two frames that create an image it has created a larger error in willing to lower the difference between the frames. In addition, the order of the images used in burst collections as great importance and our use of random selection may not be appropriate here.

Works Cited

Branch, Taylor. “The Shame of College Sports.” Atlantic 308.3 (2011): 80-110. Print.

Dorfman, Jeffrey. “Pay College Athletes? They’re Already Paid Up To $125,000 Per Year.” *Forbes.com*. n.d. Web. 29 August 2013.

Griffin, Geoff. *Should College Athletes Be Paid?* Farmington Hills: Greenhaven Press, 2008. Print.

Mitchell, Horace, and Marc Edelman. “Should College Student-Athletes Be Paid?” U.S. News Digital Weekly 5.52 (2013): 17. Print.

Sack, Allen L., and Ellen J. Staurowsky. *College Athletes for Hire: The Evolution and Legacy of the NCAA’s Amateur Myth*. Westport: Praeger, 1998. Print.

Zimbalist, Andrew. *Unpaid Professionals: Commercialism and Conflict in Big-Time College Sports*. Princeton: Princeton UP, 1999. Print.

https://seungjunnah.github.io/Datasets/gopro