



A Spatio-Temporal Exposure Correction Neural Network for Autonomous Vehicle

Albano Borba
Federal University of Rio Grande
Rio Grande, Brazil
albano@furg.br

Johann Pires
Federal University of Rio Grande
Rio Grande, Brazil
johannpires@furg.br

Rafael Escher
Federal University of Rio Grande
Rio Grande, Brazil
rafaelescher@furg.br

Rodrigo de Bem
Federal University of Rio Grande
Rio Grande, Brazil
rodrigobem@furg.br

Paulo Drews-Jr
Federal University of Rio Grande
Rio Grande, Brazil
paulodrews@furg.br

ABSTRACT

Overexposed and underexposed digital images may occur either by excess or deficiency of lighting during acquisition. These problems are common in uncontrolled environments, specially affecting the visual sensory of autonomous robotic vehicles outdoors. Exposure correction is a traditional task in image processing and computer vision which has been tackled with deep learning techniques lately. In the realm of videos, the naive frame-by-frame application of such methods may cause exposure inconsistencies and flickering between subsequent frames. Therefore, we introduce the V2INet network, a spatio-temporal encoder-decoder architecture which aims for coherent exposure correction in videos. Our architecture synthesises a restored frame taking into account their temporal neighbours in the degraded video. We evaluate the approach on a real dataset acquired by an autonomous vehicle called Berkeley Deep Drive (BDD). We use the SSIM and PSNR metrics, and also introducing a surrogate measurement to assess temporal coherence. The V2INet has shown competitive results in both quantitative and qualitative comparisons with closely related methods, being capable of enforcing temporal consistency in the restored videos.

CCS CONCEPTS

• **Computing methodologies** → **Image processing; Vision for robotics; Neural networks;**

KEYWORDS

Exposure Restoration, Autonomous Vehicles, Video Processing, Convolutional Neural Networks, Deep Learning

ACM Reference Format:

Albano Borba, Johann Pires, Rafael Escher, Rodrigo de Bem, and Paulo Drews-Jr. 2021. A Spatio-Temporal Exposure Correction Neural Network for Autonomous Vehicle. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21)*, March 22–26, 2021, Virtual Event, Republic of Korea. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3412841.3441955>

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8104-8/21/03...\$15.00

<https://doi.org/10.1145/3412841.3441955>

1 INTRODUCTION

The restoration of digital images is a classical task in image processing and computer vision. It consists of recovering an image from a degraded observed version of itself [4]. In recent years, the widespread deep learning techniques have been employed to successfully tackle restoration problems such as noise suppression [47], resolution augmentation [41], scattering effects [11, 13, 38], and exposure correction [34, 40].

The exposure correction problem, focus of the present work, undertakes a degradation which rises mainly from limitations in the process of capturing digital images. Specifically, in environments with inadequate illumination, excess of light on the sensor can generate the saturation effect known as overexposure. On the other hand, in low light, the sensor may generate darker images, a process called underexposure [16].

The efficient restoration of badly exposed images has applications in many areas, particularly under uncontrolled lighting conditions, as in autonomous driving cars [3], outdoors mobile robots [31], unconstrained human pose estimation [8], and robotics [10, 35]. In many of these applications, videos are captured as the sensory data. However, most of exposure correction methods found in the literature approach the problem in a *frame-by-frame* manner [34, 40]. This strategy may generate temporal inconsistencies and flickering in the restored videos since subsequent frames are recovered independently and might present incoherent exposure variations. Besides that, the temporal information may also benefit the restoration process. Neighbouring frames in time may support the correct restoration of each other, as suggested by related works [45].

In this context, the main goal of the present work is to perform exposure correction in videos aiming to simultaneously preserve spatial and temporal coherence. This objective crucially involves the adoption of a strategy to incorporate the temporal information contained in image sequences. Regarding this particular aspect, we analyse related work in the literature and suggest a taxonomy dividing the approaches into direct, modular, and indirect techniques, as detailed in the next section. Considering this literature review, we developed a new deep network, dubbed V2INet which performs the restoration of badly-exposed videos. The proposed architecture, illustrated in Fig. 1, is based on the UNet model [30]. It is adapted to receive a sequence (temporal window) of degraded

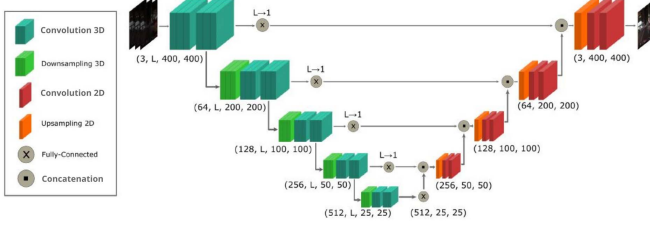


Figure 1: The V2INet architecture is composed of an encoder (green modules) and a decoder (red modules). The input data consists of a temporal window, from a video, which is used to synthesise each restored frame produced as output. In the encoder, max-pooling layers perform spatial downsampling. In the decoder, transposed convolutions [12] execute spatial upsampling. Temporal window’s feature maps at each encoder stage are compressed through fully-connected layers, which reduce them from L to 1 frame ($L \rightarrow 1$), where L is the temporal window size (number of frames). The compressed feature maps are concatenated to their counterparts at the correspondent stages of the decoder through skip-connections [30]. Further details in Sec. 3.

images as input and synthesise an image with its exposure corrected as output. We train and evaluate our model on the *Berkeley Deep Drive* dataset [44], obtaining competitive results in comparison with the state-of-the-art. Additionally to the standard image quality metrics, e.g. Structural Similarity Index Measure (SSIM) [42] and Peak-Signal-to-Noise Ratio (PSNR) [18], we also introduce a surrogate metric aiming to overcome the limitations in the literature regarding the evaluation of temporal coherence of exposure in videos. To the best of our knowledge, the proposed model is the first deep learning method to tackle the restoration of overexposed and underexposed videos in a spatio-temporal manner. In summary, our main contributions are:

- (1) The introduction of a simple taxonomy of deep learning methods for video processing.
- (2) The proposal of a new neural network model, dubbed *V2INet*, for exposure correction in videos aiming for spatio-temporal coherence among frames.
- (3) The achievement of competitive results on the *Berkeley Deep Drive* [44] dataset in comparison with the state-of-the-art methods and the introduction of a simple surrogate metric for evaluating temporal coherence of exposure in videos.

2 RELATED WORKS

2.1 Deep Convolutional Networks for Video Processing

Video processing is one relevant application of deep learning in computer vision and it has received more attention from the research community lately. In this context, some problems are inherent to videos, such as frame interpolation [21], frame forecasting [21], optical flow estimation [9], and events spotting [39]. In contrast, other problems are extensions of their still images counterparts to the domain of videos. In this case, temporal information between

frames constitutes an extra dimension which needs to be additionally considered by the deep learning models. Among these works, we can mention tracking-by-detection [22], semantic segmentation in videos [37], and noise suppression [43].

Aiming to assist in the understanding of the diverse strategies employed to tackle video processing with modern deep learning methods, we introduce a simple taxonomy focused mainly on deep convolutional neural networks. We divided the related works into three groups detailed below. Although most of the main deep learning methods for video processing tend to fit into one of these three groups, it is not an exhaustive categorisation of the literature.

2.1.1 Direct Approaches. These are methods that deal with temporal information as an additional input dimension. Such methods adopt tensors ($C \times L \times W \times H$) as inputs, with dimensions corresponding to images colour channels C , number of frames (time) L , width W , and height H , and apply three-dimensional (3D) convolutions to such volumes. These network architectures simultaneously learn spatio-temporal features relevant to solve tasks. This advantage is obtained at the cost of a significant increase in the number of the model’s trainable parameters, which additionally increase the demand for data and the computational cost for training.

For instance, Karpathy et al. [23] propose a three-dimensional convolutional network model with two channels assisting context learning for classifying videos from the Youtube-8M and the UCF101 datasets [2, 32]. Varol et al. [37] also use a 3D CNN for evaluating large temporal windows to recognize and classify actions on videos in the UCF101 and HMDB51 datasets [25, 32]. Tran et al. [36] introduce 3D convolutional kernels to perform classification across multiple datasets. In general, the approaches which employ this kind of direct strategy focus on tasks such as classification and detection.

2.1.2 Modular Approaches. Traditional 2D convolutions are used to extract features of each image individually and process them jointly in a distinct network. At this second module, responsible for the temporal analysis, some type of recurrent neural network is used, such as *Long Short-Term Memory* (LSTM) [17] or *Gated Recurrent Unit* (GRU) [7]. Although several methods for video processing adopt this strategy, the data dependency between the spatial and temporal modules may represent a bottleneck in the learning process. Moreover, the two-step process is also time-consuming.

Kang et al. [22] propose a modular architecture using CNNs for spatial features extraction and a recurrent LSTM neural network for processing temporal information, applied to the detection and tracking of objects in videos. Zhang et al. [46] present a model based on *Generative Adversarial Networks* (GAN) [14] to train a discriminator composed of LSTMs to detect key frames in a video to be summarized from dilated convolutions. Aafaq et al. [1] use a 2D convolution model connected through a dense layer to a set of GRUs to perform the description of videos. Although they achieve state-of-the-art performance in certain applications, this kind of approach is still limited in terms of performance and convergence.

2.1.3 Indirect Approaches. These methods adopt an implicit interpretation of time, using auxiliary inputs or filters for the 2D convolutional networks. In this case, low-level features, such as optical flow, are computed offline and used as inputs. CNNs are used

quite efficiently here since at runtime it is similar to the traditional convolutions. Although it creates the need for additional datasets to learn the auxiliary features.

For example, Jiang et al. [21] use an encoder-decoder network to perform frame interpolation in videos. The authors use as input to the network the bi-directional optical flow previously calculated in addition to the RGB frames. Zhu et al. [48] also use the optical flow as additional information to the network training process, focusing on object detection in videos.

2.2 Exposure Correction in Videos

Several techniques exist in the literature to correct the exposure of images and videos. For instance, modern digital cameras already have embedded luminance adjustment processes for image capturing, such as the High Dynamic Range (HDR) [27]. It aims for the optimal level of lighting through a composition of various images of the same scene captured with diverse levels of exposure. Another technique is based on manually adjusting *gamma* luminance component. However, these processes are not capable to generalise, depending on the scene and on its original exposure levels. Usually, they also rely heavily on human evaluation to reach acceptable results.

Considering this, there are more efficient and automated algorithms that are concerned with recovering information of an image to reconstitute it at a more appropriate exposure level. Some of them use classical deterministic methods based mainly on contrast analysis and histogram equalisation [28]. Recently, Chen et al. [6] take into account the analysis of HSV in the restoration process. Wang et al. [40] estimate an *image-to-illumination mapping* for modelling varying-lighting conditions. The illumination map lights up an underexposed image, improving not only its exposure but also its contrast, details, and colour. Mertens et al. [27] use the idea of composing images with different exposure levels and analysing factors such as saturation and contrast.

Regarding modern machine learning techniques, Chen et al. [5] achieves remarkable results with a simple convolution network using an elaborated pre-processing methodology. Radford et al. [29] employ GANs to train an unsupervised model that generates indoor images with adjusted exposure. Guo et al. [15] exempts any paired or unpaired data during the training phase by using designed non-reference loss functions. Steffens et al. [34] adopt dilated convolutions and a customised objective function, penalising poorly lit regions and achieving state-of-the-art results for still images. To the best of our knowledge, spatio-temporal deep learning methods for exposure correction in videos were not previously explored in literature.

3 METHODOLOGY

The main challenge for our method is to restore badly-exposed videos while simultaneously preserving temporal coherence. Aiming to achieve this goal, a frame in a video is recovered taking into account not only its degraded version but also its neighbours in time. We adopt a direct end-to-end approach. The video is scrolled with a temporal sliding window of size L , which corresponds to the number of frames used as input. Therefore, each temporal window consists of a target frame and $L - 1$ auxiliary ones, which

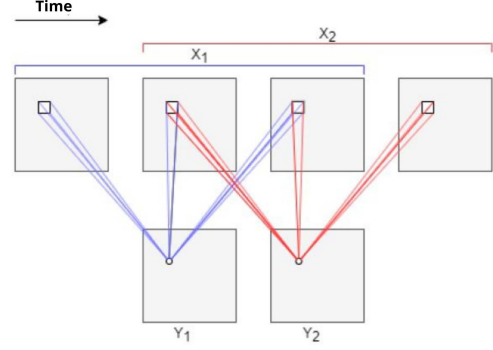


Figure 2: Illustration of a temporal sliding window of size $L = 3$ going through a sequence of frames. The X_n window is used as input to synthesise the Y_n frame.

represent the past and future states of the target scene. The extra contextual information tends to contribute to the restoration of the target frame [45]. See the illustration of the temporal sliding window process in Fig. 2.

3.1 Network Architecture

The aforementioned strategy requires an architecture that maps a three-dimensional input (time $\times$ width $\times$ height) to a two-dimensional output (width $\times$ height)¹. The proposed model is based on the well-established UNet architecture [30], an *image-to-image* network widely used for image segmentation. Our architecture, named V2INet, consists mainly of encoder and decoder blocks connected through a set of skip-connections which receives a sequence of frames from a video as input and synthesises corrected images as outputs, as illustrated in Fig. 1. The main characteristics of the V2INet architecture are detailed in the next sections.

3.1.1 Encoder. Let us consider a sequence of frames as input, represented by a tensor X of dimensions $C \times L \times W \times H$, corresponding to colour channels, number of frames, width, and height, respectively. The sequence X is input to the encoder, which is composed of a series of five stages, each one consisting of two 3D convolution layers [20], with $3 \times 3 \times 3$ kernels, and one batch normalisation layer [19]. Max-pooling layers are also applied from the 2nd to the 5th stages, each of them downsampling by half the spatial dimensions $W \times H$. The temporal dimension remains the same throughout the encoding process, being reduced only in the last layer of the encoder. At that point, the convolutional volume is compressed by a fully-connected layer to the dimension of a single image. In this way, the encoder module learns spatial-temporal features from the input sequence.

3.1.2 Decoder. The decoder receives a tensor of dimensions $C \times W \times H$. This module augments the dimension of the tensor through four stages, each one composed of two transposed convolutional layers [12], with 3×3 kernels, and one batch normalization layer [19]. Each transposed convolutional layer upsamples by two the tensor spatial dimensions $W \times H$. Unlike the original UNet decoder, we added transposed convolutions to allow the network to have

¹For simplicity, we neglect the number of channels which does not change.

trainable parameters also in the decoding process. By the end, a restored frame is obtained with the same spatial dimension as the input frames.

3.1.3 Skip-Connections and Activation Function. Following the structure proposed by [30], skip-connections between stages at the same level are adopted. This strategy preserves the structure and details of the input. In this process, temporal window’s feature maps from layers in the encoder are concatenated to their counterparts in the decoder, upon the required adjustments regarding their temporal dimension. For that, they are compressed through fully-connected layers, which reduce them from L to 1 frame ($L \rightarrow 1$), as shown in Fig. 1. Finally, the rectified linear unit (ReLU) is used as activation function throughout the architecture.

3.2 Learning Criteria

The output image $\tilde{Y}$ is compared the corresponding target image Y using an objective function. Given the complexity inherent in the exposure correction task, we seek for a learning criterion capable of measuring not only the mapping of structure and luminance between two images. Thus, our objective function is a composite of two other functions, following [33, 35]. The first function is the mean square error (MSE), which calculates the Euclidean distance between the pixel intensities of each image. The second function is the Structural Dissimilarity (DSSIM) [26], that calculates the dissimilarity between two images, taking into account a reference image. Our loss function is a weighted sum of both functions defined as:

$$\mathcal{F}(Y, \tilde{Y}) = \beta_0 \text{MSE}(Y, \tilde{Y}) + \beta_1 \text{DSSIM}(Y, \tilde{Y}), \quad (1)$$

with the scale factors β_0 and β_1 adjusted accordingly to the functions’ dimensions to balance their influence on the model.

4 EXPERIMENTAL RESULTS

In this section, we describe the dataset employed in our experiments, as well as the training process, the hyperparameters tuning, and the evaluation metrics. Finally, we compare our method with previous art by using quantitative and qualitative evaluation.

4.1 Dataset

To the best of our knowledge, there is not a dataset in the literature specifically designed for training models tackling the problem of exposure correction in videos. Naturally, capturing videos with the correct level of exposure along with their corresponding overexposed and underexposed versions is a costly effort if undertaken in an uncontrolled environment. Considering this, we synthetically change the exposure in the training data aiming to reproduce natural conditions of exposure.

The original videos for training and testing our network are from the *Berkeley Deep Drive* (BDD) dataset [44]. It contains videos acquired from drivers’ point of view of cars riding in a large variety of environments, weather conditions, and times in the day. The dataset consists of 100,000 videos, each one presenting approximately 1000 frames of dimension 1240×720 pixels. Given this expressive amount of data and aiming to understand how it is distributed according to luminance, we have built an exposure histogram. For each video, we compute the intensity mean, which is based on the mean luminance

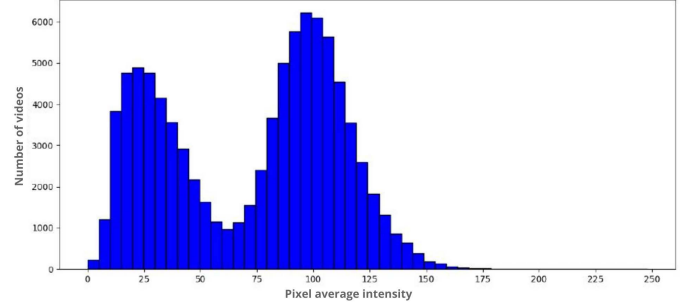


Figure 3: Average luminance histogram for the BDD dataset videos.

of its frames. As shown in Fig. 3, the data presents a bimodal distribution, in which intensity means around 100 correspond to daylight videos. On the other hand, videos with intensity means around 25 were acquired during the night or indoors. We also observed that the videos in the extreme of the histogram show specific situations, e.g. parking a car inside a garage.

From the analysis of the data, we randomly select a subset of 10,000 videos in which the average luminance ranges between 90 and 110, corresponding to a better exposure in comparison to the rest of the data. This subset is synthetically underexposed and overexposed and used as input data X during training, while the original frames Y are kept as ground-truth or the target frames for exposure correction.

We change the intensity of pixels by modifying the *gamma* curve to synthetically produce badly-exposed frames [33], defined as:

$$I_s = I_o^\gamma, \quad (2)$$

where I_o is the intensity of a pixel in the original image, and I_s is the corresponding pixel intensity in the synthetic badly-exposed image. With $\gamma = 1$, the original image is not changed. Values greater than 1 decrease the luminance of the image, causing it to darken simulating the effect of underexposure. Analogously, values smaller than 1 cause the effect of overexposure, generating a much brighter image. Thus, this methodology allows the precise control of the luminance level applied over the synthetic images. Following [34], the synthetic frames are generated by a random *gamma* value between 4, 6, and 8 for underexposure, and 1/4, 1/6, and 1/8 for overexposure. Fig. 4 illustrates the synthetic badly-exposed frames produced with this technique.

4.2 Training, Hyperparameters and Baselines

From our subset of 10,000 videos from the BDD dataset, 7,000 videos are used for training while 3,000 are for testing. We use the first 50 frames of each video which are spatially downscaled by 0.55 and centrally cropped ending up with 400×400 pixels. RGB values are normalised through min-max scaling and data augmentation is not employed. We alternate batch sizes between 2 and 6, varying it inversely to the temporal window size L . Given the particularities of the underexposure and overexposure effects, each problem is tackled by one instance of our architecture. Both identical instances are trained separately for restoring either overexposed or underexposed videos but following the same protocols. All variations of

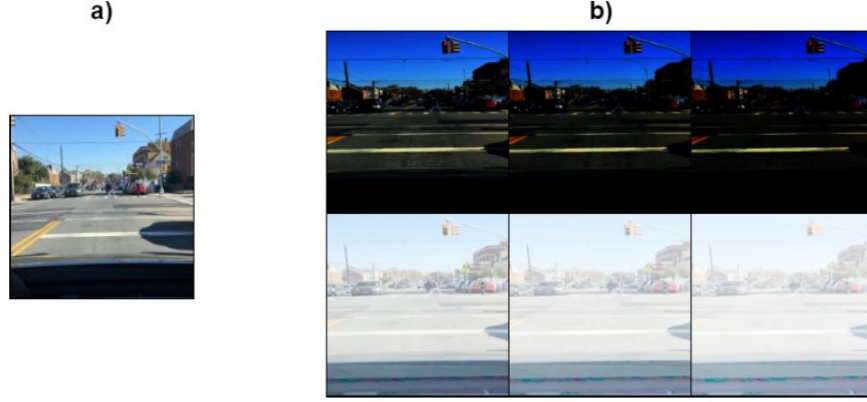


Figure 4: Sample frame from BDD with exposure modified synthetically through the *gamma* component. a) Original image. b) First row: underexposure effect, with γ respectively equals to 4, 6, and 8. Second row: overexposure effect, with γ respectively equals to 1/4, 1/6, and 1/8.

the models are implemented in the PyTorch framework and trained using Adam optimization [24] with default hyperparameters from the framework. Training each model takes approximately 5 days in a NVIDIA Titan X and it is stopped following the early-stop criterion based on the error stabilisation in the test set. OpenCV and the scikit-image libraries are employed for data preprocessing.

Regarding hyperparameters tuning, the most relevant task is to define the temporal window size L . We recall that the network input consists of a temporal window with L frames, composed of a target image to be restored and $L - 1$ auxiliary ones. Therefore, larger windows are expected to add more temporal information at the cost of runtime and memory consumption. Thus, it is desirable to explore the impact of the temporal window size on the exposure correction accuracy. To assess the performance of the architecture, four variations of the model are evaluated, using temporal windows of 3 and 7 frames for both, underexposure and overexposure correction. The results from all variations are shown in the next sections.

Finally, we adopt the UCan [34] and the U-Net [30] models as baselines. The UCan presents state-of-the-art results on exposure correction for still images, comparing their performance with several related works from the literature. The U-Net is the underlying model of our architecture and serves as a reference to the proposed modifications to deal with videos. The comparisons with the baselines are detailed in the following sections. They consider both spatial image quality and temporal coherence.

4.3 Spatial Evaluation

We adopt Peak-Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) metrics [42] for spatial image quality evaluation. Such metrics are widely used in the signal and image processing literature. The PSNR quantifies the difference between two images on a decibels scale, in which higher values represent higher similarities between them. The SSIM metric uses variances and covariances to analyse factors such as the luminance, contrast

distortion, and the loss of correlation between two images. This metric aims to produce a measurement closer to the human perception of a scene. Similar images present SSIM closer to one.

In Tab. 1, we show the mean and the standard deviation of the SSIM and PSNR scores for three synthetic exposure levels (4, 6 and 8) on our test set. We also show results for two different temporal window sizes (3 and 7). All the results are compared with our baselines, i.e. pre-trained and finetuned UCan models [34], as well as a finetuned U-Net architecture [30].

Table 1: Evaluation of the spatial metrics for the underexposure effect. The best results are highlighted in bold.

Gamma	Model	SSIM (mean)	SSIM (sd)	PSNR (mean)	PSNR (sd)
4	V2INet (3)	0.936	0.027	25.414	1.920
	V2INet (7)	0.948	0.025	25.981	2.500
	UCan original	0.610	0.107	15.614	2.150
	UCan finetuned	0.894	0.054	24.662	4.025
	U-Net	0.919	0.035	23.976	2.523
6	V2INet (3)	0.876	0.030	22.150	1.335
	V2INet (7)	0.900	0.027	24.812	1.529
	UCan original	0.611	0.099	15.656	2.028
	UCan finetuned	0.899	0.050	24.904	3.899
	U-Net	0.839	0.040	20.245	1.284
8	V2INet (3)	0.815	0.039	19.277	1.694
	V2INet (7)	0.846	0.033	21.556	1.427
	UCan original	0.605	0.104	15.536	2.159
	UCan finetuned	0.900	0.051	25.138	4.005
	U-Net	0.777	0.046	18.303	1.084

The results show a marginal improvement related to a larger temporal window since the window of size 7 allowed for higher SSIM and PSNR scores. Our V2INet model presents results compatible with the UCan after finetuning, generating better scores for levels 4 of *gamma* and values slightly below for level 8, with smaller standard deviation in both cases suggesting better temporal coherence. Regarding the exposure level 6, both models present very similar performances, alternating better scores of SSIM and PSNR with small differences. The results also validate our training methodology since finetuning and transfer learning were successfully performed with both the U-Net and the UCan models.

Table 2: Evaluation of the spatial metrics for the overexposure effect. The best results are highlighted in bold.

Gamma	Model	SSIM (mean)	SSIM (sd)	PSNR (mean)	PSNR (sd)
1/4	V2INet(3)	0.985	0.004	28.506	1.026
	V2INet(7)	0.988	0.006	30.489	1.893
	UCan original	0.771	0.051	14.170	1.413
	UCan finetuned	0.986	0.012	30.329	3.370
	U-Net	0.930	0.022	19.394	1.945
1/6	V2INet(3)	0.971	0.007	24.005	1.316
	V2INet(7)	0.987	0.004	28.506	1.026
	UCan original	0.769	0.052	14.176	1.412
	UCan finetuned	0.987	0.011	30.377	3.366
	U-Net	0.968	0.014	23.702	2.105
1/8	V2INet(3)	0.948	0.029	22.102	2.706
	V2INet(7)	0.988	0.005	28.056	1.288
	UCan original	0.769	0.052	14.176	1.412
	UCan finetuned	0.987	0.011	30.377	3.366
	U-Net	0.978	0.014	24.919	1.624

Analogously, Tab. 2 shows comparisons for the overexposure effect, with *gamma* values of 1/4, 1/6 and 1/8. In this experiment, overall results from all models are better, showing that the overexposure is less degrading on this dataset. Again, our V2INet model with temporal windows of size 7 and retrained UCan obtain the best SSIM and PSNR scores in all scenarios. Despite showing similar results, the UCan model presents standard deviations two to three times greater than our V2INet, possibly due to temporal incoherence caused by the frame-by-frame independent restoration.

4.4 Temporal Evaluation

The evaluating of the temporal coherence in the restored videos presents a certain difficulty. To the best of our knowledge, there is no specific metric in the literature to tackle such an issue. Thus, a surrogate metric is proposed for indirect assessment of temporal restoration. The experiments are carried out on the synthetically badly-exposed test set and compared with the original videos. Only the best performing models regarding the spatial metrics are tested, namely the V2INet with temporal window size of 7 and the finetuned UCan model.

Our surrogate metric seeks to measure the luminance oscillation between the frames of a video, evaluating the variation of the average intensity. It corresponds to the derivative in time of the average luminance of frames. In doing so, we expect to visualise the coherence of the restorations throughout the video and to check if there are significant discrepancies. Figs. 5 and 6 show the resulting variations for underexposure and overexposure, respectively. In both cases, the results show a significant change in the mean variation between the original (blue lines) and the synthetically exposed (red lines) videos, showing the impact caused by this degradation. Regarding the restored videos, similar values are obtained by the V2INet (yellow lines) and the UCan (green lines) models. Our network produces a curve slightly closer to the original sequence at some points, suggesting a good restoration of exposition.

4.5 Qualitative Results

In this section, we show qualitative results highlighting the temporal coherence of exposure levels in our restorations. In Fig. 7, we show a sequence of consecutive frames from a BDD dataset video sample. The original frames are shown in the top row, while the

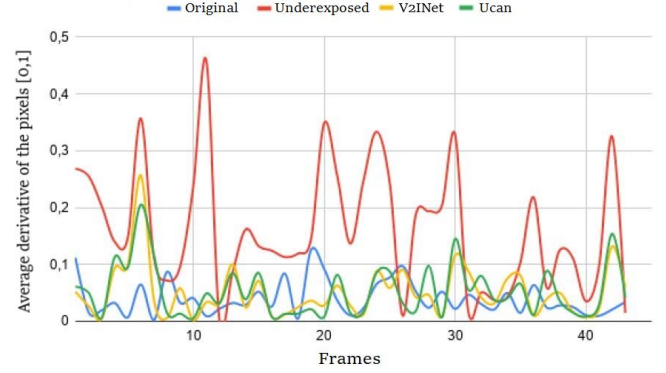


Figure 5: Variation of the mean luminance of frames throughout the test set of synthetically underexposed videos. We compare the following videos: originals (blue lines), synthetically overexposed (red lines), restored by our V2INet (yellow lines), and restored by the UCan [34] (green lines).

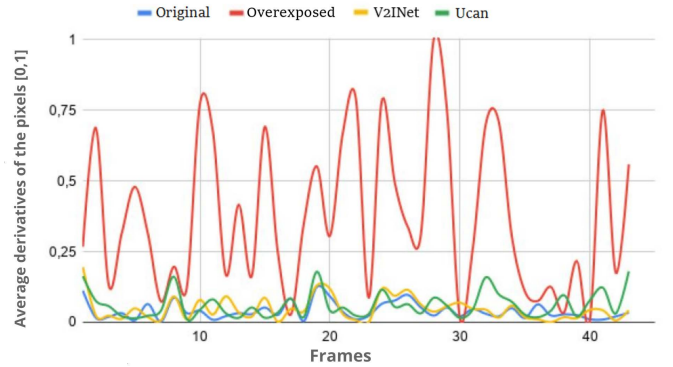


Figure 6: Variation of the mean luminance of frames throughout the test set of synthetically overexposed videos. We compare the following videos: originals (blue lines), synthetically overexposed (red lines), restored by our V2INet (yellow lines), and restored by the UCan [34] (green lines).

following rows show, respectively, the synthetically underexposed frames, the ones restored by our V2INet, and the ones restored by UCan [34]. It is possible to observe a higher variation in the exposure levels of the frames restored by the UCan method (last row). These visual artifacts are particularly noticeable in the car dashboard panel which clearly presents variations of luminance along the sequence. In the other hand, the sequence restored by our V2INet method presents more coherent exposure along the sequence. Finally, we show analogous results for overexposed frames in Fig. 8. In this case, the frames recovered by the V2INet are slightly brighter than the ones from UCan, although differences in the variations of the exposure levels are not so evident.

5 CONCLUSION

We introduce the V2INet, a deep neural network for exposure correction in videos. Our architecture is a novel video-to-image network based on the image-to-image U-Net model [30]. It aims for restoring the exposure of videos preserving the temporal coherence between consecutive frames. To evaluate our model, the SSIM and PSNR metrics were employed to measure the quality of the restored images. As these metrics are limited to the spatial scope, we introduce a surrogate metric to assess the temporal coherence of the restoration. We compare our results with the closely related methods in the literature, namely the U-Net and the UCan [34] models, on the BDD dataset. Our approach has shown competitive results regarding image quality and temporal coherence, presenting better performance in some experiments, particularly for underexposed frames. As future work, we intend to explore diverse and non-linear strides between frames in the temporal window.

REFERENCES

- [1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*.
- [2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
- [3] José M Álvarez Alvarez and Antonio M López. 2010. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems* (2010).
- [4] Harry C Andrews and Bobby Ray Hunt. 1977. *Digital image restoration*. Prentice-Hall.
- [5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *CVPR*.
- [6] Ying-Kang Chen, Shaou-Gang Miao, Ying-Cheng Lin, and Yi-Ming Chen. 2018. Color Image Enhancement Using Luminance Histogram Equalization and Two-Factor Saturation Control. In *Joint 7th ICIEV and 2nd icVPR*.
- [7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [8] Rodrigo de Bem, Anurag Arnab, Stuart Golodetz, Michael Sapienza, and Philip Torr. 2018. Deep Fully-Connected Part-Based Models for Human Pose Estimation. In *ACML*.
- [9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. FlowNet: Learning optical flow with convolutional networks. In *ICCV*.
- [10] P. Drews-Jr, E. R. Nascimento, M. F. M. Campos, and A. Elfes. 2015. Automatic restoration of underwater monocular sequences of images. In *IEEE/RSJ IROS*.
- [11] P. L. J. Drews-Jr, E. R. Nascimento, S. S. C. Botelho, and M. F. Montenegro Campos. 2016. Underwater Depth Estimation and Image Restoration Based on Single Images. *IEEE Computer Graphics and Applications* (2016).
- [12] Vincent Dumoulin and Francesco Visin. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).
- [13] J. F. Gaya, A. Duarte, F. Codevilla, P. L. J. Drews-Jr, and S. S. C. Botelho. 2019. Single Image Restoration for Participating Media Based on Prior Fusion. *IEEE Computer Graphics and Applications* (2019).
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.
- [15] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*.
- [16] Samuel W Hasinoff. 2014. Saturation (imaging). *Computer Vision: A Reference Guide* (2014).
- [17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).
- [18] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters* (2008).
- [19] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.
- [20] Shuiwang Ji, Wei Xu, Ming Wei Yang, and Kai Yu. 2010. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE TPAMI* (2010).
- [21] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. 2018. Super slo-mo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*.
- [22] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. 2017. Object detection in videos with tubelet proposal networks. In *CVPR*.
- [23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- [24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *ICCV*.
- [26] A. Loza, L. Mihaylova, N. Canagarajah, and D. Bull. 2006. Structural Similarity-Based Object Tracking in Video Sequences. In *IEEE Fusion*.
- [27] Tom Mertens, Jan Kautz, and Frank Van Reeth. 2009. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*.
- [28] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. 1987. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing* (1987).
- [29] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.
- [31] Inwook Shim, Joon-Young Lee, and In So Kweon. 2014. Auto-adjusting camera exposure for outdoor robotics using gradient information. In *IEEE/RSJ IROS*.
- [32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [33] Cristiano Steffens, Paulo Lilles Jorge Drews-Jr, and Silvia Silva Botelho. 2018. Deep Learning Based Exposure Correction for Image Exposure Correction with Application in Computer Vision for Robotics. In *IEEE LARS/SBR*.
- [34] Cristiano Rafael Steffens, Valquiria Huttner, Lucas Ricardo Vieira Messias, P L J Drews-Jr, Silvia SC Botelho, and RS Guerra. 2019. CNN-Based Luminance And Color Correction For ILL-Exposed Images. In *ICIP*.
- [35] Cristiano R Steffens, Lucas RV Messias, Paulo L J Drews-Jr, and Silvia S d C Botelho. 2020. CNN Based Image Restoration. *Journal of Intelligent & Robotic Systems* (2020).
- [36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.
- [37] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2017. Long-term temporal convolutions for action recognition. *IEEE TPAMI* (2017).
- [38] E. S. Vaz Junior, Paulo L J Drews-Jr, and Cristiano R Steffens. 2018. Restoration of Images Affected by Welding Fume. In *SIBGRAPI*.
- [39] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TNet: Real-time temporal and spatial video analysis of table tennis. In *CVPR Workshops*.
- [40] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. 2019. Underexposed photo enhancement using deep illumination estimation. In *CVPR*.
- [41] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. 2018. A fully progressive approach to single-image super-resolution. In *CVPR Workshops*.
- [42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* (2004).
- [43] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2017. Video enhancement with task-oriented flow. *IJCV* (2017).
- [44] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Mahdavan, and Trevor Darrell. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687* (2018).
- [45] Jiawan Zhang, Liang Li, Yi Zhang, Guoqiang Yang, Xiaochun Cao, and Jizhou Sun. 2011. Video dehazing with spatial and temporal coherence. *The Visual Computer* (2011).
- [46] Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. 2019. Dtrgan: Dilated temporal relational adversarial network for video summarization. In *Proceedings of the ACM Turing Celebration Conference-China*.
- [47] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *CVPR*.
- [48] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *ICCV*.



Figure 7: Comparison showing the original video, the synthetically underexposed video, and the outputs from the V2INet (ours) and the UCan [34] networks. It is possible to observe a higher variation in the exposure levels of the frames restored by the UCan method (last row), particularly in the car dashboard panel. In the other hand, the sequence restored by our V2INet method presents more coherent exposure along the sequence.



Figure 8: Comparison showing the original video, the synthetically overexposed video, and the outputs from the V2INet (ours) and the UCan [34] networks. The frames recovered by the V2INet are slightly brighter than the ones from UCan, although differences in the variations of the exposure levels are not so evident.