

MdVRNet: Deep Video Restoration under Multiple Distortions

Claudio Rota¹ ^a and Marco Buzzelli¹ ^b

¹*Department of Informatics Systems and Communication, University of Milano – Bicocca, Italy*
c.rota30@campus.unimib.it, marco.buzzelli@unimib.it

Keywords: Video Restoration, Video Enhancement, Multiple Distortions, Denoising, Compression Artifacts

Abstract: Video restoration techniques aim to remove artifacts, such as noise, blur, and compression, introduced at various levels within and outside the camera imaging pipeline during video acquisition. Although excellent results can be achieved by considering one artifact at a time, in real applications a given video sequence can be affected by multiple artifacts, whose appearance is mutually influenced. In this paper, we present Multi-distorted Video Restoration Network (MdVRNet), a deep neural network specifically designed to handle multiple distortions simultaneously. Our model includes an original Distortion Parameter Estimation sub-Network (DPEN) to automatically infer the intensity of various types of distortions affecting the input sequence, novel Multi-scale Restoration Blocks (MRB) to extract complementary features at different scales using two parallel streams, and implements a two-stage restoration process to focus on different levels of detail. We document the accuracy of the DPEN module in estimating the intensity of multiple distortions, and present an ablation study that quantifies the impact of the DPEN and MRB modules. Finally, we show the advantages of the proposed MdVRNet in a direct comparison with another existing state-of-the-art approach for video restoration. The code is available at <https://github.com/clauidiom4sir/MdVRNet>.

1 INTRODUCTION


During the last decade, the number of multimedia contents produced every day has considerably increased due to the growing diffusion of digital devices, such as digital cameras and smartphones. Although modern cameras are able to capture high-quality videos, there are some situations in which the quality of these contents is significantly reduced. For example, when videos are captured in poor light conditions or they are compressed to reduce memory occupation, their quality is reduced because of artifacts damaging their contents, causing problems to both user experience and machine vision applications.


Due to the remarkable results that Convolutional Neural Networks (CNNs) have shown in many vision tasks, several deep learning approaches to restore the quality of degraded videos have been introduced in the literature under the name of deep video restoration methods. Based on the degradation operators affecting the sequence, different video restoration tasks are usually addressed, such as video denoising, video deblurring and video compression artifact reduction.

Despite many methods to restore videos affected

by different artifacts have been proposed in the literature, the vast majority of them are designed to deal with a specific distortion type. Such methods produce excellent results on videos affected by the considered artifacts, but they might fail in the restoration process when multiple artifacts are present. Therefore, having a single framework able to restore videos even when they are simultaneously corrupted by multiple artifacts can be very useful, finding applications in many domains ranging from videoconferencing software to surveillance cameras.

In this paper, we present a framework to restore multi-distorted videos, that is, videos simultaneously corrupted by multiple degradation operators. The proposed approach, named Multi-distorted Video Restoration Network (MdVRNet) and visualized in Figure 1, is a two-stage restoration architecture that progressively aligns adjacent frames, allowing to extract both spatial and temporal information from the target frame and its adjacent ones. MdVRNet includes an original Distortion Parameter Estimation sub-Network (DPEN) specifically devised to obtain information about degradation operators affecting the video sequence, and make the restoration process more robust. The proposed framework uses novel Multi-scale Restoration Blocks (MRB) to ex-

^a  <https://orcid.org/0000-0002-6086-9838>

^b  <https://orcid.org/0000-0003-1138-3345>

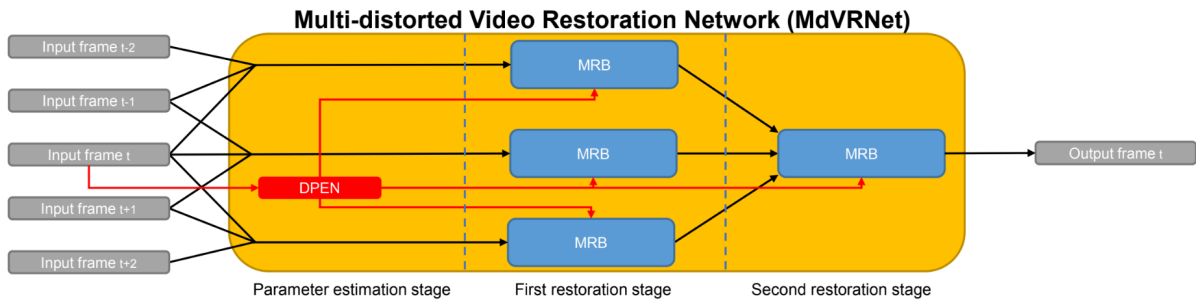


Figure 1: Architecture of Multi-distorted Video Restoration Network (MdVRNet) proposed to restore videos simultaneously affected by multiple distortions, using a custom Distortion Parameter Estimation sub-Network (DPEN), several Multi-scale Restoration Blocks (MRB), and implementing a two-stage restoration process.

tract features at different scales using two parallel streams, one for detail reconstruction and the other to take the semantics into account.

We carried out an extensive experimentation with different purposes, including but not limited to assessing the effectiveness of the proposed MdVRNet in restoring videos simultaneously affected by multiple distortions, noise and compression artifacts.

Our contributions can be summarized as follows:

- We present a novel deep learning approach to restore videos simultaneously corrupted by multiple distortions, named Multi-distorted Video Restoration Network (MdVRNet).
- We demonstrate that the main components of MdVRNet are all essential to obtain the best restoration performance.
- We show the effectiveness of MdVRNet in restoring multi-distorted videos by comparing it with another existing state-of-the-art approach for video restoration, using a benchmark datasets.

2 RELATED WORK

Video restoration is an active area of research, and many methods have been proposed in the literature to address different restoration tasks.

TOFlow (Xue et al., 2019) is a framework designed to deal with four independent restoration tasks: temporal frame interpolation, super resolution, denoising and compression artifact removal. DeBlurNet (Su et al., 2017) was proposed to address blur produced by camera shaking. Unlike TOFlow, DeBlurNet is able to exploit spatial and temporal information coming from multiple frames to restore the target one without using specific modules for explicit motion estimation and compensation.

VESPCN (Caballero et al., 2017) combines the efficiency of sub-pixel convolutions (Shi et al., 2016)

with the performance of spatial transformer networks (Jaderberg et al., 2015) to obtain a fast and accurate framework for video super resolution. Another contribution to video super resolution was given by DUF (Jo et al., 2018), which implicitly uses motion information between consecutive frames to generate dynamic upscaling filters to upsample the target frame.

EDVR (Wang et al., 2019) won all the four independent tracks of the NTIRE19 video restoration and enhancement challenge (Nah et al., 2019), i.e. video super resolution, deblurring and compression artifact removal. The cores of the network are the alignment module, known as PCD (Pyramid, Cascading and Deformable convolutions), and the fusion module, known as TSA (Temporal and Spatial Attention). EDVR achieves excellent performance in different restoration tasks, but its main limitation is represented by its high computational complexity. Instead, EVRNet (Mehta et al., 2021) is a method proposed for real-time video restoration, using a very lightweight network able to deal with various tasks, such as denoising and super resolution.

STDF (Deng et al., 2020) was proposed to remove compression artifacts from videos using a new spatio-temporal deformable fusion schema based on the idea of deforming the spatio-temporal sampling positions of standard convolutions, making them able to capture more relevant information. MFQE2.0 (Guan et al., 2019) is another solution to restore compressed videos, based on the idea of exploiting quality fluctuation among adjacent frames and using only high quality frames to restore the target one.

DVDNet (Tassano et al., 2019) is a framework for video denoising composed of three explicit steps: single image denoising, pixel motion estimation and warping, and multiple image denoising. More recently, the authors proposed an improved version, called FastDVDNet (Tassano et al., 2020), which performs implicit motion estimation and compensation

between frames to avoid artifacts caused by wrong motion estimation, also increasing its efficiency. Similarly to DVDNet, FastDVDNet uses the noise map of the target frame to obtain information related to the level of noise, and removes noise from videos in two steps. Despite the method is effective in removing noise from videos, it has two main limitations: the noise map used to help the denoising process must be provided with the true noise information, which is hardly available at inference time, and the use of a simple encoder-decoder architecture for denoising makes it difficult to reconstruct finer details when noise is strong.

3 PROPOSED METHOD

Multi-distorted Video Restoration Network (MdVRNet) is a two-step cascaded architecture taking five consecutive frames as input, plus a degradation map that encodes the intensity of the artifacts and thus provides the necessary information to treat a specific level of distortion. Inspired by FastDVDNet (Tassano et al., 2020), our method overcomes its main limitations by internally estimating the distortion intensity and by better handling the information extracted from frames.

An overview of the proposed MdVRNet is shown in Figure 1. It includes an original Distortion Parameter Estimation sub-Network (DPEN) properly devised to automatically infer the characteristics of multiple degradation operators affecting video sequences, and a novel Multi-scale Restoration Block (MRB) characterized by the following properties: a full-resolution branch, used to extract features without decreasing the spatial resolution to learn fine pixel dependencies and accurately reconstruct details, a low-resolution branch, used to extract semantic features and learn coarse pixel dependencies, and a channel attention mechanism, used to weight the features extracted by the two feature branches according to the importance they have in reconstructing the target frame. Overall, the MdVRNet framework contains about 3M parameters.

It is worth noting that, in contrast to other methods that can only deal with a specific distortion level at a time, such as STDF (Deng et al., 2020) and MFQE2.0 (Guan et al., 2019), our MdVRNet can handle different levels of distortion using a single model.

3.1 Distortion Parameter Estimation

Degradation operators commonly considered by restoration methods include additive white Gaussian

noise (AWGN) and JPEG compression (Yu et al., 2018). Each of these operators is characterized by some parameters: AWGN is defined by the standard deviation σ_N (since the mean is usually considered as zero), whereas JPEG compression requires to specify the quality factor q . Estimating such parameters is equivalent to estimating the intensity of the artifacts because there is a correlation between them: the higher the value of σ_N , the higher the intensity of noise; the lower the value of q , the higher the intensity of the blocking artifacts.

Although different methods to estimate the parameters of different degradation operators exist, they can accurately estimate the parameters of the considered distortion and they may produce inaccurate estimations when multiple distortions are present. Therefore, we devised a new CNN called Distortion Parameter Estimation sub-Network (DPEN) and we integrated it into the MdVRNet framework.

DPEN is a feedforward neural network consisting of five convolutional blocks and three fully connected blocks, as shown in Figure 2. It is a very shallow net-

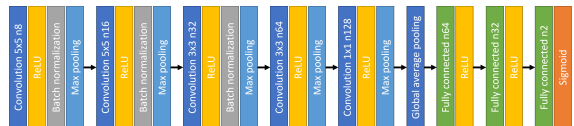


Figure 2: Distortion Parameter Estimation sub-Network (DPEN) devised to estimate the intensity of artifacts affecting the input sequence and integrated into the MdVRNet framework. N represents the number of kernels for convolutional layers and the number of neurons for fully connected layers.

work, as it has just about 53K parameters, hence it can be integrated into MdVRNet introducing very little overhead. The parameter values inferred by DPEN are expanded as feature maps so that they can be easily used by each MRB.

3.2 Multi-scale Restoration Block

The effectiveness of MdVRNet in restoring multi-distorted videos lies on the Multi-scale Restoration Block, which is a two-stream network that allows to extract spatial and temporal features at different scales, weight them according to their importance using an attention mechanism and fuse them to obtain a map, containing the artifacts detected, that is finally removed from the degraded target frame to restore it.

The detailed representation of the Multi-scale Restoration Block is shown in Figure 3. A stack of three consecutive frames, along with the degradation map estimated by DPEN, are used as input. After a set of two convolutions, each followed by batch nor-

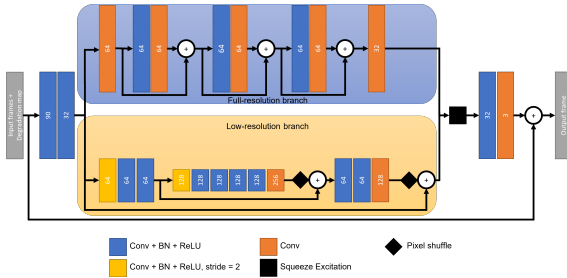


Figure 3: Multi-scale restoration block (MRB) used by MdVRNet to restore multi-distorted videos. The values in convolutional layers represent the number of kernels.

malization (Ioffe and Szegedy, 2015) and ReLU (Nair and Hinton, 2010), the computation is broken into two parallel branches working at different resolutions.

The first branch works at full resolution to extract fine pixel dependencies, capturing spatially accurate details. This branch is important to obtain detail-rich features, which allow to restore the target frame accurately reconstructing high-frequency components, such as edges. The first convolutional layer is used to increase the number of feature maps from 32 to 64. Then, a set of three residual blocks are applied to detect the artifacts at full resolution, paying more attention to finer details. Finally, the number of feature maps is reduced from 64 back to 32 using a convolutional layer. The full-resolution branch contains a total of 8 convolutional layers, which allow to extract useful information without excessively increasing the computational cost.

The second branch allows to extract coarse pixel dependencies in local areas to obtain semantically-rich features using an encoder-decoder architecture working at low resolution. As the input passes through this branch, a set of convolutional layers, batch normalization and ReLU decreases the spatial resolution while increasing the number of feature maps. Skip connections forward the output of each encoder layer directly to the input of the corresponding decoder layer using pixel-wise addition to ease and speed up the training process. Downsampling is performed using strided convolutions, each one halving the spatial dimension. There are a total of two downscaling operations so that, at the bottleneck, the spatial dimension corresponds to a quarter of the input spatial dimension, and the receptive field is large enough to capture semantic contents. Upsampling is performed using Pixel Shuffle layers (Shi et al., 2016) to reduce gridding artifacts.

The features extracted by the two branches are then concatenated and passed through a Squeeze-Excitation block (Hu et al., 2018), which performs channel attention to weight each feature map according to its importance in reconstructing the target

frame. The weighted feature maps are then fused together using a final set of convolutional layers, batch normalization and ReLU, to obtain the map containing the artifacts detected, considering both spatial details and semantics of the objects in the scene, that is finally subtracted from the degraded target frame to restore it.

Motion handling is a crucial aspect that characterizes all video restoration approaches. When multiple artifacts are present in a video sequence, the correlation among the values of the same pixel in adjacent frames may be broken, making the motion estimation process very challenging. For this reason, a MRB also has the burden of implicitly estimating pixel motion and aligning adjacent frames to properly extract temporal information, fundamental to avoid flickering artifacts.

3.3 Two-stage Restoration

Splitting the restoration process into two steps is a known strategy to make the most of the temporal information coming from adjacent frames and produce temporally stable results (Tassano et al., 2020). We adopted a two-stage restoration process both to improve temporal consistency and to allow MdVRNet to focus on different levels of detail, since our method deals with multiple distortions simultaneously.

Ideally, the first restoration stage should pay more attention to single pixel restoration, removing the artifacts introduced by punctual degradation operators, and the second restoration stage should pay more attention to restoring local areas, removing the artifacts introduced by local degradation operators to produce the final result. To provide evidence of this, we reported in Figure 4 an example of maps generated by each stage when the network restores frames affected by noise and compression artifacts.

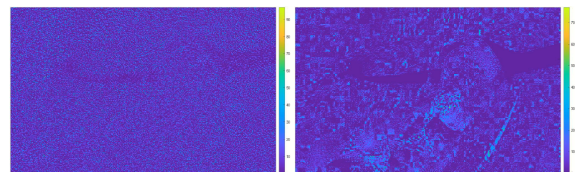


Figure 4: Maps containing the artifacts detected by the first and the second restoration stage of MdVRNet on a frame affected by additive white Gaussian noise and JPEG compression artifacts. Left: output of the first stage. Right: output of the second stage.

It is clear that the distortions contained in the map generated by the first restoration stage (on the left) are related to artifacts introduced at pixel level, which correspond to noise. Instead, the distortions contained

in the map produced by the second restoration stage (on the right) are different, and they are related to coarser artifacts introduced by the compression algorithm, as the presence of visible blocks suggests.

The first restoration stage is composed of three Multi-scale Restoration Blocks placed in parallel, as shown in Figure 1, each of which processes a stack of three adjacent frames with the purpose of improving the quality of the central one. Note that the weights of the three MRBs in the first stage are shared, hence they perform the same identical operations, also allowing to reduce the overall number of parameters. The second restoration stage is composed of just a single MRB, which further processes the three improved frames coming from the first stage to produce a clear and temporally consistent version of the target frame.

4 EXPERIMENTS

4.1 Experimental Setup

We used the Densely-Annotated Video Segmentation 2017 dataset (Pont-Tuset et al., 2017), containing 120 480p video sequences (90 for training and 30 for testing), for all the experiments.

We generated synthetic samples using two different degradation operators commonly used to assess the performance of restoration methods (Yu et al., 2018): additive white Gaussian noise (AWGN) and JPEG compression. We obtained multi-distorted frames by adding AWGN to the clean frames and then applying the JPEG compression. More in detail, we used the following parameters to degrade the input frames: $\sigma_N \in [5, 55]$ for AWGN and $q \in [15, 35]$ for JPEG compression. To speed up the training process and increase the number of training samples, we used patches of size 64×64 randomly cropped, for a total of 256000 samples.

We trained all the DPEN models for 500 epochs, using a learning rate initially set to $1e-4$, L_1 as loss function and Adam (Kingma and Ba, 2014) as optimizer. We reduced the learning rate by a factor of 10 whenever the loss function did not decrease for 20 consecutive epochs. We carried out all the MdVRNet experiments using models trained until convergence for a maximum of 64000 steps, using a batch size set to 32. We set the learning rate to $1e-3$ for the first five epochs, and to $1e-4$ for the remaining ones. We fixed the temporal neighborhood to five frames (two previous and two successive). We optimized our models using Adam as optimizer and MSE as loss function.

The results have been quantitatively assessed in terms of peak signal-to-noise ratio (PSNR) and struc-

tural similarity index (SSIM) (Wang et al., 2004).

4.2 Distortion Parameter Estimation

We conducted a set of experiments to evaluate the accuracy of DPEN in predicting the intensity of the artifacts affecting the video sequences, both in the case of single and multiple distortions.

Table 1: Mean absolute error (MAE) of two different DPEN models in estimating the distortion parameters of additive white Gaussian noise (AWGN) and JPEG compression. The lower the better.

AWGN	σ_N MAE	10	20	30	40	50
		0.70	0.82	0.94	1.10	1.24
JPEG	q MAE	15	20	25	30	35
		1.89	1.87	2.18	2.26	3.75

Table 1 shows the performance, in terms of mean absolute error (MAE), achieved by two different DPEN models trained and evaluated in predicting the distortion parameters of the considered degradation operators in the case of single distortions. As shown, DPEN is able to infer quite accurate values of the σ_N parameter for AWGN, and the error increases as the noise intensity increases. Concerning the quality factor q used by the JPEG compression algorithm, DPEN infers values with an error of about 2. Moreover, the error increases as the value of q increases. This is due to the fact that, when the quality factor is quite high, the blocking artifacts are not as pronounced as they are when the value is low.

To evaluate the effectiveness of DPEN in predicting distortion parameters in the case of multiple distortions, we trained and evaluated a single DPEN model considering the artifacts introduced by AWGN and JPEG compression simultaneously. The performance measured in MAE obtained by DPEN in predicting the distortion parameters of the considered distortion combination is reported in Table 2. It is

Table 2: Mean absolute error (MAE) of DPEN in estimating the distortion parameters on videos simultaneously affected by additive white Gaussian noise (AWGN) and JPEG compression artifacts. The lower the better.

σ_N	MAE for σ_N (AWGN) / MAE for q (JPEG)				
	$q = 15$	$q = 20$	$q = 25$	$q = 30$	$q = 35$
10	4.57/3.02	4.12/1.84	3.75/2.18	3.50/2.67	3.47/4.28
20	4.13/1.80	3.61/1.57	3.29/1.94	3.10/2.07	3.12/3.01
30	3.74/1.41	3.37/1.25	3.06/1.64	2.90/1.77	2.90/2.31
40	3.57/1.27	3.28/1.25	3.10/1.62	2.78/1.60	2.88/2.14
50	3.99/1.32	2.89/1.16	2.70/1.63	2.45/1.55	2.46/1.71

possible to notice that the error made in estimating the σ_N value is higher than the error made when the frame is corrupted just by noise, as reported in Table 1. Indeed, the maximum error increased from 1.24

to 4.57. Interestingly, while in the previous case the error made increases as the degradation intensity increases, here the opposite happens, i.e. the stronger the noise level, the lower the error made by DPEN. In addition, the error decreases as the quality factor q increases. The estimated q related to JPEG artifacts is quite precise, especially in the presence of strong noise, and the MAE is very similar to the MAE reported in Table 1. This means that DPEN is not sensitive to noise when inferring the distortion parameter related to compression artifacts. In addition, as it happens for single distortions, the higher the compression, the lower the error made in predicting the q parameter.

Additional experiments pointed out that using frames of different size from the one used during training increases the error. To solve this problem, we decompose the target frame into non-overlapping patches, estimating the distortion parameters on each patch and finally averaging the obtained estimations.

4.3 Comparison with state-of-the-art FastDVDNet

In order to evaluate the effectiveness of MdVRNet in restoring videos simultaneously affected by multiple distortions, we performed a direct comparison with FastDVDNet (Tassano et al., 2020), considered a state-of-the-art solution for video restoration with applications to denoising. We compared the capability of the models to remove artifacts introduced by additive white Gaussian noise and JPEG compression, considering three degradation intensities, on the DAVIS 2017 testset and on the Set8 dataset, as described within the FastDVDNet experimental setup.

Experimental results measured in PSNR and SSIM are reported in Table 3. As shown, MdVRNet

Table 3: Quantitative comparison between MdVRNet and baseline FastDVDNet in restoring multi-distorted videos, considering three distortion levels: Low ($\sigma_N = 10$, $q = 35$), Medium ($\sigma_N = 30$, $q = 25$) and High ($\sigma_N = 50$, $q = 15$). The higher the better.

Metric	Method	DAVIS 2017 testset			Set8 dataset		
		Low	Med.	High	Low	Med.	High
PSNR	FastDVDNet	33.90	31.50	29.37	29.71	28.52	26.82
	MdVRNet	34.48	32.05	29.78	31.69	29.40	27.51
SSIM	FastDVDNet	0.908	0.857	0.802	0.824	0.791	0.735
	MdVRNet	0.924	0.874	0.816	0.895	0.830	0.784

outperforms FastDVDNet both in PSNR and SSIM, regardless of the intensity of the artifacts. More in detail, MdVRNet is able to restore multi-distorted videos with a lower reconstruction error than FastDVDNet, as the difference in PSNR is about 0.51 dB on DAVIS 2017 and 1.18 dB on Set8. The same consideration is also valid considering the perceptual

similarity, as the difference in SSIM is about 0.02 on DAVIS 2017 and 0.05 on Set8.

Examples of qualitative comparison between the proposed MdVRNet and FastDVDNet are presented in Figure 5, which shows different video frames simultaneously corrupted by noise and compression artifacts (first column), restored by FastDVDNet (second column), restored by MdVRNet (third column) and the ground truth frames (fourth column). MdVRNet produces better results than FastDVDNet, whose outputs still contain visible artifacts. By looking at the cropped patches, it is possible to see that MdVRNet is able to remove the vast majority of artifacts from the distorted frames and to better reconstruct details. In addition, MdVRNet turns out to be more effective than FastDVDNet also in removing artifacts from flat regions (the sky in the first and second example and the wall in the third one).

These outcomes suggest that the novel Multi-scale Restoration Block used by MdVRNet, provided with information about distortion intensity inferred by DPEN, is effectively able to increase the quality of restored frames.

5 ABLATION STUDY

In this ablation study, we quantify the contributions of the main components of MdVRNet by alternatively removing one of them, demonstrating that they are all essential to achieve the best restoration performance. The results of the different experiments are reported in Table 4.

Table 4: Ablation study on the components of MdVRNet. Results on videos simultaneously affected by additive white Gaussian noise (with standard deviation σ_N) and JPEG compression artifacts (with quality factor q), reported as PSNR (left) and SSIM (right). The higher the better. First row: Simplified MdVRNet (experiment A). Second row: Blind MdVRNet (experiment B). Third row: One-stage MdVRNet (experiment C). Fourth row: MdVRNet.

	PSNR			SSIM				
	σ_N	$q = 15$	$q = 25$	$q = 35$	σ_N	$q = 15$	$q = 25$	$q = 35$
10		31.67	33.10	34.22		0.876	0.899	0.915
		31.77	33.52	34.37		0.874	0.904	0.916
		31.50	33.21	34.12		0.870	0.895	0.911
		31.86	33.59	34.48		0.881	0.912	0.924
30		30.71	31.64	32.00		0.848	0.863	0.869
		30.88	31.78	32.10		0.852	0.866	0.873
		30.82	31.76	32.09		0.849	0.866	0.872
		31.08	32.05	32.43		0.857	0.874	0.881
50		29.46	30.04	30.20		0.808	0.820	0.823
		29.56	30.00	30.16		0.810	0.820	0.824
		29.56	30.05	30.19		0.808	0.817	0.819
		29.78	30.29	30.50		0.816	0.826	0.831

We measured the contribution of the Multi-scale

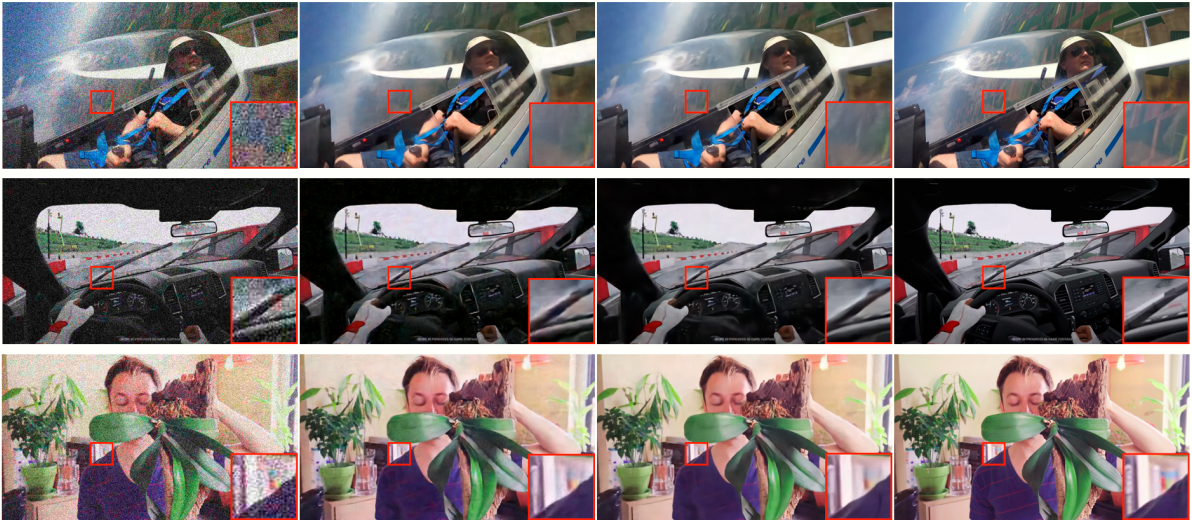


Figure 5: Qualitative comparison between MdVRNet and baseline FastDVDNet in restoring videos simultaneously affected by additive white Gaussian noise ($\sigma_N = 50$) and JPEG compression artifacts ($q = 15$). First column: distorted frames. Second column: FastDVDNet results. Third column: MdVRNet results. Fourth column: ground truth.

Restoration Block (experiment A) by comparing MdVRNet with the model that uses a simplified version of the MRBs, consisting of a simple encoder-decoder architecture obtained by removing the full-resolution branch and the Squeeze-Excitation block (Hu et al., 2018) from each MRB. We called this model Simplified MdVRNet. Both the models use the degradation map generated by the same DPEN model, thus preventing the difference in performance to be attributed to errors in predicting the intensity of artifacts. As shown in Table 4, the novel design of the MRB allows MdVRNet to improve the restoration performance by about 0.35 dB and 0.01 in terms of PSNR and SSIM, respectively, with respect to using a simple encoder-decoder architecture. This improvement is quite constant for all the values of σ_N and q tested.

To assess the contribution of the information provided by DPEN (experiment B), we compared MdVRNet with the blind model using the degradation map filled with zeros (both at training and test time), so that no information about degradation operators is available. We called this model Blind MdVRNet. As shown in Table 4, using DPEN to provide MdVRNet with information about the distortion intensity improves the performance. Indeed, PSNR improves by about 0.2 dB and SSIM by about 0.01 on average. More in detail, the improvement in PSNR is higher when the artifacts are stronger, since in this case the average improvement is about 0.3 dB. This means that the use of DPEN is particularly useful to reduce reconstruction errors when the distortions are severe. Concerning SSIM, the performance improvement is constant for all the tested values of σ_N and q ,

representing the distortion intensity.

Finally, to evaluate the impact of the two-stage restoration process (experiment C) of MdVRNet when dealing with multi-distorted videos, we compared MdVRNet with a single-step architecture, that we called One-stage MdVRNet, obtained by removing one stage and modifying the Multi-scale Restoration Block to receive five frames as input instead of three, so that the temporal dimension of the input does not change. As reported in Table 4, MdVRNet outperforms One-stage MdVRNet in restoring multi-distorted videos at any level of distortion, demonstrating that the two-stage restoration process of MdVRNet is more effective than the single-stage restoration process of One-stage MdVRNet. On average, the difference in PSNR and SSIM is about 0.3 dB and 0.01, respectively.

In all our experiments, MdVRNet obtained better restoration performance with respect to its variants, confirming that each of the main components of our method has an important contribution in improving the effectiveness of the restoration process.

6 CONCLUSIONS

In this paper, we presented Multi-distorted Video Restoration Network (MdVRNet), a novel approach to restore multi-distorted videos, that is, videos simultaneously corrupted by multiple artifacts.

MdVRNet is a two-step cascaded architecture that includes an original Distortion Parameter Estima-

tion sub-Network to increase the robustness of the restoration process and several Multi-scale Restoration Blocks to properly reconstruct finer details even when the artifacts are very strong.

We demonstrated that DPEN is able to accurately infer the intensity of the distortions affecting the input sequences, and compared MdVRNet with another existing state-of-the-art method for video restoration, showing both quantitatively and qualitatively the superiority of the proposed approach in restoring multi-distorted videos. Additionally, we provided an ablation study in which we demonstrated that the DPEN and MRB modules, as well as the two-stage restoration process of MdVRNet, are all essential to obtain the best restoration performance.

As future developments, we plan to investigate other types of degradation operators, such as blur caused by motion, and to improve the model via neural architecture search (Bianco et al., 2020).

REFERENCES

- Bianco, S., Buzzelli, M., Ciocca, G., and Schettini, R. (2020). Neural architecture search for image saliency fusion. *Information Fusion*, 57:89–101.
- Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., and Shi, W. (2017). Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857.
- Deng, J., Wang, L., Pu, S., and Zhuo, C. (2020). Spatio-temporal deformable convolution for compressed video quality enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10696–10703.
- Guan, Z., Xing, Q., Xu, M., Yang, R., Liu, T., and Wang, Z. (2019). Mfqc 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1.
- Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 448–456.
- Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial transformer networks. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*.
- Jo, Y., Oh, S. W., Kang, J., and Kim, S. J. (2018). Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3224–3232.
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Mehta, S., Kumar, A., Reda, F., Nasery, V., Mulukutla, V., Ranjan, R., and Chandra, V. (2021). Evrnet: Efficient video restoration on edge devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 983–992.
- Nah, S., Timofte, R., Gu, S., Baik, S., Hong, S., Moon, G., Son, S., and Mu Lee, K. (2019). Ntire 2019 challenge on video super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *ICML’10*, page 807–814, Madison, WI, USA. Omnipress.
- Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. (2017). The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*.
- Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883.
- Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., and Wang, O. (2017). Deep video deblurring for hand-held cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 237–246.
- Tassano, M., Delon, J., and Veit, T. (2019). Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809.
- Tassano, M., Delon, J., and Veit, T. (2020). Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1354–1363.
- Wang, X., Chan, K. C., Yu, K., Dong, C., and Change Loy, C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1954–1963.
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612.
- Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127.
- Yu, K., Dong, C., Lin, L., and Loy, C. C. (2018). Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2443–2452.