Video Denoising Using Cascaded skip connection Feedforward UNets

Abhijeet M. Pimpale[1] and Kishor M. Bhurchandi[2]

Abstract— **Quality video services have already gained high technical and commercial importance. The published work so far in this domain proposed mathematically and computationally complex algorithms, followed by the recent training-greedy deep learning-based denoising algorithms. This work proposes a video-denoising algorithm based on multiple UNet networks. The proposed video-denoising algorithm uses multiple encoder-decoder networks for video noise residual frame estimation, unlike the single encoder-decoder used by the published denoising algorithms. Using multiple skip connection UNets, we increase the residual noise modeling accuracy while restricting the signal features, which helps to improve denoising performance. The proposed network is trained end-to-end without motion compensation to reduce its complexity. The proposed network outperforms all the video denoising algorithms in terms of SSIM metric while it yields comparable performance in terms of PSNR.**

I. INTRODUCTION

While capturing a video, it frequently gets affected by ambient and systemic noise due to poor lighting and other unfavourable conditions. There is always a demand for quality video services that demand good perceptual and noise-free videos. Videos are recorded using a high ISO level camera setting in poor light, which results in capturing of noisy videos. Video denoising is, therefore, essential in lowering such noise. The majority of the video-denoising techniques published so far address denoising in presence of Gaussian noise as it most commonly available.

Video denoising is a more complex and challenging task than image denoising due to the third dimension i.e. temporal of the video signal. Thus, video-denoising algorithms need to take care of temporal information along with the spatial information. Video denoising without considering temporal information will introduce temporal artefacts and make the perceptual experience very uncomfortable. In case of machine vision applications, the video noise contain may even result in wrong decisions or activities. As temporal sequences of frames, video denoising needs to preserve temporal coherence for achieving better denoising performance. Videos frequently have moving subjects or cameras, which can cause motion blur or misalignment between frames. Accurate estimation and compensation for motion are essential for preventing the motion artefacts and maintaining crisp details during denoising.

Traditional and deep learning denoising algorithms have been proposed to address the aforementioned challenges. The methods that employ the non-local similarity of image patches to denoise without explicit motion compensation include VBM3D[1] and VBM4D[2], which resulted from natural extensions of BM3D[3]. The BM3D extensions resulted in poorer results compared to Deep learning-based results. The Deep learning-based algorithms for video denoising have recently outperformed the conventional techniques, producing cutting-edge results. Deep-learning patch-based video denoising techniques Patch Craft[4] and VNLNet[5] used 2-dimensional frame patches to implicitly model the video noise residue. Subsequently, the residue is subtracted from the noisy frame. Video denoising techniques based on patches are computationally expensive because patch-based techniques search for n similar patches from spatial search areas and the neighbouring frames, followed by the denoising process.

The State-of-the-art deep learning-based denoising techniques FastDVDnet[6] and DVDnet[7] use UNet[8] as the fundamental architectural block. A UNet-based architecture, on the other hand, consists of a single encoder-decoder structure with two skip connections for information flow, while the residue is modelled by the CNN UNet architecture. As a result, there are limited paths for noise residual to be transferred from input to output. We propose multiple cascaded UNets network with multiple skip connections for video denoising to address this issue. Thus we improve the noise residual flow and modelling of the residue at various stages from input to output using the proposed modified UNet architecture. This enhances retrieval of distorted image content structures due to the noise content resulting in improved denoising performance in terms of SSIM. The main contributions of the work are as follows.

- We introduce multiple multi-level cascaded encoder-decoder structures for video denoising.
- For the feature map to propagate from input to output, we also provided multiple skip connections, allowing the network to maximise the noise residual flow from input to output.
- The proposed method is trained without motion compensation to reduce its computational complexity.

II. RELATED WORK

In the early years, traditional Video denoising algorithms mostly relied on patched-based methods. By introducing a patch-based method that searches for non-local, self-similar 2D patches in the spatial and temporal domains, VBM3D [1] extends the image denoising implemented in BM3D[3] for video denoising. VBM4D[2] extends VBM3D[1] to 3D patches and follows the same block transform-based denoising approach leading to moderate PSNR denoising performance around 32 dB. In recent years, many deep learning approaches [4], [5], [6], [7] for video denoising have produced state-of-the-art results and outperformed tra-

[1]Abhijeet M. Pimpale is a Ph.D. student with department of Electronics and Communication Engineering, VNIT Nagpur, India apimpale06@gmail.com

[2]Kishor M. Bhurchandi is Professor with faculty of Electronics and Communication Engineering, VNIT Nagpur, India bhurchandikm@yahoo.co.in

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

ditional video denoising algorithms yielding PSNR performance around 34 dB.The denoising performance in terms of SSIM of all the state-of-the-art algorithms could not touch 0.92. SSIM is responsible for representing preservation of fine data structures in the denoised image compared to the original image.

The patch-based deep learning approaches are patchCraft[4] and VNLNet[5] . By assembling the n-nearest neighbours of non-overlapping patches of real frames, PatchCraft created artificial frames, which were then given to the network for training. On the other hand, the non-local patch search module that processes VNLNet creates an n-dimensional non-local feature vector that is used as the network's input. The patch-based strategy increased the computation cost by looking for n similar closest-neighbour patches in both the spatial and temporal directions. Explicit motion compensation is used in video denoising methods in [9], [7], [10] and [11]. The optical flow between the reference frame and the following frames is initially determined via motion-compensated methods. The computational complexity of video-denoising techniques increases as a result. U-Net has attained cutting-edge performance in video denoising techniques like FastDVDnet[6] and DVDnet[7]. The UNet implementations so far include single-level intra-block skip connections for implementing multiple noise residual flow paths and single encoder and decoder structures.

III. Proposed Method

Most video-denoising algorithms performed video denoising by considering noise as an additive Gaussian. The noisy video is represented as follows:

$$v(x,y,t) = f(x,y,t) + n(x,y,t) \tag{1}$$

where $v(x,y,t)$ represent noisy video, $f(x,y,t)$ represent clean video and $n(x,y,t)$ represent additive gaussian noise. (x,y) and t are the spatial and temporal coordinates, respectively. The denoising algorithms assume the form of a low pass filter $h(x,y,t)$ such that,

$$\hat{f}(x,y,t) = h(x,y,t) * f(x,y,t) \tag{2}$$

Where * and $\hat{f}(x,y,t)$ represents convolution and a denoised image, respectively. Noise is assumed to occupy higher frequencies in the spectrum of noisy signal, while the original clean signal is assumed to occupy lower frequencies.

A. Problem Definition

Deep learning-based video denoising algorithms formulated as:

$$\hat{f} = F(v;\theta) \tag{3}$$

F and v represent the model or network with trainable parameters and noisy video, respectively. θ and $\hat{f}$ are a set of trainable parameters and the denoised video, respectively. Thus in deep learning, we are finding a set of optimum network parameters $\hat{\theta}$ that minimizes the error,

$$\hat{\theta} = arg \min_{\theta} L\left(\hat{f}, v\right) \tag{4}$$

The loss function used in the proposed method is Mean Squared Error(MSE). Where $L\left(\hat{f}, v\right)$ is the loss function.

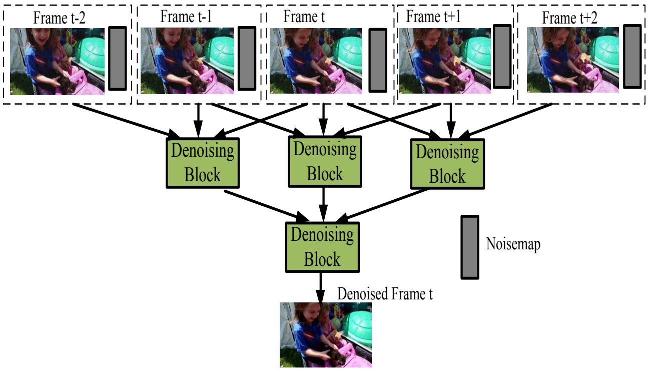


Fig. 1. Basic scheme of the Proposed Method

B. Proposed Architecture

The proposed method models noise in each frame using the previous and next frames, a noise map and a trained network for noise estimation in spatial and temporal directions. The estimated noise is further subtracted from the noise frame to obtain the respective denoised frame. Obviously, the denoising performance depends upon how accurately the noise is modelled without any signal content of the frame. Thus, the proposed network for video denoising aims to remove significant details and structures while modelling noise. It achieves this by employing multiple skip connections that successively remove traces of fine features from earlier layers to obtain a finer noise model and prevent the loss of fine-scale information in denoised frame. Indirectly, these skip connections enhance the network's ability to handle varying noise patterns over signal content by fusing noise features from different levels. This enables adaptation to various noise patterns overlying the signal content. Furthermore, skip connections address the vanishing gradient issue by providing a direct path for gradients to flow from deeper layers to shallower levels, improving convergence during training. The efficient transfer of gradients through skip connections ensures effective utilization of data from various network depths during both forward and backward passes. This contributes to the network's denoising performance without distorting the signal contents. The proposed architecture is derived by combining multiple UNet proposed in FastDVDnet[6] with added skip connection. To maximize noise residual flow, we replace a single UNet proposed in [6] with multiple UNets. The proposed method performs denoising in two steps, as depicted in Fig. 1. Each stage consists of denoising blocks. Each denoising block can consists of several cascaded UNets.

In this work, typically, we have used two cascaded denoising blocks. The input for a complete architecture is five consecutive frames and their respective noise maps containing noise standard deviation values. The input for a denoising block is three frames and a respective noise map. Each denoising block with three input frames and the noise maps yields a partially denoised frame at the first step. The three denoised frames are further given to another similar denoising block in step two, which yields the final denoised frame. This denoised frame corresponds to the middle frame of the five frames. Thus five successive frames are used to train the complete architecture to predict the denoised

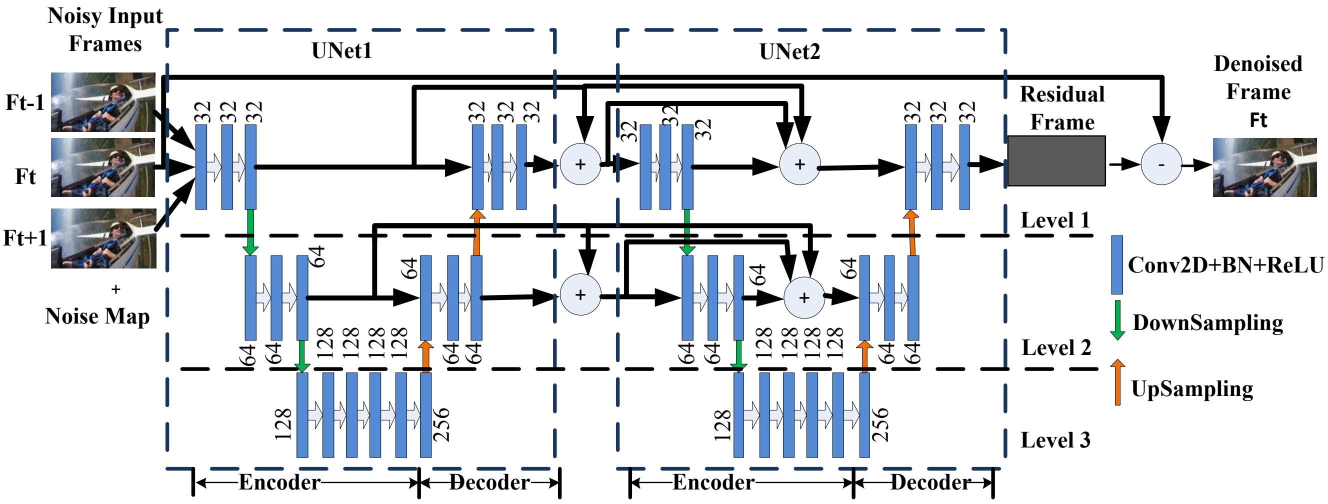


Fig. 2. Proposed Video Denoising block using multiple UNets and added skip connection

middle frame. Each denoising block contains multiple UNets with added skip connections, as shown in Fig. 2. Thus the proposed modified UNet has two novelties I) Cascading UNets and II) Added skip connections. The UNet in [6] has only one path, i.e. skip connections for noise residual flow, while the proposed UNet-based architecture has three skip connections, i.e. path for better noise residual flow from input to output to preserve frame features.

C. Multiple UNets in a denoising block

We proposed multiple cascaded UNets in a noise estimation block that replaces UNet in FastDVDnet[6] to preserve more information while denoising and obtaining a more accurate residual frame that represents noise. As shown in Fig. 2, we cascaded two UNets connected by adding two-point skip connections in each level for proving additional paths to transfer more noise features. The features map of the first module obtained using skip connections is added in similar layers(in size) of the next UNet module. This forms a ladder between the two cascaded UNets, providing paths to transfer maximum noise residual map from input to output. The proposed network accepts three consecutive frames and a noise map as inputs the noise map is a matrix of the same size as the input frames and contains the noise standard deviation value. Each UNet can compromise several convolution layers, though we have used typically three layers in the first and second levels and six at the third levels. Batch normalization[12] and the ReLU[13] activation function follows the multiple filter convolutions in each convolutions layer. We also used a dropout layer to avoid over-fitting during training. Each UNet block has an encoding operation across three levels followed by a decoding operation across two subsequent levels. Using downsampling, the network encoder gradually reduced the spatial resolution. We used a convolution layer with stride = 2 to accomplish downsampling, and the network's decoder gradually increased the spatial resolution using upsampling. We used the PixelShuffle[14] layer to upsample.

In fact, by starting with noisy input frames and progressively modelling noise via successive stages, cascaded U-Nets in the noise modelling provide a progressive accurate noise model. This method improves noise estimation at each pixel because each U-Net concentrates on estimating residual noise, if any. Furthermore, the cascaded architecture enables hierarchical feature learning across frames, with initial U-Net modelled noisy input frames and later stages removing complex temporal features, improving both; the removal of the underlying structures and the subsequent estimation of accurate residual noise content for each frame. This residual noise frame is subtracted from the noisy frame to obtain the denoised frame.

IV. EXPERIMENTS AND RESULTS

A. Dataset

The Proposed model is trained and validated using the DAVIS[15] database, consisting of 90 videos of resolution 854x480. Out of 90 videos, 72 are used for training, and 18 are used for validation. We have tested our model on the DAVIS test set and the Set8[7] dataset. The Davis test set consists of 30 videos of resolution 845 x 480. Set8 consists of 4 colour sequences from the Derfs Test Media collection and 4 colour sequences captured with a GoPro camera.

B. Training Details

The proposed model receives five noisy consecutive frames and a noise map as input. It gives the central denoised frame as output. We extracted spatiotemporal patches of spatial size 96x96 and temporal size 5 from the same locations of noisy and original frames to form noisy and ground truth sample pairs for training. There are 393582 spatiotemporal samples of size 96x96x5, extracted for training from a training set of the DAVIS database. Gaussian Noise of $\sigma = [0,50]$ is added in extracted samples during training to form a noisy dataset. Out of the five consecutive patches central patch of each clear video patch acts as a ground truth. A proposed model is trained to predict the residual map of the central frame. The predicted residual map is then subtracted to get a denoised central frame.

Mean Squared Error (MSE) between the clear frame patch and the reconstructed patch is used as a loss function for training. Peak Signal to Noise Ratio(PSNR), and Structural Similarity Index(SSIM) are two metrics used to evaluate the performance of the proposed model. PSNR is derived from MSE and measures average overall similarity between the two frames. The SSIM measures structural similarity between the corresponding physical location in the original clear frame and the denoised frame. Thus SSIM is a more rigorous measure for representing point-to-point similarity and reconstruction of distorted features between the original frame and the denoised frame. The proposed methods are implemented using popular Python library PyTorch[16]. The ADAM [17] algorithm was employed to optimise, learning weights and minimise the loss function. A proposed model is trained for 100 epochs with batch size=32. The initial learning rate is 1e-3 for the first 50 epochs and then decreases to 1e-4 between 50 to 60 epochs. Finally, the learning rate for the rest of the epochs is 1e-6. For the dropout layer, we used a dropout rate of 0.01.

C. Results

To train our model, we have conducted extensive experiments with various hyperparameters. On the NVIDIA TITAN RTX GPU setup with 128 GB RAM, our model takes about 10 days to train. Our model has been compared with various cutting-edge video denoising techniques, including conventional[2] and deep learning-based methods[4], [5], [6], [7]. Two metrics—PSNR and SSIM have been used to validate the performance of the proposed model.

On the DAVIS and Set8 datasets, Table I presents a quantitative comparison of various video-denoising methods based on PSNR and SSIM. We have validated and compared our algorithm with various values of sigmas, such as 10,20,30,40 and 50. VNLNet[5] and FastDVDNet[6] perform well in terms of PSNR for the lower sigma values on the DAVIS dataset, whereas DVDNet performs well for the higher sigma values. VNLB[18] has the highest average PSNR of any video-denoising technique for the Set8 dataset. The SSIM is a more robust and popular metric for assessing the perceived similarity between two images or video frames. It determines how similar images are structurally close to each other by examining the images' Luminescence, contrast, and structural aspects of the images more realistically. In terms of, SSIM performance of the proposed algorithm surpasses all the state-of-the-art benchmarking algorithms on both datasets. Thus, SSIM is more important for video denoising. A higher SSIM indicates more fine structural similarity between the original and denoised video. Our proposed model yielded the highest SSIM among all the video-denoising algorithms. Thus, the fundamental image structures, textures, and features are preserved more by the proposed algorithm. The PSNR best results are mainly achieved by VNLNet while the best SSIM results are achieved consistently by the proposed algorithm and indicated in bold as shown in Table I.

While PSNR is still widely used due to its simplicity and computational efficiency, SSIM has become the preferred metric for denoising evaluations[19], [20]. This is because it provides a more accurate and perceptually meaningful assessment of the denoising performance. In term of SSIM we outperform all other methods.

Fig. 3 shows the MSE vs Epochs graph for training and validation. From this, we can observe that our model is trained without overfitting as validation loss closely follows training loss. We used dropout as a regularisation method to avoid overfitting with a dropout rate of 0.01. Fig. 4 shows PSNR vs Epochs graph for training and validation. Average training PSNR increases up to 35.30 dB, whereas validation PSNR closely follows training PSNR.

Fig. 5 and Fig. 6 show a qualitative comparison of the visual performance of video denoising algorithms on zoomed blocks of Set8 and DAVIS test images with the proposed method, respectively, for σ=30. In Fig. 6 the proposed result appears to be slightly sharper than others. In terms of quality, our method is comparable to state-of-the-art methods. It can be qualitatively observed that the result of the proposed work is slightly better for visualization compared to the FastDVDNet.

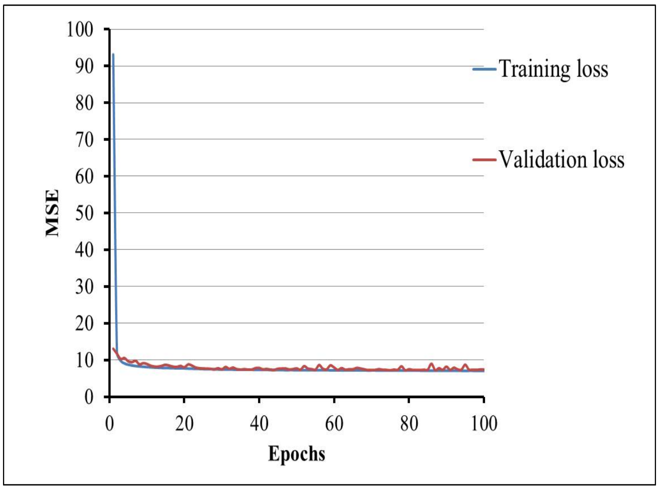


Fig. 3. MSE Vs EPOCH

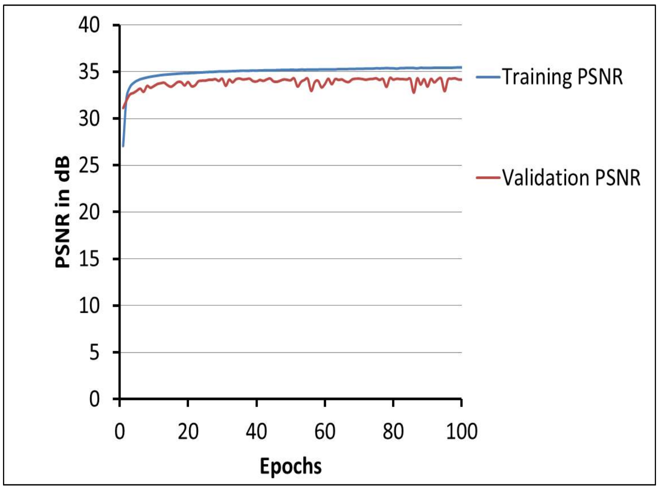


Fig. 4. PSNR Vs EPOCH

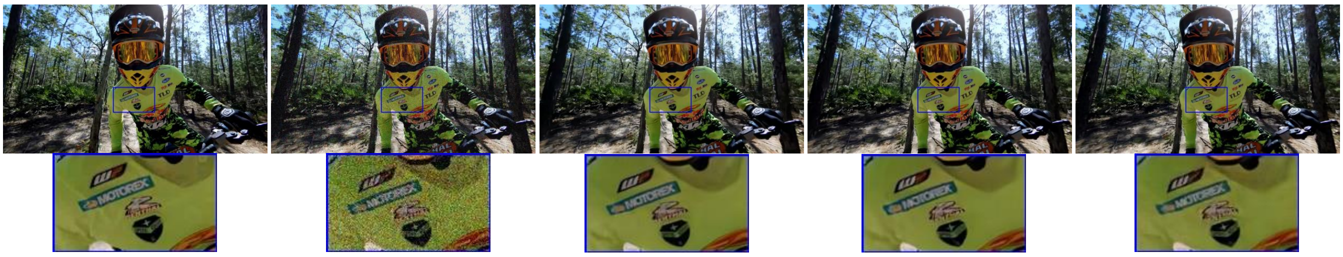
D. Running time

Comparable to FastDVDNet which takes 0.1 for each frame our approach denoises a frame with the same 960 x 540 resolution in 0.23 seconds on the above said platform.

TABLE I

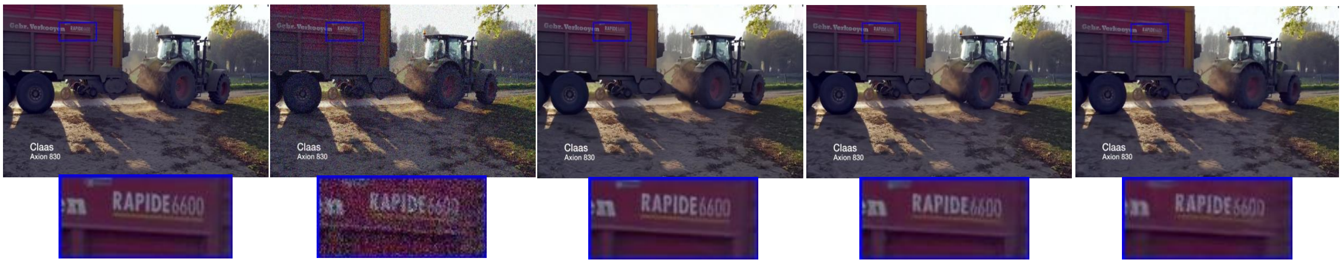
COMPARISON OF PSNR / SSIM ON THE DAVIS AND SET8 TEST SET. FOR PSNR: LARGER IS BETTER; THE BEST RESULTS ARE SHOWN IN BOLD. FOR SSIM: LARGER IS BETTER; THE BEST RESULTS ARE SHOWN IN BOLD

DAVIS	Noisy	VBM4D	DVDNet	VNLNet	VNLB	FastDVDNet	Proposed
$\sigma = 10$	28.13	37.58/-	38.13/0.9657	**39.56**/0.9707	38.85/-	38.71/0.9672	38.32/**0.9842**
$\sigma = 20$	22.11	33.88/-	35.7/0.9422	**36.53**/0.9464	35.68/-	35.77/0.9405	35.09/**0.9660**
$\sigma = 30$	18.58	31.65/-	**34.08**/0.9188	-	33.73/-	34.04/0.9167	33.18/**0.9480**
$\sigma = 40$	16.08	30.05/-	**32.86**/0.8962	33.32/0.8996	32.32/-	32.82/0.8949	31.85/**0.9291**
$\sigma = 50$	14.15	28.8/-	31.85/0.8745	-	31.13/-	**31.86**/0.8747	30.80/**0.9095**
Average	19.81	32.39/-	34.52/0.9195	-	34.34/-	**34.64**/0.9188	32.73/**0.9478**
Set8	Noisy	VBM4D	DVDNet	VNLNet	VNLB	FastDVDNet	Proposed
$\sigma = 10$	28.13	36.05/-	36.08/0.9510	**37.28**/0.9606	37.26/-	36.44/0.9540	36.06/**0.9842**
$\sigma = 20$	22.11	32.19/-	33.49/0.9182	**34.02**/0.9273	33.72/-	33.43/0.9196	32.89/**0.9631**
$\sigma = 30$	18.58	30.00/-	**31.79**/0.8862	-	31.74/-	31.68/0.8889	31.04/**0.9454**
$\sigma = 40$	16.08	28.48/-	30.55/0.8564	**30.72**/0.8622	30.39/-	30.46/0.8608	29.73/**0.9103**
$\sigma = 50$	14.15	27.33/-	**29.56**/0.8289	-	29.24/-	29.53/0.8351	28.71/**0.8969**
Average	19.81	30.81/-	32.29/0.8881	-	**32.47**/-	32.308/0.8917	31.68/**0.9394**



Ground Truth Noisy DVDNet FastDVDNet Proposed

Fig. 5. Qualitative comparison on motorbike sequence from Set8 dataset for σ=30



Ground Truth Noisy DVDNet FastDVDNet Proposed

Fig. 6. Qualitative comparison on tractor sequence from DAVIS dataset for σ=30

V. Conclusion

This work proposes a video denoising method by residual noise modeling for each frame and the subsequent subtraction of the residual noise from the noisy frame to obtain a denoised frame. The proposed network is trained for residual noise estimation rather than the conventional denoising approach. Proposed cascaded UNet networks for noise modeling minimize the signal information flow into the residual noise model and preserve structural similarity in the denoised image due to the added skip connection at different levels. Using multiple paths for the noise model construction learns more critical noise features and removes signal features from the noise model, resulting in improved denoising performance in terms of the more stringent denoising performance measure; SSIM. In terms of PSNR, the performance of the proposed method is comparable to the state-of-the-art. The proposed work focuses on Gaussian noise and is trained using gaussian noise added images. In reality, noise in videos may be more complex than Gaussian. The future scope of work can focus on more complex real noise types. More UNet blocks can be added to the architecture by cascading. Each UNet block can have more CNN layers. Also, more levels can be added to the UNet architecture to obtain more accurate residue to obtain better denoising performance.

References

[1] K. Dabov, A. Foi, and K. Egiazarian, “Video denoising by sparse 3d transform-domain collaborative filtering [c],” in *Proc. 15th European Signal Processing Conference*, vol. 1, no. 2, 2007, p. 7.

[2] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, “Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms,” *IEEE Transactions on image processing*, vol. 21, no. 9, pp. 3952–3966, 2012.

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-d transform-domain collaborative filtering,” *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[4] G. Vaksman, M. Elad, and P. Milanfar, “Patch craft: Video denoising by deep modeling and patch matching,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2157–2166.

[5] A. Davy, T. Ehret, J.-M. Morel, P. Arias, and G. Facciolo, “A non-local cnn for video denoising,” in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 2409–2413.

[6] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1354–1363.

[7] ——, "Dvdnet: A fast network for deep video denoising," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1805–1809.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[9] H. Chen, Y. Jin, K. Xu, Y. Chen, and C. Zhu, "Multiframe-to-multiframe network for video denoising," *IEEE Transactions on Multimedia*, vol. 24, pp. 2164–2178, 2021.

[10] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[11] A. Buades and J.-L. Lisani, "Enhancement of noisy and compressed videos by optical flow and non-local denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1960–1974, 2019.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[14] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "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*, 2016, pp. 1874–1883.

[15] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675*, 2017.

[16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] P. Arias and J.-M. Morel, "Video denoising via empirical bayesian estimation of space-time patches," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 1, pp. 70–93, 2018.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] Z. Kotevski and P. Mitrevski, "Experimental comparison of psnr and ssim metrics for video quality estimation," in *International Conference on ICT Innovations*. Springer, 2009, pp. 357–366.