IEICE TRANS. FUNDAMENTALS, VOL.E107–A, NO.8 AUGUST 2024

LETTER

Video Reflection Removal by Modified EDVR and 3D Convolution*

Sota MORIYAMA[†], ***Student Member***, **Koichi ICHIGE**[†a)], **Yuichi HORI**[††], ***and*** **Masayuki TACHI**[††], ***Members***

SUMMARY In this paper, we propose a method for video reflection removal using a video restoration framework with enhanced deformable networks (EDVR). We examine the effect of each module in EDVR on video reflection removal and modify the models using 3D convolutions. The performance of each modified model is evaluated in terms of the RMSE between the structural similarity (SSIM) and the smoothed SSIM representing temporal consistency.

key words: *video reflection removal, video restoration, machine learning, convolutional neural network, deformable convolutions, 3D convolutions*

1. Introduction

Analysis of video images captured by cameras and sensors has attracted attention in recent research on automated driving systems and traffic monitoring systems. One of the problems in analyzing videos is undesirable reflections. Thus, reflection removal using methods of deep learning like convolutional neural networks (CNN) has been studied to remove reflections and restore natural images. In the field of single-image reflection removal, several methods have been proposed [1]–[3] and have shown a certain level of accuracy on real images.

On the other hand, for videos, a method without using deep learning has been proposed for video reflection removal [4]. However, this conventional method has the following problems.

- The single scale method has difficulty with flow estimation and frame alignment when the motion is large.
- The estimation method for a center frame does not take into account temporal features from neighboring frames.

In frame alignment, a fusion of temporal features is effective in aggregating temporal information of neighboring frames [5], [6]. We consider deep learning methods to extract temporal features in video reflection removal to address these problems.

A Recent work on high-resolution video processing (e.g., video de-blurring) using deep learning contains explicit motion estimation with flow-based motion compensation modules to perform frame alignment [7]. However, it is more effective for reflection removal to segment reflection regions as well as frame alignment as in [4]. Thus, we focus on video restoration with enhanced deformable networks (EDVR) [9], [10], which has been proposed as a video restoration framework for super-resolution and de-blurring. We chose EDVR as a video processing method using deep learning for the following reasons.

- The multi-scale pyramid structure enables flow estimation and frame alignment even for large motion of reflection.
- Deformable convolutions in EDVR have been applied to object detection [8] and semantic segmentation [14]. They are effective for reflection removal to capture reflection regions dynamically and segment them implicitly.

In this paper, we apply EDVR to reflection removal [11] and create our own synthetic reflection video dataset from the realistic and dynamic scenes (REDS) dataset [12]. We compare the usefulness of each module through ablations. Furthermore, we propose modified EDVR models using 3D convolutions. The performance of each model is evaluated not only in terms of structural similarity (SSIM) [13] but also in terms of the RMSE between the smoothed SSIM and the original SSIM, which represents temporal consistency.

2. Video Restoration with Enhanced Deformable Networks (EDVR) [10]

EDVR is a video restoration framework for video super-resolution and de-blurring. It consists of four modules: a pre-deblur module, pyramid cascading deformable (PCD) align module, temporal and spatial attention (TSA) fusion module, and reconstruction module. In the case of super-resolution, the input is transmitted directly to these modules, whereas in the case of de-blurring, it is transmitted in the low-resolution domain through the downsampling layer.

Pre-deblur module. The pre-deblur module has a pyramid structure with residual blocks, which enables global and local features to be extracted by processing in high- and low-resolution regions. This module is applicable only for de-blurring.

PCD align module. The pyramid cascading deformable align module consists of deformable convolutional networks (DCN) [14], [15] with a pyramid structure. Deformable convolutions implicitly align the neighbor frames (feature maps)

Manuscript received August 17, 2023.
Manuscript revised November 2, 2023.
Manuscript publicized December 11, 2023.

[†]Department of Electrical and Computer Engineering, Yokohama National University, Yokohama-shi, 240-8501 Japan.
[††]Huawei Technologies Japan K.K., Tokyo, 108-0075 Japan.

*Part of this paper has been presented in [11].

a) E-mail: koichi@ynu.ac.jp

DOI: 10.1587/transfun.2023EAL2078

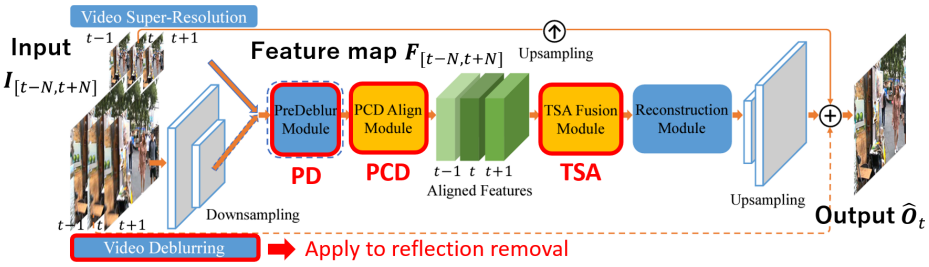

Fig. 1 EDVR network architecture [10].

to the center frame by using the channel-directed combination of the neighbor frame and the center frame as an offset [16]. Furthermore, it is expected to detecting reflection regions for reflection removal.

TSA fusion module. The temporal and spatial attention module focuses on temporal and spatial attention, and aligned feature maps are fused by dynamically aggregating neighboring frames on a pixel-by-pixel basis to interpolate areas not aligned in the previous PCD align module.

Reconstruction module. The reconstruction module consists of a series connection of residual blocks and some upsampling layers. A pixel shuffler is used for upsampling.

3. Proposed Approach

We apply a de-blurring flow using EDVR to remove reflections due to the high resolution of inputs. The EDVR network for reflection removal is shown in Fig. 1. Note that the predeblur module, PCD align module, and TSA fusion module are hereafter shortened to PD, PCD, and TSA, respectively.

3.1 Ablation of Each EDVR Module

We examine the effect of each EDVR module on reflection removal. To this end, we create several EDVR models without PD, PCD, and TSA and perform ablation experiments. The reconstruction module is not subject to ablation because it is necessary for upsampling frames.

3.2 Modified Models Introducing 3D Convolution

3D convolutional neural networks are widely used for 3D data such as 3D images and videos [17] although all EDVR convolutions are 2D convolutions including deformable convolutions. Thus, we propose two models introducing 3D convolutions to extract temporal features between frames: a pre-3D convolution (P3DC) model and 3D convolutional alignment (3DCA) model. The 3D convolution is effective in restoring sharp pixels in neighboring frames by extracting temporal features in video processing [18]. We verify the effectiveness of extracting temporal features as an auxiliary process to deformable convolution by introducing 3D convolution before and after the PCD module.

P3DC model. The pre-3D convolution model replaces the PD module of EDVR with 3D convolutional layers to aggregate temporal features before aligning frames.

Table 1 Modified EDVR models.

	module		
	PD	PCD	TSA
EDVR [10]	○	○	○
woTSA	○	○	×
woPCD	○	×	○
woPD	×	○	○
woPD_PCD	×	×	○
woPD_TSA	×	○	×
woPCD_TSA	○	×	×
woPD_PCD_TSA	×	×	×
P3DC (proposed)	3D Conv.	○	○
3DCA (proposed)	○	3D Conv.	○

3DCA model. The 3D convolutional align model replaces the PCD module with 3D convolutional layers to align frames instead of deformable convolutions.

We do not consider replacing the TSA module extracting spatial features with 3D convolutions because we examine the effect of 3D convolutions on temporal features in this paper. Table 1 lists our modified EDVR models.

3.3 Metrics for Evaluating Temporal Consistency

It is important to consider the temporal consistency of neighboring frames in order to improve the quality of video reflection removal. Flicker-like noise usually occurs between frames if temporal consistency is not maintained. Therefore, we propose a metric that evaluates the intensity of flickering between frames in order to quantitatively express temporal consistency.

We consider SSIM [13] oscillation through frames to evaluate the intensity of flickering. This is based on the assumption that SSIM oscillates from frame to frame if an output video is flickering due to temporal inconsistency. Next, we consider smoothed SSIM using a moving average filter with a filter size of 3. The difference between the original SSIM and the smoothed SSIM becomes small when the original SSIM has only slight oscillations. Therefore, we also propose using the RMSE between the original and smoothed SSIM calculated by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \tag{1}$$

$$y_t = \mathrm{SSIM}(\hat{\boldsymbol{O}}_t, \boldsymbol{O}_t) \tag{2}$$

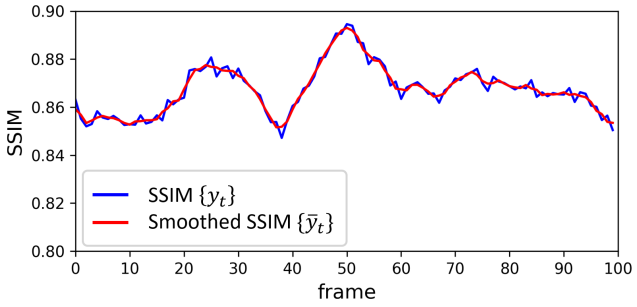

Fig. 2 Examples of oscillating SSIM ($T = 100$).

$$\bar{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3} \tag{3}$$

where T represents the last frame, and SSIM($\hat{\boldsymbol{O}}_t, \boldsymbol{O}_t$) represents the SSIM between the t-th estimated frame $\hat{\boldsymbol{O}}_t$ and the t-th GT image $\boldsymbol{O}_t$. In (2) and (3), the SSIM and the smoothed SSIM are represented by y_t and $\bar{y}_t$, respectively. Examples of $\{y_t\}$ and $\{\bar{y}_t\}$ for $T = 100$ are shown in Fig. 2. A large value for RMSE indicates that the SSIM has large oscillations and flickers, i.e., temporal consistency is not maintained. On the other hand, a small value indicates that the SSIM is smooth and does not flicker, meaning that temporal consistency is maintained. Therefore, the RMSE reflects relative temporal inconsistency. In this paper, we use the RMSE of (1) to compare the accuracy of the temporal consistency of each proposed model.

4. Experiments

We evaluated the performance of the modified EDVR models for reflection removal through simulation. We created a synthetic reflection video dataset by combining two natural images from the REDS dataset and used this dataset to train each modified model. The training results were evaluated in terms of the mean of the SSIM and the RMSE of the smoothed SSIM, which represents temporal consistency. Table 2 shows the various parameters for each model in this experiment. We used Charbonnier loss [19].

4.1 Example of Estimation Results

We estimated 20 test videos for each model that had been trained. Figure 3 shows the SSIM through 100 frames of a test video. Figure 4 shows estimation results for frame #44 in Fig. 3, where the red, blue, and yellow regions respectively show remarkable artifacts and remaining reflections.

From Fig. 3, we confirmed that P3DC had a higher SSIM overall, which means that the outputs were restored to a higher quality. At around frame #44, the SSIM was generally depressed due to remaining reflections in the lower part of the image, but P3DC and woPD_TSA were able to maintain a high level of SSIM. From Fig. 4, we can see that some modified models failed to restore partially better than EDVR. In particular, the SSIM is degraded for woTSA and woPD_PCD in Fig. 3, which is due to residual red reflection

Table 2 Parameter specifications.

No. of videos [100 frame/pair]	Train	105
	Valid	10
	Test	20
No. of input frames		5
Size of an input frame		256×256
No. of training images		10,500
Batch size		2
Iteration / Epoch		5,250
Total iteration		2,000,000
Epochs		381
Loss function		Charbonnier Loss
Optimization		ADAM
Learning rate		0.0001

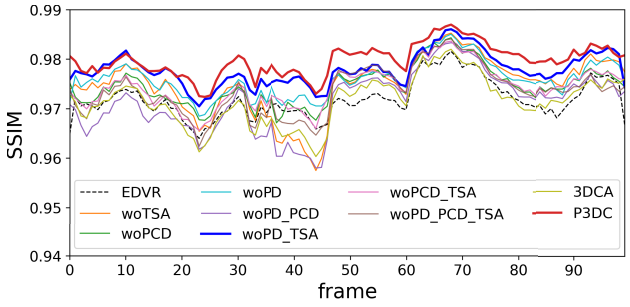

Fig. 3 SSIM through 100 frames.

in the yellow region. On the other hand, woPD_TSA and P3DC could well reduce reflections both in the yellow and blue regions. In addition, detailed results (red region) for EDVR and P3DC are shown in Fig. 5. At first glance, both EDVR and P3DC seem to be accurate in Fig. 4, but Fig. 5 shows that EDVR had artifacts and color degradation. This is due to EDVR trying to remove reflections excessively, which is why EDVR had a lower SSIM.

4.2 Overall Test Results

We also evaluated the average results across all of the test data. Table 3 shows a quantitative evaluation of all test data and the results of STDANet [7] as a comparison of a recent method for video processing using deep learning. The first column shows the average of the 100-frame SSIMs averaged over 20 test videos, the second column shows the average RMSE of the smoothed SSIM over 20 test videos, and the third column shows the number of parameters for each model. The best values in each column are shown in bold, and the second best values are underlined.

Table 3 confirms that the latest STDANet has the highest accuracy and is temporally consistent, comparing EDVR with the latest method. However, STDANet has a large amount of parameters and the disadvantage of high model cost for the accuracy. In EDVR and its modified models, woPD_TSA had the best SSIM accuracy for all of the test data. This indicates that this model with only the PCD module is the most effective for reflection removal. Additionally,

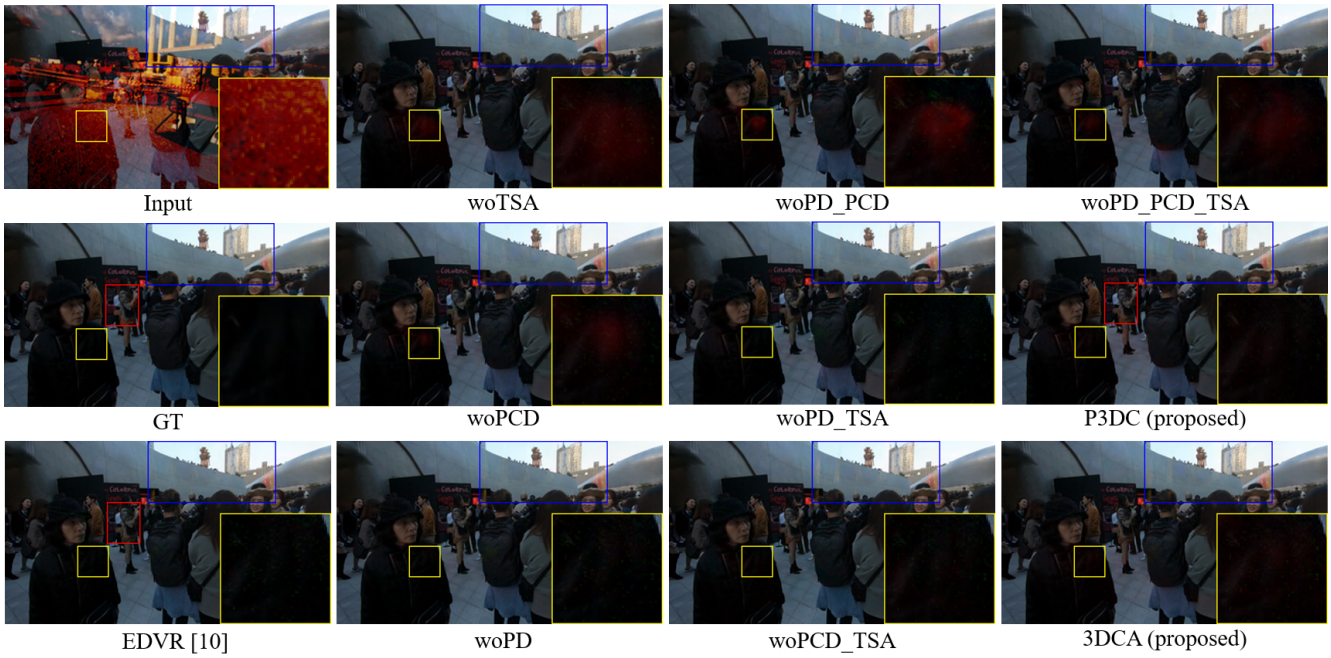

Fig. 4 Qualitative results for frame #44.

Fig. 5 Detailed comparison of EDVR [10] and P3DC.

Table 3 Quantitative results for all test data.

	SSIM	RMSE ($\times 10^{-4}$)	No. of parameters
STDANet [7]	**0.9886**	**3.804**	13,837,029
EDVR [10]	0.9818	6.683	4,042,851
woTSA	0.9840	5.874	3,738,979
woPCD	0.9811	6.195	2,658,563
woPD	0.9836	6.003	3,373,987
woPD_PCD	0.9796	6.923	1,989,699
woPD_TSA	0.9843	6.161	3,070,115
woPCD_TSA	0.9810	6.016	2,354,691
woPD_PCD_TSA	0.9799	6.735	**1,685,827**
P3DC (proposed)	0.9833	6.437	3,377,443
3DCA (proposed)	0.9807	7.067	2,990,531

a comparison of the proposed P3DC and 3DCA showed better accuracy for P3DC and worse accuracy for 3DCA. This indicates that 3D convolutions can be an alternative for PD with strong effects but are not a sufficient alternative for alignment by PCD. From RMSE, although there was generally an inverse correlation with the accuracy of SSIM, the woTSA model had the best temporal consistency accuracy. This is because the TSA module fuses the temporal attention followed by the spatial attention, which makes it difficult to maintain temporal consistency in the final output. Therefore, we have to remove the spatial fusion of TSA when we place emphasis on maintaining temporal consistency. From the number of parameters, woPD_PCD_TSA has the smallest number of parameters, but the number increases for models that include PD and PCD modules due to the large size of the modules. Therefore, PCD has a large number of parameters, although not as large as STDANet.

In short, all of the models that tended to have higher SSIMs for all test data included PCD. Therefore, frame alignment by deformable convolutions in PCD is also effective. The results of this experiment demonstrate that the woPD_TSA model including only PCD is the most effective. However, our proposed P3DC model that replaces the PD module with 3D convolutions is also partially effective since incorporating all EDVR modules would result in excessive artifacts.

5. Conclusion

In this paper, we proposed modified EDVR models with 3D convolutions for video reflection removal. We performed comparative experiments and found the alignment frames by deformable convolutions included in the PCD module to be the most effective. We also found our P3DC model that introduces 3D convolutions to the suppression of artifacts to be effective. In addition, we proposed metrics for evaluating temporal consistency using the RMSE between the smoothed SSIM and the original SSIM. Future prospects include proposing new networks using deformable convolutions and 3D convolutions.

References

[1] C. Li, Y. Yang, K. He, S. Lin, and J.E. Hopcroft, "Single image reflection removal through cascaded refinement," Proc. Computer Vision and Pattern Recognition (CVPR), pp.3565–3574, June 2020.

[2] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," Proc. Computer Vision and Pattern Recognition (CVPR), pp.8178–8187, June 2019.

[3] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," Proc. Computer Vision and Pattern Recognition (CVPR), pp.4786–4794, June 2019.

[4] A. Nandoriya, M. Elgharib, C. Kim, M. Hefeeda, and W. Matusik, "Video reflection removal through spatio-temporal optimization," Proc. Int'l Conf. on Computer Vision (ICCV), pp.2411–2419, Oct. 2017.

[5] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," Proc. Computer Vision and Pattern Recognition (CVPR), pp.2848–2857, July 2017.

[6] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," Proc. Computer Vision and Pattern Recognition (CVPR), pp.3897–3906, June 2019.

[7] H. Zhang, H. Xie, and H. Yao, "Spatio-temporal deformable attention network for video deblurring," Proc. European Conf. on Computer Vision (ECCV), pp.581–596, Oct. 2022.

[8] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," Proc. European Conf. on Computer Vision (ECCV), pp.331–346, Sept. 2018.

[9] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally deformable alignment network for video super-resolution," Proc. Computer Vision and Pattern Recognition (CVPR), pp.3360–3369, June 2020.

[10] X. Wang, K.C.K. Chan, K. Yu, C. Dong, and C.C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," Proc. Computer Vision and Pattern Recognition (CVPR), June 2019.

[11] S. Moriyama, Y. Toyoda, K. Ichige, Y. Hori, and M. Tachi, "Application of video restoration model EDVR to reflection removal and its evaluation," IEICE General Conf., no.A-8-5, March 2023.

[12] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K.M. Lee, "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," Proc. Computer Vision and Pattern Recognition (CVPR), June 2019.

[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, April 2004.

[14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," Proc. Int'l Conf. on Computer Vision (ICCV), pp.764–773, Oct. 2017.

[15] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," Proc. Computer Vision and Pattern Recognition (CVPR), pp.9308–9316, June 2019.

[16] K.C.K. Chan, X. Wang, K. Yu, C. Dong, and C.C. Loy, "Understanding deformable alignment in video super-resolution," Proc. AAAI Conf. on Artificial Intelligence, vol.35, no.2, pp.973–981, May 2021.

[17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.1, pp.221–231, Jan. 2013.

[18] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," IEEE Trans. Image Process., vol.28, no.1, pp.291–301, June 2019.

[19] W.S. Lai, J.B. Huang, N. Ahuja, and M.H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," Proc. Computer Vision and Pattern Recognition (CVPR), pp.624–632, July 2017.