MULTI-FIELD DE-INTERLACING USING DEFORMABLE CONVOLUTION RESIDUAL BLOCKS AND SELF-ATTENTION

Ronglei Ji and A. Murat Tekalp

Koc University, Dept. of Electrical and Electronics Engineering, Istanbul, Turkey

ABSTRACT

Although deep learning has made significant impact on image/video restoration and super-resolution, learned deinterlacing has so far received less attention in academia or industry. This is despite deinterlacing is well-suited for supervised learning from synthetic data since the degradation model is known and fixed. In this paper, we propose a novel multi-field full frame-rate deinterlacing network, which adapts the state-of-the-art superresolution approaches to the deinterlacing task. Our model aligns features from adjacent fields to a reference field (to be deinterlaced) using both deformable convolution residual blocks and self attention. Our extensive experimental results demonstrate that the proposed method provides state-of-the-art deinterlacing results in terms of both numerical and perceptual performance. At the time of writing, our model ranks first in the Full FrameRate LeaderBoard at `https://videoprocessing.ai/benchmarks/deinterlacer.html`

***Index Terms*—** deep learning, deinterlacing, deformable convolution, feature alignment, self attention

1. INTRODUCTION

Interlaced scanning was invented to strike a balance between spatial and temporal video resolution in analog TV broadcasting to overcome insufficient bandwidth to transmit at a high enough frame rate. It corresponds to scanning odd and even fields alternately, which results in doubling the field rate at the expense of vertical spatial resolution without increasing bandwidth, and enhances visual perception of fast motion scenes.

In order to convert interlaced video content to progressive content for displaying interlaced videos on progressive monitors, deinterlacing continues to be a significant problem of interest since there still exists a large amount of interlaced and telecined catalogue content despite the fact that almost all video cameras are progressive nowadays.

The industry employs simple traditional methods for deinterlacing, telecine, and 50 Hz to 60 Hz and vice versa conversion [1]. Many alternative traditional approaches have also been proposed by the academia for intra-field deinterlacing [2, 3] and inter-field deinterlacing [4], which, however, introduce either blurring or artifacts such as flicker and jaggedness due to the time difference between adjacent fields.

This paper proposes a novel multi-field deinterlacing network in Section 3 featuring two stages: a field-alignment stage consisting of novel deformable convolution residual blocks in parallel with a self-attention module, and a reconstruction module, which generates two progressive frames for each pair of input interlaced fields. Section 4 presents extensive experimental results and Section 5 concludes the paper.

2. RELATED WORKS AND CONTRIBUTIONS

Related Works. There are only few prior works that have attempted deinterlacing using deep learning [5, 6, 7]. Yet, [5, 6] consider only intra-frame deinterlacing and [7] does not employ feature-level alignment of supporting fields; hence, their results are suboptimal in the presence of large motion. As deformable convolution has been adopted in many video processing tasks [8, 9, 10, 11], we presented one of the first works on feature-level field alignment using deformable convolutions specialized for the deinterlacing task [12]. This paper is an improved edition of [12], and achieves comparable performance to recent papers [13, 14] that also adopt deformable convolution for feature-level alignment for the deinterlacing. However, different from [12], with the excellent results in computer vision applications achieved by self attention (SA) [15], this paper proposes to combine deformable convolution residual (DfRes) block and self-attention block in parallel to align field feature for full-frame rate deinterlacing.

Contributions. This paper advances the state-of-the-art in learned deinterlacing as follows:

1) Different from our early work [12], we employ a self-attention module in parallel to the DfRes blocks and show that this new architecture provides improved performance by enabling better multi-frame feature alignment and fusion.

2) We propose to utilize separate reconstruction module to restore even and odd fields separately, with which the overall performance of all methods presented here are significantly improved over those in [12].

3) We provide ablation and generalization results to show the effectiveness of proposed model over competing learned deinterlacing methods.

This work was supported in part by TUBITAK 2247-A Award No. 120C156 and KUIS AI Center funded by Turkish Is Bank. A.M. Tekalp also acknowledges support from Turkish Academy of Sciences (TUBA). Ronglei Ji would like to acknowledge a Fung Scholarship.

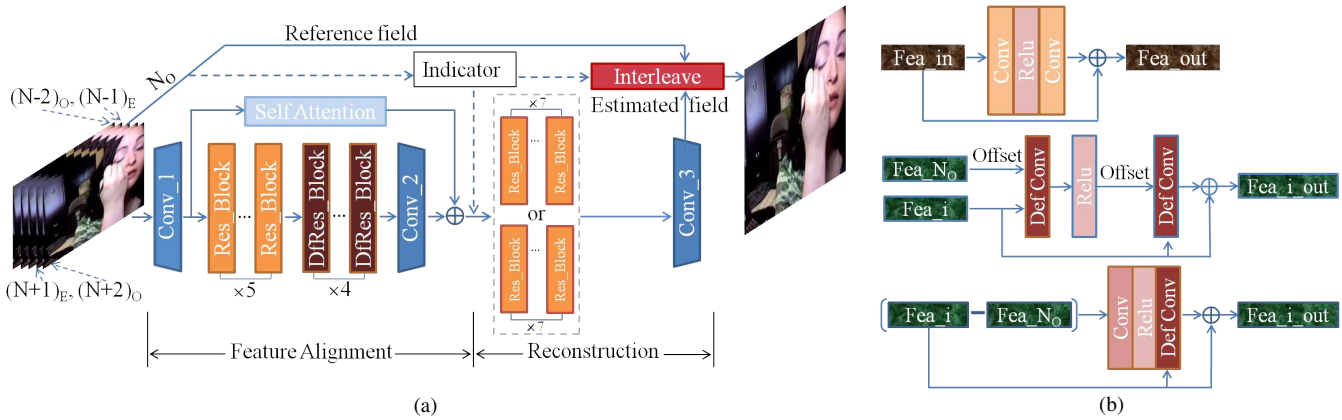


Fig. 1. (a) The proposed deinterlacing network with five input fields $(N-2)_O, (N-1)_E, N_O, (N+1)_E, (N+2)_O$, where the reference field for alignment is N_O, the field to be estimated is N_E. (b) Top, block diagram of a standard residual block depicted by orange boxes in (a). Middle, block diagram of a DfRes block shown by the brown box in (a). Bottom, block diagram of a differential DfRes (ΔDfRes) block. Fea_i represents each input field. Fea_i_out is the corresponding output feature after one DfRes (ΔDfRes) block and it will replace Fea_i to be aligned through next DfRes (ΔDfRes) block.

3. PROPOSED DEINTERLACING ARCHITECTURE

3.1. Overview of the Proposed Architecture

The proposed overall architecture is depicted in Fig. 1 (a). The input to the field feature alignment stage consists of an odd number of fields centered about the reference field. We use an indicator bit, which is 1 when the reference field is an even field; or 0 when the reference field is an odd field, when forming the output progressive frames.

Each data field is first mapped into 64 feature channels via the convolution layer Conv_1 and next processed by five regular residual blocks. These features are then fed into four separate alignment modules to align the features of each supporting field to those of the reference field using deformable convolution residual blocks (DfRes_Block $\times 4$). The aligned features are then stacked and they are fused and reduced to 64 channels through Conv_2.

In addition, we employ an SA module in parallel to the DfRes blocks to enhance the performance of feature registration. The 64-channel feature tensors processed by the SA module and DfRes module in parallel are added together to form the final 64-channel aligned feature tensor. Experiments demonstrate the effectiveness of the proposed strategy.

The 64-channel aligned feature tensor is input to the reconstruction stage, where it is processed by two branches of 7 regular residual blocks each. The last Conv_3 layer maps 64 channels to 3 channels to generate the estimated field.

Finally, we interleave the available reference field and estimated opposite parity field using the indicator bit to form progressive output frames as shown in Fig. 2 (d) (f) (g). If the indicator field is 0 (1), we place the estimated field in the even (odd) lines of the output progressive frame.

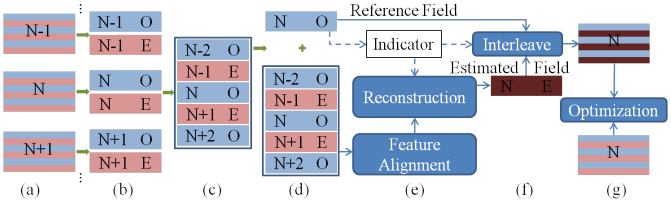


Fig. 2. Overview of data processing during training. (a)-(b) synthetic interlaced videos are generated by extracting odd and even fields of video frames, where N_O and N_E represent odd and even fields of frame N. (c) input fields. (d)-(g) deinterlacing process.

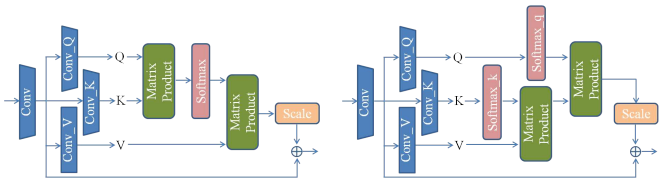
3.2. Field Alignment using Deformable Convolution Blocks

Deformable convolution residual (DfRes) block [12], shown in Fig. 1 (b) middle diagram employs two deformable convolution layers instead of regular convolutions in order to align features of supporting fields with those of the reference field.

ΔDfRes Block [12] has similar performance but less parameters compared to DfRes block, and it is depicted in Fig. 1 (b) bottom diagram.

3.3. Field Alignment using Efficient Self-Attention

The SA module shown in dashed light blue box in Fig. 1 (a) and depicted in detail in Fig. 3 (a), is parallel to feature extraction Res_Block $\times 5$ module and feature alignment DfRes module. The data fields with n pixels in each field are first mapped to 64 channel feature space via regular convolution operations (Conv_1 in Fig. 1 (a) and Conv in Fig. 3 (a)). Then, queries (Q $\in R^{8\times n}$), keys (K $\in R^{8\times n}$) and values (V $\in R^{64\times n}$) are computed by regular convolution operations



(a) (b)

Fig. 3. (a) Block diagram of the self-attention module, where Softmax is defined in Equation 2. (b) Block diagram of the efficient self-attention module, where Softmax_k and Softmax_q are defined in Equation 4.

Conv_Q, Conv_K and Con_V separately, where the output feature channels of Conv_Q and Conv_K are $\frac{1}{8}$ times of those input feature channels as shown with blue trapezoids in Fig. 3 (a), and the input and output channels of Con_V are the same as shown with blue rectangle in Fig. 3 (a). After obtaining Q, K and V matrix, we multiply Q and K by transposing Q and feed the result into softmax operation $\rho(Q^T \times K)$ for normalization. Finally, we multiply V and the transpose of the normalized softmax output to gain an initial output of Self Attention Module. The final output of Self Attention Module is the original 64 channel feature resulted from Conv in Fig. 3 (a) plus the scaled initial output as shown in plus sign in Fig. 3 (a). Hence, the Self Attention Module operation can be expressed in the following equation:

$$SA(Q, K, V) = V \times (\rho(Q^T \times K))^T \cdot Scale, \quad (1)$$

where T denotes matrix transpose operation and $\times$ matrix product. Scale is a learnable factor for scaling initial output. We use softmax function for normalization:

$$Softmax: \ \rho(Y) = \sigma_{row}(Y), \quad (2)$$

where σ_{row} applys softmax rowwisely.

In dealing with the huge computation cost of self attention, we utilize the following approximate equation in test [16]:

$$ESA(Q, K, V) = V \times \rho_k(K^T) \times \rho_q(Q) \cdot Scale, \quad (3)$$

where ρ_k and ρ_q are softmax functions utilized to normalize K and Q features. They are defined as:

$$\begin{aligned} Softmax_k: \ \rho_k(Y) &= \sigma_{col}(Y), \\ Softmax_q: \ \rho_q(Y) &= \sigma_{row}(Y), \end{aligned} \quad (4)$$

where σ_{col} and σ_{row} apply softmax columnwisely and rowwisely, respectively.

Efficient Self Attention Module as shown in Fig. 3 (b) reduces test time to the level of models without self attention module with only about 0.02 dB PSNR loss, which generates a close approximation due to the non-linearity of softmax function. We call the DfRes module based deinterlacing network with this parallel Self Attention Module DfRes_SA.

Table 1. Comparison of PSNR (top) and SSIM (bottom) for different methods on multiple datasets. The best and the second best are marked with red and blue color respectively.

Datasets	**DfRes_SA**	**DfRes**	EDVR	TDAN	DUF
UCF101	51.92	51.48	51.23	47.74	49.75
	0.99931	0.9993	0.9992	0.998	0.999
Vimeo	44.35	43.45	42.15	40.73	40.66
	0.9949	0.994	0.991	0.9889	0.989
REDS	37.00	36.51	33.20	30.09	31.36
	0.984	0.982	0.964	0.944	0.955

Table 2. Comparison of regular offset estimation and proposed DfRes offset estimation on UCF101 dataset.

Methods	Regular + SA	DfRes_SA
PSNR	51.89	51.92
SSIM	0.99929	0.99931

3.4. Frame Reconstruction Module

We propose to employ two same reconstruction blocks in parallel (i.e., Res_Block $\times 7$, right two orange box in Fig. 1 (a), and details are in Fig. 1 (c)) to reconstruct even and odd fields separately in each data batch, which can restore fields more pertinently and directionally and thus produce better performance than the model with only one reconstruction block. We call these two parallel reconstruction blocks Separate Reconstruction Module. We utilize this proposed module in all three proposed DfRes designs, which are DfRes, ΔDfRes and DfRes_SA, and In Section 4, we will compare deinterlacing results from these three designs.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Settings

Dataset. We used UCF101 [17], REDS [10, 18] and Vimeo-90K [19] datasets with the same separation in [12] to demonstrate the effectiveness of our method and compare it with other methods. We also used Vid4 [20] dataset for testing.

Evaluation Methods. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are used to quantitatively evaluate results and zooming in on details for subjective evaluation of the results.

Training Procedure. All the training settings are the same with [12] except that we train our model for 150,000 iterations using Pytorch over tesla_v100 and that in the loss function in [12] we replace $SSIM(Y^{pred}, Y^{GT})$ with $0.1 \cdot Cb(Y^{pred}, Y^{GT})$, where $Cb(\cdot)$ is Charbonnier Loss.

4.2. Comparison with the State-of-the-Art methods

We compare our proposed DfRes, ΔDfRes and DfRes_SA deinterlacing methods with three state-of-the-art SR methods: EDVR [10], TDAN [9] and DUF [21]. We plug modules of

Table 3. Ablation study of proposed method.

Modules	DfRes+SA	only DfRes	only SA
PSNR	51.92	51.48	50.23
SSIM	0.99931	0.99925	0.99907

these three comparison networks into our deinterlacing architecture by replacing modules in Fig. 2 (e) with corresponding modules of the comparison networks. We do not modify these comparison methods except that we change the upsampling step to direct combination of reference field and estimated field specifically for deinterlacing as shown in Fig. 2 (f)-(g).

The quantitative comparison results in Table 1 show that on three datasets, our proposed DfRes and DfRes_SA methods outperform all other methods in terms of PSNR and SSIM.

Visual comparisons in Fig. 4, which include EDVR method without attention module (EDVR_woTSA), show that our proposed DfRes_SA, DfRes and ΔDfRes methods produce sharper frames and EDVR method plugged into our deinterlacing architecture can also yield comparable visual effect.

In [8, 9, 10], offsets are estimated by regular convolution operation on the concatenation of neighboring field and reference field. In Table 2, we demonstrate the improvement of proposed DfRes offset estimation over regular one.

Ablation study in Table 3 shows that Deformable Convolution and Self Attention module can complement each other for feature alignment to improve performance.

Visual comparisons shown in Fig. 4 illustrate that our proposed methods produce sharper results than [5], which is also neural network based full frame rate deinterlacing methods.

More quantitative and qualitative comparison results concerning proposed methods for same category deinterlacing methods can be found on MSU benchmark website[22].

4.3. Evaluation of Generalization Performance

Table 4 shows that our method in general generalizes well but the performance varies by the amount of motion in the training and test sets. Extensive experimental results reveal that both training and test dataset affect the performance of deinterlacing methods.

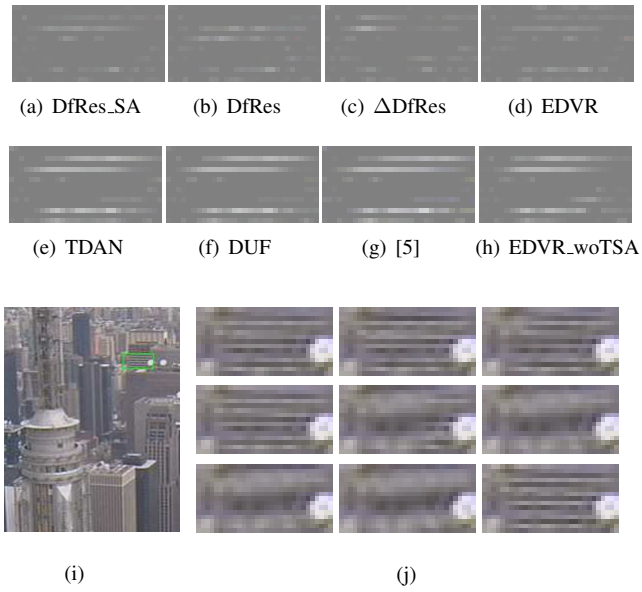


Fig. 4. Visual evaluation of a deinterlaced frame from Vid4 dataset. (a)-(h) Differences between the actual and reconstructed progressive frames. (i) Progressive ground truth frame. (j) Zooming in the green box. From upper left to lower right: DfRes_SA (PSNR: 37.29), DfRes (PSNR: 37.15), ΔDfRes (PSNR: 37.25), EDVR (PSNR: 36.93), EDVR_woTSA (PSNR: 36.21), TDAN (PSNR: 35.41), DUF (PSNR: 36.08), [5] (PSNR: 29.31), and ground truth.

5. CONCLUSION

We introduce a new network architecture for video deinterlacing with feature alignment by combining deformable convolution residual block and self attention, which can incorporate state of the art superresolution approaches for deinterlacing task. Experimental results demonstrate that the proposed network yields state-of-the-art performance and can be adopted to effectively deinterlace videos.

Table 4. Generalization of the proposed DfRes, ΔDfRes and DfRes_SA methods trained and tested on different datasets.

Train on		**DfRes_SA:** Test on			**DfRes:** Test on			**ΔDfRes:** Test on		
		UCF101	Vimeo	REDS4	UCF101	Vimeo	REDS4	UCF101	Vimeo	REDS4
UCF101	PSNR	51.92	40.78	34.86	51.48	40.59	34.40	51.46	38.99	34.39
	SSIM	0.999	0.990	0.975	0.999	0.990	0.972	0.999	0.984	0.972
Vimeo	PSNR	46.63	44.35	36.00	46.06	43.45	35.20	46.11	43.71	35.39
	SSIM	0.998	0.995	0.980	0.998	0.994	0.976	0.998	0.994	0.977
REDS	PSNR	38.04	42.39	37.00	37.26	42.04	36.51	37.78	41.82	36.49
	SSIM	0.9919	0.9933	0.9835	0.9917	0.9928	0.9817	0.9915	0.9924	0.9816

6. REFERENCES

[1] A Murat Tekalp, *Digital Video Processing*, Prentice Hall, 2015.

[2] Hoon Yoo and Jechang Jeong, "Direction-oriented interpolation and its application to de-interlacing," *IEEE Transactions on Consumer Electronics*, vol. 48, no. 4, pp. 954–962, 2002.

[3] Wonki Kim, Soonjong Jin, and Jechang Jeong, "Novel intra deinterlacing algorithm using content adaptive interpolation," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 3, pp. 1036–1043, 2007.

[4] Ohjae Kwon, Kwanghoon Sohn, and Chulhee Lee, "Deinterlacing using directional interpolation and motion compensation," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 1, pp. 198–203, 2003.

[5] Haichao Zhu, Xueting Liu, Xiangyu Mao, and Tien-Tsin Wong, "Real-time deep video deinterlacing," *arXiv preprint arXiv:1708.00187*, 2017.

[6] Ahmet Oğuz Akyüz et al., "Deep joint deinterlacing and denoising for single shot dual-iso hdr reconstruction," *IEEE Transactions on Image Processing*, vol. 29, pp. 7511–7524, 2020.

[7] Michael Bernasconi, Abdelaziz Djelouah, Sally Hattori, and Christopher Schroers, "Deep deinterlacing," in *SMPTE Annual Technical Conf. Exhibition*, 2020.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2017, pp. 764–773.

[9] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3360–3369.

[10] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[11] M Akın Yılmaz and A Murat Tekalp, "Dfpn: Deformable frame prediction network," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1944–1948.

[12] Ronglei Ji and A Murat Tekalp, "Learned multi-field deinterlacing with feature alignment via deformable residual convolution blocks," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2021, pp. 1–5.

[13] Yang Zhao, Yanbo Ma, Yuan Chen, Wei Jia, Ronggang Wang, and Xiaoping Liu, "Multi-frame joint enhancement for early interlaced videos," *arXiv preprint arXiv:2109.14151*, 2021.

[14] Yuqing Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao, "Spatial-temporal correlation learning for real-time video deinterlacing," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li, "Efficient attention: Attention with linear complexities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3531–3539.

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[18] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2019.

[19] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[20] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown, "Frame-recurrent video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6626–6634.

[21] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3224–3232.

[22] Online, "Msu deinterlacer benchmark," `https://videoprocessing.ai/benchmarks/deinterlacer.html`, 2021.