EDVR and U-Net for video quality mapping and artifacts removal

Łukasz Bala, Michał Kudelski, Dmitry Hrybov, Marcin Możejko

ML in PL
Virtual Event 2020

Problem statement

- Image denoising and artifacts removal were of big interest in scientific community in recent year, especially after deep learning methods popularization Recently, new methods were developed for video, in order to take into account relationship between subsequent frames. Especially attention based methods and modifications to standard convolutional filters contribute to advancements in this field.

- In order to speed up research in video area, NTIRE workshop was set up, with cooperation between ETH Zurich, CVPR conference and many private and public entities. We present our results from Video Quality Mapping path, which aim was to develop method to map quality from video in worse (for example more compressed) quality to the video in better one.
  - Training set with 60 video pairs
  - Validation set with 20 videos
  - Test set with 20 videos

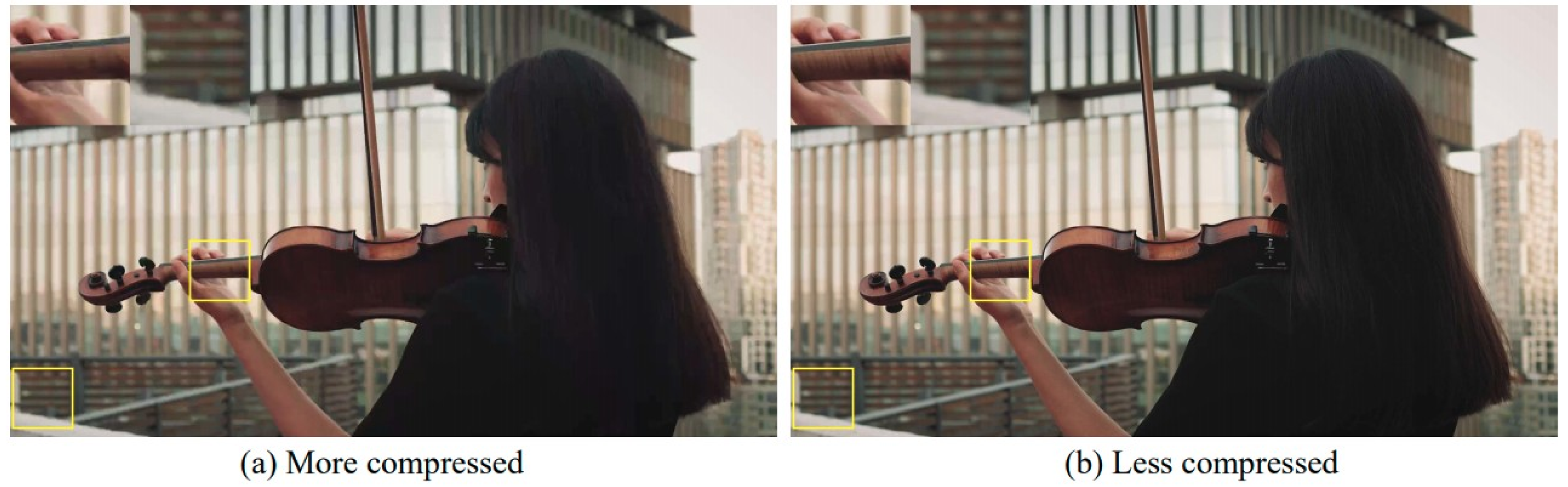

Fig. 1. Example of training (left) and ground truth (right) data from IntVid dataset.

Proposed solution

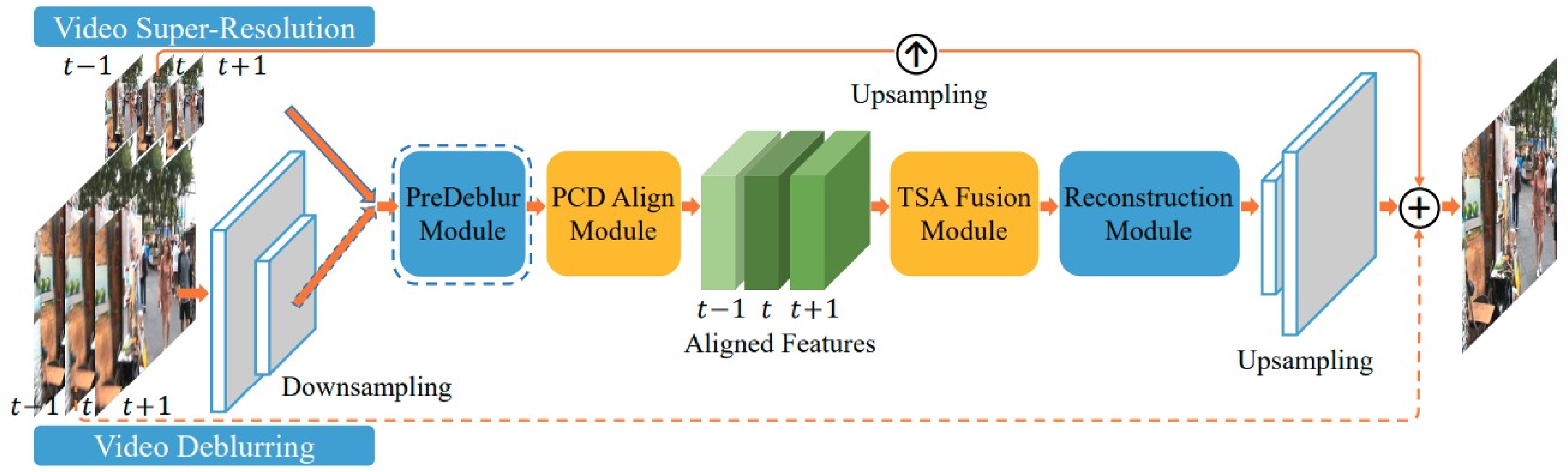

Fig 2. EDVR pipeline as presented in original paper [2].

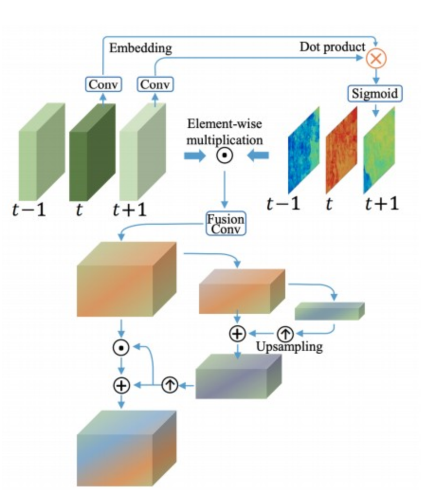

Fig 3. Schematic representation of PCD module.

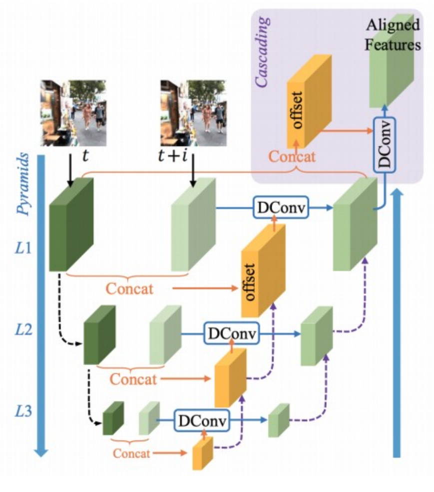

Fig 4. Schematic representation of TSA module.

The most important parts of a pipeline presented on Figure 2 are:

- PCD (Pyramid, Cascading and Deformable Convolution)
- TSA (Temporal and Spatial Attention)

The first uses deformable convolutions – intelligent kernels that learn offsets for each point of a filter. The second one takes advantage of attention mechanism in two different ways – temporal and spatial.

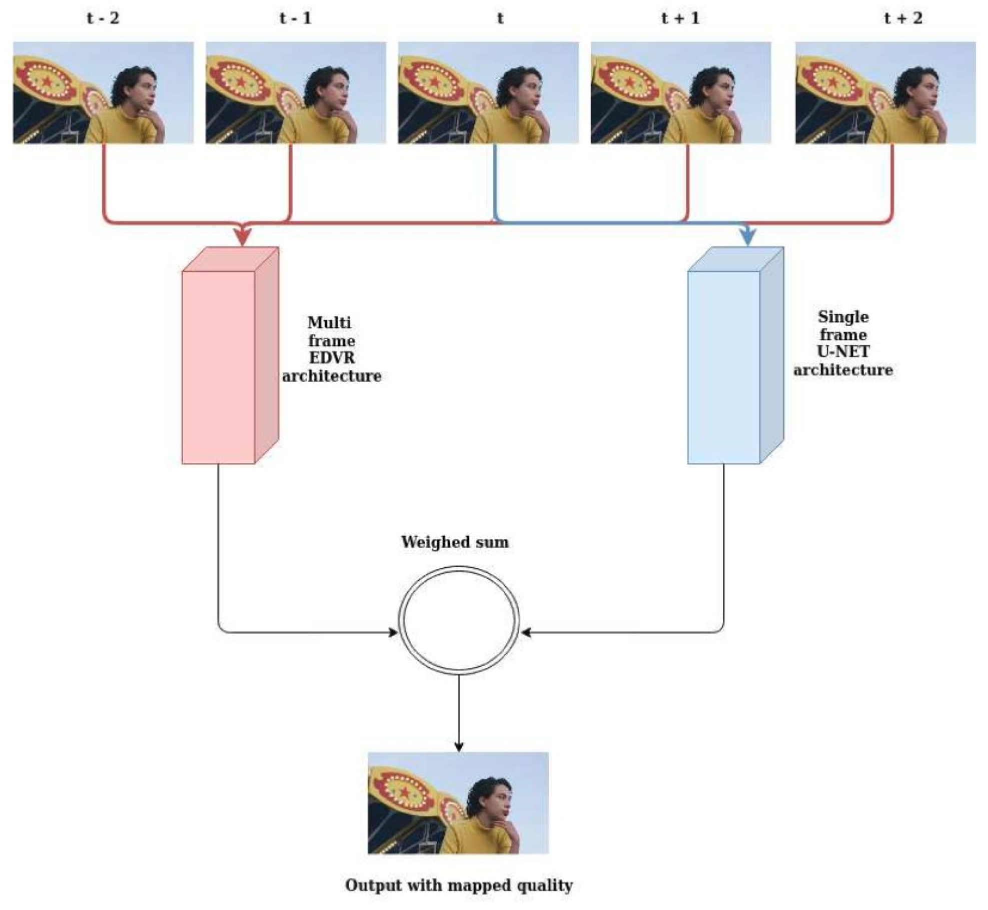

Fig 5. Our pipeline presented in final results [1].

Our final solution takes advantage not only from EDVR pipeline[2], but we strengthen intra-frame relationship by using U-Net architecture[3]. Thanks to that we have not only taken into account inter-frame dependencies with deformable convolutions and attention, but consistency inside the frame are higher as well. It's important to note that we use learning rate schedulers and freeze and un-freeze parts of the architecture in order for training to converge.

Results and final thoughts

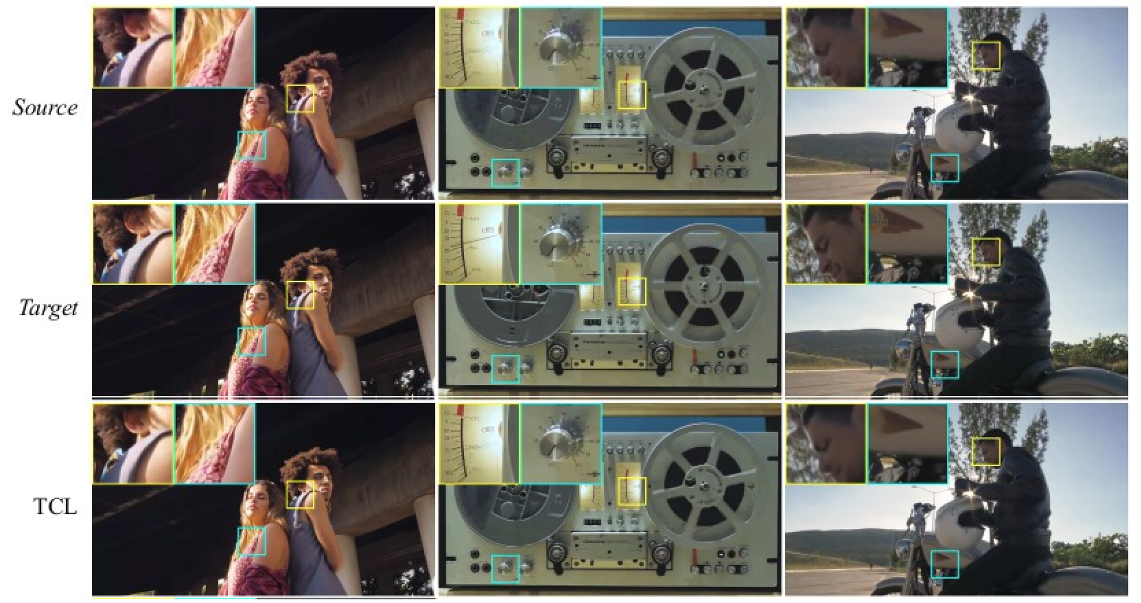

Fig 6. Examples of ours network inference as compared to source and target images.

One can see that after applying our trained network to videos from test set we are able to remove most of the artifacts inside frame as compared to source videos, however some of the details are lacking (such as captions). Moreover, there is an artificial feeling present due to too much smoothness.

	Method	↑PSNR	↑SSIM	↓LPIPS	TrainingReq	TrainingTime	TestReq	TestTime	Parameters	ExtraData
Participants	BossGao	**32.419**	**0.905**	0.177	8×V100	5-10d	1×V100	4s	n/a	No
	JOJO-MVIG	32.167	0.901	0.182	2×1080Ti	≈ 4d	1×1080Ti	2.07s	≈22.75M	No
	GTQ	32.126	0.900	0.187	2×2080Ti	≈ 5d	1×2080Ti	9.74s	19.76M	No
	ECNU	31.719	0.896	0.198	2×1080Ti	2-3d	1×1080Ti	1.1s	n/a	No
	TCL	31.701	0.897	0.193	2×1080Ti	≈ 3d	1×1080Ti	25s	≈8.92M	No
	GIL	31.579	0.894	0.195	1×970Ti	≈ 6d	1×970Ti	11.37s	3.60M	No
	7-th team	30.598	0.878	**0.176**	n/a	4d	n/a	0.5s	≈7.92M	Yes
	No processing	30.553	0.877	**0.176**						

Fig 7. Results of our network team as measured by SSIM and PSNR metrics. We achieve 5th place in PSNR metric and 4th in SSIM.

$$SSIM = \frac{\left(\left(2 * \mu_x \mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}$$

$$PSNR = 20 * \log_{10} * \left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

References

[1] Fuoli, Dario, et al. "NTIRE 2020 challenge on video quality mapping: Methods and results." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
[2]Wang, Xintao, et al. "Edvr: Video restoration with enhanced deformable convolutional networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.
[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.