Visual Media Enhancement using CNNs

Claudio Rota
Department of Informatics Systems and Communication
University of Milano-Bicocca, Italy

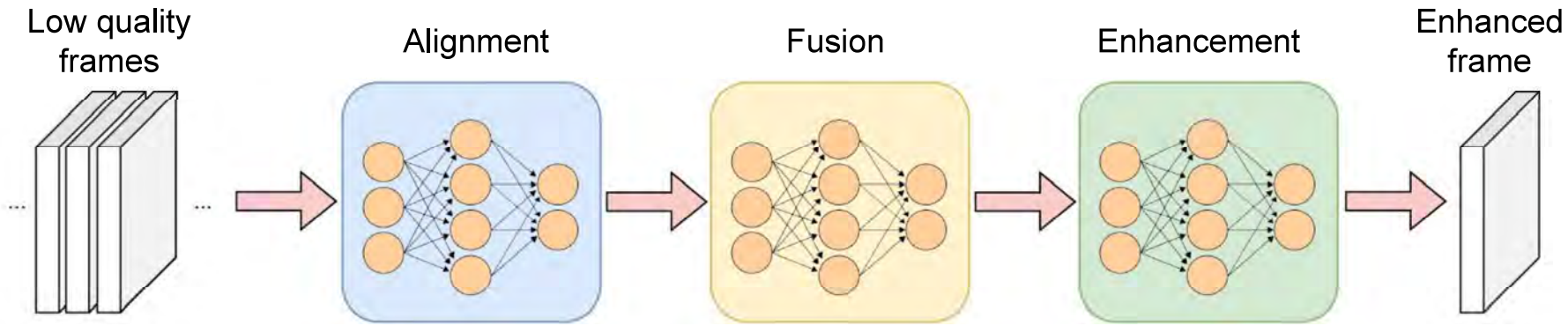
Visual media enhancement

- The process of improving the perceived quality of visual media to make them more appealing to users.
- Several refinements among noise removal, reduction of compression artifacts, deblurring, contrast enhancement, super-resolution etc.
- Previously performed using traditional methods (e.g., histogram equalization, wavelet decomposition), recent methods are now based on Convolutional Neural Networks (CNNs)

Key elements of CNNs for video enhancement

General framework:

- **Alignment**: multiple adjacent frames are spatially aligned to the target one
- **Fusion**: aligned features are fused to compensate the lack of details in the target frame
- **Enhancement**: fused features are used to enhance in the target frame and obtain the enhanced frame



Rota, Claudio, et al. "Video restoration based on deep learning: a comprehensive survey". Submitted to Artificial Intelligence Review (2022).

Alignment solutions

- Motion estimation and motion compensation
  - Per-pixel motion vectors between source and target frames are estimated using optical flow estimation methods
  - Estimated motion vectors are used to warp the source frame to the target one
- Deformable alignment
  - Alignment is performed using a deformable convolution layer
  - Features from source and target frames are processed together using a CNN to learn spatial deformable offsets
- Non-local search
  - Patch-based similarity is computed between target and a set of source frames
  - Patch similarity helps finding matching regions, which are expected to have very similar contents
- Implicit alignment
  - No specific module for frame alignment
  - The network learns the best suitable transformations to apply to a stack of input frames

Rota, Claudio, et al. “Video restoration based on deep learning: a comprehensive survey”. Submitted to Artificial Intelligence Review (2022).
Xue, Tianfan, et al. “Video enhancement with task-oriented flow.” *International Journal of Computer Vision* 127.8 (2019): 1106-1125.
Dai, Jifeng, et al. "Deformable convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2017.
Wang, Xintao, et al. "Edvr: Video restoration with enhanced deformable convolutional networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

Video enhancement under multiple distortions

Video enhancement methods usually address one enhancement task at a time, but videos may contain multiple distortions simultaneously

A CNN for noise and compression artifact removal

Frame alignment is performed implicitly

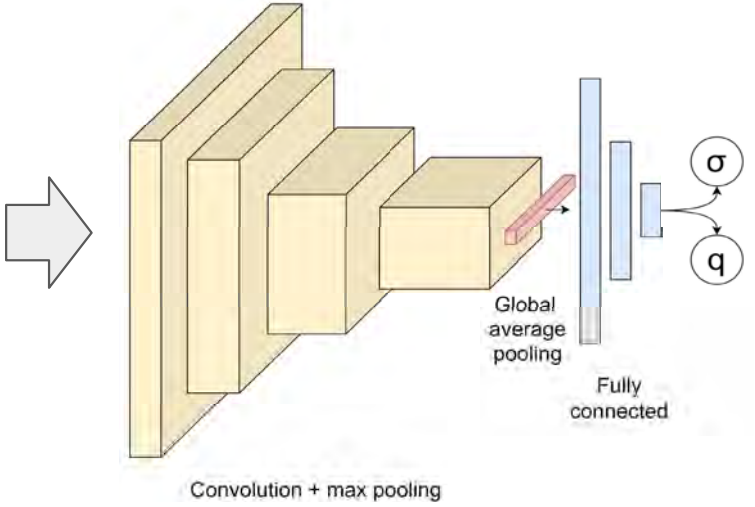
The main features are:

- Estimation of distortion intensity
- Feature processing using multiple scales
- Enhancement process in two steps

Rota, Claudio, and Marco, Buzzelli. "MdVRNet: deep video restoration under multiple distortions." *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.* 2022.

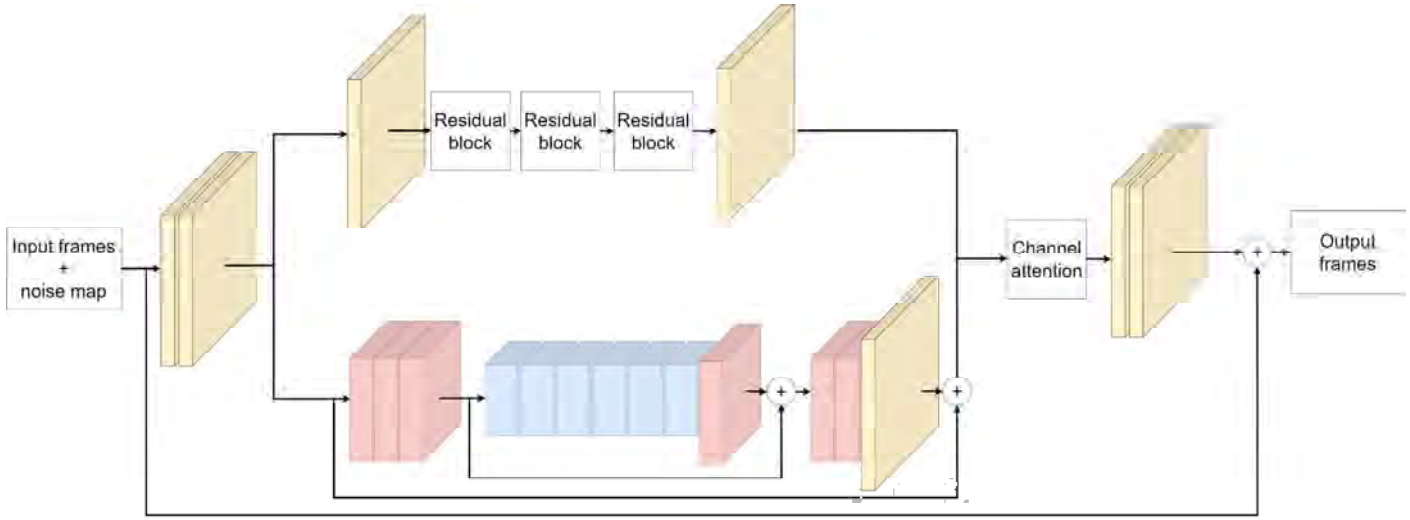
Estimation of distortion intensity

- Non-blind methods perform better than blind methods, but they require additional knowledge which limits their applicability
- If the enhancement network knows the intensity of the underlying artifacts, performance improves
- A CNN is used to estimate them
- The estimated parameters are expanded as feature maps and concatenated with input frames
- +0.2 db in PSNR, +0.01 in SSIM



Feature processing using multiple scales

- Two parallel branches with different purposes
- The high-resolution branch (upper branch) focuses on detail enhancement
- The low-resolution branch (lower branch) learns to enhance the main structures
- Channel attention is finally performed to properly fuse the complementary features
- +0.35 in PSNR, +0.01 in SSIM

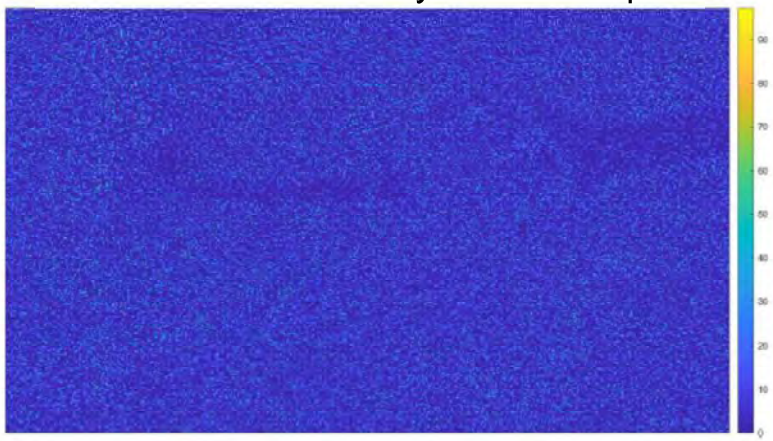


Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018
Zamir, Syed Waqas, et al. "Learning enriched features for real image restoration and enhancement." *European Conference on Computer Vision*. Springer, Cham, 2020.

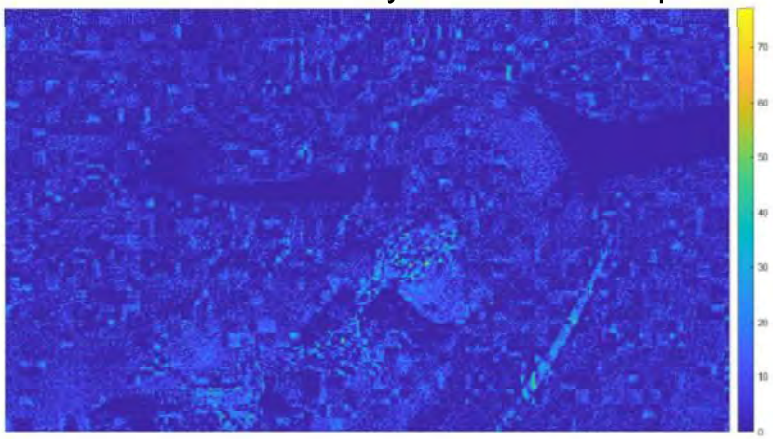
Enhancement process in two steps

- Two networks are cascaded
- The first enhancement step removes artifacts introduced at pixel level (e.g., noise)
- The second step removes artifacts introduced in local regions (e.g., compression artifacts)
- Below are the residual at each enhancement step
- +0.3 in PSNR, +0.01 in SSIM

Artifacts removed by the first step



Artifacts removed by the second step



Quantitative results

Comparison with a state-of-the-art model for video enhancement (denoising)

Training on DAVIS 2017 trainset, testing on DAVIS 2017 testset and Set8 dataset

Three levels of distortion tested:

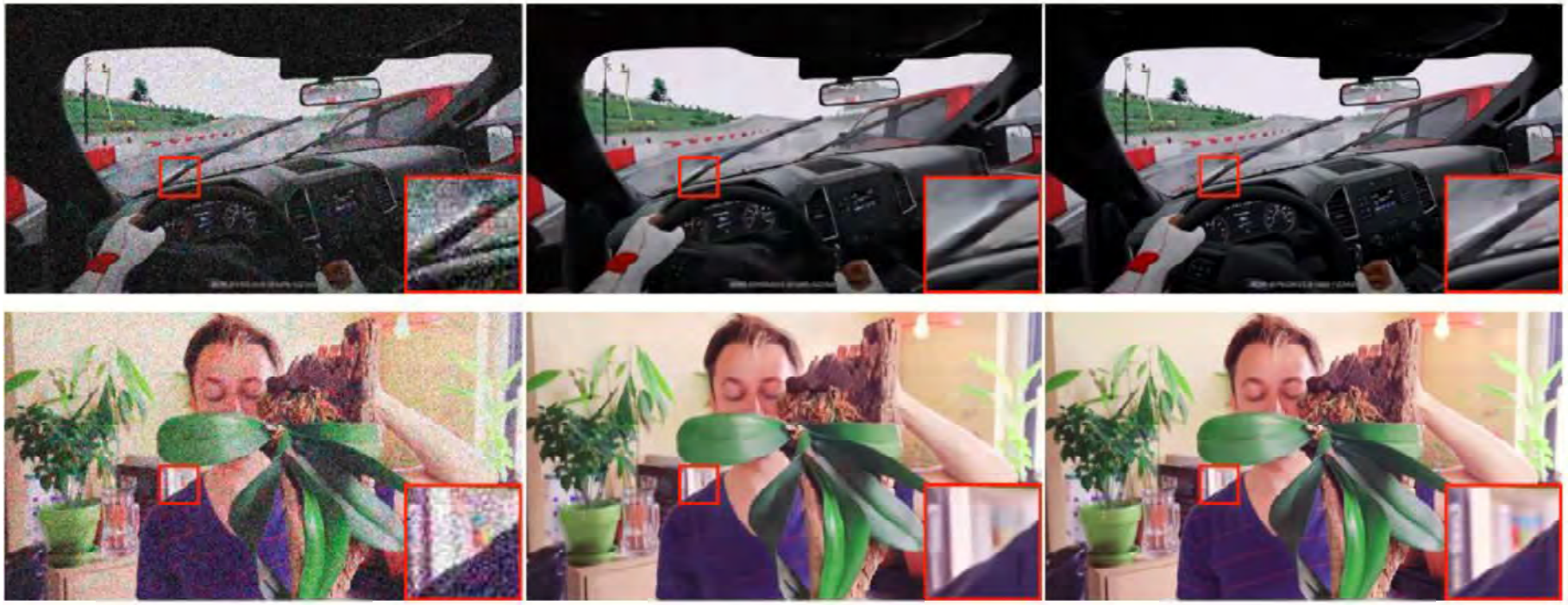
- Mild $\longrightarrow \sigma = 10, q = 35$
- Moderate $\longrightarrow \sigma = 30, q = 25$
- Severe $\longrightarrow \sigma = 50, q = 15$

+0.51 db in PSNR, +0.02 in SSIM on the DAVIS 2017 test set, +1.18 db in PSNR, +0.05 in SSIM on the Set8 dataset

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

Tassano, Matias, Julie Delon, and Thomas. "Fastdvdnet: Towards real-time deep video denoising without flow estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
Pont-Tuset, Jordi, et al. "The 2017 davis challenge on video object segmentation." *arXiv preprint arXiv:1704.00675* (2017).

Qualitative results



Distorted Enhanced GT

Challenges and future research directions

- Real-time enhancement performance
  - Methods are commonly evaluated on high-performing GPUs
  - We now expect to run these models on smartphones in real time
- Studying more robust alignment strategies
  - Existing alignment strategies have several limitations
  - Combination of these strategies may lead to result improvement
- Definition of more suitable evaluation metrics
  - PSNR and SSIM are commonly used, but they do not correlate well with human perception
  - Temporal consistency of the results is only rarely taken into account
- Realization of datasets containing real distortions
  - Most existing datasets for video enhancement are synthetically generated
  - Complex acquisition systems required to capture distorted and clear videos

Rota, Claudio, et al. “Video restoration based on deep learning: a comprehensive survey”. Submitted to Artificial Intelligence Review (2022).
Chan, Kelvin CK, et al. "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment." *arXiv preprint arXiv:2104.13371* (2021).


THANK YOU!