2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9898028

# MANET: IMPROVING VIDEO DENOISING WITH A MULTI-ALIGNMENT NETWORK

*Yaping Zhao*[1,3,*] , *Haitian Zheng*[2,*], *Zhongrui Wang*[1,3], *Jiebo Luo*[2], *Edmund Y. Lam*[1,3,†]

[1] The University of Hong Kong [2] University of Rochester
[3]ACCESS — AI Chip Center for Emerging Smart Systems


**Fig. 1**: Given a noisy video sequence as input, we propose a neural network using multiple alignments for video denoising to synthesize clean frames.

## ABSTRACT

In video denoising, the adjacent frames often provide very useful information, but accurate alignment is needed before such information can be harnassed. In this work, we present a multi-alignment network, which generates multiple flow proposals followed by attention-based averaging. It serves to mimic the non-local mechanism, suppressing noise by averaging multiple observations. Our approach can be applied to various state-of-the-art models that are based on flow estimation. Experiments on a large-scale video dataset demonstrate that our method improves the denoising baseline model by $0.2\,\mathrm{dB}$, and further reduces the parameters by 47% with model distillation. Code is available at https://github.com/IndigoPurple/MANet.

***Index Terms*—** video denoising, image alignment, video enhancement, attention, image synthesis

## 1. INTRODUCTION

Video denoising aims to restore a clean video sequence from one that is corrupted with noise. Due to thermal effects, sensor imperfections or low-light, noise inevitably degrades the quality of many captured videos, for which video denoising becomes necessary. Additionally, it is also an indispensable sub-task of the remastering of vintage films, which are corrupted with a lot of noise because of technical limitations.

Unlike image denoising, video denoising can take advantage of information from adjacent frames. However, correspondence matching and frame alignment are crucial problems. Recent works [1, 2] rely on optical flow together with backward warping to perform such frame alignment. However, the existence of occlusion, motion blur, rotation or lighting change across different frames decreases the accuracy of many alignment methods. Similarly, non-local means [3] is a powerful technique in image denoising because of its ability to suppress noise through averaging multiple observations of similar patches. In the context of video denoising, this approach would again rely on accurate frame alignment [1, 2, 4].

In this paper, we present a new video denoising method

*Equal contribution. †Corresponding author.

This work is supported in part by ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR, in part by Hong Kong Research Grant Council (Grant No. 27206321), National Natural Science Foundation of China (Grant No. 62122004).

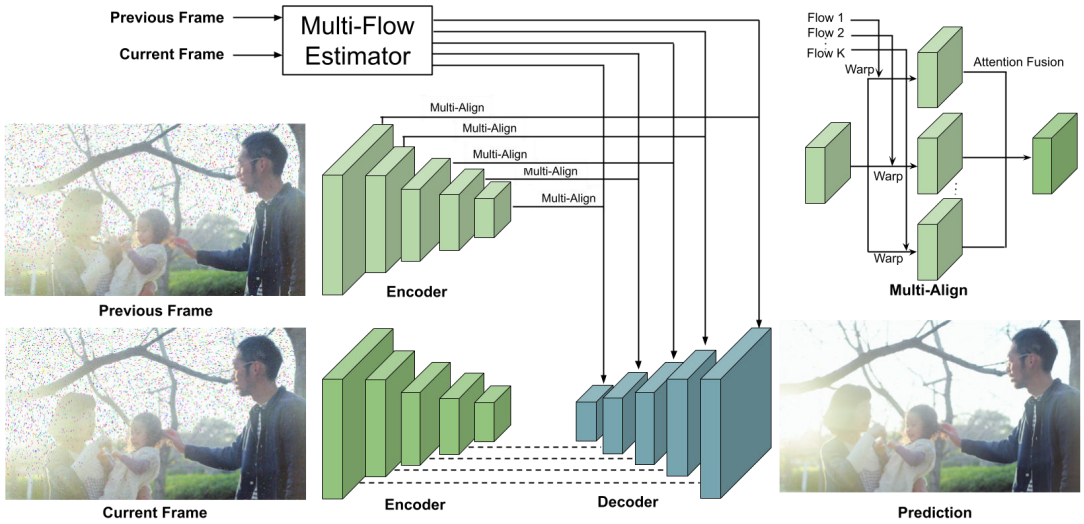


**Fig. 2**: Network structure of the video denoising framework with multiple alignment.

as Figure 1 shows, together with the multiple alignments. To address the issue of information loss from misalignment, we propose to generate multiple optical flow candidates using a learning-based flow estimator. To mimic the non-local behavior for noise cancellation, we perform an attention-based averaging on the aligned feature. Based on the designed architecture, we further perform model distillation to reduce the model size. The resulting model is lightweight, and can be easily incorporated with recent methods.

To demonstrate the effectiveness of our approach, we perform experiments on a large-scale video dataset Vimeo [2]. Compared with the recent baseline [1], our approach improves the performance by $0.2\,\mathrm{dB}$, suggesting a possible direction to improve video denoising.

## 2. RELATED WORK

Let $x_i$ be the clean frame from a video sequence at timestep $T = i$, and $y_i = c(x_i)$ be an observed video frame through a corrupted channel $c(\cdot)$. The task of video denoising is to utilize the addition information from the observed adjacent frames $Y_i(t) = \{y_j | j = i - t, \ldots, i - 1\}$ to restore a clean frame $\hat{x}_i$. Therefore, video denoising can be formalized by

$$\hat{x}_i = \mathcal{D}(y_i, Y_i). \tag{1}$$

Traditionally, the image restoration task can be formulated by the *maximum a posteriori* (MAP) estimation under a predefined image prior. Typical image priors include hyper-Laplacian [5], mixture of Gaussian [6], total variation [7], local linear embedding [8], and sparsity [9]. With the availability of image restoration datasets, the supervised learning approach has also been explored. The typical approaches include decision tree [10], random forest [11], and dictionary learning [12]. Recently, deep learning approaches [13, 14] further push the limit of the image restoration performance.

To utilize the adjacent frame information, video restoration approaches make use of the correspondence estimation to align them with the current frame. One approach is to perform patch motion compensation [15], while some others [1, 2] rely on optical flow. Specifically, let $\boldsymbol{f}_{j\to i}$ be the estimated flow field from frame $j$ to $i$, Xue *et al.* [2] performs backward warping on the adjacent frame, while Zheng *et al.* [1] additionally performs backward warping on the feature pyramid.

## 3. METHOD

### 3.1. Network Structure

To perform multiple alignments and frame synthesis for video denoising, we employ the network structure of [1] and replace its flow estimator and alignment module with our multi-alignment modules. The resulting network consists of two feature pyramid extractors, a multi-flow estimator, a multi-scale feature domain alignment module, and a UNet [16] structure for frame synthesis, as Figure 2 depicts.

### 3.2. Multiple Alignment

For simplicity, we consider denoising the current frame using observation from the current and the previous frame, *i.e.*,

$$\hat{x}_i = \mathcal{D}(y_i, y_{i-1}). \tag{2}$$

Nevertheless, it should be pointed out that our model can be easily extended to include multiple adjacent frames.

**Multiple Flow Estimation.** Similar to [1], our approach relies on a learning-based flow estimator named FlowNet [17]. However, due to possible errors in the flow estimation, information from the adjacent frame is not efficiently used. Therefore, we use FlowNet to generate $K$ optical flows $\{\boldsymbol{f}^{(1)}_{i\to i-1}, \ldots, \boldsymbol{f}^{(K)}_{i\to i-1}\}$ by increasing the channel of the flow output layer from 2 to $2K$. By generating multiple possible flow estimations, the possibility of misalignment is reduced.

**Image Alignment.** In the next step, we perform the spatial alignment $K$ times using backward warping. Specifically, we generate $K$ alignment with

$$\begin{aligned} I_w^{(1)} &= \text{warp}(\boldsymbol{f}_{i\to i-1}^{(1)}, y_{i-1}), \\ &\vdots \\ I_w^{(K)} &= \text{warp}(\boldsymbol{f}_{i\to i-1}^{(K)}, y_{i-1}), \end{aligned} \tag{3}$$

where $\text{warp}(\cdot,\cdot)$ denotes the warping operation implemented by the spatial transformer network [18].

**Attention-based Averaging.** We further propose an attention-based averaging to rule out the misalignment. It serves to suppress noise by mimicking the non-local mechanism.

Specifically, we develop two different approaches to generate the unnormalized attention map, namely, **fc** and **ip** :

- **fc** refers to fully-connected. Specifically, we add an additional $1 \times 1$ convolutional layer with $K$ output channels to the flow estimator, which generates a $K$-channel unnormalized attention map $\{b^{(1)}, \ldots, b^{(K)}\}$.
- **ip** refers to inner-product. It is based on computing the similarities between the current frame $y_i$ and the warped previous frame $I_w^{(l)}$. Specifically, $y_i$ and $I_w^{(l)}$ are individually transferred to the feature maps $\theta(y_i)$ and $\phi(I_w^{(l)})$ by a $1 \times 1$ convolutional layer. Then unnormalized attention maps are generated by inner product

$$b^{(l)} = \theta(y_i)^\top \phi(I_w^{(l)}), l = 1, \ldots, K. \tag{4}$$

The normalized attention map is generated by performing the channel-wise softmax operation at each location $(m, n)$, *i.e.*,

$$a_{mn}^{(l)} = \frac{e^{b_{mn}^{(l)}}}{e^{b_{mn}^{(1)}} + \cdots + e^{b_{mn}^{(K)}}}, \qquad l = 1, \ldots, K. \tag{5}$$

The attention weights are then used for averaging the multiple aligned feature maps, where

$$I_a = \sum_{l=1}^{K} a^{(l)} \odot I_w^{(l)}. \tag{6}$$

It is worth noting that the alignment and attention fusion in Equations 3 and 6 can be applied both on the image and the feature map, meaning that our multiple alignment framework can be incorporated into a wide range of models [1, 2, 4].

### 3.3. Model Distillation

To reduce the model size, we cut the channel size of the flow estimator model and perform model distillation based on our trained model. We use the original loss to train the slimmed model. Additionally, we leverage an $\ell_1$ loss to minimize the feature and flow prediction difference between the two models to ensure that the slimmed model has a similar performance as the trained large model.

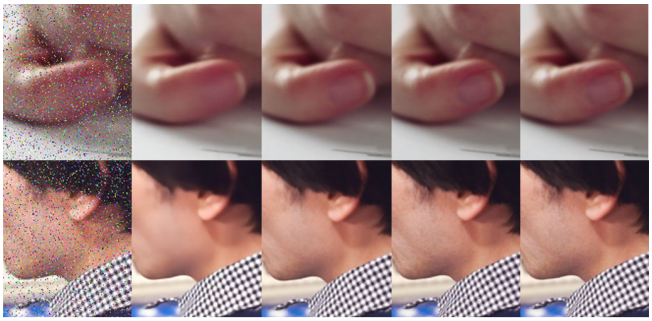


**Fig. 3**: Video denoising comparisons between different algorithms on the Vimeo dataset. Zoom in to see details.

| Method | TOFlow | CrossNet | MANet-fc | MANet-ip | MANet-ip-s |
|---|---|---|---|---|---|
| PSNR | 33.51 | 45.02 | 45.11 | 45.23 | 45.22 |
| Parameters | 17.00 M | 35.18 M | 35.46 M | 35.45 M | 18.67 M |

**Table 1**: Comparisons of denoising performance and parameter sizes among different models.

## 4. EXPERIMENTS

### 4.1. Denoising Performance Comparisons

To validate the effectiveness of our approach, we take CrossNet [1] as the baseline and replace its alignment module with our proposed multi-alignment modules, resulting in *MANet-fc* and *MANet-ip*, as mentioned in Section 3.2. Training and testing are performed on the Vimeo dataset [2] with mixed noise including a 10% salt-and-pepper noise in addition to the Gaussian noise with a standard deviation (std) of 0.1, while the performance is measured by the PSNR metric. For intuitive qualitative comparisons, we additionally take the method TOFlow [2], as Figure 3 shows.

As Table 1 shows, the *MANet-fc* improves the baseline by $0.09\,\text{dB}$ and the *MANet-ip* improves the baseline by $0.21\,\text{dB}$, while the slimmed model *MANet-ip-s* after model distillation improves the baseline by $0.2\,\text{dB}$.

### 4.2. Training Details

All the models are trained from scratch for $11$ epochs using the Adam [19] optimizer. The learning rates are set to $3 \times 10^{-5}$ and decay by a factor of $0.1$ after every $4$ epochs. In our experiment, the multiple alignment factor $K$ is set to $4$. We experimented other $K$ values and found that PSNR saturates after $K > 4$, *e.g.*, $K = 8$ increases PSNR by $0.03$. To reduce memory usage, therefore, we set $K$ to the optimal value $4$.

### 4.3. Visualization

To further understand what the multiple alignment model has learned, we visualize the generated attention individually, and

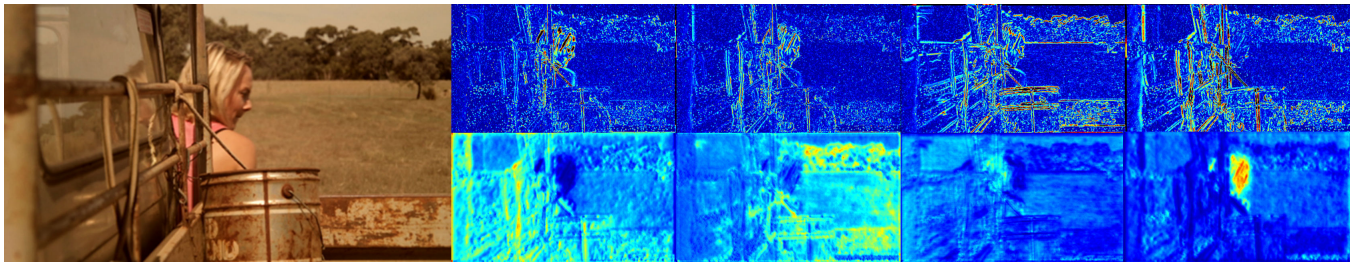

**Fig. 4**: The visualization of the ground truth (the first column), warping error (the first row on the right) and the corresponding attentional weight (the second row on the right).

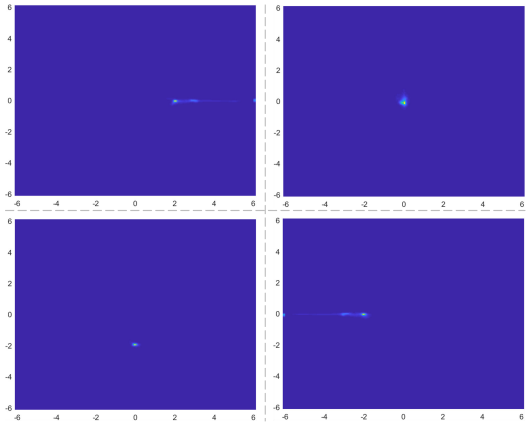

**Fig. 5**: The visualization of the offset from multiple flows generated by MANet to CrossNet-generated flow.

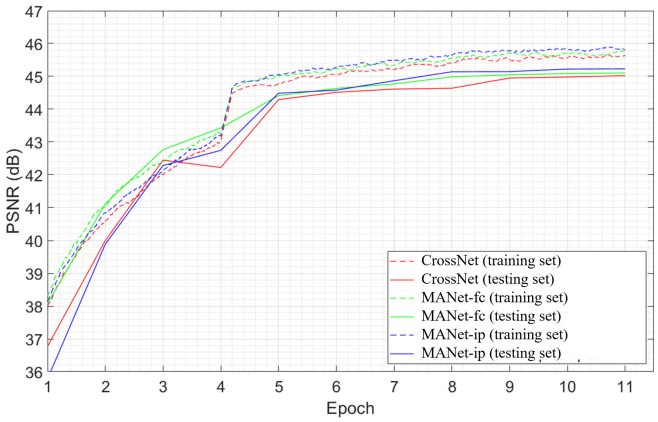


**Fig. 6**: The convergence analysis on our proposed multiple alignment methods versus the baseline model.

compare it with the error residue map by taking the difference between the reference frame and the current frame. Each row from the right side of Figure 4 depicts an attention map with its warping error residue. From the second and the third columns, we can observe that the attention map tries to avoid the center misalignment region. It can also be observed that the attention is evenly split to the $K$ attention maps for the smooth region, *e.g.*, the top left glass region.

To understand the behavior of the multiple optical flows, we additionally visualize a 2-D histogram of the offset from multiple flows generated by MANet to CrossNet-generated flow. From Figure 5, we can see that the learned multiple flow map tends to concentrate on the fixed offset. It suggests that for learning more adaptive multiple flow combinations, a more powerful flow estimator design is required.

### 4.4. Convergence

As Figure 6 shows, the *MANet-fc* converges faster than the baseline mode consistently. The *MANet-ip* converges slower than the other two models. However, it performs best at the end of the training. We speculate that the initial learning rate is too large for *MANet-ip* to converge well.

### 4.5. Network Efficiency

Our model is based on generating $K$ flow estimations for enriched feature alignment. However, the additional parameters introduced are negligible as we only increase the flow prediction layer and the convolution layer from 2 to $2K$, while adding a few more $1 \times 1$ convolutional layers for the aligned feature averaging in Eq. 6. The size of the three models are shown in Table 1. Our models *MANet-fc* and *MANet-ip* only require $0.28M(0.79\%)$ and $0.27M(0.77\%)$ additional parameters in comparison to the baseline model. Furthermore, after model distillation, the slimmed model *MANet-ip-s* reduces parameters by $47\%$.

## 5. CONCLUSION

In this paper, we present a multi-alignment network to address the misalignment issue and incorporate the non-local mean approach. The multiple flow estimation and alignment reduce the risk of misalignment. The attention-based averaging mimics the non-local component for effective denoising. Combined with knowledge distillation, model parameters are reduced by $47\%$. Experiments on a large-scale dataset show that our approach outperforms a recent baseline, suggesting a new angle for improving video denoising performance.

## 6. REFERENCES


[1] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 88–104.

[2] 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.

[3] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[4] Yaping Zhao, Mengqi Ji, Ruqi Huang, Bin Wang, and Shengjin Wang, "Efenet: Reference-based video super-resolution with enhanced flow estimation," in *CAAI International Conference on Artificial Intelligence*. Springer, 2021, pp. 371–383.

[5] Dilip Krishnan and Rob Fergus, "Fast image deconvolution using hyper-laplacian priors," *Advances in neural information processing systems*, vol. 22, pp. 1033–1041, 2009.

[6] Daniel Zoran and Yair Weiss, "From learning models of natural image patches to whole image restoration," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 479–486.

[7] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos, "Total variation super resolution using a variational approach," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 641–644.

[8] Hong Chang, Dit-Yan Yeung, and Yimin Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* IEEE, 2004, vol. 1, pp. I–I.

[9] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[10] Jordi Salvador and Eduardo Perez-Pellitero, "Naive bayes super-resolution forest," in *Proceedings of the IEEE International conference on computer vision*, 2015, pp. 325–333.

[11] Samuel Schulter, Christian Leistner, and Horst Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3791–3799.

[12] Chih-Yuan Yang and Ming-Hsuan Yang, "Fast direct super-resolution by simple functions," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 561–568.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[15] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia, "Detail-revealing deep video super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4472–4480.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.