SUPPLEMENTARY MATERIAL
Lightweight Video Denoising using Aggregated Shifted Window Attention

1. Evaluation of Memory Consumption

To demonstrate the lightweight behaviour of our method compared to the state-of-the art VRT [3], we evaluate the peak memory consumption during inference. In detail, we ran both models in inference mode for several runs using a constant input frame number of 6, but increasing the spatial image size. The maximum spatial image size was limited by VRT. Figure 1 clearly demonstrates that our method requires significantly less memory and thereby enables a more effective processing of high-resolution videos.

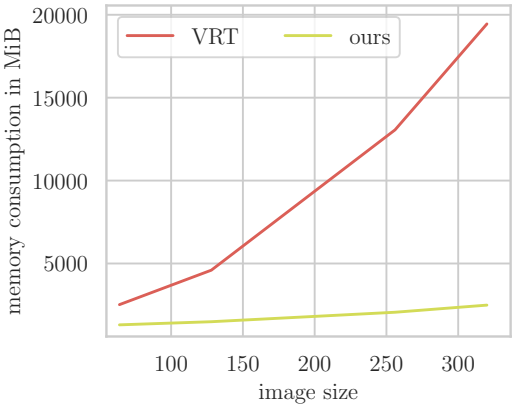


Figure 1: Evaluation of the memory consumption of VRT and our method with increasing input image size.

2. Non-reference Video Quality Assessment

In this section, we provide more detailed results for the non-reference image quality assessment (NR-IQA) we conducted, using the MUSIQ [2] score metric. We show the individual results achieved on each of the 10 real test videos along with the mean value over all videos, see Table 1. As can be seen our method performs best on 8 of the 10 videos, and second best on the remaining 2 videos. On average, our approach outperforms all other denoising methods. It is important to note that the MUSIQ network was not trained on any data used in this work. We used the pretrained model provided by the authors, trained on the KonIQ-10k dataset [4], which is the largest IQA dataset of quality-scored images.

	noisy	UDVD	MF2F	DarkEnergy	NeatVideo	ours
video 01	23.40	26.31	34.63	28.83	31.72	**38.9**
video 02	16.73	18.38	24.91	22.12	23.72	**26.97**
video 03	23.59	23.70	**33.77**	29.54	28.67	32.22
video 04	33.04	31.94	48.99	39.68	49.71	**56.14**
video 05	21.20	21.87	**32.95**	30.07	30.46	30.61
video 06	18.65	18.91	29.53	24.66	24.48	**34.62**
video 07	25.65	25.92	36.59	32.11	33.07	**37.35**
video 08	32.17	32.16	35.97	34.92	35.30	**40.42**
video 09	25.59	26.09	36.68	31.75	32.65	**40.97**
video 10	31.11	32.46	38.83	36.86	41.62	**43.40**
mean	25.11	25.77	35.29	31.05	33.14	**38.16**

Table 1: Quantitative evaluation of image quality using MUSIQ [2] for digitized analog videos. Best and second best score are printed in **bold** and blue, respectively.

3. Noise Synthesis Details

As the noise synthesis pipeline employs a double degradation, each of the following noise types, as well as the resizing operation, are applied twice in random order. Gaussian noise is applied with a probability of 1 while all other degradations are applied with a probability of 0.5.

Gaussian Noise: The noise pipeline adopts a 3D generalized zero-mean Gaussian noise model, where the correlation between the R, G, and B channels is described by a 3×3 covariance matrix, which represents the noise correlation across the color channels. The two extreme cases of this noise model are grayscale Gaussian noise and additive white Gaussian color noise. The general case and the two extreme cases are sampled with probabilities of 0.2, 0.4, and 0.4, respectively, and the noise level is uniformly sampled between $[2/255, 50/255]$.

Poisson Noise: To sample signal-dependent Poisson noise – which is generally used to represent photon shot noise – the clean image is first multiplied by 10^{γ}, then the signal-dependent Poisson noise is added, and the image is divided by 10^{γ}. Here, γ is uniformly sampled from the interval $[2, 4]$. Grayscale Poisson noise can be applied to the image by simply converting the clean image to grayscale before adding noise, resulting in the same grayscale Pois-

son noise for each RGB channel.

Camera sensor noise: Although we focus on digitized analog videos, modeling camera sensor noise is still of interest, since during the digitizing process the analog video is processed in a similar manner as in a digital in-camera image processing pipeline (ISP). This kind of noise is incorporated into the noise synthesis by applying a reverse ISP pipeline [1] to the video, resulting in raw images. Subsequently, read-and-shot-noise is added before applying the forward ISP pipeline in order to again obtain RGB images.

Speckle Noise: Multiplicative speckle noise can simply be modeled by multiplying Gaussian noise (generated by Gaussian noise synthesis as above) to a clean image.

JPEG compression noise: Since JPEG compression causes reduced image quality and can introduce strong block artifacts, also JPEG compression noise is considered in the noise synthesis pipeline. To achieve this kind of degradation the image quality factor is uniformly sampled from the interval $[30, 95]$.

Resizing: Digitized analog videos often exhibit analog film grain, which is spatially correlated noise. The resizing operation itself does not introduce any additional noise to clean videos, however, the noise distribution of a video already degraded with one of the noise models described above is altered. Spatial correlation of noise can be achieved by upsampling, while a lower signal dependency can be achieved by downsampling. Resizing is performed by using bicubic upsampling/downsampling, where the scaling factor is sampled uniformly from the interval $[0.5, 2]$.

4. User Study

We provide additional results for the conducted user study. In Figure 2, one can see the first choice of the participants for the two criteria: noise removal (blue) and temporal consistency (orange). As can be seen, our method was favoured by a large margin independent of the criterion for academic methods, see Figure 2(b), and commercial methods, see Figure 2(a). For the comparison with academic methods, MF2F was a clear second choice for both criteria evaluated. For the comparison with commercial methods, NeatVideo is the favoured second choice of the participants, again for both criteria.

Furthermore, we provide more detailed insights on how the individual participants decided for both tasks combined. Figure 3 depicts the total count of each participants 'First Choice' ratings given to the different methods for all sequences shown. Figure 3(a) shows the results for the comparison with academic methods (MF2F and UDVD), while Figure 3(b) shows the results for the commercial methods (NeatVideo) and DarkEnergy). Our method (blue) receives most 'First Choice' ratings across the different individuals. The result is even more pronounced when comparing our approach with the two academic methods MF2F and UDVD, but also still very clear when comparing to the commercial methods NeatVideo and DarkEnergy.

Before starting each test session in the user study the participants were informed about the testing procedure and were additionally given a *task instruction sheet*, see Figure 4(a), as well as a *user interface explanation*, see Figure 4(b). There was no time limit to make a decision for a video.

5. Additional Ablation Study

We investigated the influence of the positional encoding used in our ASwin block. We found that absolute positional encoding performs on par with the combination of relative and absolute positional encoding, and slightly better than just relative position encoding, see Table 2. We also analyzed the influence of the depth of our ASwin/ACSconv block. The best results were achieved with a block depth of 3 or 4, see Table 3. Since our goal was to keep the network as lightweight as possible and there was no significant performance difference between using a block depth of 3 and 4, we decided to use 3, which keeps the runtime and computational effort lower.

positional encoding			
	rel.	**abs. + rel.**	**abs.**
PSNR	36.97	37.06	37.12

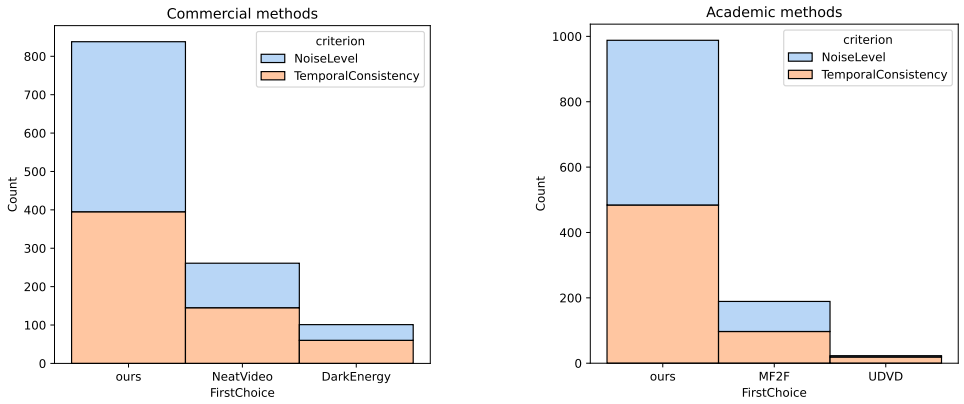
Table 2: Influence of positional encoding.

ASwin-ACSconv Block Depth			
depth	**2**	**3**	**4**
PSNR	36.83	37.12	37.13

Table 3: Influence of ASwin-ACSconv Block Depth.

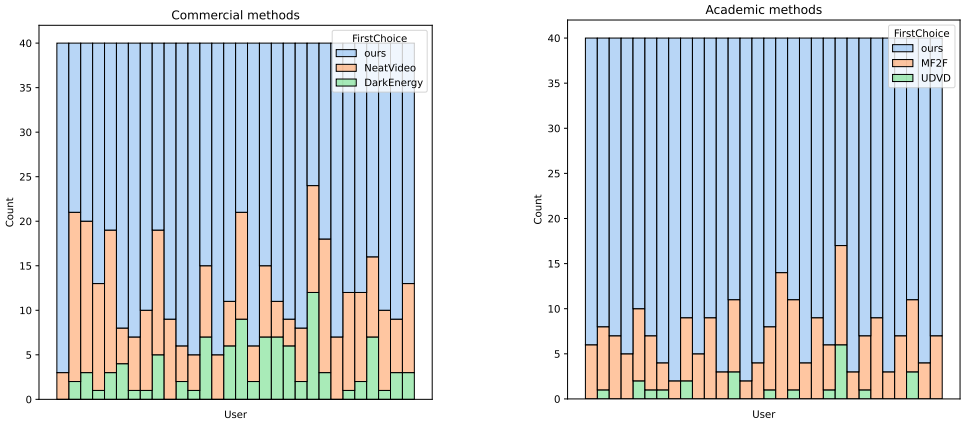
6. Test Set of Digitized Analog Videos

In this section, we provide an overview of the real noisy videos used as test set, consisting of 10 digitized analog movie scenes, see Figure 5 row 1 and Figure 6 row 1. The test sequences have a spatial resolution of 2K and consist of 20 to 30 frames. The first row shows crops of size 700x400 of the original noisy videos. The remaining rows show the denoising results of our method and all compared methods, respectively, please zoom for a better view. Additionally we provide 6 mp4 videos attached in the supplement (3 videos for each of the two sets of compared methods: academic and commercial), to show the significant superiority of our method in temporal consistency. For example, judging just on an individual frame, as shown in Figure 5 and Figure 6, the results of MF2F and sometimes even NeatVideo seem to be comparable to ours, however, when also considering



(a) Commercial Methods (b) Academic methods

Figure 2: Histogram of preferred methods of the participants for the two criteria: noise removal (blue) and temporal consistency (orange). It can be observed that our approach is favoured among all tested methods, both in terms of Denoising Level and Temporal Consistency.



(a) Commercial Methods (b) Academic methods

Figure 3: Overview of how the individual participants decided for both tasks combined and both sets of compared methods, i.e. commercial (a) and academic (b). Our method is always shown in blue.

the temporal consistency, differences become very apparent. We recommend to view the videos in loop mode, for better comparison.

7. Noisy/Clean Video Pairs with Realistic Noise

Figure 7 shows examples of the realistic noisy videos that were generated through the noise synthesis pipeline and used for training the general purpose blind denoising network. We show the results for a group of three frames to also enable an overview of the temporal component.

References

[1] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *CVPR*, pages 11036–11045. Computer Vision Foundation / IEEE, 2019.

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

[3] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022.

[4] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Koniq-10k: Towards an ecologically valid and large-scale iqa database. *CoRR*, abs/1803.08489, 2018.

User Study – Task Sheet

Background:
We investigate multiple denoising algorithms on historic digitized video footage. The noise in the video is characteristic for digitized analog film and is induced by various sources. Our goal is to reconstruct the underlying clean video and receive a visually pleasing results.

Your task:
You will be shown video sequences that have been denoised by different video denoising algorithms. Pick the video that is most visually appealing to you with respect to the current task. There are two tasks to focus on in total.

Tasks
1) Noise removal: Which denoised video removes the noise best while preserving details, i.e. shows the best visual quality with respect to the reference video after denoising?
2) Temporal consistency: Which denoised video is most consistent over time, i.e. shows the least flickering artifacts over time?

Important: Only consider the current task for your choice.

(a) The task sheet

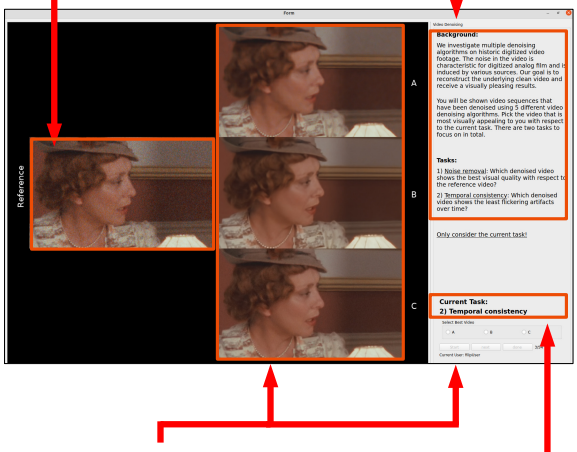
User Study – User Interface Explanation Sheet

Reference Frame:

On the left you are shown the noisy reference video.

Background & Tasks:

Here you can see the general description of the user study.



Denoised videos & choices:

You are shown three denoised videos, produced by different methods. Select your first and second choice depending on the task.

The current task:

Here you see which task you are currently performing.

(b) The user interface sheet

Figure 4: Written information given to the individual participant prior to starting a test session in the user study.

Figure 5: Visualization of the qualitative denoising performance on the first 5 sequences of the test set (please zoom).

Figure 6: Visualization of the qualitative denoising performance on the remaining 5 sequences of the test set (please zoom).

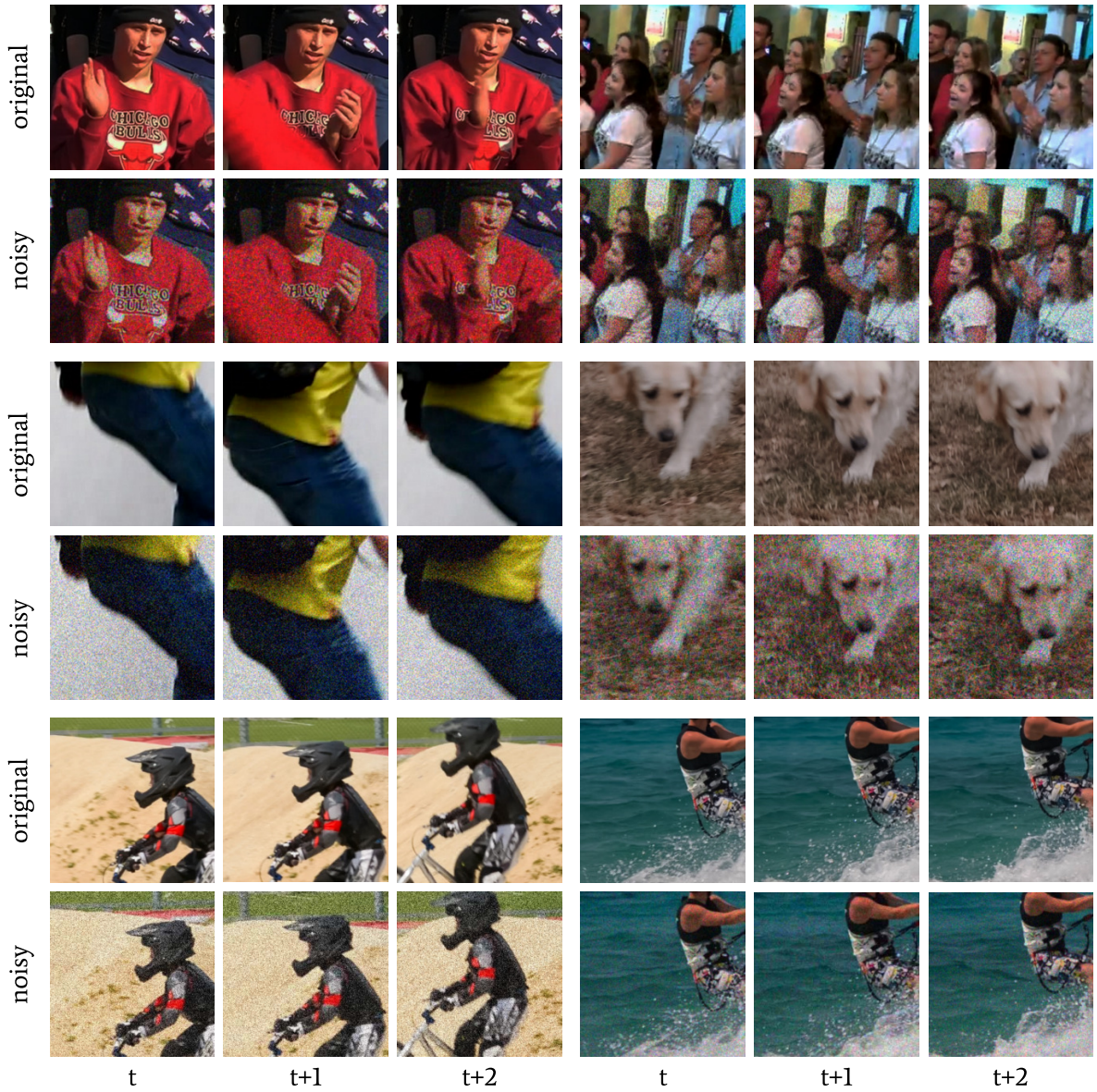


Figure 7: Examples of the realistic noisy videos that were generated through the noise synthesis pipeline.