Deblurring via Video Diffusion Models

Yan Wang[1,*] and Haoyang Long[1,*]

[1] Beijing University of Posts and Telecommunications, Beijing Haidian, China

Abstract. Video deblurring poses a significant challenge due to the intricate nature of blur, which often arises from a confluence of factors such as camera shakes, object motions, and variations in depth. While diffusion models and video diffusion models have respectively shone brightly in the fields of image and video generation, achieving remarkable results. Specifically, Diffusion Probabilistic Models (DPMs) have been successfully utilized for image deblurring, indicating the vast potential for research and development of video diffusion models in the realm of video deblurring. However, due to the significant data and training time requirements of diffusion models, the prospects of video diffusion models for video deblurring tasks remain uncertain. To investigate the feasibility of video diffusion models in video deblurring, this paper proposes a diffusion model specifically tailored for this task. Its model structure and some parameters are based on a pre-trained text-to-video diffusion model, and through a two-stage training process, it can accomplish video deblurring with a relatively small number of training parameters and data. Furthermore, this paper compares the performance of the proposed model with baseline models and achieves state-of-the-art results.

Keywords: Computer vision, Video deblurring, Diffusion model.

1 Introduction

Video deblurring poses a longstanding and intricate challenge, which entails reviving successive frames amidst spatially and temporally fluctuating blurring effects. This endeavor is exacerbated by the inherent complexities introduced by camera shakes, moving objects, and depth variations that occur within the exposure duration.

In recent years, generative tasks have attracted widespread attention in academia and industry. Among them, diffusion models have gained attention for their striking and powerful performance. On the other hand, diffusion models are also applied to image denoising and image deblurring. By utilizing the characteristics of diffusion models, researchers can effectively remove noise and blur from images, improving the quality and clarity of the images. This provides a new solution for the field of image processing, helping to improve the visual effect and interpretability of images. In recent times, research pertaining to video diffusion models has started to emerge. Primarily, these studies concentrate on the task of text-to-video generation. Their efforts have yielded

*Equally contribution.

remarkable results in video generation, highlighting the proficiency of generative models utilizing a diffusion model architecture. Notably, these models have achieved state-of-the-art (SOTA) performance across various tasks, including image generation, image deblurring, and video generation.

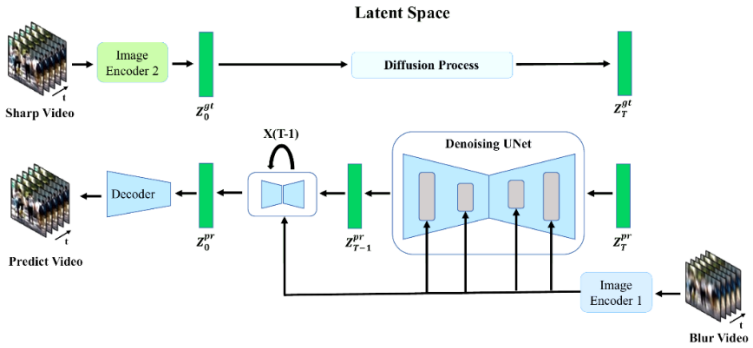

Fig. 1. Overview of our video deblurring model. The Unet, decoder model, and image encoder 1 all use pre-trained weights from the text-to-video model. The model and weight of image encoder 2 are the same as image encoder 1 in the beginning, but they are changed in the subsequent training process.

In this study, we adopt a unique perspective by framing the task of video deblurring as a conditional generative modeling problem, and innovatively attempt to utilize the video diffusion model as the fundamental basis for this endeavor. Addressing the challenges posed by the substantial data requirements and prolonged training durations inherent to video diffusion models, we implement a series of targeted optimization measures. Through these refinements, we successfully adapt the model originally intended for text-to-video tasks to the video deblurring task with remarkable efficiency and reduced cost. To the best of our knowledge, this is the first work that leverages diffusion models for video deblurring task.

2 Related Work

2.1 Video Deblurring

Video Deblurring techniques constitute a crucial area of research in computer vision, primarily emphasizing the effective utilization of information contained within multiple frames of a video sequence. Traditional methodologies in this domain typically approach the problem by aggregating temporal data, either directly or indirectly, to enhance the quality of deblurred frames.

Direct aggregation methods, as exemplified in studies such as, aim to combine information from multiple frames without explicitly modeling the blur process. These approaches often rely on techniques like frame averaging or weighted combinations to

mitigate the effects of blur. While these methods can achieve some degree of success, they may struggle to handle complex motion patterns or severe blurring artifacts.

On the other hand, indirect aggregation methods tackle the problem by formulating an inverse problem based on a blur model that incorporates multiple frames. As demonstrated in studies like [13,35], these approaches aim to recover the latent sharp frames by estimating the blur kernels and motion parameters jointly. This allows for a more accurate representation of the blur process and can lead to improved deblurring results.

To efficiently integrate temporal information in both direct and indirect aggregation methods, motion compensation techniques play a crucial role. Techniques such as tomography [18], local patch matching [4], and optical flow [7] have been widely employed to align frames and mitigate motion-induced blurring artifacts. Homography-based methods assume a global transformation between frames, while local patch matching and optical flow techniques provide more fine-grained motion estimation.

However, accurately estimating motion within blurred frames remains a significant challenge. The presence of blur can introduce ambiguities and make it difficult to determine precise motion patterns. To address this issue, several methods have been developed that alternate between estimating motion and deblurring frames. These iterative approaches, as exemplified in studies like [3,2], refine both the motion estimation and deblurring results iteratively, leading to improved overall performance.

2.2 Diffusion Probabilistic Models.

Diffusion Probabilistic Models (DPMs), a powerful class of generative models originally proposed in [24], have garnered significant attention in the field of large-scale image synthesis. These models have consistently demonstrated remarkable effectiveness, as evidenced by numerous studies. In fact, DPMs have emerged as viable alternatives to other dominant generative models, such as Generative Adversarial Networks (GANs) [9] and Variational Autoencoders (VAEs) [14]. What sets DPMs apart is their ability to achieve both high diversity and fidelity in the generated images.

Despite their impressive performance, the original DPMs face significant challenges when it comes to efficient image and video generation. Primarily, this is due to their reliance on iterative denoising processes and the need to operate in high-resolution pixel spaces. These factors contribute to the overall computational complexity and slow down the generation process.

To alleviate the first challenge of low sampling efficiency, researchers have dedicated considerable effort to developing improved sampling methods. One approach involves the use of learning-free sampling technique [26], which aim to optimize the sampling process without relying on additional learning algorithms. Alternatively, learning-based sampling strategies [23] leverage machine learning techniques to enhance the efficiency and quality of the generated samples.

On the other hand, addressing the second challenge of high-resolution pixel spaces requires a different approach. Methods such as LDM [22] have explored the use of manifolds with lower intrinsic dimensionality to reduce the computational burden associated with high-resolution images. By projecting the data onto a lower-dimensional

manifold, these methods aim to preserve the essential features while significantly reducing the computational complexity.

Our proposed model, takes inspiration from the foundational principles of ModelScopeT2V [31] but extends its capabilities specifically for video deblurring tasks.

3 Preliminaries of Video Diffusion Model

Diffusion models encompass two primary components: a forward diffusion process and a subsequent iterative denoising phase. In the forward diffusion process, clean data x_0 undergoes the gradual introduction of random noise within a Markovian chain framework:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_{t-1}}x_{t-1}, \beta_t I\right), t = 1, \dots, T \tag{1}$$

where $\beta_t \in (0,1)$ is a noise schedule and T is the total time step. When T is sufficiently large, x_T will be nearly a random Gaussian distribution $\mathcal{N}(0, I)$. The diffusion model is trying to denoise x_T and learn to iteratively estimate the reversed process:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta (x_t, t)\right) \tag{2}$$

A denoising model $\hat{x}_\theta$ parameterized with θ usually be used to approximate the original data x_0 by optimizing the following v-prediction:

$$\mathcal{L}_{\text{base}} = E_\theta\left[\|v - \hat{x}_\theta(x_t, t, c)\|_2^2\right] \tag{3}$$

where c is conditional information such as textual prompt, and v is the parameterized prediction objective. Follow the model structure in [31] we use 3D-UNet modified from its 2D version by inserting additional temporal blocks.

4 The Proposed Approach

4.1 Video Deblurring Diffusion Model

The architecture of our model is shown in Figure1. Our model incorporates components such as Unet, a decoder, and two image encoders. Notably, both Unet and the decoder leverage pre-trained weights from a text-to-video model. Image encoder 1 also utilizes these pre-trained weights, serving as a foundation for its initialization. Meanwhile, image encoder 2 starts with the same initialization as image encoder 1 but undergoes modifications during subsequent training iterations, allowing for further refinement and specialization. This approach ensures that our video deblurring model benefits from the rich representational power of pre-trained models while retaining the flexibility to adapt to specific task demands.

During the diffusion and inference process. The image encoder 1 encodes the prompt blur video into image embedding. Then the embedding is inputted into the denoising UNet to direct the denoising process. During training, a diffusion process is performed, transitioning from Z_0^{gt} to Z_T^{gt}; so the denoising UNet could be trained on these latent variables. Conversely, during inference, random noise Z_0^{pr} is sampled and utilized for the denoising procedure and then the decoder decode it from latent variables to deblur video.

Image encoder Similar to prior text-to-video endeavor [32], the model receives a video $I_{\text{video}} \in F \times H \times W \times C$, where F, H, W, and C represent the frame, height, width, and channel dimensions, respectively. This branch incorporates conditional signals to impart semantic direction to the content generation process. To guarantee that each condition can independently manipulate the generated content, we introduce random dropout of image embeddings with a designated probability during the training phase. These image embeddings are obtained from the central frame of the source video using CLIP's image encoder [21].

Denoising UNet From video diffusion model, we use 3D-UNet as the Deboising UNet. 3D-UNet [5] is an advanced deep learning architecture specifically designed for volumetric image segmentation tasks. It is an extension of the popular 2D U-Net, which revolutionized medical image segmentation by effectively learning contextual and hierarchical features from limited training data. The key innovation of 3D-UNet lies in its ability to process three-dimensional volumes directly, rather than treating them as a stack of two-dimensional slices. This enables the model to capture spatial relationships and structures in three dimensions, crucial for accurate segmentation of complex anatomies and pathologies. Recently, 3D-UNet has gained widespread application in video generation tasks due to its exceptional ability to capture spatial and temporal relationships within volumetric data, making it a versatile and effective tool for generating accurate and realistic video content.

Decoder Following previous work, we use decoder in VQGAN [6] as our decoder architecture. The decoder component of VQGAN is a critical element responsible for reconstructing high-quality images from quantized latent representations. It is designed as a deep convolutional neural network comprising multiple transposed convolutional layers that progressively increase the spatial resolution of the feature maps. These layers, coupled with non-linear activation functions, effectively capture the intricate details necessary for accurate image reconstruction. The decoder's architecture is optimized to ensure faithful reconstruction of the original image from the compressed latent space, thereby preserving visual fidelity and enhancing the overall quality of the generated images.

Training Loss Following previous work, the loss function for our model training is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{coherence}} \tag{4}$$

$$\mathcal{L}_{\text{coherence}} = E_{\theta}[\sum_{j=1}^{F-1} \| (v_{j+1} - v_j) - (o_{j+1} - o_j) \|_2^2] \tag{5}$$

where $\mathcal{L}_{\text{coherence}}$ is a temporal coherence loss that utilizes the frame difference as an additional supervisory signal, o_j and v_j are the predicted frame and corresponding ground truth. And λ is a balance coefficient that is set empirically to 0.1.

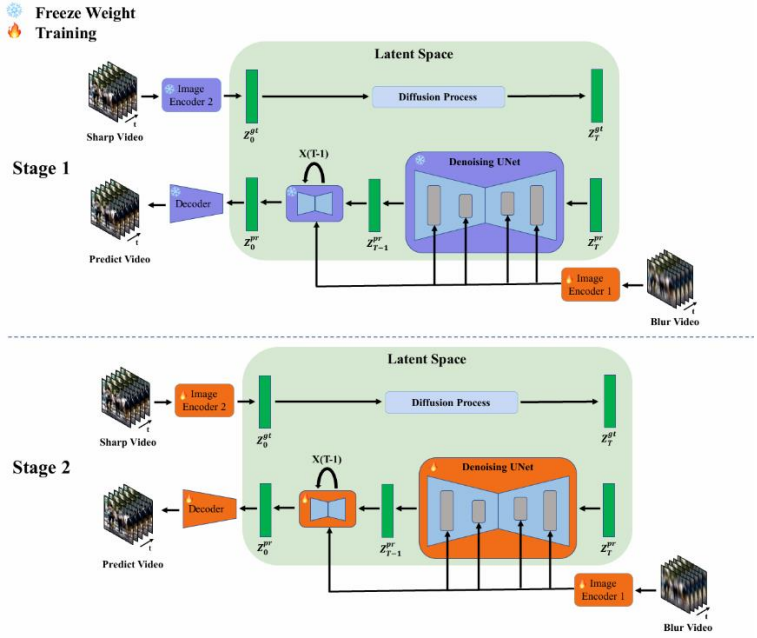

Fig. 2. Two stages training of the video deblurring diffusion model. In the first phase, only the encoder for blurred images is trained, while in the second phase, supervised fine-tuning is conducted, involving all parameters of the entire model in the training process.

4.2 Two Stages Training

Due to the significant demands of diffusion models on both data and the number of training iterations, video diffusion models exacerbate these computational resource requirements even further. To conserve resources while effectively leveraging relevant information within pre-trained models, we propose a two-stage training strategy, shown in Figure 2. This approach aims to enhance the performance of deblurring tasks by utilizing information from a pre-trained text-to-video model while minimizing computational costs.

In the first stage, only the encoder for blurred videos (i.e., Image Encoder 2) is trained. This is because all other model parameters have been previously learned through pre-training. To ensure that information from blurred images is effectively encoded and properly linked with the remaining model components, we prioritize training the encoder for blurred images during this initial phase.

After completing the first stage of training, the encoder for blurred images has learned corresponding representational information. Subsequently, we fine-tune the

entire model to ensure better integration among the newly added modules and to enhance the guidance for the deblurring task, ultimately achieving optimal performance.

5 Experiments

5.1 Datasets

GoPro [19] The GoPro dataset is a widely used synthetic dataset in the field of deep learning for deblurring, generated by blending high-speed camera-captured clear video frames. It provides a training set of 2,103 and a test set of 1,111 pairs of blurred and clear images. This dataset is significant for benchmarking deblurring methods, as it represents a synthetic yet realistic scenario for training and testing networks.

DVD [27] The DVD (Deep Video Deblurring for Hand-held Cameras) dataset plays a pivotal role in advancing the field of video deblurring. This dataset is a cornerstone in the study of motion blur reduction from videos captured by hand-held devices. The dataset is designed to support the development and evaluation of deep learning solutions for video deblurring, focusing on the significant challenge posed by motion blur due to camera shake. It contains 71 videos with 6,708 blurry-sharp image pairs, splitting into 61 training videos and 10 testing videos.

5.2 Model Instantiation and Training Details

Following previous work in [31,32], we utilize the DDPM [10] with T set to 1,000 steps. Additionally, for inference, we adopt the DDIM sampler [26] within the framework of classifier-free guidance [11], using 50 steps as the default configuration. Our model primarily consists of four modules: the blur image encoder(i.e. Image Encoder 1), the sharp image encoder(i.e. Image Encoder 2), the denoising UNet and the Deocder. The checkpoint of the four module are all obtained from TF-T2V [32], and the initial checkpoint of blur image encoder and sharp image encoder are the same. We train our model with the AdamW optimizer [16] with $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of 5×10^{-5}. For input videos, we select 16 frames per iteration from each video and crop a 256×256 region randomly.

5.3 Evaluation Metrics

In this study, three widely utilized evaluation metrics are employed to compare the synthesized images with the ground truth images. These metrics include the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), and the learned perceptual image patch similarity (LPIPS) [38]. Specifically, PSNR assesses the relative sharpness of the images, SSIM evaluates their structural similarity, and LPIPS gauges their perceptual quality.

5.4 Experimental Results

Table 1. Deblurring results on the GoPro dataset. "↑" indicates the higher the better and "↓" indicates the lower the better. Our model achieved the best results across all three metrics.

Methods	PSNR ↑	SSIM ↓	LPIPS ↓
Ground Truth	$+\infty$	1.0	0.0
STFAN [40]	28.59	0.86	0.20
SRN [28]	30.61	0.90	0.16
EDVR [33]	31.54	0.93	0.10
PVDNet [25]	31.52	0.92	0.12
Deblur-NeRF [17]	31.89	0.93	0.10
Ours	**32.22**	**0.95**	**0.07**

Results on GoPro Dataset Table 1 shows quantitative results on the GoPro dataset. We compared our model with the current state-of-the-art (SOTA) methods. And our model achieved the best results across all three metrics. In comparison to the runner-up model Deblur-NeRF, the proposed model improves PSNR, SSIM and LPIPS by 0.43dB, 0.02 and 0.03 respectively. To present the results of our model more intuitively, Figure 3 displays the input images, model outputs, and ground truth from the GoPro dataset. As shown in the figure, our model effectively performs deblurring on the inputs. Moreover, compared to Deblur-NeRF, our model demonstrates significant performance improvement in terms of image restoration authenticity, especially in scenarios with severe blurring.

Table 2. Deblurring results on the DVD dataset. "↑" indicates the higher the better and "↓" indicates the lower the better. Our model achieved the best results across all three metrics.

Methods	PSNR ↑	SSIM ↓	LPIPS ↓
Ground Truth	$+\infty$	1.0	0.0
STFAN [40]	31.24	0.92	0.11
SRN [28]	30.53	0.89	0.14
EDVR [33]	31.82	0.92	0.9
PVDNet [25]	32.31	**0.93**	0.9
Deblur-NeRF [17]	32.29	0.92	0.8
Ours	**32.59**	**0.93**	**0.06**

Results on DVD Dataset Table 2 shows quantitative results on the GoPro dataset. Upon comparing Table 1 and Table 2, it is evident that the same model generally performs better on the DVD dataset, potentially due to the simpler nature of image blurring

in this dataset. As Table 2 illustrates, our model achieves state-of-the-art results and significantly outperforms other models on the LPIPS evaluation metric. This suggests that the images generated by our model not only remove blur but also align more closely with human perception. We attribute this success to the fact that our model is a generative model based on a diffusion model, which has achieved notable success in image generation and is widely recognized for its ability to produce more realistic images.

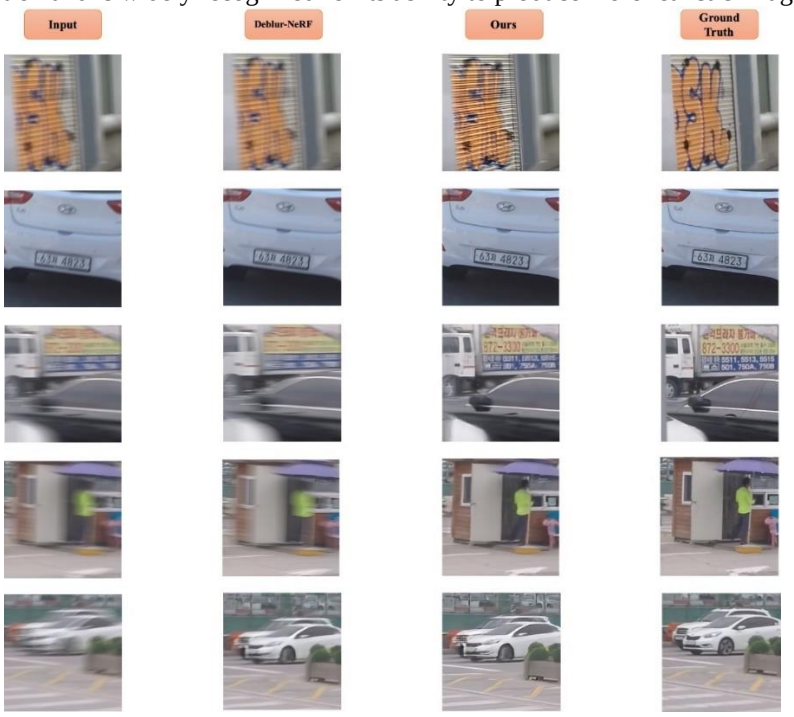

Fig. 3. Sample deblurred results from GoPro dataset.

Fig. 4. Visual comparison with error maps of our model. Regions with red box indicate emphasized regions of error map.

5.5 Error Map Analysis

Figure 4 illustrates the error map between the deblurred output of our model and the ground truth. As evident from the figure, the majority of the model's output closely resembles the ground truth. The errors predominantly concentrate around the edges of objects, particularly those spanning across the entirety of the scene. This suggests that while our model excels at generating the internal details of objects within the scene, it struggles with precisely localizing them. We hope to address this limitation in the future work.

6 Conclusions

In this paper, we extend the application of video diffusion models to the task of video deblurring. To the best of our knowledge, this is the first work to propose the use of video diffusion models for video deblurring tasks. While implementing the video deblurring diffusion model architecture, we address the challenges associated with the data-intensive and slow convergence nature of diffusion model training. Specifically, we introduce a two-stage training approach that leverages previously pre-trained checkpoints, significantly accelerating convergence while conserving computational resources. Experimental results demonstrate that our proposed model achieves state-of-the-art (SOTA) performance on both the GoPro and DVD datasets. Moreover, our model exhibits notable improvements in human perception-related evaluation metrics compared to previous models.

References

1. Anger, J., Delbracio, M., Facciolo, G.: Efficient blind deblurring under high noise levels. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA). pp. 123–128. IEEE (2019)
2. Bar, L., Berkels, B., Rumpf, M., Sapiro, G.: A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
3. Cho, S., Matsushita, Y., Lee, S.: Removing non-uniform motion blur from images. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
4. Cho, S., Wang, J., Lee, S.: Video deblurring for hand-held cameras using patch-based synthesis. ACM Transactions on Graphics (TOG) 31(4), 1–9 (2012)
5. C¸ i¸cek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
6. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: Vqgan-clip: Open domain image generation and editing with natural language guidance (2022)

7. Delbracio, M., Sapiro, G.: Hand-held video deblurring via efficient fourier aggregation. IEEE Transactions on Computational Imaging 1(4), 270–283 (2015)
8. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: Acm Siggraph 2006 Papers, pp. 787–794 (2006)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM 63(11), 139–144 (2020)
10. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Jin, M., Roth, S., Favaro, P.: Normalized blind deconvolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 668–684 (2018)
13. Kim, T.H., Lee, K.M.: Generalized video deblurring for dynamic scenes (2015)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8183–8192 (2018)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblur-nerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)
18. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. IEEE Transactions on pattern analysis and Machine Intelligence 28(7), 1150–1163 (2006)
19. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
20. Pan, J., Hu, Z., Su, Z., Yang, M.H.: Deblurring text images via l0-regularized intensity and gradient prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2908 (2014)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
24. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)

25. Son, H., Lee, J., Lee, J., Cho, S., Lee, S.: Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. ACM Transactions on Graphics (TOG) 40(5), 1–18 (2021)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
27. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1279–1288 (2017)
28. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8174–8182 (2018)
29. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: Stripformer: Strip transformer for fast image deblurring. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
30. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxim: Multi-axis mlp for image processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5769–5780 (2022)
31. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
32. Wang, X., Zhang, S., Yuan, H., Qing, Z., Gong, B., Zhang, Y., Shen, Y., Gao, C., Sang, N.: A recipe for scaling up text-to-video generation with text-free videos. arXiv preprint arXiv:2312.15770 (2023)
33. Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C.: EDVR: video restoration with enhanced deformable convolutional networks. CoRR abs/1905.02716 (2019), http://arxiv.org/abs/1905.02716
34. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018)
35. Wulff, J., Black, M.J.: Modeling blurred video with layers. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 236–252. Springer (2014)
36. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)
37. Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R.W., Yang, M.H.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2521–2529 (2018)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
39. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3770 (2019)
40. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2482–2491 (2019)