2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222579

IMPROVING VIDEO COLORIZATION BY TEST-TIME TUNING

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

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

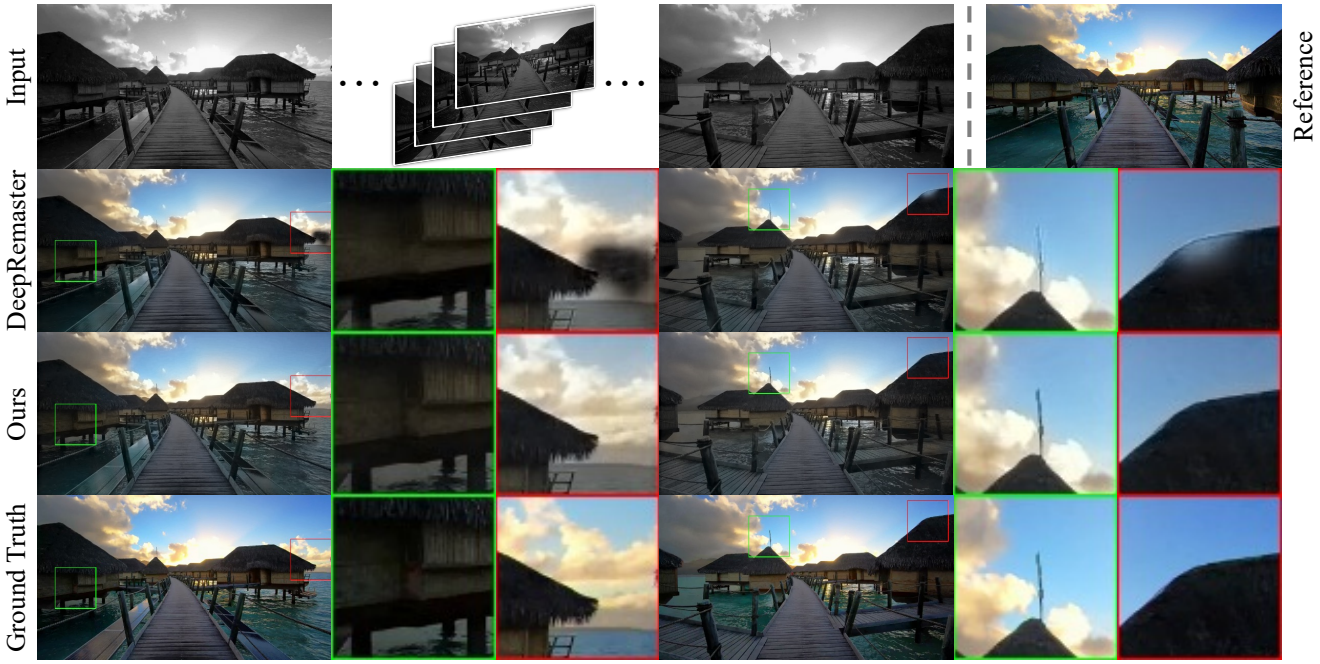

Fig. 1: Given a monochrome video sequence as input and the colorized first frame as reference, we propose a simple yet effective method using test-time tuning on pretrained model DeepRemaster [1] to perform video colorization with superior performance. Green and red boxes are zooming into details.

ABSTRACT

With the advancements in deep learning, video colorization by propagating color information from a colorized reference frame to a monochrome video sequence has been well explored. However, the existing approaches often suffer from overfitting the training dataset and sequentially lead to suboptimal performance on colorizing testing samples. To address this issue, we propose an effective method, which aims to enhance video colorization through test-time tuning. By exploiting the reference to construct additional training samples during testing, our approach achieves a performance boost of $1 \sim 3$ dB in PSNR on average compared to the baseline. Code is available at: https://github.com/IndigoPurple/T3.

Index Terms— video colorization, video restoration, image processing

1. INTRODUCTION

Nowadays, it has become common practice to leverage an ideally colorized video frame as a reference to colorize an entire monochrome video sequence. Suah a video colorization task has a wide range of applications, such as reviving vintage films. State-of-the-art methods for video colorization rely on deep learning and benefit from large-scale datasets. However, these existing approaches inevitably suffer from overfitting the training dataset, which in turn leads to suboptimal performance on colorizing testing samples in a sequential manner.

To address this issue, our paper introduces a simple yet effective method that enhances video colorization through test-time tuning. Our key insight is that the reference can be considered as ground truth and can help construct an additional

*Equal contribution. †Corresponding author.

This work is supported in part by the Research Grants Council (GRF 17201620), by the Research Postgraduate Student Innovation Award (The University of Hong Kong), and by ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR.

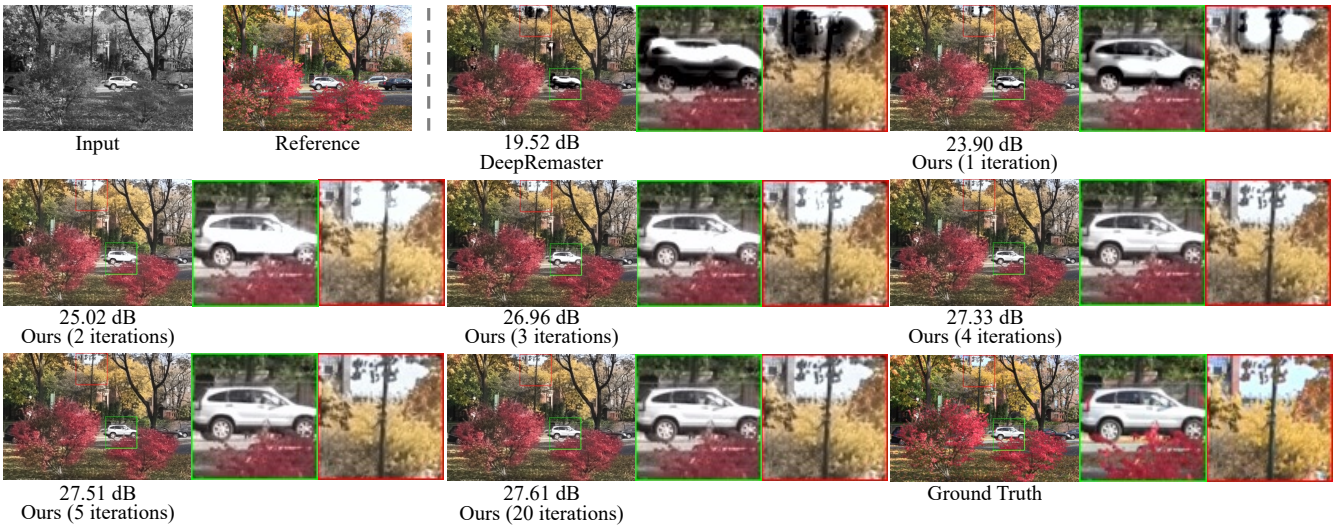

Fig. 2: Video colorization comparisons on the video sequence `foliage` of the Vid4 dataset [2]. While DeepRemaster [1] severely results in artifacts and distortion, our method boosts the performance with merely a few iterations for test-time tuning. Quantitative evaluations are provided in terms of the PSNR metric. Green and red boxes are zooming into details.

training sample during testing. Specifically, we demonstrate that using the "ref-mono" pair, which comprises the reference with color information and its corresponding monochrome video frame, is adequate to guide the pre-trained model tuning towards higher performance.

To demonstrate this, we utilize the pre-trained model from DeepRemaster [1] and fine-tune the network parameters with objective functions that only consider the ref-mono pair. By doing so, we expose testing samples that are previously unseen in the training dataset to the neural network for adaptive tuning. Our experiments on various datasets showcase the effectiveness and efficiency of our proposed method.

Our main contributions are as follows:

- A **simple yet effective** method that improves video colorization through test-time tuning, resulting in a performance boost of $1 \sim 3$ dB to the baseline.
- A **low-cost** fine-tuning paradigm that utilizes the reference to finetune pre-trained models, without requiring additional network parameters or annotated labels.
- A **highly efficient** iterative optimization that updates the network parameters to achieve high performance within a few iterations, making it feasible for real-time applications.

2. RELATED WORK

In video colorization, one or more video frames with color information are typically provided as reference to provide cues and ensure accuracy. While image colorization methods [3, 4, 5] can be directly applied to video colorization, Zhang *et al.* [6] introduced a temporal consistency loss [7] to improve video colorization with a recurrent framework. Later, video colorization is further developed by incorporating GAN encoders [8] and visual tracking [9]. DeepRemaster [1] established source-reference correspondence by finding similarities within the reference image and target image, which has been frequently used in recent years [10, 11, 12, 13, 14, 15, 16].

3. METHOD

3.1. Overall Framework

Given a monochrome video sequence as input and the colorized first frame as reference, we denote the input as $\mathbf{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{T}$ and the reference as $\boldsymbol{z}$, where t indicates the time step of the video frame, and T is the total number of video frames. Moreover, we assume $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 1}$, $\boldsymbol{z} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width, 1 and 3 are the channel numbers for grayscale (input) or RGB (reference).

To perform video colorization, deep networks are applied by implicitly learning a function $\hat{\mathbf{Y}} = f_\theta(\mathbf{X}; \boldsymbol{z})$ that maps a monochrome video sequence $\mathbf{X}$ to a colorized one $\hat{\mathbf{Y}} = \{\hat{\boldsymbol{y}}^{(t)}\}_{t=1}^{T}$, where θ is the network parameters we aim to optimize, and $\hat{\boldsymbol{y}}^{(t)} \in \mathbb{R}^{H \times W \times 3}$ is the synthesized video frame at time step t. However, state-of-the-art neural networks for video colorization are all invariably trained on datasets. While they learn priors from the dataset to achieve excellent performance, they inevitably suffer in overfitting the training samples, which usually have a gap with the testing samples. Sequentially and unfortunately, such a gap leads to suboptimal performance.

Method	tractor	motorbike	sunflower	park_joy	snowboard	touchdown	rafting	hypersmooth	Average
Zhang *et al.* [6]	18.22, 0.8518	23.05, 0.9043	15.50, 0.7376	21.36, 0.8707	19.30, 0.9348	25.44, 0.9144	22.54, 0.9105	23.73, 0.9141	21.14, 0.8798
DeepRemaster [1]	24.09, 0.8791	24.51, 0.9024	24.57, 0.8268	25.22, 0.8627	24.07, 0.9349	27.25, 0.9270	23.41, 0.9133	26.05, 0.9074	24.90, 0.8942
Ours	**25.11, 0.8867**	**25.36, 0.9077**	**24.82, 0.8290**	**26.02, 0.8878**	**24.79, 0.9440**	**30.19, 0.9329**	**24.47, 0.9151**	**26.52, 0.9190**	**25.91, 0.9028**

Table 1: Video colorization comparisons on the Set8 dataset [17] in terms of PSNR and SSIM.

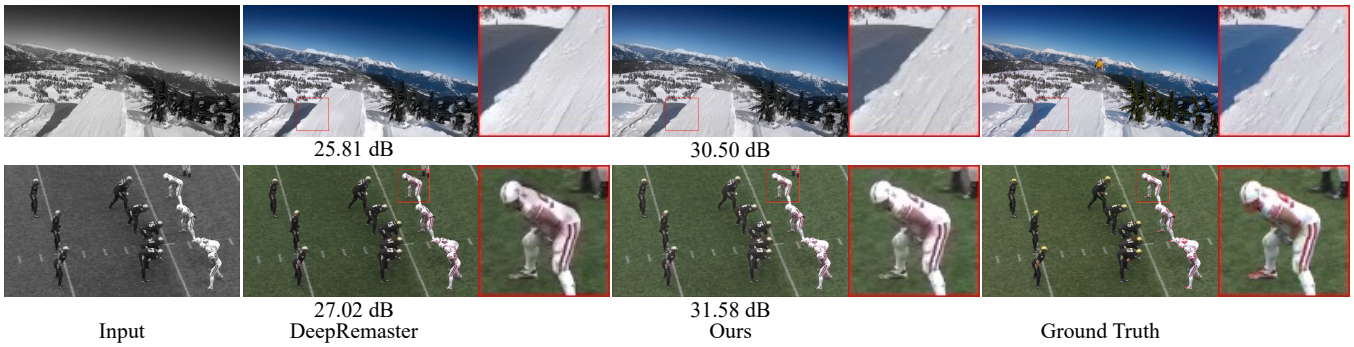

Fig. 3: Video colorization comparisons on the video sequence `snowboard` (the upper row) and `touchdown` (the bottom row) of the Set8 dataset [17]. While DeepRemaster [17] suffers in detail loss and artifacts, our method reconstructs explicit content. Quantitative evaluations are provided in terms of the PSNR metric. Red boxes are zooming into details.

To address this issue, we propose a test-time tuning approach that leverages an ideally colorized reference frame $\boldsymbol{z}$. The pre-trained network is fine-tuned using an combined objective function, which improve color transfer from the reference to the monochrome.

3.2. Objective Function

Our aim is to alleviate the overfitting problem by test-time tuning the pre-trained network parameters. To achieve this, we make full use of the reference, which can be considered ideally colorized, *i.e.*, ground truth. Our key insight is that the reference can provide a label to construct an additional training sample during testing time. Specifically, we adopt the neural network from DeepRemaster [1] and employ an objective function $\mathcal{L}$ to perform finetuning on the pre-trained model. Considering the ref-mono pair $\hat{\boldsymbol{y}}^{(1)}$ and $\boldsymbol{z}$, though we can directly apply a loss function like $\mathcal{L}_{rgb} = ||\hat{\boldsymbol{y}}^{(1)} - \boldsymbol{z}||_2^2$, to finetune the model parameters, we use a well-designed set of combined objective functions using LAB color space:

$$\mathcal{L} = \mathcal{L}_l + \mathcal{L}_{ab}, \tag{1}$$

where the overall objective $\mathcal{L}$ for finetuning regularization is consist of two loss functions $\mathcal{L}_l$ and $\mathcal{L}_{ab}$. The former prevents the neural network to generate artifacts and noise. The latter is responsible for better transferring the colors from the reference to the monochrome. Specifically, $\mathcal{L}_l$ and $\mathcal{L}_{ab}$ are calculated by:

$$\mathcal{L}_l = ||\hat{\boldsymbol{y}}_l^{(1)} - \boldsymbol{z}_l||_2^2, \tag{2}$$

$$\mathcal{L}_{ab} = ||\hat{\boldsymbol{y}}_{ab}^{(1)} - \boldsymbol{z}_{ab}||_2^2, \tag{3}$$

where $\hat{\boldsymbol{y}}_l^{(1)}$ and $\hat{\boldsymbol{y}}_{ab}^{(1)}$ are the luminance and chrominance component, respectively, of the first colorized video frame $\hat{\boldsymbol{y}}^{(1)}$ in the LAB color space. Similarly, the $\boldsymbol{z}_l$ and $\boldsymbol{z}_{ab}$ are obtained by splitting the reference $\boldsymbol{z}$ to the luminance and chrominance.

By simply using our proposed objective functions, our approach effectively improve the performance of the existing method DeepRemaster [1] in a test-time tuning manner with only a few iterations. As shown in Figure 2, while DeepRemaster [1] often results in artifacts and distortion, our method improves the PSNR performance by more than $4\,\mathrm{dB}$ after only one iteration of test-time tuning and by $8\,\mathrm{dB}$ after only 20 iterations.

4. EXPERIMENT

Experiment Setting. We evaluated our method on two commonly used video datasets, Vid4 [2] and Set8 [17], using both quantitative and qualitative metrics. For quantitative evaluation, we used two commonly used metrics, PSNR and SSIM [18]. We compared our method to two video colorization methods, Zhang *et al.*[6] and DeepRemaster[1]. In our experiments, the pre-trained model from DeepRemaster [1] is finetuned for 20 iterations using the Adam [19] optimizer, with a learning rate of 1×10^{-4}. We experimented with other iteration steps and found that increasing the number of iterations beyond 20 did not significantly improve PSNR, with a saturation effect observed. For example, increasing the number of iterations to 50 only increased PSNR by approximately 0.07 compared to 20 iterations.

Method	Calendar	City	Foliage	Walk	Average
Zhang *et al.* [6]	18.98, 0.9126	32.18, 0.9676	19.13, 0.9181	26.17, 0.9507	24.12, 0.9372
DeepRemaster [1]	23.39, 0.9127	28.49, 0.9662	19.52, 0.9220	27.16, 0.9504	24.64, 0.9378
Ours	**23.81, 0.9139**	**33.06, 0.9725**	**26.47, 0.9441**	**28.10, 0.9544**	**27.86, 0.9462**

Table 2: Video colorization comparisons on the Vid4 dataset [2] in terms of PSNR and SSIM.

Fig. 4: Video colorization comparisons on the video sequence `calendar` of the Vid4 dataset [2]. Compared to DeepRemaster [17], the colors of our results are more consistent with the reference. Quantitative evaluations are provided in terms of the PSNR metric. Red boxes are zooming into details.

Resolution	320×576	320×400	256×256
Time (5 iters)	1.50	1.13	0.53
Time (20 iters)	5.47	4.10	1.94

Table 3: Average tuning time per video sequence in minutes using identical environments, regarding different image resolutions and optimization iterations. The image resolution is denoted as $H \times W$, where H and W are height and width, respectively.

Dataset	Vid4 [2]	Set8 [17]
Baseline *w/* $\mathcal{L}_{rgb}$	27.65, 0.9421	25.63, 0.8974
Ours *w/o* $\mathcal{L}_l$	25.64, 0.9397	25.01, 0.8962
Ours *w/o* $\mathcal{L}_{ab}$	27.35, 0.9408	25.40, 0.8972
Ours	**27.86, 0.9462**	**25.91, 0.9028**

Table 4: Ablation study regarding different objective functions for test-time tuning. The performance is evaluated in terms of average PSNR and SSIM on each dataset.

Quantitative and Qualitative Results. As Table 1 and 2 show, our model outperforms previous methods Zhang *et al.* [6] and DeepRemaster [1] on all the video sequences of different datasets. To further illustrate the effectiveness of our method, we compare the qualitative performance with the competitive baseline DeepRemaster [1]. As Figure 3 shows, while DeepRemaster [17] suffers from artifacts and loss of details, our method reconstructs explicit contents. In Figure 4, compared to DeepRemaster [17], the colors of our results are more consistent with the reference.

Tuning Time. As Table 3 shows, our method requires a short time for tuning the model. Moreover, we observe that the PSNR could achieve a relatively higher level after merely 5 iterations, which is also intuitively verified in Figure 2. Therefore, to reduce tuning time, the iteration number could be set to 5.

Ablation Study. We investigate the role of the loss functions formulated in Equation 2 and 3 by two variants of our pipeline denoted as Ours *w/o* $\mathcal{L}_l$ and Ours *w/o* $\mathcal{L}_{ab}$, which respectively turn off one of the components while remaining another. In addition, we also explore the experimental results simply using $\mathcal{L}_{rgb} = ||\hat{\boldsymbol{y}}^{(1)} - \boldsymbol{z}||_2^2$, denoted as Baseline *w/* $\mathcal{L}_{rgb}$. According to Table 4, disabling either component in LAB color space or using the $\mathcal{L}_{rgb}$ in RGB color space leads to a performance drop, suggesting the soundness of our proposed method.

5. CONCLUSION

In this paper, we propose an efficient approach for enhancing video colorization through test-time tuning, leveraging the reference as a ground truth to generate an additional training sample during testing. By using the reference with color information and its corresponding monochrome frame, we show that the pre-trained model achieves higher performance after test-time tuning. We employ a pre-trained model from recent work and finetune the network parameters using a linear combination of objective functions with LAB color space. This enables adaptive tuning to previously unseen testing samples. Experimental results on various datasets validate the effectiveness and efficiency of our approach. For future work, another direction worthy of exploration is adding some noise to the network parameters before finetuning to alleviate the overfitting problem and further improve testing performance [20, 21].

6. REFERENCES

[1] Satoshi Iizuka and Edgar Simo-Serra, "DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement," *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA)*, vol. 38, no. 6, pp. 1, 2019.

[2] Ce Liu and Deqing Sun, "On bayesian adaptive video super resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 346–360, 2013.

[3] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–16, 2018.

[4] Chufeng Xiao, Chu Han, Zhuming Zhang, Jing Qin, Tien-Tsin Wong, Guoqiang Han, and Shengfeng He, "Example-based colourization via dense encoding pyramids," in *Computer Graphics Forum*. Wiley Online Library, 2020, vol. 39, pp. 20–33.

[5] Yaping Zhao, Haitian Zheng, Mengqi Ji, and Ruqi Huang, "Cross-camera deep colorization," in *Artificial Intelligence: Second CAAI International Conference, Part I*. Springer, 2022, pp. 3–17.

[6] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.

[7] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua, "Coherent online video style transfer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1105–1114.

[8] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan, "Towards vivid and diverse image colorization with generative color prior," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14377–14386.

[9] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy, "Tracking emerges by colorizing videos," in *Proceedings of the European conference on computer vision*, 2018, pp. 391–408.

[10] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5801–5810.

[11] Guanghan Li, Yaping Zhao, Mengqi Ji, Xiaoyun Yuan, and Lu Fang, "Zoom in to the details of human-centric videos," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3089–3093.

[12] Yaping Zhao, Guanghan Li, Zhongrui Wang, and Edmund Y Lam, "Cross-camera human motion transfer by time series analysis," *preprint arXiv:2109.14174*, 2021.

[13] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu, "Deep animation video interpolation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6587–6595.

[14] Yaping Zhao, Mengqi Ji, Ruqi Huang, Bin Wang, and Shengjin Wang, "EFENet: Reference-based video super-resolution with enhanced flow estimation," in *Artificial Intelligence: First CAAI International Conference, Part I 1*. Springer, 2021, pp. 371–383.

[15] Min Shi, Jia-Qi Zhang, Shu-Yu Chen, Lin Gao, Yukun Lai, and Fang-Lue Zhang, "Reference-based deep line art video colorization," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[16] Yaping Zhao, Haitian Zheng, Zhongrui Wang, Jiebo Luo, and Edmund Y Lam, "MANet: improving video denoising with a multi-alignment network," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2036–2040.

[17] Matias Tassano, Julie Delon, and Thomas Veit, "Dvdnet: A fast network for deep video denoising," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1805–1809.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[20] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie, "Noisytune: A little noise can help you finetune pretrained language models better," *arXiv preprint arXiv:2202.12024*, 2022.

[21] Yaping Zhao, Zhongrui Wang, and Edmund Y Lam, "Improving source localization by perturbing graph diffusion," in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2022, pp. 1–9.