Histogram-guided Video Colorization Structure with Spatial-Temporal Connection

Zheyuan Liu

College of Computer Science and Technology Zhejiang University of Technology Hangzhou, China zheyuanliu@zjut.edu.cn

Hanning Xu College of Computer Science and Technology Zhejiang University of Technology Hangzhou, China 202006010425@zjut.edu.cn Pan Mu*

College of Computer Science and Technology Zhejiang University of Technology Hangzhou, China panmu@zjut.edu.cn

Cong Bai College of Computer Science and Technology Zhejiang University of Technology Hangzhou, China congbai@zjut.edu.cn

Abstract-Video colorization, aiming at obtaining colorful and plausible results from grayish frames, has aroused a lot of interest recently. Nevertheless, how to maintain temporal consistency while keeping the quality of colorized results remains challenging. To tackle the above problems, we present a Histogram-guided Video Colorization with Spatial-Temporal connection structure (named ST-HVC). To fully exploit the chroma and motion information, the joint flow and histogram module is tailored to integrate the histogram and flow features. To manage the blurred and artifact, we design a combination scheme attending to temporal detail and flow feature combination. We further recombine the histogram, flow and sharpness features via a Ushape network. Extensive comparisons are conducted with several state-of-the-art image and video-based methods, demonstrating that the developed method achieves excellent performance both quantitatively and qualitatively in two video datasets.

Index Terms—Video colorization, deep learning, histogram and flow-guided

I. INTRODUCTION

Video colorization task aims to convert the gray frame sequences into colorful and plausible ones, which has wide applications in domains like old films restoration and anime creation. Video colorization can also assist other tasks like video action recognition, detection, tracking and segmentation.

Although significant progress has been achieved, automatic video colorization still poses the following three challenges: the chroma quality of produced images, the temporal consistency between frames and the possibility of producing distinct colorization result. Image-based colorization solutions [1]–[7] can achieve satisfactory visual result. Nonetheless, those methods tend to take longer time to colorize images and the temporal coherence of generated results is relatively poor. Video-based methods [8]–[14] are aimed to produce colorized frames with vivid color and minimum temporal flickering. But

to achieve a trade-off between quality of single frame and the temporal consistency of neighbor frames is hard.

The final challenge makes video colorization an ill-posed problem, since the same object can have distinct but possible colors at the same time (i.e., the color of T-shirt can be green or blue, a flower can be purple or white). To solve this problem, researchers either build models to produce multiple results all at once, or tend to seek references to guide the model to colorize the video frames. User-guided colorization like [6] seek the indication from users to guide the model. Exemplarbased methods like [11], [15], [16] utilize exemplar images as color references. Thus the generated results differs with different exemplars. Histograms can also be served as color references to provide an global color distribution for models. For instance, HistoGAN [17] employs histogram to manipulate the color of GAN-generated images, [4] uses histograms to facilitate the extraction of semantic information.

Over the past several years, the attention mechanism along with the transformer architecture has been broadly used in vision based tasks like detection [18], segmentation [19] and image restoration [20], [21]. In the field of colorization, ColTran [3] proposes a transformer model based on the theory of probability and samples colors from the learned distribution to generate diverse and plausible results. CT^2 [2] converts the colorization task as a classification problem and introduces color tokens into their model while limiting the range of *Lab* color space.

In this work, we develop a histogram-guided video colorization with spatial-temporal connection structure, along with a developed joint flow and histogram feature module, spatialtemporal connection scheme and a feature recombinationaimed U-shape network. In particular, the developed method firstly calculates the temporal sharpness and flow feature, to help the model better restore the blurred detail caused by motion and the artifact brought by optical flow. Aiming to integrate the flow and histogram information to facilitate

This work is supported by Natural Science Foundation of China under Grant No. 62202429, U20A20196 and Zhejiang Provincial Natural Science Foundation of China under Grant No. LY23F020024, LR21F020002.

^{*} The corresponding author.



Fig. 1. Model overview of the proposed ST-HVC. Temporal sharpness and flow features are computed and forwarded into U-shape network with the frames. Histogram and flow features are integrated into joint flow and histogram feature module (short for JFHM) to facilitate the colorization process.

the colorization, we present joint flow and histogram feature module and employ it in skip connections and in the bottleneck of the U-shape network. We extract the reference histogram feature in multiple level via splatting the pixel histogram along the range dimension. We feed merely the histogram of the middle frame in a video into our model, which can further boost the application of the proposed method. We make the following contributions:

- We propose a novel histogram-guided video colorization network with spatial-temporal connection structure (i.e., ST-HVC) to tackle the color assignment and temporal consistency challenges.
- For the first time we calculate the temporal sharpness and flow feature to form the inputs along with the frames. This design can better manage the blurred detail caused by motion and the artifact brought by flow.
- To ameliorate the wrongly-assigned color and improve the temporal coefficient, we integrate the histogram of the middle frame and flow features into a joint module (i.e., JFHM) to guide the colorization process.
- The experiments confirm that the proposed method significantly outshines existing video colorization approaches both quantitatively and qualitatively.

II. METHOD

A. Method Overview

Video colorization task aims at attaining colorful and plausible results from black and white frames while reducing the flickering artifacts and maintain the quality of colorized images. In Fig. 1, we illustrate the overall pipeline of our developed network ST-HVC. The proposed method takes in $2\tau + 1$ video frames as input and then computes the temporal sharpness F_s [22], [23] and flow feature F_{flow} , the two features are then concatenated with original frames $\{x_{t-\tau}, ..., x_{t+\tau}\}$ and are forwarded into a dynamic region convolution [24] based encoder-decoder architecture. To remove the artifacts commonly occurred in video restoration results and boost the consistency between frames, four JFHMs are placed in skip connections and in bottleneck. We decouple the



Fig. 2. Structure of JFHM with flow and histogram feature integrated. ϕ_x stands for the input feature. M_1 means the middle feature map and M_2 denotes the output of JFHM module.

original multi-head self-attention in JFHM along spatial and temporal dimension to better utilize the spatially histogram color hint and temporally flow feature. We extract the histogram reference feature H_f in multiple levels by introducing bilateral grid [25] and via splatting the pixel histogram along the range dimension. It is noteworthy that the reference we feed into the model is merely the middle frame of the video. We don't need any other references to guide the colorization process.

B. Temporal Detail and Flow Combining

Temporal sharpness and flow features are firstly computed according to the frames before they are forwarded into Ushape network. This prior is based on the observation that the same object in different frames has sharp and blurred pixels simultaneously. Since many colorization works cannot perform well in blurred details, the temporal sharpness can prompt our model to sense the sharpness and help to restore those details.

The temporal sharpness F_s calculated from frames $\{x_{t-\tau}, ..., x_{t+\tau}\}$ indicates the sharpness area of frames. The overall input F_o can be formed below:

$$F_o = Concat(x_i, (F_s \otimes x_i), F_{flow}), i \in [t - \tau, t - \tau] \quad (1)$$

where Concat and \otimes represents the concatenation and multiplication. We use RAFT [26] to form the optical flow F_{flow} .



Fig. 3. Chroma comparison with representing and state-of-the-art approaches. From top to bottom, the figure demonstrates outstanding performance of the proposed method regarding the front color, background color, blurred pixels and the detail.

C. Joint Flow and Histogram Feature Module

The designed joint flow and histogram feature module (JFHM) is tailored to utilize the chroma and motion feature condensed in histogram and flow as depicted in Fig. 2, frames features are first put into the module. Flow and histogram information is integrated in SA and feature refinement.

Inspired by [27]–[30], we decouple the attention schema along the spatial and temporal dimension, thus forming the temporal attention (TA) and spatial attention (SA). Specifically, given the input feature $F^i, i \in [t - \tau, t + \tau]$, we split it into s^2 non-overlapped windows along the height and width dimension where each window of F^i is denoted as F_{jk}^i , $j, k \in [1, s]$. We take different window size for each attention and $s_{temporal}$ is smaller than $s_{spatial}$.

The temporal approach group windows along temporal dimension, i.e., $G_{temporal} = \{F_{j,k}^{t-\tau}, ..., F_{j,k}^{t+\tau}\}, j, k \in [1, s_{temporal}]$ so that temporal attention is performed across tokens in $G_{temporal}$. Thus, continuous movement of the object inside a small spatial window can be detected.

In spatial branch, we gather them along spatial dimension $G_{spatial} = \{F_{1,1}^i, F_{1,2}^i, ..., F_{s_{spatial}}^i\}, i \in [t - \tau, t + \tau],$ and attention is performed across tokens in it. The operation helps our model to understand similar textures in spatially-neighboring pixels.

Giving a sequence of frame features ϕ_x , the output after concatenating the TA and SA features M_1 can be described:

$$M_1 = \phi_x \oplus TA(LN(\phi_x) \oplus SA(LN(\phi_x), F_{flow})$$
 (2)

where LN means the Layer Normalization and F_{flow} stands for flow features of the frames. The sign \oplus denotes the concatenate operation. After concatenation, we integrate the histogram feature and we introduce the skip connection after the feature refinement and feed forward layer.

The histogram reference serves as a multi-scale color distribution indicator. Following [25], [31] and by splicing the pixel histogram along the range dimension, we can obtain different histogram patch and attain distinct feature map at multiple scale. The HistConv in Fig. 1 denotes the 2×2 convolution

without bias and with its weight shrunk to downscale the histmap. The output of joint feature module M_2 is:

$$M_2 = M_1 \oplus FFN(LN(M_1 \oplus M_{FR})) \tag{3}$$

where FFN stands for feed forward network. M_{FR} means the middle output after integrating histogram and feature refinement. The structure of feature refinement module is composed of three times duplicated sequence of 3×3 convolution and LeakyReLU, where we inject color information of reference histogram into grayscale frame features.

D. Feature Recombination via U-shape Network

The concatenation results from Sec.II-B are put into the designed U-shape network which consists of basic encoder-decoder, JFHM module and a histogram exploitation module.

Motivated by [4], [17], [31], we design to exploit the reference histogram from the middle frame of video. The JFHM aims to integrate the flow and histogram information, and help the reconstruction process. As is shown in Fig. 1, we apply JFHM in skip connection and replace the bottleneck with it as well. Besides, we employ dynamic region-aware convolution [24] in our decoder to learn different kernels according to local illumination features. The DRBlock in the legend of Fig. 1 is composed of a sequence of original DR convolution, batch normalization and ReLU, repeated twice.

E. Loss Functions

Inspired by [8], [9], our model aims to reduce the temporal flickers via adopting the warping loss L_w as:

$$L_{w} = \sum_{d=\{1,2\}} \sum_{t=1}^{N-d} \left\| M_{t+d \Rightarrow t} \odot \left(y_{t} - y_{t+d}^{\text{warp}} \right) \right\|_{2}$$
(4)

where d stands for the frame interval, thus the loss under larger interval means longer temporal coefficient. y_t means the t_{th} frame and $M_{t+d \Rightarrow t} = \exp\left(-\alpha \|y_t - y_{t+d}^{\text{warp}}\|_2^2\right)$ represents the visibility mask [8]. $y_{t+d}^{\text{warp}} = \mathcal{W}(y_{t+d}, f_{t+d \to t})$ where \mathcal{W} warps the $(t+d)_{th}$ frame under the indication of flow $f_{t+d \Rightarrow t}$ from $(t+d)_{th}$ frame to t_{th} frame.

 TABLE I

 QUANTITATIVE COMPARISONS ON DAVIS AND VIDEVO DATASETS. WE PICK BOTH IMAGE AND VIDEO-BASED APPROACHES FOR COMPARISONS. THE

 BEST RESULT IS HIGHLIGHTED IN BLACK BOLD WHILE THE SECOND BEST IS MARKED UNDERLINE.

Methods	DAVIS			Videvo				
	PSNR ↑	SSIM ↑	Warp Error \downarrow	L2 Error \downarrow	PSNR ↑	SSIM ↑	Warp Error↓	L2 Error \downarrow
CIC [5]	22.77	0.9431	0.06055	15.88	22.56	0.9417	0.03317	17.11
IDC [6]	24.85	0.9436	0.05377	11.99	25.17	0.9568	0.02997	11.69
InstColor [32]	24.51	0.9411	0.07828	13.20	24.80	0.9458	0.04917	20.37
GCP [7]	23.53	0.9309	0.04978	12.45	24.25	0.9369	0.02864	12.08
CIC [5]+BTC [8]	22.11	0.9298	0.05170	16.67	21.77	0.9343	0.02891	17.68
IDC [6]+BTC [8]	23.91	0.9006	0.04498	12.91	23.74	0.9383	0.02786	12.60
FAVC [10]	22.98	0.9055	0.06002	13.26	23.47	0.9183	0.03236	12.21
TCVC [9]	25.49	0.9550	0.04819	11.86	25.43	0.9570	0.03589	11.59
Ours	26.68	0.9612	0.04626	10.38	26.95	0.9623	0.02612	10.13
Ours w/o histogram	<u>25.62</u>	<u>0.9590</u>	0.05052	<u>11.49</u>	25.22	0.9587	0.02963	11.37

Besides, we adopt the Charbonnier loss $L_c = \sum_{t=1}^{N} \sqrt{(\hat{y}_t - y_t) + \epsilon^2}$ and smooth loss L_s in Lab space to avoid the gradient exploding and produce smooth result. The overall loss function is shown below:

$$L_{total} = \lambda_1 L_w + \lambda_2 L_c + L_s \tag{5}$$

where λ_1 and λ_2 denote the weights of loss functions.

III. EXPERIMENT

A. Experimental Procedure

Dataset and Metrics. Following previous works [8], [10], [13], we adopt DAVIS dataset [33] and Videvo dataset [8] for training and testing. Originally designed for video segmentation, DAVIS dataset includes a variety of moving objects and distinct motion types. It contains 60 short videos for training and 30 for testing, with 100 frames in each video. The Videvo dataset has 80 long videos for training and 20 for testing. There are about 300 frames in each video clips. To evaluate the quality of generated video frames, we use PSNR, SSIM, warp error [8] and L2 error. The warp error is to measure the temporal continuity of the generated frames.

Implementation Details. We utilize the pytorch framework to implement our model with Adam [34] its optimizer. A single GeForce RTX 3090 graphic card is used to train and test the models. During the training, the size of input frames are resized to 256 * 256. The learning rate is initialized to 5e-5 and we set the α in loss functions to 9 according to [8].



Fig. 4. Illustrations of each feature connection schema where (c) is employed in our proposed ST-HVC.

TABLE II Ablation study on feature connection schema. Detailed structures of (A), (b) and (c) are defined in Fig 4.

Settings	(a)	(b)	(c)
PSNR	26.12	26.49	26.95
drop rate	3.08%	1.71%	-

TABLE III Ablation study on key components. Experiments are conducted in Videvo dataset.

Settings	PSNR↑	SSIM↑	Warp Error \downarrow
w/o histogram guide	25.22	0.9590	0.02993
w/o flow integration	26.31	0.9577	0.03836
w/o spatial attention	24.39	0.9370	0.03295
w/o temporal attention	26.16	0.9551	0.04122
ours	26.95	0.9623	0.02612

B. Ablation Studies

Illustrating Different Connection Schema. We firstly conduct ablation studies on connection schema as is depicted in Tab. II. Detailed structure of each schema is shown in Fig. 4 where (c) is employed in our ST-HVC. The PSNR scores and drop rates from experiments demonstrate that temporal sharpness contributes to the reconstruction process and the direct concatenate operation is less effective than multiplication.

Effectiveness of Different Modules. We perform ablation studies on key components (i.e., the histogram guidance module, flow feature integration, temporal and spatial attention.) in the Videvo dataset. The statistical results are summarized in Tab. I and Tab. III. Without histogram, the proposed model still outperform other methods, and in many cases is merely second to the full model. Besides, the drop rates of PSNR are 6.42% and 2.37%, which demonstrate the great improvement brought by the histogram exploitation schema. The warp error grows nearly a half without flow, meaning the significance of flow integration paradigm.



Fig. 5. Visual results of ablation on flow integration. \rightarrow denotes the movement of frames. Caused by inherent attributes of optical flow, artifacts occur without flow integration.



Fig. 6. Comparison under warp error. We denote the corresponding warp error among the depicted frames in the right. We achieve the best results in both perception and statistics.

Without temporal attention, warp error increases over a half, indicating the contribution of temporal branch in reducing the frames inconsistency. Without spatial attention, the pixel-level scores drop heavily, showing that the spatial branch can better handle the texture information.

Effectiveness of Flow Integration. In Fig. 5, we demonstrate how will flow integration in JFHM affect the visual result. The first column shows the overall frame of dogs jumping and the following three figures show the details of green and red rectangles area while frames move forward. Without flow integration, undersaturated area occurs. It is interesting to find that the area are most obvious in the first frame of one batch and will diminish along the frame sequences. We argue that the phenomenon is stemmed from the inherent attributes of optical flow. With the flow integration, the undersaturated area vanishes and color becomes plausible and vivid again.

C. Comparing with State-of-the-Arts

We conduct thorough experiments between our ST-HVC and several state-of-the-art approaches: image-based methods (i.e., CIC [5], IDC [6], InstColor [32] and GCP [7]) and video-based methods (i.e., FAVC [10], TCVC [9]. Moreover, we apply the blind temporal consistency [8] on CIC [5] and IDC [6] to form another two groups of comparison.

Quantitative Comparison. We conduct quantitative comparisons on DAVIS and Videvo datasets where the results are summarized in Tab. I. The top four rows show the performance of image colorization methods and middle four rows present the result of video-based models. In general, image-based methods can achieve relatively higher PSNR and SSIM scores, but their temporal coherence is poor. GCP ranks the first among image-based methods regarding warp error, since it utilizes generative color prior and the adjacent frames may have similar result when put into GAN encoder. Videobased approaches like BTC is prone to boost the temporal consistency, consequently the warp error is improved. But the cost is singe image colorization quality when compared to original results of CIC and IDC. TCVC strike a balance between temporal consistency and colorization performance due to its bidirectional propagation of frame-level features.

As can be noticed in the bottom row of Tab. I, our method achieved the best score in most cases. Due to our designed architecture, we rank the first in PSNR, SSIM and L2 error in both datasets. It demonstrates that the proposed method can generate the best textual and pixel-wise result closest to the ground truths. Besides, thanks to our flow integration, we outshine most models regarding the temporal consistency.

Comparison in Color. We conduct comparison regarding the chroma with representing and state-of-the-art approaches in Fig. 3. From top to bottom, the figure demonstrates outstanding performance of the proposed method regarding the front color, background color, blurred pixels and the detail. Thanks to designed reference exploitation schema in JFHM, our proposed method can exploit the indication of color distribution. We can surprisingly assign the correct color in details as shown in the mouth of black-swan in bottom row.

Temporal Consistency Comparison. In Fig. 6, we show the temporal consistency performance of our model. Severe temporal flickering with inconsistent colors is occurred in InstColor. As for GCP and TCVC, the overall results are great, but the color change of the body of cars between frames shows that they are still suffering from temporal inconsistency. Our proposed ST-HVC can generate satisfactory images with the least flickering and the lowest warp loss as is depicted in the figure.

IV. CONCLUSION

We proposed a novel network with spatial-temporal connection schema to tackle the chroma assignment and temporal coefficient challenges in video colorization. To demonstrate its effectiveness, we conduct comparisons with the several stateof-the-art image and video-based methods. The experiment results show that the developed method achieves excellent scores on four quality metrics in two classic datasets.

Limitations Despite the competitive performance, the proposed network cannot colorize well if scene changes rapidly, since histogram of the middle frame is no longer representative. Future work will take these cases into account.

REFERENCES

- Patricia Vitoria, Lara Raad, and Coloma Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *Proceed*ings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2445–2454.
- [2] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi, "Ct2: Colorization transformer via color tokens," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–16.
- [3] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner, "Colorization transformer," arXiv preprint arXiv:2102.04432, 2021.
- [4] Jie Zhang, Yi Xiao, Guo Chen, Qingping Sun, Fangqiang Xu, and Chi-Sing Leung, "Histogram-guided semantic-aware colorization," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 2549–2553.
- [5] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 649–666.
- [6] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros, "Real-time user-guided image colorization with learned deep priors," ACM Trans. Graph., 2017.
- [7] 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.
- [8] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang, "Learning blind video temporal consistency," in *Proceedings of the European conference on computer vision* (ECCV), 2018, pp. 170–185.
- [9] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong, "Temporally consistent video colorization with deep feature propagation and self-regularization learning," arXiv preprint arXiv:2110.04562, 2021.
- [10] Chenyang Lei and Qifeng Chen, "Fully automatic video colorization with self-regularization and diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3753–3761.
- [11] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.
- [12] Chenyang Lei, Yazhou Xing, and Qifeng Chen, "Blind video temporal consistency via deep video prior," Advances in Neural Information Processing Systems, vol. 33, pp. 1083–1093, 2020.
- [13] Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou, "Vcgan: Video colorization with hybrid generative adversarial network," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.
- [14] Pan Mu, Zhu Liu, Yaohua Liu, Risheng Liu, and Xin Fan, "Triple-level model inferred collaborative network architecture for video deraining," *IEEE Transactions on Image Processing*, pp. 239–250, 2021.
- [15] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan, "Deep exemplar-based colorization," ACM Transactions on Graphics (TOG), pp. 1–16, 2018.
- [16] Hengyuan Zhao, Wenhao Wu, Yihao Liu, and Dongliang He, "Color2embed: Fast exemplar-based image colorization using color embeddings," arXiv preprint arXiv:2106.08017, 2021.
- [17] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown, "Histogan: Controlling colors of gan-generated and real images via color histograms," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 7941–7950.
- [18] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.
- [19] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.
- [20] Pan Mu, Haotian Qian, and Cong Bai, "Structure-inferred bi-level model for underwater image enhancement," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2286–2295.

- [21] Risheng Liu, Pan Mu, Jian Chen, Xin Fan, and Zhongxuan Luo, "Investigating task-driven latent feasibility for nonconvex image modeling," *IEEE Transactions on Image Processing*, pp. 7629–7640, 2020.
- [22] Jinshan Pan, Haoran Bai, and Jinhui Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3043–3051.
- [23] Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai, "Explore image deblurring via encoded blur kernel space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11956–11965.
- [24] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun, "Dynamic region-aware convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8064–8073.
- [25] Jiawen Chen, Sylvain Paris, and Frédo Durand, "Real-time edge-aware image processing with the bilateral grid," ACM Transactions on Graphics (TOG), pp. 103–es, 2007.
- [26] Zachary Teed and Jia Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 402–419.
- [27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference* on Computer Vision, 2015, pp. 4489–4497.
- [28] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.
- [29] Kaidong Zhang, Jingjing Fu, and Dong Liu, "Flow-guided transformer for video inpainting," in *European Conference on Computer Vision*. Springer, 2022, pp. 74–90.
- [30] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen, "Decoupled dynamic spatial-temporal graph neural network for traffic forecasting," arXiv preprint arXiv:2206.09112, 2022.
- [31] Haoyuan Wang, Ke Xu, and Rynson WH Lau, "Local color distributions prior for image enhancement," in *European Conference on Computer Vision.* Springer, 2022, pp. 343–359.
- [32] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang, "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7968–7977.
- [33] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.
- [34] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.