ColorMNet: A Memory-based Deep Spatial-Temporal Feature Propagation Network for Video Colorization -Supplementary Material-

Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan†

Nanjing University of Science and Technology
{yangyixin, jxdong, jinhuitang, jspan}@njust.edu.cn

Overview

In this document, we first present the network details in Section 1. Then, we analyze the effectiveness of the proposed large-pretrained visual model guided feature estimation (PVGFE) module and the memory-based feature propagation (MFP) module in Section 2 and Section 3. To examine the effectiveness of the proposed local attention (LA) module on video colorization, we further analyze it in Section 4. We then compare with closely-related methods in Section 5. In addition, we conduct a user study to investigate the subjective preference by human observers of each colorization method in Section 6. Finally, we show more visual comparisons on both synthetic datasets and real-world videos in Section 7.

1 Network Details

As stated in Section 3 of the main manuscript, our method contains a large-pretrained visual model guided feature estimation module, a memory-based feature propagation module, and a local attention module for video colorization. We also show the network details of the proposed memory-based deep spatial-temporal feature propagation network for video colorization in Figures 2 of the main manuscript. In this document, we list the detailed architecture of our proposed ColorMNet in Table 1. The spatial resolution of the input image is 448×448 pixels. To handle test videos of varying dimensions, we first follow the widely used protocol to pad images so that they are divisible by 112 as the usage of ViT-S/14 from DINOv2 and stage-4 features from ResNet50 requires images to be divisible by 14 and 16, respectively. Then we interpolate the features generated by DINOv2 to match the resolution of those generated by ResNet.

2 Effectiveness of the Large-Pretrained Visual Model Guided Feature Estimation Module

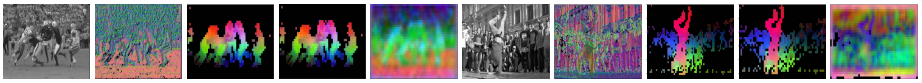
As stated in Section 5 of the manuscript, we have analyzed the effectiveness of the large-pretrained visual model guided feature estimation (PVGFE) module. In this supplemental material, we further show more visual comparisons

Table 1: Detailed architecture of our proposed ColorMNet. [Conv. 7×7, 64, stride 2] denotes a convolution with the filter size of 7×7 pixels with the filter number of 64 with stride 2, Embed dim. denotes the dimension of embedding, [Interpolation, ×2] denotes an interpolation operation with a scale factor equal to 2, [ResBlock, 256] denotes a ResBlock consisting of convolutions with the filter size of 3×3 pixels with the filter number of 256.

		ColorMNet	
	Output size	ResNet50 [3]	DINOv2 [9]
Key feature extractor (PVGFE)	28×28×1024	Conv. 7×7, 64, stride 2 MaxPool, 3×3, stride 2 [Conv. 1×1, 64 Conv. 3×3, 64 Conv. 1×1, 256] × 3 [Conv. 1×1, 128 Conv. 3×3, 128 Conv. 1×1, 512] × 4 [Conv. 1×1, 256 Conv. 3×3, 256 Conv. 1×1, 1024] × 6	Patch Embedding [Transformer block Patch size = 14 Embed dim. = 384 Heads = 6 Blocks = 12 FFN layer = MLP] × 12 Concat features from last 4 layers Conv. 1×1, 1536 Interpolation, ×14/16 Conv. 3×3, 1024
	28×28×64	Conv. 3×3, 64	Conv. 3×3, 64
	28×28×64	Cross-channel attention [13]	
Value feature extractor (ResNet18 [3])	28×28×256	ResNet18 [3] Conv. 7×7, 64, stride 2 MaxPool, 3×3, stride 2 [Conv. 1×1, 64 Conv. 3×3, 64] × 2 [Conv. 1×1, 128 Conv. 3×3, 128] × 2 [Conv. 1×1, 256 Conv. 3×3, 256] × 2	
	28×28×512	Conv. 3×3, 512	
Decoder	112×112×256	Interpolation, ×2 ResBlock, 256 Interpolation, ×2 ResBlock, 256	
	112×112×2	Conv. 1×1, 2	
	112×112×2	GRU [1]	
	448×448×2	Interpolation, ×4	

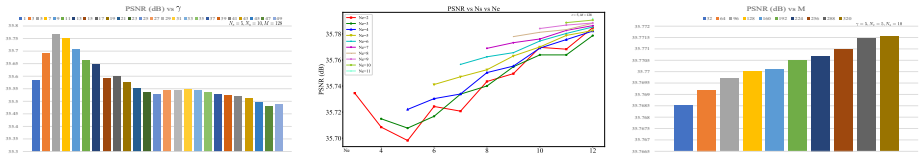
to demonstrate the effectiveness of the PVGFE module. In *'ComparisonsWithSOTA.mp4'*, we show that our proposed ColorMNet with using the PVGFE is able to generate better-colorized videos.

To better understand the feature estimators mentioned above, we use the PCA tools by [9] to visualize the features generated by them. Figure 1(b) shows that ResNet50 cannot generate features that are aware of semantic structures. Although DINOv2 generates more semantic features in Figure 1(c) and (d), it lacks the local details vital for colorization tasks, which explains why DINOv2 performs favorably in high-level vision tasks, *i.e.*, classification and segmentation (see [9] for details), but fails in colorization as the exact colors for pixels of objects are crucial considerations in colorization, unlike in segmentation where the primary decision is whether or not a pixel belongs to a human. Figure 1(e) shows that our proposed PVGFE module is capable of generating better features optimized for colorization, retaining both semantic relevance and local details.



(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

Fig. 1: Visualization of features. We use the PCA tools by [9]. (a) Input frame. (b)-(e) are the features generated by the feature extractors of ColorMNet$_{\text{w/ ResNet50}}$, ColorMNet$_{\text{w/ DINOv2}}$, ColorMNet$_{\text{w/ Concatenation}}$ and ColorMNet (Ours), respectively. Compared with (b), (c) and (d), our proposed PVGFE can generate features that are not only semantic-aware (*i.e.*, the players and the dancer in the foreground) but also sensitive to local details (*i.e.*, a crowd of spectators in the background) in (e).



(a) Ablation on different γ (b) Ablation on N_e and N_s (c) Ablation on different M

Fig. 2: Extensive ablation study on the detailed design of the proposed MFP module.

3 Effectiveness of the Memory-based Feature Propagation Module

As stated in Section 5 of the manuscript, we have analyzed the effectiveness of the memory-based feature propagation (MFP) module. In this supplemental material, we further show more visual comparisons to demonstrate the effectiveness of the MFP module. In *'ComparisonsWithSOTA.mp4'*, we show that our proposed ColorMNet with using the MFP is able to generate better-colorized videos compared with the method without using the MFP.

In addition, we conduct an extensive ablation study on the parameters of the proposed MFP module. Figure 2(a) shows that our method achieves its peak PSNR when γ equals 5, as a higher γ risks potential information loss, while a lower γ could contribute redundant data. Figure 2(b) and (c) show that our method generally achieves slightly higher PSNR values with N_e, N_s and M increasing, respectively. However, note that the GPU memory usage escalates correspondingly with larger values of N_e, N_s and M.

4 Effectiveness of the Local Attention Module

As stated in Section 5 of the manuscript, we have analyzed the effectiveness of the local attention (LA) module. We empirically set $\lambda = 7$ for $\lambda \times \lambda$ patch $\mathcal{N}(p)$. In this supplemental material, we further show more visual comparisons to demonstrate the effectiveness of the LA module. In *'ComparisonsWithSOTA.mp4'*, we show that our proposed ColorMNet with using the LA is able to generate better-colorized videos.

(a) Inputs (b) MAMBA [11] (c) MeMOTR [2] (d) Ours

Fig. 3: Comparison results with closely-related methods on the DAVIS [10].

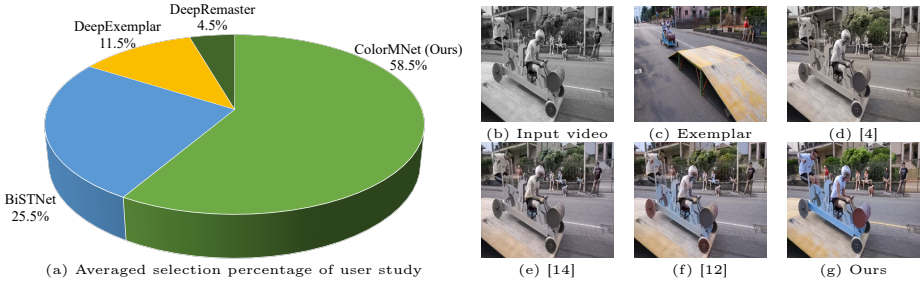


Fig. 4: User study result and an example of a group of results displayed to human observers in the user study. (a) shows that our proposed ColorMNet achieves obviously higher score than other state-of-the-art methods, which demonstrates its subjective advantages. (b)-(g) are the input video, the exemplar image, the colorized videos by DeepRemaster [4], DeepExemplar [14], BiSTNet† [12] and ColorMNet (Ours), respectively. We make the methods anonymous and randomly sort the videos in (d)-(g) to ensure fairness. † denotes that two exemplars are used.

5 Closely-related methods.

To the best of our knowledge, we are the first to optimize a memory bank strategy suitable for colorization, yet it should be acknowledged that related strategies have been explored in some video processing works, *e.g.*, MAMBA [11] constructs a memory bank to solve video object detection by employing random selection strategy, MeMOTR [2] introduces a long-term memory to solve video object tracking by assigning exponentially decaying weights to it. Unlike MAMBA which applies a randomized selection approach, treating every feature on par, or MeMOTR which updates past memorized features via exponentially decaying weights, our proposed MFP module stores features based on their importance which is determined by the frequency of usage, thus empowering the ability of global relation mining.

We further adopt the random selection in MAMBA and the decaying weights in MeMOTR to replace our MFP for comparison. To ensure a fair comparison, the same training settings are kept for model testing. Figure 3 shows that our method can generate better colors for the dancing girl and the green grass.

6 User Study

To evaluate whether our results are favored by human observers, we further conduct user study experiments. Specifically, we compare our method with exemplar-based methods, i.e., BiSTNet [12], DeepExemplar [14] and DeepRemaster [4]. We randomly select 10 input videos from the DAVIS [10] validation set, the Videvo [7] validation set and the NVCC2023 [5] validation set together with the colorization results and the exemplar images displayed to 20 online observers without constraints. We make the methods anonymous and randomly sort the videos in each group to ensure fairness. Observers are asked to choose the most visually pleasing results from a group of videos. Figure 4 shows that our method is preferred by a wider range of users than other state-of-the-art methods.

7 More Experimental Results

In this section, we provide more visual comparisons with state-of-the-art methods on both synthetic and real-world videos. Figures 5-16 show the comparisons, where our method generates better colorized frames. In *'ComparisonsWithSOTA.mp4'*, we show that the proposed method generates vivid and realistic videos.

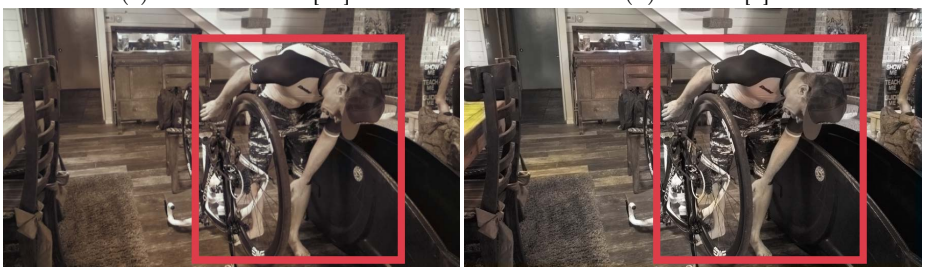
References

1. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
2. Gao, R., Wang, L.: MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In: ICCV (2023)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
4. Iizuka, S., Simo-Serra, E.: Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. ACM TOG **38**(6), 1–13 (2019)
5. Kang, X., Lin, X., Zhang, K., et al.: Ntire 2023 video colorization challenge. In: CVPRW (2023)
6. Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X.: Ddcolor: Towards photo-realistic image colorization via dual decoders. In: ICCV (2023)
7. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: ECCV (2018)
8. Liu, Y., Zhao, H., Chan, K.C., Wang, X., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. arXiv preprint arXiv:2110.04562 (2021)
9. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
10. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)

11. Sun, G., Hua, Y., Hu, G., Robertson, N.: Mamba: Multi-level aggregation via memory bank for video object detection. In: AAAI (2021)
12. Yang, Y., Peng, Z., Du, X., Tao, Z., Tang, J., Pan, J.: Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. IEEE TPAMI pp. 1–14 (2024)
13. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
14. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: CVPR (2019)
15. Zhao, H., Wu, W., Liu, Y., He, D.: Color2embed: Fast exemplar-based image colorization using color embeddings. arXiv preprint arXiv:2106.08017 (2021)
16. Zhao, Y., Po, L.M., Yu, W.Y., Rehman, Y.A.U., Liu, M., Zhang, Y., Ou, W.: Vcgan: video colorization with hybrid generative adversarial network. IEEE TMM (2022)

(a) Input frame and exemplar image (b) DDColor [6]

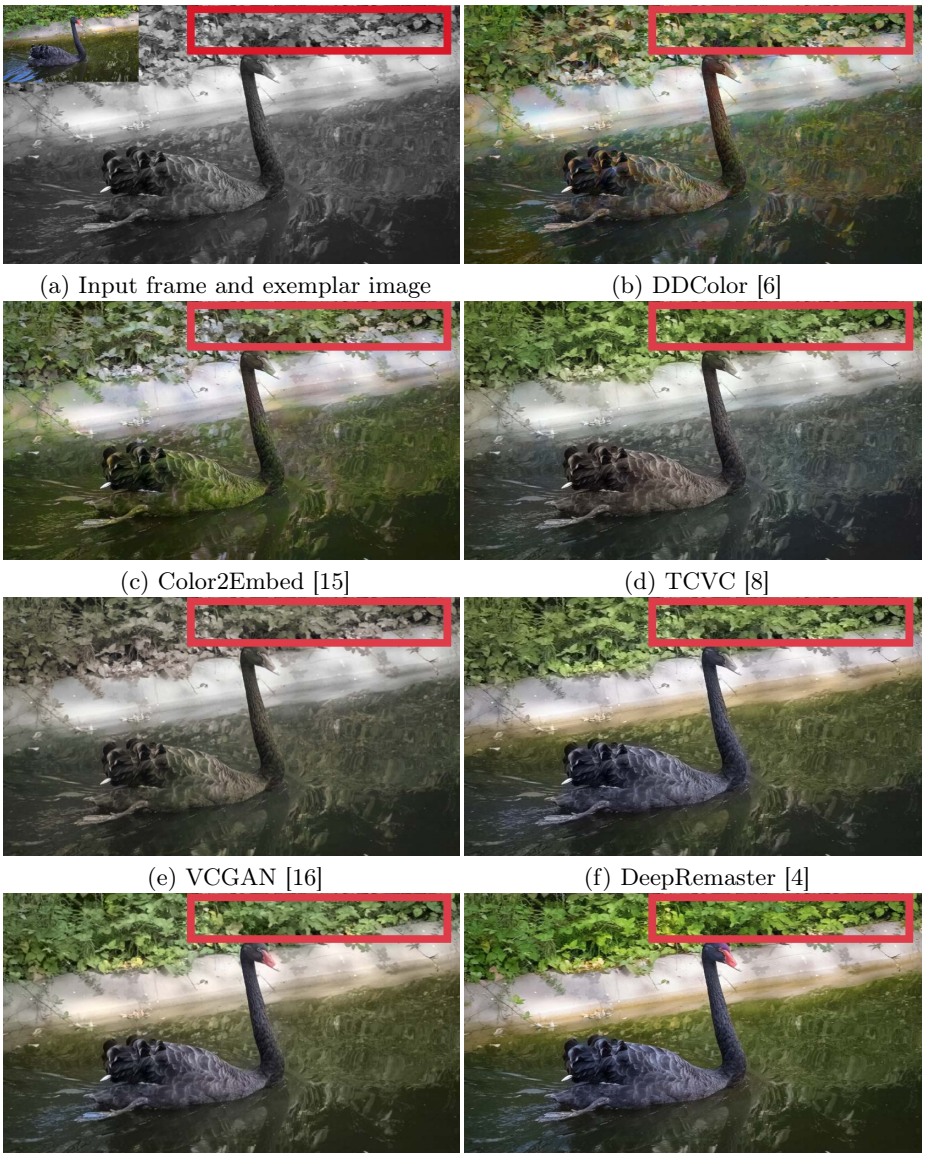
(c) Color2Embed [15] (d) TCVC [8]



(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 5: Colorization results on clip *bike-packing* from the DAVIS validation dataset [10]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. [4, 8, 14, 16] do not recover the man well. In contrast, our proposed method generates a better-colorized frame, where the man is restored well and the colors look better.



(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 6: Colorization results on clip *blackswan* from the DAVIS validation dataset [10]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a better-colorized frame.

(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 7: Colorization results on clip *breakdance* from the DAVIS validation dataset [10]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a better-colorized frame, where the colors of the dancer are restored well.

(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 8: Colorization results on clip *car-roundabout* from the DAVIS validation dataset [10]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method restores the colors of the flowerbed and generates a better-colorized frame.

(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 9: Colorization results on clip *loading* from the DAVIS validation dataset [10]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a vivid and realistic frame, where the colors of the box and the man's hands are better restored.

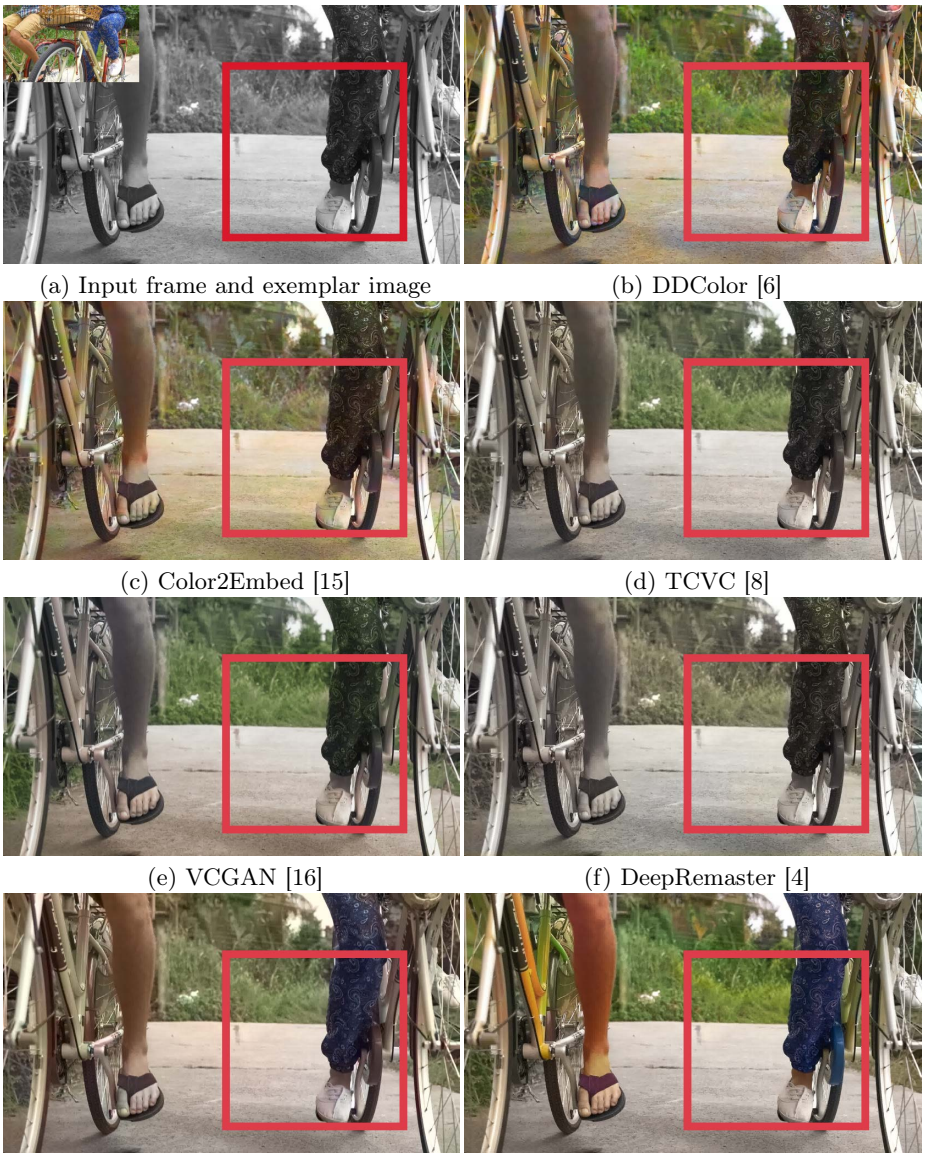
(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 10: Colorization results on clip *CoupleRidingMotorbike* from the Videvo validation dataset [7]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a realistic frame that is faithful to the exemplar image.



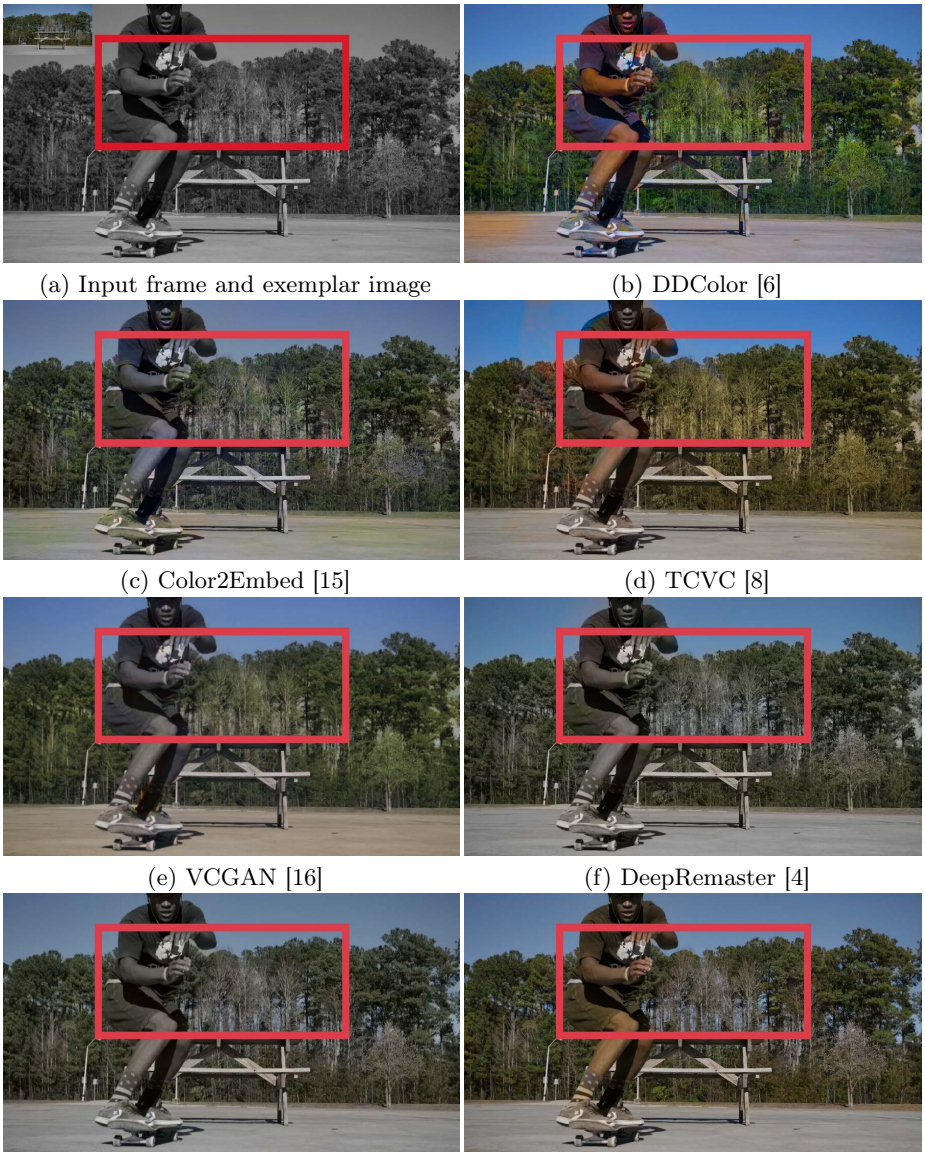
(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 11: Colorization results on clip *Cycling* from the Videvo validation dataset [7]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a vivid frame than other stage-of-the-art methods.



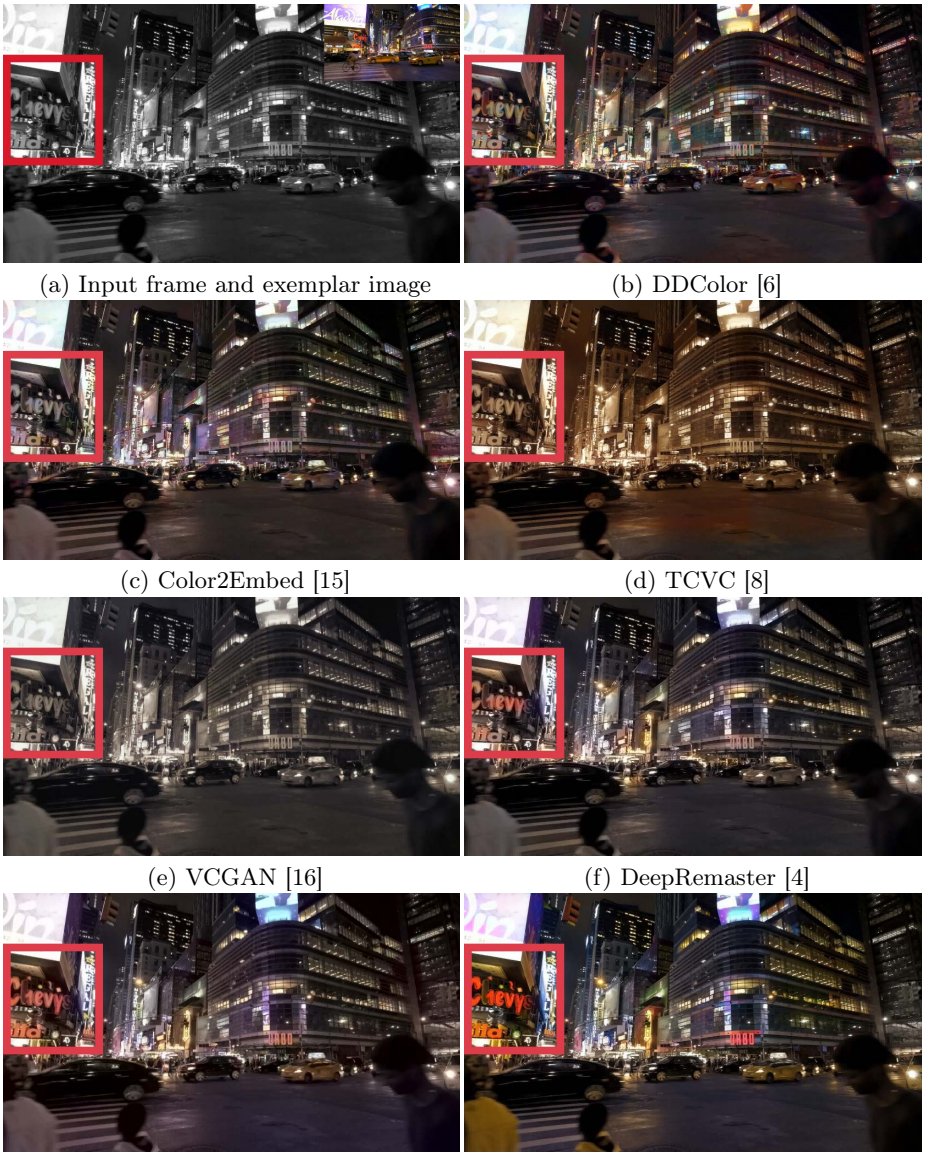
(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 12: Colorization results on clip *SkateboarderTableJump* from the Videvo validation dataset [7]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a realistic frame, where the colors of the skateboard man and the trees are better restored.



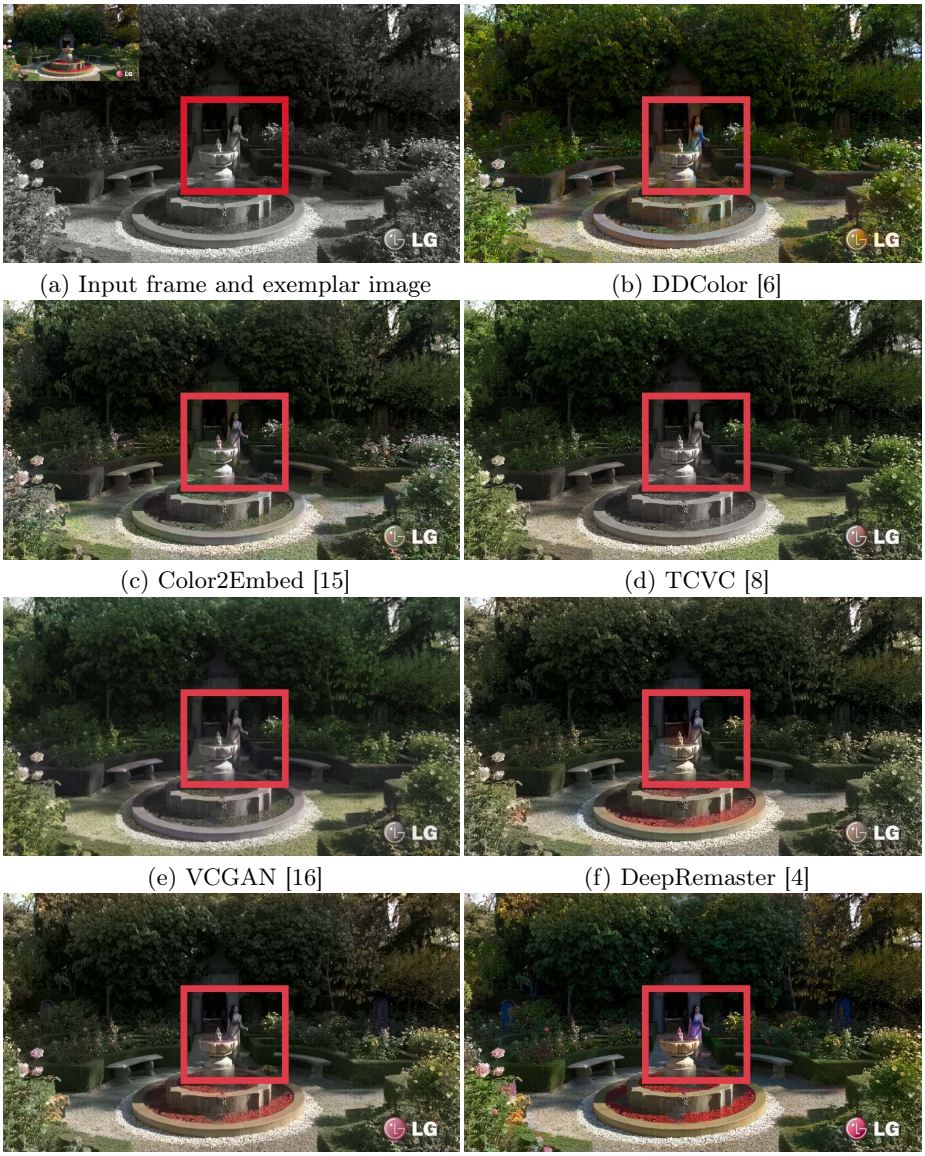
(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 13: Colorization results on clip *TimeSquareTraffic* from the Videvo validation dataset [7]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a vivid and realistic frame against other stage-of-the-art methods.



(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 14: Colorization results on clip *001* from the NVCC2023 validation dataset [5]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a better-colorized frame, where the colors of the woman and the leaves are better restored.

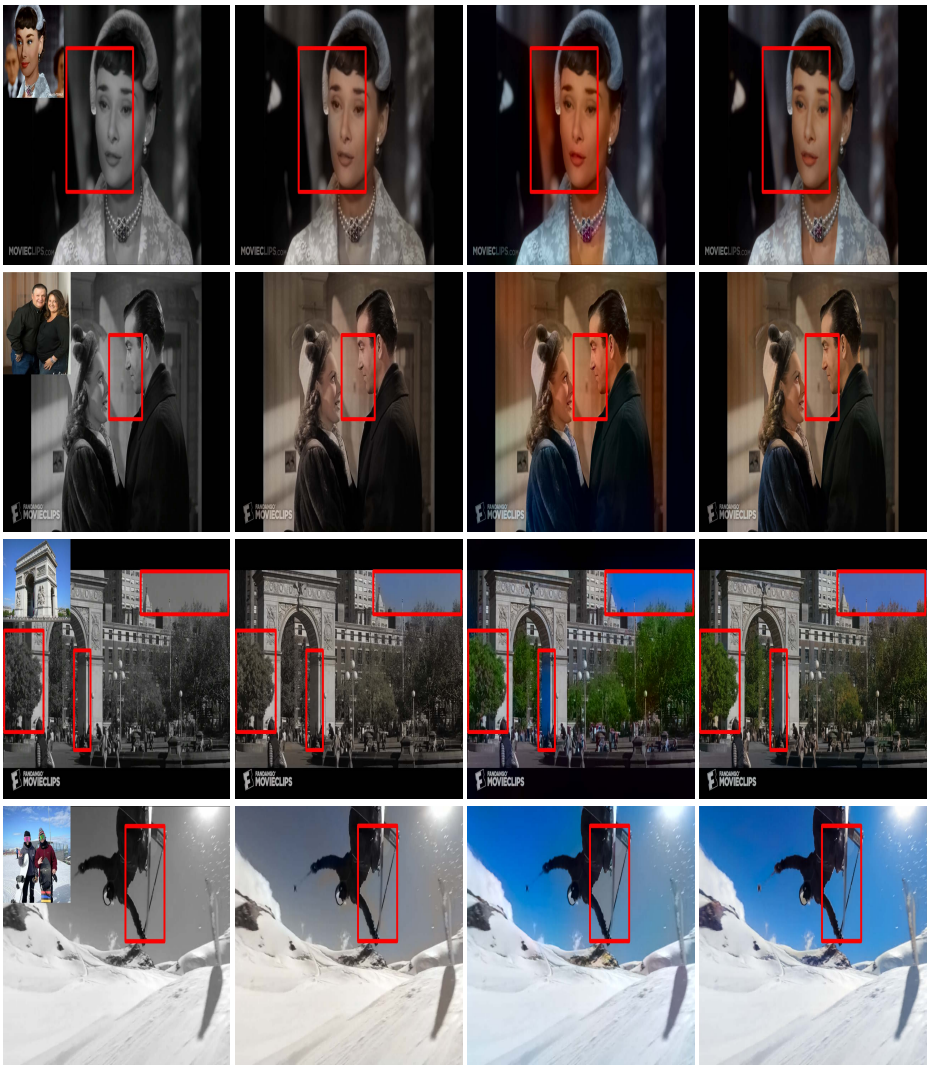
(a) Input frame and exemplar image (b) DDColor [6]

(c) Color2Embed [15] (d) TCVC [8]

(e) VCGAN [16] (f) DeepRemaster [4]

(g) DeepExemplar [14] (h) ColorMNet (Ours)

Fig. 15: Colorization results on clip *014* from the NVCC2023 validation dataset [5]. The results shown in (b) and (c) still contain significant color-bleeding artifacts. In contrast, our proposed method generates a vivid frame in (h) that is not only more colorful compared with (d-g) but also faithful to the exemplar image in (a).



(a) Inputs (b) DeepRemaster [4] (c) DeepExemplar [14] (d) ColorMNet (Ours)

Fig. 16: Colorization results on real-world videos. From top to bottom are respectively the film clip from *Roman Holiday (1953)*, the film clip from *Miracle on 34th Street (1947)*, the film clip from *Manhattan (1979)* and a real-world video collected from the internet. We obtain the exemplars by searching the internet to find the most visually similar images to the input video frames. DeepRemaster [4] cannot gerenate vivid frames. The results shown in (c) generated by DeepExemplar [14] still contain significant color-bleeding artifacts (the wall of the building and the skiing man) and cannot maintain faithfulness to the given exemplar images (over-saturated colors on the both the face of the man and the face of the woman). In contrast, our proposed method generates vivid and realistic frames.