Colorblind-Shareable Videos by Synthesizing Temporal-Coherent Polynomial Coefficients

XINGHONG HU, The Chinese University of Hong Kong
XUETING LIU, Caritas Institute of Higher Education
ZHUMING ZHANG, The Chinese University of Hong Kong
MENGHAN XIA, The Chinese University of Hong Kong
CHENGZE LI, The Chinese University of Hong Kong and Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
TIEN-TSIN WONG, The Chinese University of Hong Kong and Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

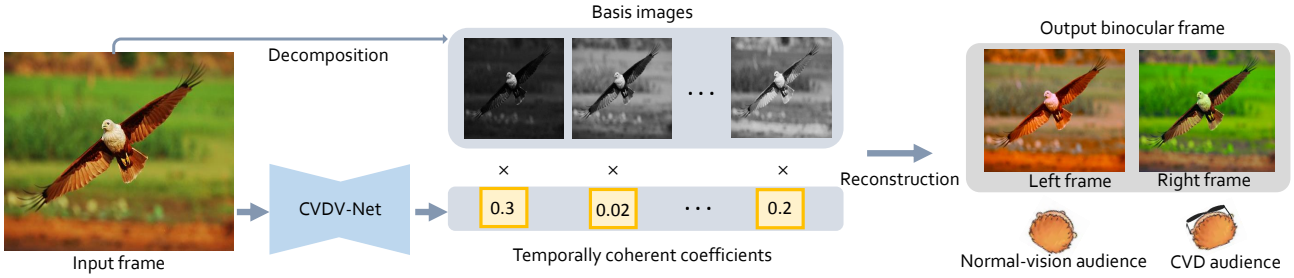

Fig. 1. Given an input video that contains color-confusing regions to CVD audience (e.g. the brown eagle and the green grass), we train a deep neural network to generate a binocular video pair that can be simultaneously shared by CVD and normal-vision audiences. Instead of generating the left and right videos independently which may exhibit color inconsistency and violation of binocular constraints, we propose an indirect approach to maintain the color consistency, the temporal coherence, as well as all binocular constraints. We generate the temporal-coherent polynomial coefficients and linearly combine with basis images to construct the final left and right videos.

To share the same visual content between color vision deficiencies (CVD) and normal-vision people, attempts have been made to allocate the two visual experiences of a binocular display (wearing and not wearing glasses) to CVD and normal-vision audiences. However, existing approaches only work for still images. Although state-of-the-art temporal filtering techniques can be applied to smooth the per-frame generated content, they may fail to maintain the multiple binocular constraints needed in our applications, and even worse, sometimes introduce color inconsistency (same color regions map to different colors). In this paper, we propose to train a neural network to predict the temporal coherent polynomial coefficients in the domain of global color decomposition. This indirect formulation solves the color inconsistency problem. Our key challenge is to design a neural network to predict the temporal coherent coefficients, while maintaining all required binocular constraints. Our method is evaluated on various videos and all metrics confirm that it outperforms all existing solutions.

CCS Concepts: • **Computing methodologies → Image processing**.

Additional Key Words and Phrases: Color vision deficiency, temporal coherence, machine learning

ACM Reference Format:
Xinghong Hu, Xueting Liu, Zhuming Zhang, Menghan Xia, Chengze Li, and Tien-Tsin Wong. 2019. Colorblind-Shareable Videos by Synthesizing Temporal-Coherent Polynomial Coefficients. *ACM Trans. Graph.* 38, 6, Article 174 (November 2019), 12 pages. https://doi.org/10.1145/3355089.3356534

1 INTRODUCTION

There are about 4.5% of the entire human population who suffer from various degrees of color vision deficiencies (CVD). It is common for CVD people to share the same visual content with normal-vision individuals. Researchers have made attempts to enhance the sharing of visual content between CVD and normal-vision audiences by using stereoscopic displays [Chua et al. 2015; Shen et al. 2016]. The idea is to generate an output image pair from an input image, and feed the output to a stereoscopic display. Only the CVD audiences will wear stereoscopic glasses, which can help them better distinguish the colors that are originally confusing to them. In the meantime, the visual experiences of normal-vision audiences will not be affected.

Authors' addresses: Hu Xinghong, Zhang Zhuming, Xia Menghan, The Chinese University of Hong Kong, {huxh,zhangzm,mhxia}@cse.cuhk.edu.hk; Liu Xueting, Caritas Institute of Higher Education, tliu@cihe.edu.hk; Li Chengze, WONG Tien-Tsin, The Chinese University of Hong Kong and Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, {czli, ttwong}@cse.cuhk.edu.hk.

0730-0301/2019/11-ART174 $15.00
https://doi.org/10.1145/3355089.3356534

Existing approaches cannot be directly extended to video due to their temporal incoherency and computational inefficiency, and hence their practical usage is significantly limited. In this paper, we propose the first practical solution for fast synthesis of temporal-coherent colorblind-sharable video. Our fast solution is accomplished by a convolutional neural network (CNN) approach. Although the state-of-the-art video-to-video translation network [Wang et al. 2018] can suppress the flickering, it does not ensure the color consistency, i.e. the same-color regions remain the same in color after recoloring. For example, both the floor and the back of chair in Figure 2(c) exhibit color inconsistency artifact. The underlying reason is the local nature of CNN during the direct image generation, as well as the insufficiency of color constraint.

To avoid the color inconsistency, one must constrain the recoloring process so that similar colors remain similar after the process. Due to this global constraint, we propose to use a global color transformation formulation. It first decomposes the color channels of each input video frame into several basis images and then linearly recombines them with the corresponding coefficients (Figure 1, middle). Instead of directly generating the pixels, we generate these indirect coefficients. This indirect formulation offers sufficient constraint on the global color consistency, by trading off the flexibility of per-pixel processing. Such polynomial transformation has been applied in other applications [Lu et al. 2012; Wolf 2003]. This transformation can decompose any single image, without relying on a set of images. It mainly operates in the color space and is independent of the image content. This matches our recoloring application well, as we do not want to alternate the image content. Our major challenge is how to generate the temporal-coherent polynomial coefficients. In this paper, we propose to train a convolutional neural network to predict these temporal-coherent coefficients (Figure 1, right).

Besides the color consistency, our generated colorblind-sharable video also needs to satisfy four constraints, including the color distinguishability (CVD audiences can better distinguish colors), the binocular fusibility (left and right views of CVD audiences can perceptually fuse), the color preservation (normal-vision audiences hardly observe difference to the original video), and temporal coherence (no flickering for both CVD and normal-vision audiences). Instead of generating the left and right videos separately, we train our network to predict temporal-coherent coefficients for generating a single difference video (between left and right views), which in turn to generate the binocular pair (Figure 3). Through this formulation, our model implicitly learns the binocular constraints required. Our training is unsupervised and directly optimizes the above four constraints, and hence relieves the burden of preparing paired training data.

To evaluate our method, we conduct qualitative and quantitative experiments as well as the user study. All results confirm that our method can generate convincing colorblind-sharable videos. Our contributions can be summarized as follows:

- We propose the first method to rapidly generate temporal-coherent and visually-shareable videos for CVD and normal-vision audiences, using a CNN-based approach.

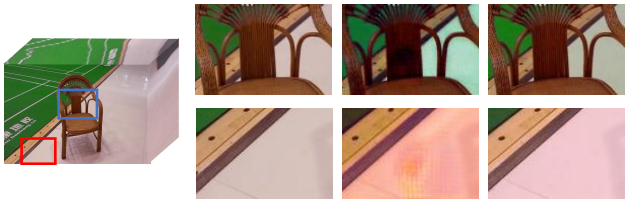

Fig. 2. Given an input video (a), we generate a binocular video pair (d). Compared to Vid2Vid result (c), our result maintains both the temporal consistency and the spatial color consistency.

- We propose an unsupervised learning scheme to train the model by optimizing an objective function of all above visual-sharing constraints. So the burden of collecting paired training data is relieved.
- We propose a novel indirect solution that generates temporal-coherent parameters to a difference video, instead of directly generating pixels of left and right videos. This formulation ensures the global color consistency.

2 RELATED WORK

2.1 Visual Sharing with CVD

Tinted glasses have been proposed to enhance the color distinguishability for CVD audiences by filtering light in certain wavelengths. Wearing the glasses allows CVD people to share the same visual content with normal-vision people. However, such color filtering is not visually comfortable, and it hurts the depth perception of CVD audiences. For sharing the digital content, various methods have been proposed to recolor the image/video [Spe 2015; Huang et al. 2007] or overlay texture patterns on color-confusing regions [Sajadi et al. 2013] for enhancing CVD color distinguishability while minimizing the color change. However, even though the color change or overlayed patterns are optimized, visual experience of normal-vision people is still affected.

Recently, researchers attempted to enhance the sharing of visual content between CVD and normal-vision audiences by utilizing stereoscopic display [Chua et al. 2015; Shen et al. 2016]. The idea is to allocate the two visual experiences (with and without wearing glasses) to CVD and normal-vision audiences. The CVD audiences could perceive better distinguishability by wearing glasses, while the normal-vision audiences can still enjoy the original visual experiences and no glasses is needed. In particular, Chua et al. [2015] proposed to identify the color-confusing regions via segmentation and only modify the colors of those regions. Shen method [Shen et al. 2016] is proposed to generate the image pair via an optimization-based approach. Their method starts with a random initialization and optimizes in a gradient descent fashion. While Shen method can generate good quality results, the initialization is randomly guessed so that a bad initialization may lead to a result that cannot meet the constraints. More importantly, both Chua and Shen methods are only designed for still images. Both methods cannot be directly extended to video due to their temporal incoherency and computational inefficiency, which significantly limits their practical usage.

More recently, Hu et al. [2019] utilized CNN to generate visually-shareable still image pairs. However, their model is supervised by the results from Shen method, and their performance is still not promising. Their results also suffer from the color inconsistency as their method imposes no constraint on the color consistency. Furthermore, their method is only designed for still images.

In contrast, our learning-based approach can stably generate high-quality results, as learning-based method optimizes for the distribution of a vast amount of training data, while traditional optimization methods, e.g. Shen method, optimize solely based on one particular input instance and may be highly affected by the initialization. Comparing to Hu method [Hu et al. 2019], our method is unsupervised by nature, and hence our performance is not bounded. Such unsupervised learning scheme has also been demonstrated as an useful tool for image manipulation tasks [Fan et al. 2018]. Moreover, our polynomial transformation effectively avoids the color inconsistency. Furthermore, our method is recurrently trained for predicting temporal coherent coefficients, and hence solves the temporal inconsistency problem of previous methods.

2.2 Temporal Coherence in Video

While there exists various image processing filters, several methods have been proposed to extend these image-based filters to video with temporal coherence. Many of them are only applicable to certain image filters, including both traditional methods [Aydin et al. 2014; Bonneel et al. 2013, 2014; Kong et al. 2014; Wang et al. 2006; Ye et al. 2014] and deep learning based methods [Chen et al. 2017; Jiang et al. 2017; Sajjadi et al. 2018]. Although more general temporal coherence enhancement methods have been proposed [Lang et al. 2012; Paris 2008], the formulation of applied image processing filters has to satisfy certain assumption. Hence, on limited set of image filters, they cannot be directly applied to our *binocular* case in which we have a completely different problem setting.

Recently, several methods have been proposed to enhance the temporal coherence for arbitrary image processing filters [Bonneel et al. 2015; Lai et al. 2018; Wang et al. 2018]. While these methods may produce temporally coherent videos, they do not consider the dual video channels with binocular constraints as in our visual sharing application. Individually smoothing the left and right videos may violate the binocular constraints required in visual sharing. For stereo/multi-view videos, methods have been developed to satisfy the constraint among the views [Bonneel et al. 2017]. However, different from our binocular fusibility and color preservation requirements, they enforce the consistency of color values between corresponding pixels of different views. Moreover, due to the local adjustment of color during the smoothing, the above methods may fail to preserve the color consistency over the whole image. In contrast, our color transformation approach effectively avoids the color inconsistency problem of above methods.

Manchado et al. [2010] and Huang et al. [2011] also proposed to synthesize temporally coherent videos where color distinguishability is enhanced for CVD audiences. However, both methods are not designed for the visual sharing purpose.

3 OVERVIEW

Given an input video that may contain color-confusing regions to CVD audiences, our goal is to generate a colorblind-sharable pair of left and right videos that satisfies four constraints, including the color distinguishability (to help CVD audiences distinguish confusing colors), the binocular fusibility (to ensure the two views form a stable binocular single vision), the color preservation (to ensure no change to the visual experience of normal-vision audiences), and the temporal coherence. To do so, we build our system by two major components: The CVDI-Net works on still images and the video processing component, the CVDV-Net. We start by describing our still image component CVDI-Net, followed by the video component CVDV-Net.

Figure 3 overviews our CVDI-Net. Given an input image (or frame) I, CVDI-Net generates a pair of left and right images $\{O^l, O^r\}$ that satisfies three out of four constraints above, excluding the temporal coherence. To better learn the binocular contraints, we train our network to generate a difference image $D = O^r - O^l$ between left and right images, instead of generating O^l and O^r separately. The key idea to ensure the color consistency over the whole image is to decompose the input image into a linear combination of basis images and the corresponding k coefficients (Section 4). Instead of directly generating the image pixels, we train our neural network to predict the k coefficients. The output difference image D, and hence $\{O^l, O^r\}$, are then indirectly computed as the linear combination of the basis images with the predicted coefficients. This approach effectively avoids the local inconsistency color adjustment, and hence ensures the color consistency over the whole image. The details of our CVDI-Net, including the dataset preparation, the network architecture, and the loss function are discussed in Section 5.

Our CVDV-Net derived from the CVDI-Net supports video pair synthesis. The network overview is shown in Figure 4 . We design the network so that when given an input video $\{I_t | t = 0, 1, ...\}$, CVDV-Net generates a difference video $\{D_t | t = 0, 1, ...\}$, such that the pair of inferred left video $\{O^l_t | t = 0, 1, ...\}$ and right video $\{O^r_t | t = 0, 1, ...\}$ satisfies all four constraints above. The major difference between CVDV-Net and CVDI-Net, is that the CVDV-Net additionally takes the warped previous difference image D_{t-1} and a confidence map M_t as input to account for the temporal coherence. The warping is performed by computing the optical flow between the previous I_{t-1} and the current frame I_t with a pre-trained FlowNet2 model [Ilg et al. 2017]. The video-based CVDV-Net makes use of the image-based CVDI-Net as the basic building block. Whenever there is disocclusion, the CVDI-Net is utilized to generate the disoccluded pixels. The details of our CVDV-Net are presented in Section 6.

4 IMAGE DECOMPOSITION AND CONSTRUCTION

For global transformation, we adopt the polynomial representation as it is an intuitive mathematical formulation. Our approach decomposes the input image (or frame) into a linear combination of k basis images and the corresponding k coefficients to ensure the color consistency over the whole image. The network is afterwards trained to predict the coefficients and generate the difference image based on the linear combination of the basis images and the coefficients.

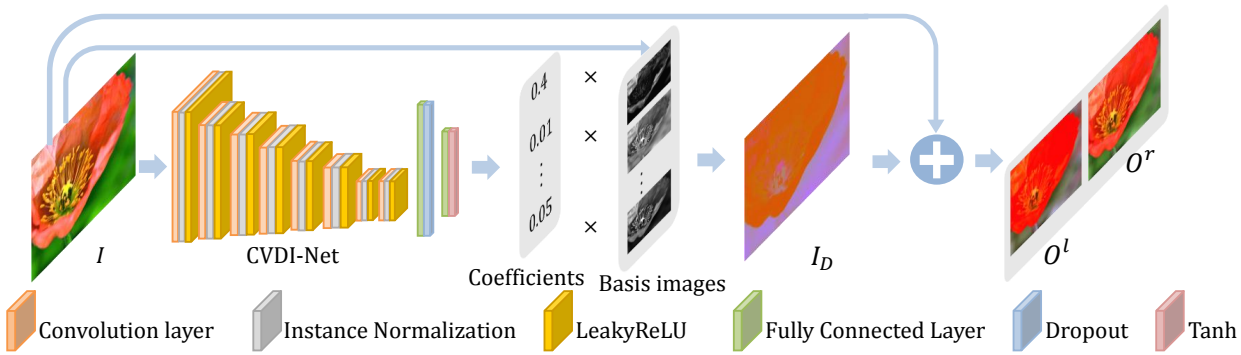

Fig. 3. System overview for still image pair generation. Given an image I, CVDI-Net generates a difference image D between the left and right output images O^l and O^r. Our key idea is to decompose the input image into a linear combination of k basis images and the corresponding k coefficients. Instead of directly generating the image pixels, we train our neural network to predict the k coefficients. In the network, we use 8 convolution layers. Except for the last convolution layer with stride 1, all convolution layers use 4×4 filters and stride 2. The first layer contains 32 feature maps. As the following layer being downsampled, the number of feature maps doubles. The number of neurons for the two fully connected layers are 2,048 and 27, respectively.

We utilize the second-order polynomial formulation [Johnson 1996] to decompose our input image. Given an input image I represented in the RGB color space containing red I_r, green I_g, and blue I_b layers, the second-order polynomial gives rise to 9 basis images altogether, including I_r, I_g, I_b, I_rI_g, I_bI_g, I_rI_b, I_r^2, I_g^2, and I_b^2. With these 9 basis images, we can reconstruct the D_r, D_g,D_b layers of the output difference image D individually as the weighted combination of these basis images. Therefore, this leads to $k = 9 \times 3 = 27$ coefficients in total to compute the difference image at one time instance. The sequence of 27 coefficients over time is going to be learnt and predicted by our network.

With the predicted difference image D, we can compute the output binocular image pair as $O^l = I - D/2$ and $O^r = I + D/2$. However, the color values in the constructed left and right images may exceed the color range $[0, 1]$. Simply truncating the values to [0,1] with $O^{\{l,r\}} = \max(0, \min(1, O^{\{l,r\}}))$ may violate the color preservation constraint $O^l + O^r = 2I$. Therefore, we propose to truncate the color values in the difference image D, such that the constructed images do not exceed the color range [0,1]. Mathematically, for each pixel p in the input image, the maximal value of the difference image for this pixel p is

$$|M(p)| = 2 \cdot \min\left(I(p), 1 - I(p)\right). \tag{1}$$

So we truncate the difference image based on the calculated maximal values for each pixel as

$$D^* = \max(\min(D(p), M(p)), -M(p). \tag{2}$$

Finally, the output image pair can be constructed as $O^l = I - D^*/2$ and $O^r = I + D^*/2$.

5 STILL IMAGE PAIR GENERATION

5.1 Network Design

We start by introducing the CVDI-Net. Given a single color-confusing image, our CVDI-Net generates a colorblind-sharable still image pair without considering the temporal coherence. Our network takes an image as input and outputs 27 floating number coefficients. Inspired by the image-to-parameter network [Hu et al. 2018], our network consists of 7 convolutional layers and 2 fully connected layers (Figure 3). A downscaling factor of 2 is applied to every consecutive convolutional layers except for the last one. While the resolution of the input image has to be resampled to 256×256 before feeding to CVDI-Net, our system can actually handle input images of any resolution because we only predict the 27 coefficients to generate the output instead of generating raw pixels and the synthesis works on the original resolution. Hence, there should be no detail loss in the final output image pair.

We design the loss function with the color preservation term, the color distinguishability term, and the binocular fusibility term to guide the training. The mathematical form of these three terms are similar to [Shen et al. 2016], but is slightly modified to meet our constraints. The color preservation term maximizes the dissimilarity between the input image and the output image. It is defined as:

$$\mathcal{L}_p = |(O^l + O^r) - 2I|. \tag{3}$$

The color distinguishability term maximizes the structural similarity between the input image and the CVD-simulated output images. It is defined as:

$$\mathcal{L}_d = -\frac{1}{N}\sum^{N}(S(T \cdot O^l(p), I(p)) + S(T \cdot O^r(p), I(p))), \tag{4}$$

where N is the number of pixels in the input image, p denotes a specific pixel, T is the projection matrix from normal-vision color space to CVD-vision color space. Since there are difference type and severities of CVD individuals, the matrix varies among CVD individuals. During all our experiments, we adopt the matrix proposed by [Machado et al. 2009] for the most severe type of deuteranopia. S is a function to compute the local structural similarity, which is defined as:

$$S(x, y) = \frac{\frac{2}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) + \epsilon}{\sigma_x^2 + \sigma_y^2 + \epsilon}, \tag{5}$$

where σ_x and μ_x are the standard deviation and the mean of the local window centered at x, respectively. ϵ is a very small value to avoid division-by-zero. The binocular fusibility term is to ensure the left and right images can be stably fused into a single percept by CVD audiences. Different from [Shen et al. 2016] that utilizes the binocular fusibility measurement proposed by [Yang et al. 2012], we adopt the formulation from [Zhang et al. 2019], and define it as:

$$\begin{aligned}\mathcal{L}_f = &\frac{1}{N}\sum_{N}(\max(\alpha G(T \cdot O^l(p)) - G(T \cdot O^r(p)), 0) \\ &+ \max(\alpha G(T \cdot O^r(p)) - G(T \cdot O^l(p)), 0)),\end{aligned} \tag{6}$$

where

$$G(p) = \sqrt{G_x^2(p) + G_y^2(p)}, \tag{7}$$

and $G_x(p) = A \circledast p$, $G_y(p) = A' \circledast p$. Here, $\circledast$ is the convolution operator on a 3×3 window centered at p, A is the Scharr gradient operator [Jähne et al. 1999]:

$$A = \frac{1}{16}\begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix}. \quad (8)$$

α is the parameter for fusibility enhancement and is set to 0.6 in our experiments.

Finally, we have our loss function:

$$\mathcal{L}_{\text{img}} = \mathcal{L}_p + \lambda_d \mathcal{L}_d + \lambda_f \mathcal{L}_f \quad (9)$$

During the training, λ_d and λ_f are set to 100 and 1000, respectively.

5.2 Training

To prepare the training data, a large number of input images and corresponding output difference image pairs are needed. Since our system is tailored to improve the color distinguishability for CVD audiences, the training data should at least contain a certain amount of CVD-confusing colors. In particular, we collect 2 million input images, where half of them are collected from the Flickr1M dataset [Huiskes and Lew 2008] and the other half are collected from the Places365 dataset [Zhou et al. 2018]. To determine whether an input image contains confusing colors or not, we compute the color contrast preservation ratio (CCPR) [Lu et al. 2012] between the CVD-simulated input image and the original input image. Here, the CVD-simulated image can be obtained by applying a CVD projection matrix [Machado et al. 2009] on the original image. The CCPR metric ranges from 0 to 1. Higher value indicates better color contrast preservation in the CVD-simulated image. In particular, we set a CCPR threshold of 0.7 to select the CVD-indistinguishable images. The final dataset contains 38,650 CVD-indistinguishable images.

Nevertheless, the collected real images may only contain a limited combination of confusing colors. To enrich the coverage of confusing colors, we further synthesize 200,000 images to cover more combinations of confusing colors. We first compute a set of confusing colors by identifying whether they are confusing to CVD people but distinguishable to normal-vision people. In particular, we measure the L_2 difference between the two colors in both the normal-vision and CVD-simulated domains. If the color difference is large enough in the normal-vision domain but quite small in CVD simulation, we define them as confusing colors. We search through the RGB color spaces and identify over 900k color pairs according to the previous color difference estimation. Then we synthesize the color confusing images by randomly scattering 2-10 color pairs on the canvas and then fill the whole image with multiple interpolation algorithms, including nearest neighbor, bilinear, bicubic and quadratic (Figure 5).

The network can then be trained in an unsupervised manner by optimizing the $\mathcal{L}_{\text{img}}$ loss. However, the training takes a very long time to convergence due to the large solution space. To accelerate the training, we initialize the training with the pre-computed image pairs to guide the training. Specifically, for each input, we generate the colorblind-sharable image pairs with the existing image-based visual-sharing solution [Shen et al. 2016]. Then we compute the signed per-pixel difference to produce the guided difference image.

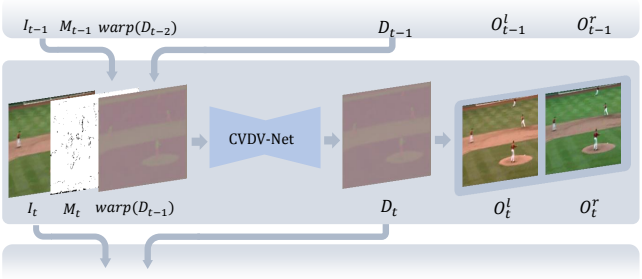

Fig. 4. System overview for video pair generation. © UCF101 is copyright owner of the input frame.

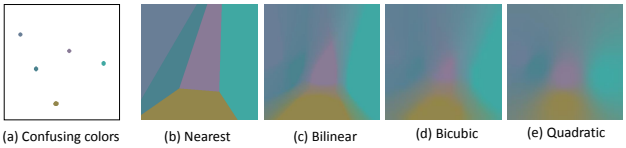

Fig. 5. Different interpolants of synthetic color-confusing images.

During the first 30 epochs of training, we minimize the dissimilarity between the output difference and the guided difference images. So the temporary loss function for the first 30 epochs is defined as the sum of L_1 distance and the SSIM [Wang et al. 2004], $\mathcal{L}_{\text{init}} = \mathcal{L}_{L_1} + \mathcal{L}_{\text{SSIM}}$. After 30 epochs, we discard the temporary loss function $\mathcal{L}_{\text{init}}$ and optimize $\mathcal{L}_{\text{img}}$. With this approach, our training then converges in 150 epochs.

6 VIDEO PAIR GENERATION

6.1 Network Design

The video pair generator CVDV-Net is based on the still image counterpart. It has a similar network architecture as the CVDI-Net except that the CVDV-Net additionally takes the warped previous difference image D_{t-1} and the previous confidence map M_{t-1} as input to account for the temporal coherence (Figure 4). With this feeding, our CVDV-Net is able to take the previous frame into account, so that the predicted coefficients, for constructing the pair of current frames, can be temporally coherent.

Given an input video sequence $\{I_t | t = 0, 1, ...\}$, we first compute the optical flow between the previous frame I_{t-1} and the current frame I_t with a pre-trained FlowNet2 [Ilg et al. 2017] model to warp across adjacent frames. Therefore, we can warp the previous difference frame D_{t-1} to obtain its warped frame $\text{warp}(D_{t-1})$ as an additional input to the CVDV-Net. Here, $\text{warp}(\cdot)$ denotes the warping operator. For the first frame I_0, we set the values of the warped previous difference frame and confidence map to 0 since there is no temporal information that can be utilized.

However, when disocclusion happens, the disoccluded pixels cannot be predicted from the previous frame. In that case, we directly predict the disoccluded pixels without relying on the temporal information. In other words, we directly used the the result of the CVDI-Net. To identify the disoccluded pixels, we compute a confidence map M_t based on the warped previous frame $\text{warp}(I_t)$:

$$M_t(p) = \begin{cases} 1, & \text{if } \|I_t(p) - \text{warp}(I_{t-1})(p)\| \le 0.02, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Here, $\|.\|$ denotes the L2 norm of the RGB color channels. Intuitively, we identify pixels as occluded where the warped previous frame are

dissimilar to the current frame. If the pixel dissimilarity exceeds a certain threshold (0.02, as in Eq. 10), we mark it as unreliable, and vice versa. This confidence map M_t is also fed to the CVDV-Net as input.

Hence, loss function consists of three losses, which are the temporal loss, the disocclusion loss, and the image loss. The temporal loss ensures that the *non-disoccluded* pixels in the current frame must be temporally coherent to the previous frame, and is defined as:

$$\mathcal{L}_t = |M_t \odot (D_t - \text{warp}(D_{t-1})|. \tag{11}$$

where D_t and D_{t-1} are the current and previous difference images, respectively; $\odot$ denotes the per-pixel multiplication operator; and $|.|$ denotes the L_1 norm. The disocclusion loss regularizes that the *disoccluded* pixels in the current frame is similar to the results generated by the CVDI-Net, and is defined as:

$$\mathcal{L}_s = |(1 - M_t) \odot (D_t - D_t^*)|. \tag{12}$$

where D_t^* is the difference image generated by our still image generator CVDI-Net. This also implies the CVDI-Net must be trained before the CVDV-Net. The overall loss function is defined as the weighted sum of the above two losses and the image loss $\mathcal{L}_{\text{img}}$:

$$\mathcal{L}_{\text{vid}} = \mathcal{L}_{\text{img}} + \lambda_t \mathcal{L}_t + \lambda_s \mathcal{L}_s \tag{13}$$

λ_t and λ_s are empirically set to 150 and 5000 in all our experiments, respectively.

6.2 Training

Unlike the training data of CVDI-Net which only contain still images, the CVDV-Net requires video samples as the training data to learn the temporal coherence. We collect 15 videos that contain CVD-confusing color regions from Youtube, UCF101 dataset [Soomro et al. 2012], and e-VDS dataset [Culurciello and Canziani 2017]. For each video, we randomly select 57 consecutive frames as the training input. No groundtruth output video is needed for training CVDV-Net.

Both CVDI-Net and CVDV-Net are initialized with the Xavier method [Glorot and Bengio 2010], and optimized using the Adam optimizer [Kinga and Adam 2015] with $(\beta_1, \beta_2) = (1e{-}3, 0.9)$. The learning rate starts at 0.0002 and is decayed exponentially. All our training and testing are performed in an NVIDIA TITAN 1070 Ti GPU.

As mentioned above, CVDI-Net must be trained before training CVDV-Net, as we need the output of CVDI-Net to compute the disocclusion loss for CVDV-Net. CVDI-Net is trained with a batch size of 16. CVDV-Net is trained for 100k iterations with a batch size of 1, where each iteration evaluates 6 consecutive frames. The training takes about 2 days for CVDI-Net and 36 hours for CVDV-Net to converge. All images and video frames are resized to 256×256 before feeding to the networks. During the training, the output image/video pair construction is performed with 256×256 resolution, while the original resolution is used during testing.

7 RESULTS AND DISCUSSION

To validate the effectiveness of our method, we tested it on several videos of different genres. The testing videos are collected from the UCF101 dataset. We compared our method to the existing methods in terms of visual quality, quantitative statistics, and user study. To compare the still image (synthesis) quality, we compare our method to the state of art visually-shareable still image synthesis method [Shen et al. 2016]. For video comparison, we compare our method to the state of art video-to-video translation method [Wang et al. 2018], and the recent neural-based blind filter approach [Lai et al. 2018].

While our model is learned in an unsupervised manner, Vid2vid requires corresponding image pairs to supervise the training. So, we trained Vid2vid with the same input video as our CVDV-Net. Their output video pairs for their training are generated by Shen method. For preparing the result of the CNN-based blind filtering, we first recolor the input videos using Shen method and then applied the CNN-based blind filtering on the left and right video sequences independently. Due to the page limit, we can only show part of the results and comparison in the paper. Comprehensive video results can be found in the supplementary materials.

7.1 Visual Comparison

We first evaluate the methods in terms of temporal coherence for both CVD and normal-vision audiences. To visualize the temporal coherence for CVD, we pick two frames for comparison. Figure 6 shows the right views of the 56th and 81th frames from all resultant videos and visualizes them by CVD simulation. Both Shen method and our still-image CVDI-Net fail to stabilize the banner color even for a short duration of less than 1.5 seconds. Although Vid2vid and CNN-based blind filtering alleviate the temporal incoherence, the color jitters across frames are still observable when one has a closer look. The blowup area tracks the same banner region over time, and clearly shows the change of color. On the other hand, our CVDV-Net successfully preserves the temporal coherence and makes the banner color stablized throughout the sequence.

To evaluate the temporal coherence for normal-vision audiences, the linear blending of the left and right videos are considered. Since all visual-sharing solutions have constrained on the color preservation, the temporal coherence of all methods are satisfactory and similar. However, since Shen method requires a initialization, the result is also dependent on the random seed, and thus cannot be stable across the temporal domain.

Vid2vid and the CNN-based blind filtering may sometimes introduce spatially inconsistent colors for CVD audiences. The right columns of Figures 7(b)&(c) present the CVD-simulation of their right frames. We can observe the halo artifact on the white wall due to the lack of global color consistency constraint in the Vid2vid and the blind filtering network. So same-color pixels might be translated to substantially different colors. Furthermore, these two methods may also fail to preserve color for normal-vision audiences, as demonstrated in the left subfigures of Figures 7(b)&(c), since no binocular constraint is imposed during the network training. The color inconsistency over time is more significant, as we can observe an obvious movement of the halos and their boundaries. In contrast, our method is free from those halos in both spatial and temporal domain. And moreover, our methods can achieve temporal coherence and global color consistency for both normal-vision and CVD simulation (Figure 7(d)), thank to the indirect coefficient formulation.

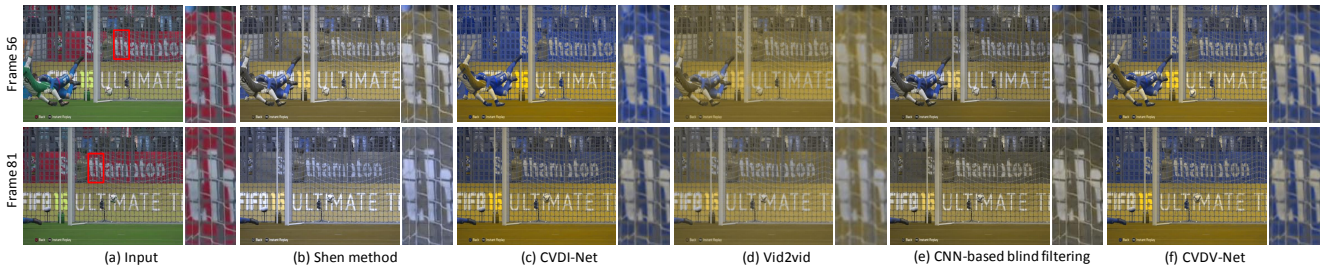

Fig. 6. Visual comparison on temporal coherence. (a) Input frames 56 & 81. (b)-(f) CVD simulation of right views of competitors' results. The name of the competitor method is labeled underneath each figure. To better visualize the color drift, we track the same region over time, and blow up for up-down comparison.

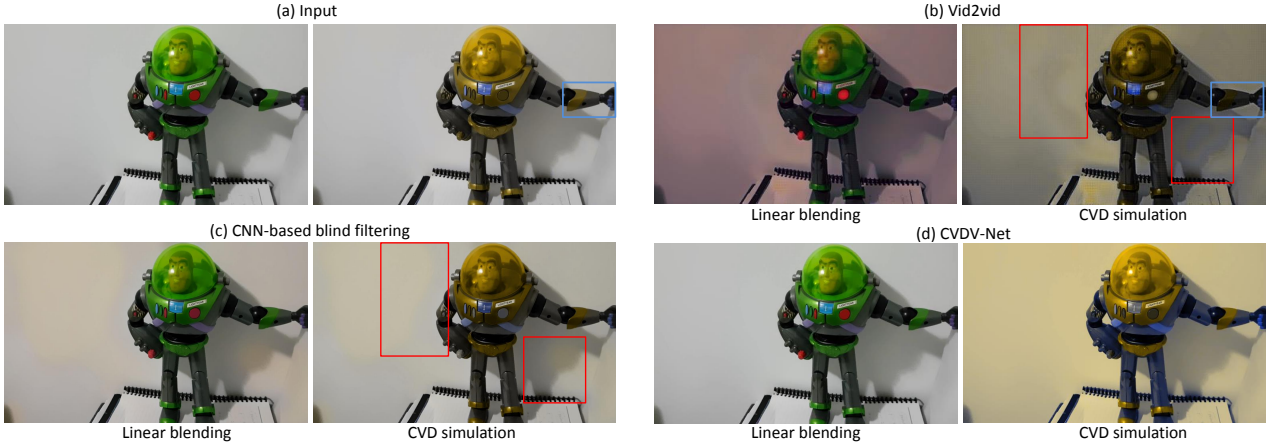

Fig. 7. Visual comparison on global color consistency. There are four pairs (a)-(d) in the figure. In each pair, the left one is in the normal vision and the right one is the CVD simulation. (a) Input (© e-VDS 2017). (b)-(d): the name of the competitor method is labeled on top of each figure pair.

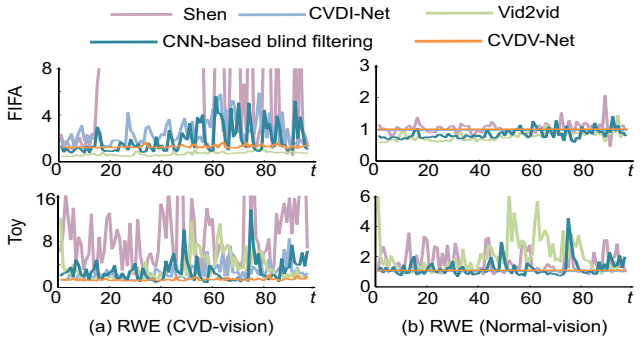

Fig. 8. Temporal coherence. (a) Relative warping error (RWE) between CVD-simulated right frames and CVD-simulated input. (b) RWE between linearly blended frames and input.

7.2 Quantitative Evaluation

We then perform quantitative evaluations in terms of the temporal consistency, the color preservation, the color distinguishability, and the binocular fusibility. Two video sequences, "FIFA" and "Toy," presented in Figures 6 and 7, respectively, are analyzed in our experiments.

Temporal Coherence. The temporal coherence for CVD and normal-vision audiences have to be evaluated with different metrics, due to their different visual experiences. For CVD audiences, we propose the relative warping error (RWE), which calculates the ratio between the temporal frame change δ of the output video O and that of the input video I, and is defined as

$$\mathrm{RWE}_t = \frac{|\delta(O_t^{\mathrm{CVD}})|}{|\delta(I_t^{\mathrm{CVD}})| + \epsilon}, \tag{14}$$

where

$$\delta(x_t) = |\mathrm{warp}(x_{t-1}) - x_t|, \tag{15}$$

x^{CVD} is the CVD simulation of the image x, and ϵ is a very small value to avoid division-by-zero. If we plot the RWE against the time and see a straight horizontal line situates close to 1, it means the video is highly temporal coherent. On the other hand, if there exists deviation or fluctuation from the straight line, the video is less temporally coherent. Figure 8(a) plot the RWE curves of all compared methods (right frames of all videos). Among all compared methods, our CVDV-Net gives the best temporal coherence as it maintains the straightest lines around 1. Vid2vid is the first runner-up. Though the CNN-based blind filter is designed for removing temporal incoherence, its performance is worse than Vid2vid (see "FIFA example"). Shen method and our CVDI-Net are inferior due to their per-frame computation nature. Shen method performs the worst due to its random optimization initializations.

The temporal consistency for normal-vision audiences is evaluated by plotting a similar RWE curves, but for the linearly blended frames (Figure 8(b)). We can see that our CVDI-Net and CVDV-Net achieves the best temporal coherence (straight lines around 1), because we have a hard constraint that our linearly blended output videos are always equal to the input video. Shen method, Vid2vid, and blind filter are temporal incoherent. Among them, blind filter is better (see "Toy" example).

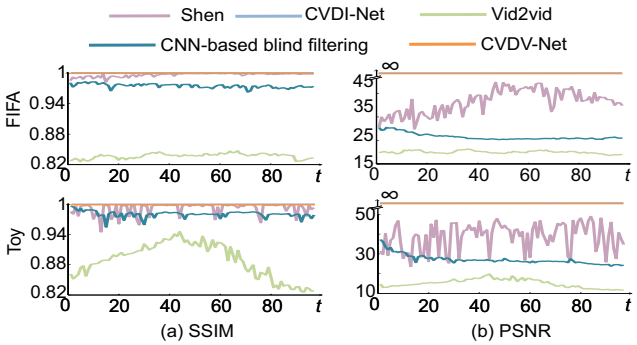

Fig. 9. Color preservation. (a) SSIM between linearly blended frames and input (b) PSNR between linearly blended frames and input.

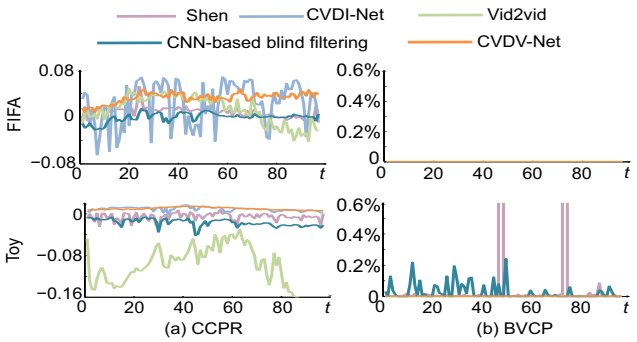

Fig. 10. Color distinguishability and binocular fusability. (a) CCPR difference between the output and input frames (both in CVD simulation) (b) Percentage of infusible pixels.

Color Preservation. We measure the color preservation for normal-vision audiences by evaluating the SSIM and PSNR between the linearly blended frames and the corresponding input frames. Figures 9(a) and (b) plot the results over time. The higher the SSIM and PSNR scores are, the more similar the two frames are, and the normal-vision audiences are less likely to be aware of any color changes. Note that the maximal score of SSIM is 1, while the maximal score of PSNR is positive infinite. The SSIM and PSNR of our CVDI-Net and CVDV-Net achieve maximal scores. Shen method, Vid2vid, and blind filter introduce certain degree of color deviation for normal-vision audiences. We can observe a great color deviation on the Vid2vid results since there is no constraint on color preservation during the network training. Blind filter introduces less color deviation than Vid2vid due to the constraints by Shen results, but it is still not visually plausible as ours.

Color Distinguishability. For evaluating the color distinguishability for CVD audiences, we calculate the CCPR [Lu et al. 2012] difference between the output and input frames (both in CVD simulation), and plot it in Figure 10(a). Again, we only show the CCPR of the right videos. Positive CCPR differences indicates that the color distinguishability has been improved, and vice versa. Higher CCPR score represents a higher improvement in color distinguishability. The larger the CCPR score is, the larger the improvement is. In contrast, negative CCPR means that the color distinguishability is reduced. Among all, Our CVDV-Net achieves better scores than Shen method and blind filter. Vid2vid generally output bad results, even worse than the CVD simulation of the input (negative values of "Toy" example) because it fails to preserve the color contrast that exists in the original CVD simulation (e.g., the yellow and gray color regions inside the blue box in Figure 7).

Binocular Fusibility. The binocular fusibility for CVD audiences is measured using the binocular vision fusibility predictor (BVCP) [Yang et al. 2012] and we calculate the percentage of pixels that are infusible. Figure 10(b) plots the results. In the upper plot of Figure 10(b), all four methods generate complete fusible results. Shen method sometimes generates infusible regions due to its unstable optimization as shown in the lower plot of Figure 10(b). CNN-based blind filtering inherits the unstable problem of Shen method.

Timing Statistics. We tested CVDV-Net and Vid2vid methods on an nVidia TITAN 1070 Ti GPU, and Shen method on a PC with Intel Xeon 3.7GHz CPU and 32GB memory. On average, our CVDV-Net method takes less than 0.08s to process a 512×512 image frame, which is much faster than 62.2s of the Shen method (CPU implementation only) and 0.5s of Vid2vid.

7.3 User Study

We also conducted a user study to evaluate our method from the user perspective. We invited 10 normal-vision individuals and 8 CVD individuals (7 deuteranomalous males and 1 deuteranomalous female) in the study. For each CVD participant, we first determine their CVD type via the Ishihara test, for presenting results that tailored for each participant. Six test videos are used in our user study. Each of them contains confusing colors that are distinguishable to normal-vision audiences but indistinguishable to CVD audiences. During the study, videos are presented on a ASUS G750JX laptop with 3D display, under the ambient illuminance around 200 lux. The distance between the eyes of the participants and the screen is around half a meter. CVD participants have to wear NVIDIA P1431 stereoscopic glasses during the experiment, while normal-vision audiences wear nothing. In this user study, we also compare our method to Shen method [Shen et al. 2016], CVDI-Net, Vid2vid, and CNN-based blind filtering [Lai et al. 2018].

Temporal Coherence. We first study the temporal coherence for both CVD and normal-vision audiences. Firstly, we evaluate how smooth the videos are as viewed by normal-vision audiences. Videos generated by different methods are randomly chosen for playing to the participants. Then participants are asked to rate how smooth each video is in the scale of [0, 5], with 5 indicating the smoothest. Figure 11(a) plots the mean (color bar) and the 95% confidence interval (indicated as black vertical ranges). All methods that consider temporal coherence obtain higher scores, while Shen method receives the lowest score due to the occasional visual flickering. Though CVDI-Net does not account for the temporal coherence explicitly, it receives rather higher score than Shen method. This is because our difference-image formulation preserves the color for normal-vision audiences all the time. One-way analysis of variance (ANOVA) with a significant level of 0.05 also shows that CVDV-Net is statistically better than still-image solutions, blind filtering, and Vid2vid method.

The user study of temporal coherence for CVD audiences is conducted similarly, except that CVD participants have to wear the

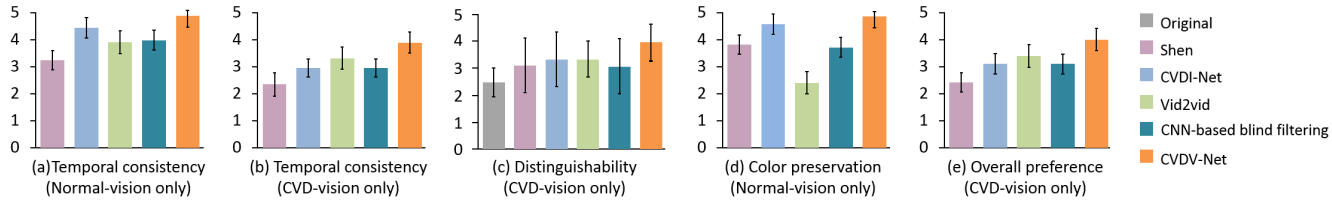

Fig. 11. Statistics for user study. (a)Temporal coherence (Normal-vision only) (b)Temporal coherence (CVD-vision only) (c) Distinguishability (CVD-vision only) (d) Color preservation (Normal-vision only) (e) Overall preference (CVD-vision only)

stereoscopic glasses. Figure 11(b) plots the statistics. CVDV-Net receives the highest scores among all methods. Though Vid2vid generates temporally coherent results for some of the cases, the observeable changes of color inconsistent regions lower its score. According to ANOVA, the outperformance of CVDV-Net is statistically significant.

Color Distinguishability for CVD audiences. We then evaluate how well each method improves the color distinguishability of CVD audiences. To do so, we first identify two specific objects or regions in the video that are confusing to CVD audiences. Then we ask the participants whether they can differentiate the colors of these two regions. It has been found that even if CVD audiences can differentiate the colors, they still have different degrees of certainty. Instead of asking the participants to answer yes or no, we ask them to rate the certainty level of distinguishability in the scale of [0, 5], with 0 being indistinguishable, and 5 means certainly distinguishable. To avoid bias, we randomly introduce pairs of same-color regions into the test set. Figure 11(c) plots the statistics. The input video receives the lowest score due to the indistinguishable color pairs. Though still-image methods can make these pairs distinguishable in each frame, it is a different story when the whole video is played back. The serious flickering obscures the users vision and, hence, the still-image methods receive rather low score and large deviation in score. The situation is similar for blind filtering since it cannot completely fix the flickering problem inherited from the still image methods. Though Vid2vid generates temporally smooth videos, it attenuates the color contrast in certain regions where the colors are originally distinguishable, which lowers their score of distinguishability. According to ANOVA, CVDV-Net method is significantly better than the input. There is no statistical difference between all the recoloring methods.

Color Preservation for Normal Vision Audiences. We further evaluate how well the resultant videos in preserving colors for normal-vision audiences. No CVD audience takes part in this user study. In each round, the input video and one of the four resultant videos are shown to the participants side-by-side on the same screen. Participants are asked to rate how each resultant video is similar to the input, in the scale of [0, 5], with 5 means the most similar and 0 means not similar at all. Figure 11(d) plots the statistics. Our CVDI-Net and CVDV-Net are of the best rating because of our hard constraint on color preservation and difference-image formulation. Vid2vid method is the worst as it lacks of binocular constraints during the training. Blind filtering is worse than Shen method as there is no contraint on color preservation expect for the first frame. According to ANOVA, our two methods statistically outperform Vid2vid, Shen method and blind filtering.

Overall Preference for CVD audiences. Lastly, we are also interested in the overall preference of methods for CVD audiences. Before rating, we brief the participants on the video content and the colors used so that they have a better understanding of the video. Then each input video and all its resultant videos are placed on the same screen for better comparison. We ask CVD participants to rate the result video in the scale of [0, 5], with 5 being the most preferred. The statistics are plotted in Figure 11(e). Note that temporal consistency usually plays an important role in user visual experience. Our CVDV-Net is the most preferred by CVD audiences, and is statistically better than other solutions since we explicitly take the temporal coherence and the binocular constraints into account.

7.4 Comparison between CVDI-Net and Existing Still-image Method

While we have evaluated how our CVDV-Net outperforms the existing methods towards generating colorblind-shareable videos, the still-image component plays an important role to achieve this desirable performance. The reason we choose our own still-image network model CVDI-Net, instead of the state-of-the-art Shen method as our still-image component, is because of the unstable optimization of Shen method. Although Hu method [Hu et al. 2019] solves the instability, their performance is limited by the performance of Shen method. Furthermore, their results suffer from color inconsistency. To evaluate how our CVDI-Net outperforms the existing methods, we conduct both qualitative and quantitative evaluations again. But in this phase the test data is still images, instead of videos. Figure 12 shows an example in which Shen method fails to fulfill the color preservation and binocular fusibility requirements due to its unstable optimization. The linear blending of the output image pair and the binocular fusibility map are shown in the figure. The binocular fusibility map is computed via BVCP, where pixels marked in green are binocularly infusible. For some of the cases, Hu method introduces certain portion of infusible regions, while all pixels in our CVDI-Net result are fusible (Figure 13).

To compare them quantitatively, we collect 1,000 natural images, 5 colored charts, and 10 Ishihara test images as the testing images. Similar to our previous quantitative evaluation above, we also measure the color preservation (SSIM and PSNR), color distinguishability (CCPR), and binocular fusibility (BVCP) for each test image, and the average scores are presented in Table 2. Note that this time we are evaluating on still images, hence we only show the average scores instead of time-varying plots. From the table, our method and Hu method perform comparably well in terms of color distinguishability, and both outperform Shen method in terms of all evaluation metrics. In terms of binocular fusibility (BVCP), our method receives a better score than Hu method. Though it seems that the improvement is

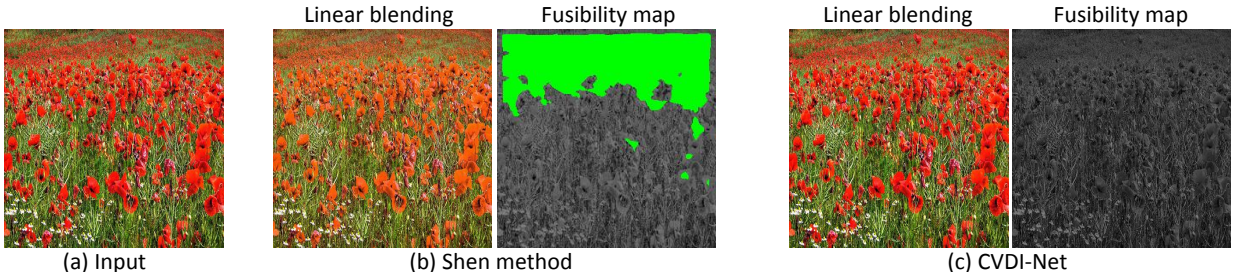

Fig. 12. CVDI-Net vs. Shen method. (a) Input image (© Flickr1M 2018). (b) Shen method. (c) CVDI-Net method.

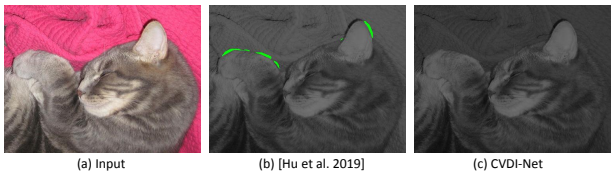

Fig. 13. CVDI-Net vs. [Hu et al. 2019] method. (a) Input image (© Flickr1M 2018). (b) Fusibility map of Hu method. (c) Fusibility map of CVDI-Net method.

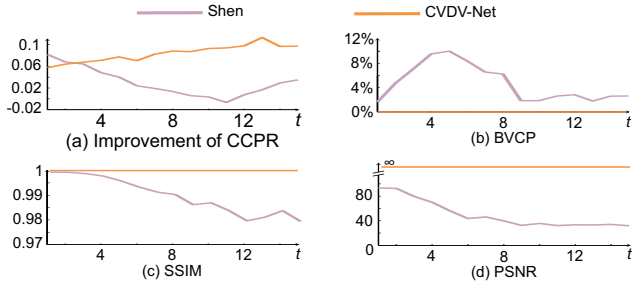

Fig. 14. Quantitative evaluation on different number of confusing color pairs. (a) Improvement of CCPR. (b) Percentage of infusible pixels. (c) SSIM between linearly blended frames and input. (d) PSNR between linearly blended frames and the input.

small (from 0.04% to from 0.01%), the percentage is computed by averaging across the test set. According to our invited CVD participants, the perceived visual contents of our results are much more stable for those cases like Figure 13. The superiority also helps to improve the visual experience for the CVD viewers. All these justify our employment of CVDI-Net as our still-image component.

We further evaluate the performance of CVDI-Net and Shen method when the amount of confusing color pairs (CFP) increases. To do so, we synthesize testing images of 15 configurations. Images of configuration 1 contain 1 CFP, and images in configuration 2 contain 2 CFPs and so on. Each configuration contains 100 images. So, we have 1,500 testing images in total. Figure 14 plots the four metrics, CCPR, BVCP, SSIM, and PSNR. In each plot, we plot the metric scores against the CFP count. In general, our CVDI-Net significantly outperforms Shen method in all metrics. For BVCP, SSIM, and PSNR, our CVDI-Net is almost unaffected by the CFP count, while Shen method exhibits instability (BVCP) or consistent degradation (SSIM and PSNR) as the CFP count increases. For CCPR, Shen method also degrades as the CFP count increases. In contrast, our CVDI-Net gradually improves in CCPR (color distinguishability) as the CFP count increases. We believe that the training with synthetic data further improves the performance of our CVDI-Net .

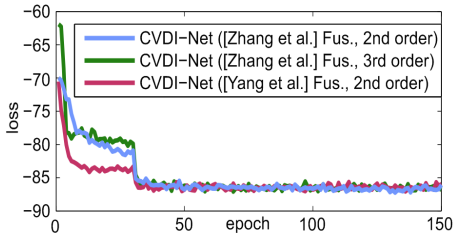

Fig. 15. Loss curve of different choices of order and fusibility term.

7.5 Ablation Study

The Choice of Polynomial Order and Color Space. To determine the order of our polynomial formulation, we conduct comparisons among the first, the second, and the third order polynomials, respectively. The statistics is shown in Table 2. We find that the first order formulation performs less satisfactorily on enhancing the color distinguishability and the binocular fusibility due to the limited solution space. Theoretically, higher order representations have larger solution spaces and can always improve the results. However, the second order outperforms the third order formulation in terms of both CCPR and BVCP. One possible reason is that higher order representations have more coefficients to train and increases the difficulty of convergence during training. According to Figure 15, the loss curve for training CVDI-Net with the third order (the green curve) is choppier than that for the second order one (the blue curve), especially during the initialization stage. Note that there is a drop at around the 30th epoch for both curves. The reason is that some of the training pairs generated by Shen method fail to preserve the binocular fusibility, so the fusibility loss remains high in the first 30 epochs. After the 30th epoch, the network starts to optimize for the fusibility term, so the loss value drops quickly. In the above experiments, input images are decomposed in the RGB color space. According to Table 4, there is no significant difference between using Lab and RGB color spaces in terms of color distinguishability and binocular fusibility.

Influence of Resizing. As all images and video frames are resized to 256×256 before feeding to the networks, while the learned coefficients are applied in the input images/frames in the original resolution for construction. As image resizing may remove high-frequency details in the original input, we conduct experiment to evaluate how the resizing affects the quality of CVDI-Net output. To eliminate the effect of resizing, we resize the 1,015 testing images to 256×256, and regard them as the testing input. We then apply the same predicted coefficients to them to obtain the 256×256 output image pairs. Table 1 shows the quality statistics. Comparing Tables 2 and 1, our CVDI-Net gives a similar (CCPR) or equal scores

Table 1. Quantitative comparison on input images (256 × 256 resolution) without resizing.

	CCPR	BVCP	SSIM	PSNR
Input (256 × 256)	0.547	N/A	N/A	N/A
Shen method (256 × 256)	0.672	3.1%	0.97	59.3
CVDI-Net (256 × 256)	0.688	0.01%	1	∞

Table 2. Statistical results of Shen method, Hu method, and CVDI-Net with different choice of order, color space and training data.

Method	CCPR	BVCP	SSIM	PSNR
Input	0.576	N/A	N/A	N/A
Shen method	0.687	3.0%	0.98	63.7
Hu method	0.704	0.04%	1	∞
CVDI-Net (first order)	0.635	1.3%	1	∞
CVDI-Net (second order)	0.701	0.01%	1	∞
CVDI-Net (third order)	0.691	0.2%	1	∞
CVDI-Net w/o syn	0.694	0.02%	1	∞
CVDI-Net w/o nat	0.640	0.003%	1	∞

Table 3. Quantitative comparisons between Yang formulation and Zhang formulation of binocular fusibility.

	CCPR	BVCP
[Zhang et al. 2019] Fus.	0.701	0.01%
[Yang et al. 2012] Fus.	0.672	0.09%

(BVCP, SSIM, and PSNR). It still outperforms the Shen method by similar magnitude. In general, there is no degradation caused by the resizing.

Natural and Synthetic Training Images. We trained our network with both natural and synthetic images. Table 2 compares the performances when our network is trained with different training sets. We can see that, taking away synthetic or natural images, both decrease the color distinguishability (lower CCPR). Training without natural images may improve the fusibility (lower BVCP), but with a significant drop of color distinguishability. Therefore, training with both natural and synthetic images is justifiable.

Different Formulations of the Binocular Fusibility Term. We have evaluated two formulations of binocular fusibility, Yang formulation [Yang et al. 2012] (the original BVCP formulation) and Zhang formulation [Zhang et al. 2019]. Figure 15 plots the training loss of both formulations (the purple and the blue curves). Note that, the two curves are not directly comparable, since the two losses are computed differently, one with Yang formulation and the other with Zhang formulation. To compare them fairly, we evaluate their CCPR and BVCP. It turns out that Zhang formulation achieves better results as shown in Table 3. Even BVCP is actually the Yang formulation, the network trained with Yang formulation cannot achieve a better evaluation score. This may be because Yang formulation is harder to train. Hence, we choose to adopt Zhang formulation in our work.

Supervised v.s. Unsupervised. While results of Shen method help our CVDI-Net network during training for early convergence, the network still fails to generate satisfactory results if the whole training process is supervised as shown in Table 4. The reason is that the quality of results of Shen method is unstable.

Table 4. Statistical results of CVDI-Net with different color spaces and training scheme.

	RGB, Unsup.	Lab, Unsup.	RGB, Sup.
CCPR	0.701	0.705	0.676
BVCP	0.01%	0.03%	2.3%

Table 5. Quantitative comparison between CVDV-Net and a direct optimization with the same loss function.

Method	RWE (CVD)	CCPR Improv.	BVCP
CVDV-Net	0.865	0.034	0.001%
Direct Opt.	0.971	0.032	0.524%

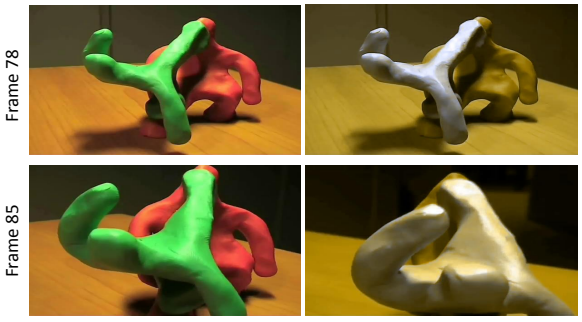

Fig. 16. Failure case due to large motion (a) Input frames. (b) CVD simulation of the right frame generated by the CVDV-Net.

Training v.s. Direct Optimization. The reason we propose a deep learning method instead of a standard optimization technique to generate videos is that the random initialization in direct optimization may produce results of instable quality due to the non-convexity of the loss function. With the guidance of the CVDI-Net, we can learn a desired mapping from the input videos to high-quality output videos. Here, we also conduct comparisons between a direct optimization and the CVDV-Net. Except for the first frame whose output is generated by the CVDI-Net, the direct optimization optimizes for the same loss function as we train the CVDV-Net to synthesize each frame. To reduce the impact of bad initialization, we initialize the polynomial coefficients of each frame with the output coefficients of the previous frame. Table 5 shows the quantitative comparisons on the 6 videos we used in the user study. Note that we do not show the metrics for normal-vision viewers since the color preservation is a hard constraint using the setting of the difference image. Our CVDV-Net outperforms the direct optimization in terms of all evaluated metrics.

7.6 Limitation

Due to the lack of CVD-confusing videos, our current training dataset only contains 15 videos, which limits the ability of our CVDV-Net to handle confusing color combinations that are not present in the dataset. Moreover, we use the optical flow generated by FlowNet2 to identify pixel correspondence. Thus, our method may fail to preserve the temporal coherence when the optical flow is mistakenly estimated, e.g. the large motion in Figure 16 leads to incorrect optical flow estimation, which in turn leads to the failure to maintain the color inconsistency. Currently, we assume the BVCP can be applied to measure the binocular fusibility of CVD audiences. However, BVCP is originally designed for normal-vision people, and it may not be accurate for CVD audiences.

8 CONCLUSION

In this paper, we propose a framework for generating temporal-coherent and visually shareable videos between CVD and normal-vision audiences. To achieve the color consistency, we propose to indirectly synthesize the video frames through generating the temporal-coherent polynomial coefficients, instead of directly generating the video frame pixels. This approach effectively maintains the color consistency as well as the temporal coherence. We utilize a convolutional neural network to generate the temporal-coherent polynomial coefficients, and achieve high quality in all metrics as well as high-speed performance.

Current network architecture only tackles a specific type and severity of CVD audiences. Different types and severities of CVD audiences need to be trained individually. In our future work, we can explore how to incorporate various types and severities into network training. Moreover, our current network still cannot achieve real-time performance, we are interested in further accelerating the speed for real-time colorblind-shareable video generation.

ACKNOWLEDGMENTS

This project is supported by the Research Grants Council of the Hong Kong Special Administrative Region, under RGC General Research Fund (Project No. CUHK 14201017), and Shenzhen Science and Technology Program (No. JCYJ20180507182410327 and JCYJ20180507182415428).

REFERENCES

2015. Spectral Edge for displays. https://www.spectraledge.co.uk/spectral-edge-for-displays7. Online; accessed 26-April-2018.

Tunç Ozan Aydin, Nikolce Stefanoski, Simone Croci, Markus Gross, and Aljoscha Smolic. 2014. Temporally Coherent Local Tone Mapping of HDR Video. *ACM Trans. Graph.* 33, 6, Article 196 (Nov. 2014), 13 pages. https://doi.org/10.1145/2661229.2661268

Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. 2013. Example-based Video Color Grading. *ACM Trans. Graph.* 32, 4, Article 39 (July 2013), 12 pages. https://doi.org/10.1145/2461912.2461939

Nicolas Bonneel, Kalyan Sunkavalli, James Tompkin, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2014. Interactive Intrinsic Video Editing. *ACM Trans. Graph.* 33, 6, Article 197 (Nov. 2014), 10 pages. https://doi.org/10.1145/2661229.2661253

Nicolas Bonneel, James Tompkin, Deqing Sun, Oliver Wang, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. 2017. Consistent video filtering for camera arrays. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 397–407.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2015. Blind Video Temporal Consistency. *ACM Trans. Graph.* 34, 6, Article 196 (Oct. 2015), 9 pages. https://doi.org/10.1145/2816795.2818107

Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. In *Proc. Intl. Conf. Computer Vision (ICCV)*.

Soon Hau Chua, Haimo Zhang, Muhammad Hammad, Shengdong Zhao, Sahil Goyal, and Karan Singh. 2015. ColorBless: Augmenting Visual Information for Colorblind People with Binocular Luster Effect. *ACM Trans. Comput.-Hum. Interact.* 21, 6, Article 32 (Jan. 2015), 20 pages. https://doi.org/10.1145/2687923

Eugenio Culurciello and Alfredo Canziani. 2017. e-Lab Video Data Set. https://engineering.purdue.edu/elab/eVDS/.

Qingnan Fan, Jiaolong Yang, David P. Wipf, Baoquan Chen, and Xin Tong. 2018. Image Smoothing via Unsupervised Learning. *CoRR* abs/1811.02804 (2018). arXiv:1811.02804 http://arxiv.org/abs/1811.02804

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

Xinghong Hu, Zhuming Zhang, Xueting Liu, and Tien-Tsin Wong. 2019. Deep Visual Sharing with Colorblind. *IEEE Transactions on Computational Imaging* (2019).

Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A White-Box Photo Post-Processing Framework. *ACM Transactions on Graphics (TOG)* 37, 2 (2018), 26.

Chun-Rong Huang, Kuo-Chuan Chiu, and Chu-Song Chen. 2011. Temporal color consistency-based video reproduction for dichromats. *IEEE Transactions on Multimedia* 13, 5 (2011), 950–960.

Jia-Bin Huang, Yu-Cheng Tseng, Se-In Wu, and Sheng-Jyh Wang. 2007. Information preserving color transformation for protanopia and deuteranopia. *Signal Processing Letters, IEEE* 14, 10 (2007), 711–714.

Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 39–43.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1647–1655.

Bernd Jähne, Horst Haussecker, and Peter Geissler. 1999. *Handbook of computer vision and applications*. Vol. 2. Citeseer.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. 2017. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *arXiv preprint arXiv:1712.00080* (2017).

Tony Johnson. 1996. Methods for characterizing colour scanners and digital cameras. *Displays* 16, 4 (1996), 183–191.

D Kinga and J Ba Adam. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, Vol. 5.

Naejin Kong, Peter V. Gehler, and Michael J. Black. 2014. Intrinsic Video. In *Computer Vision – ECCV 2014 (Lecture Notes in Computer Science)*, Vol. 8690. Springer International Publishing, 360–375.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 170–185.

Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. 2012. Practical Temporal Consistency for Image-based Graphics Applications. *ACM Trans. Graph.* 31, 4, Article 34 (July 2012), 8 pages. https://doi.org/10.1145/2185520.2185530

Cewu Lu, Li Xu, and Jiaya Jia. 2012. Contrast preserving decolorization. In *Computational Photography (ICCP), 2012 IEEE International Conference on*. IEEE, 1–7.

Gustavo M Machado and Manuel M Oliveira. 2010. Real-Time Temporal-Coherent Color Contrast Enhancement for Dichromats. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 933–942.

Gustavo M Machado, Manuel M Oliveira, and Leandro AF Fernandes. 2009. A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1291–1298.

Sylvain Paris. 2008. *Edge-Preserving Smoothing and Mean-Shift Segmentation of Video Streams*. Springer Berlin Heidelberg, Berlin, Heidelberg, 460–473. https://doi.org/10.1007/978-3-540-88688-4_34

Behzad Sajadi, Aditi Majumder, Manuel M Oliveira, Rosália G Schneider, and Ramesh Raskar. 2013. Using patterns to encode color information for dichromats. *IEEE transactions on visualization and computer graphics* 19, 1 (2013), 118–129.

Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. 2018. Frame-Recurrent Video Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6626–6634.

Wuyao Shen, Xiangyu Mao, Xinghong Hu, and Tien-Tsin Wong. 2016. Seamless Visual Sharing with Color Vision Deficiencies. *ACM Trans. Graph.* 35, 4, Article 70 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925878

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

Chung-Ming Wang, Yao-Hsien Huang, and Ming-Long Huang. 2006. An effective algorithm for image sequence color transfer. *Mathematical and Computer Modelling* 44, 7âĂŞ8 (2006), 608 – 627. https://doi.org/10.1016/j.mcm.2006.01.029

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601* (2018).

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Stephen Wolf. 2003. *Color correction matrix for digital still and video imaging systems*. National Telecommunications and Information Administration Washington, DC.

Xuan Yang, Linling Zhang, Tien-Tsin Wong, and Pheng-Ann Heng. 2012. Binocular Tone Mapping. *ACM Transactions on Graphics* 31, 4 (2012), 93:1–93:10.

Genzhi Ye, Elena Garces, Yebin Liu, Qionghai Dai, and Diego Gutierrez. 2014. Intrinsic Video and Applications. *ACM Trans. Graph.* 33, 4, Article 80 (July 2014), 11 pages. https://doi.org/10.1145/2601097.2601135

Zhuming Zhang, Chu Han, Shengfeng He, Xueting Liu, Haichao Zhu, Xinghong Hu, and Tien-Tsin Wong. 2019. Deep binocular tone mapping. *The Visual Computer* (2019), 1–15. https://link.springer.com/article/10.1007/s00371-019-01669-8

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2018), 1452–1464.