Perceptual Temporal Incoherence-Guided Stereo Video Retargeting

Bing Li, *Member, IEEE*, Chia-Wen Lin, *Fellow, IEEE*, Shan Liu, *Member, IEEE*, Tiejun Huang, *Senior Member, IEEE*, Wen Gao, *Fellow, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

***Abstract*— Stereo video retargeting aims at minimizing shape and depth distortions with temporal coherence in resizing a stereo video content to a desired size. Existing methods extend stereo image retargeting schemes to stereo video retargeting by adding additional temporal constraints that demand temporal coherence in all corresponding regions. However, such a straightforward extension incurs conflicts among multiple requirements (i.e., shape and depth preservation and their temporal coherence), thus failing to meet one or more of these requirements satisfactorily. To mitigate conflicts among depth, shape, and temporal constraints and avoid degrading temporal coherence perceptually, we relax temporal constraints for non-paired regions at frame boundaries, derive new temporal constraints to improve human viewing experience of a 3D scene, and propose an efficient grid-based implementation for stereo video retargeting. Experimental results demonstrate that our method achieves superior visual quality over existing methods.**

***Index Terms*— Stereo video retargeting, shape preservation, depth preservation, nonuniform warping, temporal coherence.**

I. Introduction

THE 3D videos become popular in our daily media consumption as they offer rich and joyful real-world viewing experience. Content-aware stereo image editing methods have been proposed such as depth remapping for stereo image [1]–[3], and stereo image retargeting [4]–[6]. Since the human visual system (HVS) is not sensitive to spatial distortions of different image contents [7] uniformly, content-aware image editing methods can adopt nonuniform spatial warping to preserve the shapes of important objects. Nevertheless, nonuniform spatial warping used in video retargeting often leads to temporal inconsistency such as jittering and flickering in retargeted video. To address this problem, state-of-the-art 2D video retargeting methods impose strong temporal coherence constraints which tend to resize all corresponding regions among frames consistently [8]–[11]. Such temporal constraints are called global temporal constraints. Although global temporal constraints are effective in maintaining temporal coherence of retargeted video, it is difficult to fulfill all temporal and shape constraints simultaneously, thus degrading shape preservation significantly.

In this work, we focus on perceptual-based stereo video retargeting. It is a more challenging problem than 2D video retargeting for several reasons. First, retargeting of the additional depth dimension has to be considered. Second, to preserve the object shape, depth and temporal coherence, we need to impose the shape, depth and temporal constraints on the nonuniform spatial warping scheme. Third, all constraints are resource-demanding, leading to resource contention. For example, shape constraints preserve the shapes of important objects at the cost of resizing unimportant regions. However, when these unimportant regions are constrained not to be resized by temporal/depth constraints, conflicts occur among these constraints, as will be demonstrated in the experiments.

A straightforward generalization of 2D video retargeting methods to stereo video often results in conflicts of multiple requirements. To address these challenges, we will show that the temporal coherence constraint can be relaxed to some extent without incurring visually noticeable artifacts. Specifically, visual contents that are more tolerant to temporal inconsistent resizing should be identified. We conduct user study on two stimuli types to investigate human perception on temporal incoherence against different stereo visual contents caused by temporally inconsistent resizing (see Fig.1 and Fig.2). Based on this new finding, we propose a perceptual temporal incoherence-aware stereo video retargeting scheme that achieves high retargeting performance by imposing effective temporal, shape and depth constraints.

As compared with our previous work in [12], our current paper has been significantly extended in several aspects. First, we quantitatively analyze the correlation between visual content and temporal incoherence caused by temporally inconsistent resizing. This analysis provides valuable insights into content-aware stereo video editing tools (such as depth editing

Manuscript received November 5, 2018; revised June 24, 2019 and October 30, 2019; accepted March 15, 2020. Date of publication April 7, 2020; date of current version April 23, 2020. This work was supported in part by Tencent America, and in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2634-F-007-013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ce Zhu. *(Corresponding author: Bing Li.)*

Bing Li is with the Visual Computing Center, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia (e-mail: bli417@usc.edu).

Chia-Wen Lin is with the Department of Electrical Engineering National Tsing Hua University, Hsinchu 30013, Taiwan, and also with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Shan Liu is with the Tencent Media Lab, Tencent, Polo Alto, CA 94301 USA (e-mail: shanliu@tencent.com).

Tiejun Huang and Wen Gao are with the School of Electronic and Computer Engineering, Peking University, Beijing 100871, China (e-mail: tjhuang@pku.edu.cn; wgao@pku.edu.cn).

C.-C. Jay Kuo is with the Department of Electrical Engineering, University of Southern California, CA 90007 USA (e-mail: cckuo@ee.usc.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.2984899

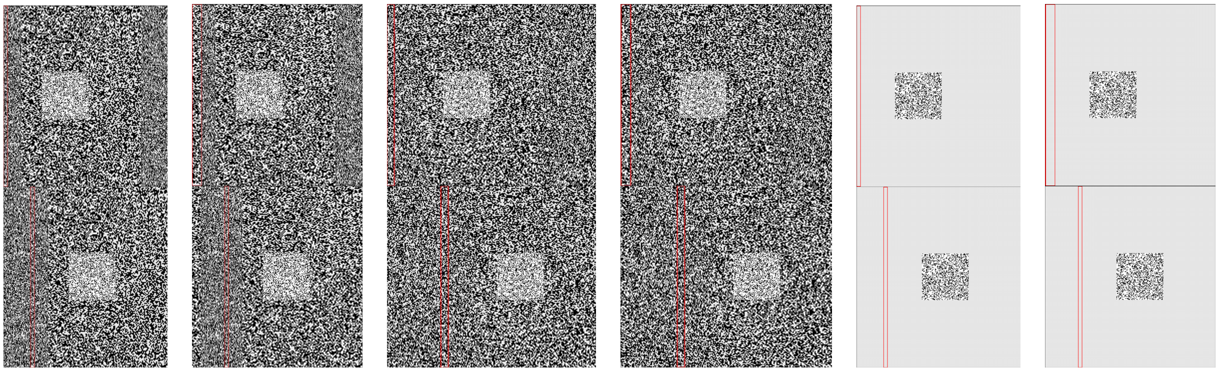

(a) Ref. Video 1 (b) Test video 1 (c) Ref. Video 2 (d) Test video 2 (e) Ref. Video 3 (f) Test video 3

Fig. 1. Illustration of stimuli constructed of random dot stereogram: left-view frames at two time instances, denoted by I^{L,t_1} (top row) and I^{L,t_2} (bottom row), for three reference and test stereo sequences. The red blocks mark a region in I^{L,t_1} and its correspondence in I^{L,t_2}. The region and its correspondence have different widths in test sequences. When the scaling factors are set to 0.35, 0.75 and 0.5 for I^{L,t_1} and 1 for I^{L,t_2}, their temporal incoherence is noticeable, visually unnoticeable and none for sequences 1, 2 and 3, respectively (see the supplementary material).

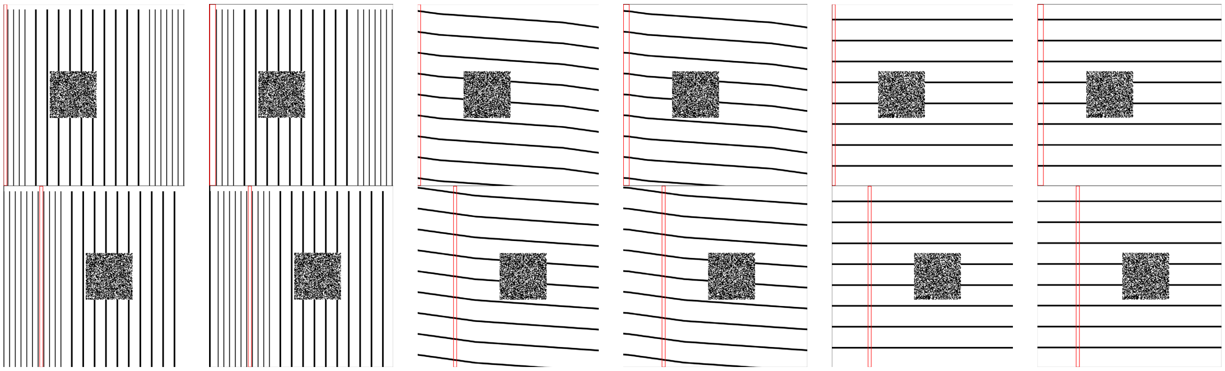

(a) Ref. Video 4 (b) Test video 4 (c) Ref. Video 5 (d) Test video 5 (e) Ref. Video 6 (f) Test video 6

Fig. 2. Illustrations of stimuli constructed of orientation-specific textures: left-view frames I^{L,t_1} (top) and I^{L,t_2} (bottom) at two time instances. The orientation angles of video sequences 4, 5 and 6 are set to 90°, 4° and 0°, respectively. The red blocks mark a region in I^{L,t_1} and its correspondence in I^{L,t_2}. The scaling factors for the marked regions are 0.5 and 1 in I^{L,t_1} and I^{L,t_2}, respectively, in test sequences. The temporal incoherence is noticeable, visually unnoticeable and none, in test video 4, 5 and 6, respectively (see the supplementary material).

and video stabilization) with respect to temporal coherence. Second, we propose a non-paired region detection algorithm that detects non-paired region in frame boundaries (especially texture-less regions) effectively. This further improves the performance of the proposed retargeting algorithm. Third, more qualitative and quantitative tests (additional subjective user study, combination of warping and cropping, etc) are conducted in the experiment section, where video shots from movies are collected as test videos. The superiority of the proposed method is extensively validated. Fourth, a keyframe-based optimization approach is proposed to significantly improve the computation efficiency of the proposed method without sacrificing the retargeting quality. Fifth, we quantify and better explain the conflicts among shape, depth, and temporal constraints, and more clearly showing the effectiveness of our method in mitigating such conflicts. We also quantitatively analyze the gain of our proposed temporal constraints on depth distortion.

There are three main contributions of our research as summarized below.

- We are among the first to explore perceptual temporal incoherence to guide stereo video retargeting so as to mitigate conflicts among multiple constraints (e.g., preserving the shape, the depth and their temporal coherence) effectively.
- In contrast with existing schemes that enforce consistent resizing on all temporal corresponding regions, we propose novel temporal constraints that allow non-paired boundary regions to undergo inconsistent resizing operations temporally.
- A grid warping framework with key-frame-based optimization is presented to implement stereo video retargeting efficiently.

The rest of this paper is organized as follows. Sec. II gives a brief survey of related 2D and 3D image and video retargeting methods. Then, the effect of temporal incoherence resulted from different visual contents on human perception is studied in Sec. III. The perceptual temporal incoherence-aware stereo video retargeting scheme is proposed in Sec. IV. Experimental results are given in Sec. V. Finally, concluding remarks are drawn in Sec. VII.

II. Related Work

A. Stereo Image Retargeting

By following [13], we classify content-aware retargeting methods into discrete and continuous two categories. Discrete methods iteratively remove or insert a group of pixels (e.g., seams) to resize image, while continuous methods warp pixels/regions non-uniformly. For stereo images, discrete methods [4], [14] resize a stereo image pair by removing seams from its left and right images. However, the method in [14] cannot preserve the depth information well due to the lack of depth preserving constraints. The method in [4] can preserve the depth information well by requiring that carved seams do not cross non-paired regions. Due to the discrete nature, these methods often introduce shape deformation to structural objects. Most continuous methods [5], [15]–[18] extended warping-based 2D image retargeting methods (e.g. [19]–[21]) to stereo image retargeting. Depth-preserving constraints that enforce the disparity of a few correspondences to be consistent with their original values were proposed in [5], [15], [16]. However, these constraints cannot preserve the depth of the whole 3D scene. In contrast, the method in [17] can faithfully preserve the scene depth by maintaining widths of individual non-paired regions and consistently warping each paired region and its correspondence.

B. Video Retargeting

2D video retargeting is more challenging than 2D image retargeting due to camera and object motions. Applying a 2D image retargeting method to each frame independently introduces temporal incoherence. To address it, 2D video retargeting methods exploit the temporal information to resize a retargeted frame based on 2D image retargeting methods. Based on developed temporal constraints, 2D video retargeting methods can be categorized into local and global approaches [22]. Local methods [23]–[26] imposes the constraint on a frame and a local time window of neighboring frames to ensure a coherent transformation. For example, temporally adjacent pixels are constrained to transform a limited number of neighboring frames consistently in [23]–[25]. Niu *et al.* [3] resized a frame sequentially, where the retargeting result of the current frame is propagated to the next one based on estimated camera motion. Local methods are more efficient computationally since only a few frames are processed at each time. However, local methods are not able to maintain temporal coherence if the duration of camera/object motions is longer than the local time window. Global methods [8]–[11], [22], [27] exploit the temporal information of the entire video. For example, methods in [9]–[11] employ the motion estimation algorithm to align corresponding grids/regions between frames and, then, demand these grids to be coherently resized throughout the entire video. Global methods achieve better temporal coherence than local methods since they adopt a longer time window.

C. Stereo Video Retargeting

Stereo video retargeting is challenging due to multiple requirements, including temporal coherence, shape preservation and depth preservation. There is little previous work. Li *et al.* [28] proposed effective spatio-temporal depth constraints by preserving spatial depth magnitude and temporal depth changes of key-points on 3 D objects. Lin *et al.* [29] and Kopf *et al.* [30] extended a grid-based 2D video retargeting method to stereo video since stereo video consisting of two 2D video sequences. Severe depth distortions are often observed in retargeted stereo videos since they do not consider depth preservation explicitly. To ensure temporal coherence, they directly employ the temporal coherence constraints derived from 2D video retargeting. However, the above methods ignore the fact that different video contents have different perceptual characteristics in the temporal dimension. For video with large motions, the imposed temporal constraints tend to have a severe conflict with the shape and the depth preservation constraints.

III. Perceptual Temporal Incoherence

Our hypothesis is that, given a non-paired region in a stereo video frame, the region and its temporal correspondences in neighboring frames can be inconsistently resized to some extent without introducing visually noticeable temporal incoherence artifacts. To validate our hypothesis, we design stimuli and conduct the following user studies.

A. Impact of Non-Paired Regions on Depth Construction

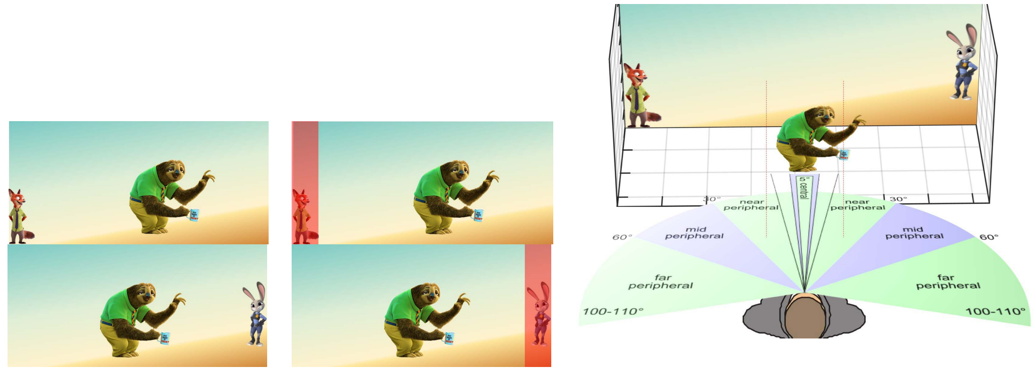
A non-paired region (also called a half-occluded region) in one view of a stereo video pair is a region that cannot find its correspondence in the other view, as illustrated in Fig. 3. As revealed in the literature, non-paired regions play an important role in depth construction. With the existence of non-paired regions in a stereo image/video, paired regions and their correspondences can together constitute the disparities in the stereo pair, thereby forming 3D objects with various depths in the perceived 3D scene. Hence, existing stereo image retargeting methods [4], [22] pointed out the importance of building additional constraints for non-paired regions, in order to faithfully maintain the depths of 3D objects.

On the other hand, HVS is non-uniformly sensitive to different visual fields. According to the studies in the psychological literature, peripheral acuity is worse than fovea acuity [31]. As a result, the speed of a moving object is perceived slower in the periphery than in the fovea [32]. When one watches a 3D video, some 3D objects in the scene are fixed by his/her eyes and lie in the fovea, while the non-paired regions at the frame boundaries lie in the periphery. In other words, HVS is less sensitive to non-paired regions at frame boundaries, compared with fixed 3D objects (see Fig. 3). Hence, we argue that some sorts of non-paired regions at frame boundaries can be inconsistently resized to some extent without introducing noticeable distortions to human eyes.

B. Stimuli

We design stimuli by taking into account the following factors.

1) Visual Content: The stimuli include two types: one is random dot stereogram (RDS) and the other is orientation-specific textures. Random dot stereogram is widely used in literatures on stereo perception. As for the random dot stereogram, the size and density of dots vary among videos to reflect various cues of visual contents. Besides, we put an additional small square as a foreground object in the videos for subjects to fix on (see Fig.1). This square is to simulate a salient object in content-aware retargeting.



(a) Original frames (b) Marked frames (c) Field of view of the HVS

Fig. 3. Illustration of non-paired boundary regions that are colored red in (b). The left and right frames are shown in the top and bottom rows of (a) and (b). (c) gives the HVS view fields, where the 3D object monkey lies in the central field of view (the fovea) while non-paired boundary regions lie in the peripheral visual field.

Therefore, the model that generates a random-dot stimuli is given as follows:

$$rds = f(a_F, \sigma_F, a_B, \sigma_B) \tag{1}$$

where a_F and σ_F respectively denote the size and density of dots in the foreground regions, and a_B and σ_B are the size and density of dots in the background regions. In this paper, we mainly explore the relationship between temporal incoherence and the resizing of background regions. Therefore, for foreground regions, both a_F and σ_F are set to be the same value in all stimuli videos, where $a_F = 8$ and $\sigma_F = 0.7$.

Compared with random dots, orientation-specific textures offer additional higher level structure information. In the orientation-specific textures, the orientation of lines also alters among videos (see Fig.2), where the range of line angle is $[0°, 90°]$. We also put an additional small square as a foreground object in all stimuli videos.

2) Motion: we simulate camera motions in a stimulus that change a non-paired/paired region to a paired/non-paired region from frame to frame. The speed of camera motions also varies in the stimuli.

3) Depth: Since a too large depth range would usually cause uncomfortable 3D experience, we limit the disparity of stimuli to fall in the range of [0 20], such that the perceptual depth does not exceed the comfort zone of HVS for most viewers.

4) Temporal Incoherence: The vision test was conducted to study whether subjects can perceive temporal incoherence artifacts caused by temporally inconsistent warping of corresponding regions in a stereo video. However, temporal incoherence cannot be directly obtained by comparing a resized video with its original version, because temporally inconsistent warping changes not only temporal information, but also other factors such as the shape and size of an object. As a result, those subjects who are not experts in image/video processing are likely to wrongly classify shape distortions as temporal incoherence. Hence, in the user study we should avoid the interference from other factors, such that subjects can focus on evaluating temporal incoherence artifacts. We therefore employ a full-reference quality assessment manner to evaluate temporal incoherence artifacts. More specifically, for each original stimulus, we build multiple pairs of retargeted videos, each consisting of a reference video and a test video for comparison. In each pair, the reference video is constructed by consistently resizing all temporal correspondences, regardless of paired/non-paired regions, across neighboring frames, while inconsistently resizing different regions in each frame. In contrast, all paired regions in the test video undergo consistent resizing with that of their temporal correspondences in the reference video, whereas the non-paired regions at frame boundaries undergo a specific degree of temporally inconsistent resizing with their correspondences across neighboring frames.

For the test video, we define the degree of temporal resizing inconsistency (TRI) as the resizing difference of two corresponding regions between neighboring frames.[1] In particular, given a region r^{t_1} in frame I^{t_1} of the original video, let r^{t_2} denote its correspondence in I^{t_2}, and s^{t_1} and s^{t_2} the scaling factors of r^{t_1} and r^{t_2}, respectively. The temporal resizing inconsistency TRI is defined as follows:

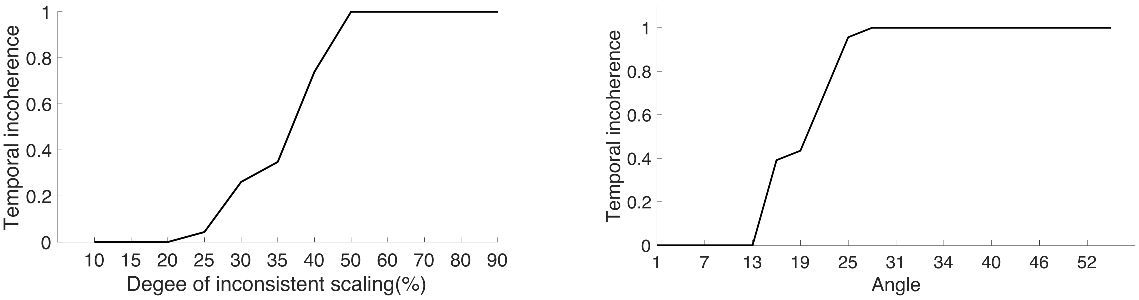
$$TRI = |s^{t_1} - s^{t_2}|. \tag{2}$$

5) Subjective Test: In the subjective user study, the reference and test videos are displayed side by side on the monitor. Subjects are asked to look at the fixed square in the test/reference video, and then answer the following question:

Q: Do you notice the motion incoherence between the two video clips (1: Yes; 2: no)?

There are a few parameters (e.g., TRI and line angle) each is of a large range. Fully evaluating all values for each parameter would be a laborious task and take huge time cost. Recently, some quality assessment methods [33], [34] proposed a binary search procedure to efficiently search the optimal value of a quantization parameter for video compression. Inspired by these methods, we propose to sample the values of these parameters from coarse to fine, so as to hierarchically evaluate value for each parameter. Take the stimuli of lines as an example, the values of angle are set to be $0°, 10°, 20°, \ldots, 90°$, in the coarse stage. After the first-round subjective evaluation, the results show that all subjects can notice motion incoherency for stimuli whose angle is greater than 30°. Without loss of generality, we argue that temporally

[1] The corresponding regions between two neighboring frames have the same width in our original videos.



(a)

(b)

Fig. 4. User study results: (a) evaluation on test video sequences retargeted from their original ones, where each test video sequence undergoes temporally inconsistent transformation to some degree (see Fig. 1(a)–(d)); and (b) evaluation on test video sequences retargeted from different original videos, where original videos are shot at different angles (see Fig. 2) and all test videos undergo the same degree of temporally inconsistent transformation. The perceptual temporal incoherence increases with the degree of temporally inconsistent in (a) and the degree of orientation angle (i.e., texture characteristics) in (b).

inconsistent resizing would cause temporal incoherency on stimuli with an angle greater than 30°. Therefore, in the fine stage, we only evaluate angle$<$ 30°, and adjust the angle in a finer range. That is, the values of angle are set to be $1°, 4°, 7°, \ldots, 28°$. We then conduct the second-round evaluation.

C. Observations and Analyses

1) Subjects: We invite 23 subjects to participate in the user study. All subjects have normal stereopsis perception.

2) Apparatus: Visual stimuli are displayed on a 53.15cm × 29.90cm 3D monitor (ASUS Vg248qe 144 Hz, 1920 × 1080 pixels). The subjects watch the stimuli through an Nvidia shuttered glass at a viewing distance of 80cm for the stereo vision test. We then ask the subjects to assess their viewing experience of watching individual stimuli.

3) Observations: According to the user study result, temporally inconsistent resizing within a range of degrees does not introduce noticeable temporal incoherence for certain types of stimuli. For example, as shown in Fig.1 and Fig. 2, perceptual temporal incoherence artifacts in test video 2, 3, 5 and 6 are negligible.

Moreover, the perceptual temporal incoherence, incurred by inconsistently resizing a region and its temporal correspondences, is related to the textural characteristics of this region. In particular, the temporal incoherence is dependent on the texture difference between the resized versions of the region and its temporal correspondence, as illustrated in Fig. 4, Fig. 1 and Fig. 2. More specifically, let r^{t_1} and r^{t_2} denote the marked region and its correspondence in I^{L,t_1} and I^{L,t_2} in Fig. 1 and Fig. 2. The temporal resizing inconsistency degree TWI_D for r^{t_1} and r^{t_2} in *test video 1* is the same as that in *test video 3*. However, the temporal incoherence of *test video 3* is perceptually unnoticeable, but that in *test video 1* is noticeable. This is because the textures of the resized r^{t_1} and r^{t_2} are visually the same in *test video 3*, but those in *test video 1* look significantly different. In contrast, *test video 1* and *test video 2* are resized from the original video, but the temporally resizing inconsistency degrees for r^{t_1} and r^{t_2} in *test video 1* are higher than their counterparts in *test video 2*. Consequently, compared with *test video 2*, the texture difference between r^{t_1} and r^{t_2} in *test video 1* is higher, thereby introducing visually more annoying temporal incoherence artifacts in *test video 1*.

As for the result with orientation-specific textures, as shown in Fig. 2, the temporal incoherence perceived by viewers is also related to the texture orientation, which is measured by the angle of the texture lines corresponding to the horizontal lines. Roughly speaking when the angle is smaller than around 45°, the larger the angle is, the higher the texture difference becomes. When the angle is smaller than around 45°, most subjects can notice the temporal incoherence.

4) Dependency of Temporal Incoherence on Texture: We further quantitatively validate the above observations. We evaluate the dependency by measuring the correlation between temporal incoherence scores and the textural feature of regions in resized videos, where the textural feature is represented by the texture difference between a resized region and its correspondence temporally.

We use the Pearson correlation, ρ, to measure the correlation, following the literature on quality assessment [35], [36]. The higher absolute value of ρ, the stronger correlation between texture and temporal coherence. Hence, the correlation is measured as

$$\rho = \frac{E[(\delta - \mu_\delta)(Y - \mu_Y)]}{\sigma_\delta \sigma_Y}, \tag{3}$$

where δ is the list of values of texture difference measured for each resized video, Y the list of the temporal incoherence scores corresponding to all test videos, $E(\cdot)$ the expectation function, and σ is the standard deviation. We adopt LBP [37] to represent the textural feature, due to its well-proven performance in recognition, retrieval, etc. We obtain $\rho = 0.705$ in the correlation estimation, which shows there does exist strong correlation between temporal coherence and texture differences.

To sum up, the perceptual temporal incoherence caused by temporally inconsistent resizing is dependent on the textural characteristics of regions. Based on this observation, we propose texture-dependent temporal constraints to adaptively control the degrees of inconsistent resizing for a non-paired region and its temporal correspondences across neighboring frames according to the regions' textural characteristics.

IV. Temporal Incoherence-Aware Retargeting

Based on the user study results on perceptual temporal incoherence, we propose a stereo video retargeting framework that employs texture-dependent temporal constraints.

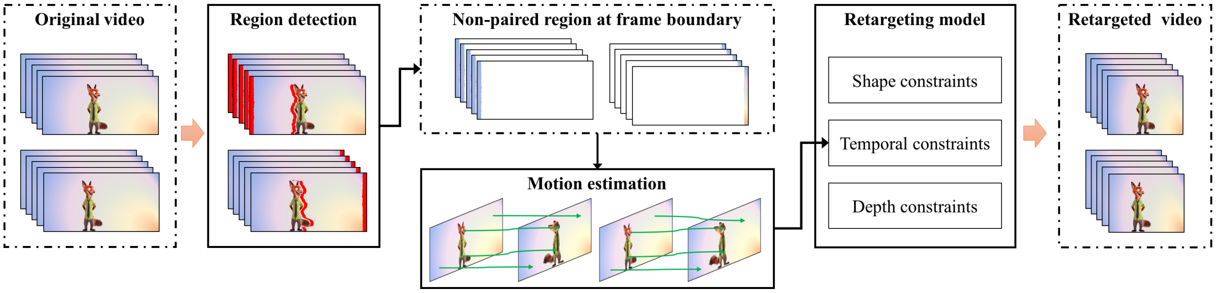


Fig. 5. Proposed framework of stereo video retargeting.

As shown in Fig. 5, our method first detects corresponding regions between the left-views and right-views of the input stereo video pair, and then classifies the regions into paired regions and non-paired ones. For each paired region, besides imposing appropriate shape and depth constraints, all its temporal correspondences across neighboring frames are further constrained to be consistently resized. In contrast, for non-paired regions, the imposed temporal constraints are relaxed such that temporal correspondences of a non-paired region can be inconsistently warped to some extent without introducing perceptual temporal incoherence artifacts. Such relaxation on temporal coherence for non-paired regions can effectively mitigate the conflicts among the requirements for preserving shape, depth, and temporal coherence. All these steps are elaborated below.

A. Problem Formulation

Suppose a stereo video is resized from $W \times H$ to $W' \times H'$. To maximize 3D viewing experience, our stereo video retargeting method aims to simultaneously preserve the shapes of salient objects, the depths of 3D scenes, and their temporal coherence by minimizing the following overall distortion:

$$D^{total} = \min\{\alpha \cdot D^S + \beta \cdot D^D + D^T\}, \tag{4}$$

where D^S denotes the shape distortion of salient objects, and D^D the depth distortion of 3D scenes, D^T the amount of perceived temporal incoherence distortion, α, and β the weights for the three distortion terms.

We adopt grid-based warping [17], [22] to solve the above optimization problem, since it has proven to be an efficient and effective means for 2D and 3D visual retargeting. We first uniformly divide the frames in a stereo video pair into grids, and then formulate an optimization model which imposes shape, depth, and temporal constraints in (4) to constrain the warping of grids. Thus, the optimal retargeted version is obtained by finding the optimal set of grid warping functions that minimize the overall distortion D^{total}.

B. Non-Paired Region Detection

We propose relax temporal constraints on part of non-paired regions, which can be detected from the disparity maps between the left-view and right-view frames estimated by stereo matching [38], [17]. However, although stereo matching can accurately estimate disparity for most regions, it often wrongly assigns zero disparity to some kinds of regions, thereby introducing large holes for these regions in the estimated disparity maps. Such regions would be wrongly classified as non-paired regions, thereby degrading retargeting performance. To address the problem, we combine stereo matching with optical flow estimation to improve the accuracy of disparity estimation, as optical flows can help avoid such incorrect holes. Specifically, we first employ stereo matching to estimate initial disparity maps, then perform optical flow estimation on those regions with significant no-correspondence holes in the initial disparity maps and refine the maps by filling the holes based on the estimated optical flows. In our implementation, we mainly employ the algorithms proposed in [38]–[40] for stereo matching and optical flow estimation.

After obtaining the refined disparity map, we detect non-paired frame-boundary regions at frame boundaries and the other non-paired regions as follows:

1) Non-Paired Boundary Regions: Given a non-paired region at a frame boundary in a left/right-view frame, its correspondence is out of view in the other view. Therefore, given a boundary pixel $p_k^{z,t}$, it is classified as a non-paired boundary pixel if it satisfies the following condition:

$$x_k^{z,t} + d_k^{z,t} < 0, \quad \text{or} \quad x_k^{z,t} + d_k^{z,t} > W, \tag{5}$$

where $x_k^{z,t}$ denotes the x coordinate of $p_k^{z,t}$ in frame $I^{z,t}$, $d_k^{z,t}$ the disparity value of pixel p_k, W the width of $I^{z,t}$.

2) Other Non-Paired Regions: For the remaining non-paired regions in left/right-view frames, we detect them in a way similar to [4]. Specifically, given a pixel in one view, we find its correspondence in the other view. If more than one pixel corresponds to the same pixel in the other view, the pixel with the smallest disparity value is the non-paired one.

C. Shape Preservation

We define the shape distortion energy of a stereo video as a weighted sum of all grids' distortion energy. To reduce computation, we adopt grid-edge-based warping proposed in [41], [42], that sets the shapes of all retargeted grids to be rectangular. In this way, the grids in each column/row are enforced to have the same width/height, thereby significantly reducing the number of variables, so as the complexity of solving the variables. Since the retargeted grids remain rectangular, the shape distortion energy of a grid can be simply characterized by the difference between its original aspect ratio and the retargeted ratio [22], [41]. Let $g_i^{z,t}$ denote a grid in frame $I^{z,t}$ where $z \in \{L, R\}$, the shape distortion is formulated as

$$\begin{aligned} D^S &= \sum_z \sum_t \sum_i D(g_i^{z,t}) \\ &= \sum_z \sum_t \sum_i \|w(g_i^{z,t}) \cdot \tilde{h}(g_i^{z,t}) - \tilde{w}(g_i^{z,t}) \cdot h(g_i^{z,t})\|^2 \cdot \eta_i^{z,t}, \end{aligned} \tag{6}$$

where $w(g_i^{z,t}), h(g_i^{z,t}))$ and $(\tilde{w}(g_i^{z,t}), \tilde{h}(g_i^{z,t}))$ denote the widths and heights of grid $g_i^{z,t}$ before and after retargeting, respectively, $\eta_i^{z,t}$ is the saliency value of $g_i^{z,t}$, calculated by averaging the saliency values of all pixels in $g_i^{z,t}$.

1) Perceptual Importance Map: Given a grid $g_i^{z,t}$, we compute its perceptual importance $\eta_i^{z,t}$ by averaging the spatial importance values of it and its temporal correspondences across frames, similar to existing video retargeting methods (e.g. [10], [22]).

$$\eta_i^{z,t} = \sum_{k \in \mathbf{c}} \eta_k^{z,t'}, \tag{7}$$

where $\mathbf{c}$ is the set containing $g_i^{z,t}$ and its temporal correspondences across frames.

As revealed in [43], depth cue indicates occlusion information which is favorable for shape preservation. Similarly, in [4], [44], depth information provides valuable cues to measure the perceptual importance of visual content in a stereo image pair. The spatial perceptual importance map of each frame is estimated by image-based saliency map [45] and depth-based saliency map.

D. Depth Preservation

We adopt the depth-preserving constraints proposed in our previous work [17] for depth preservation. The depth-preserving constraints are used to constrain the warping of grids on paired regions and non-paired regions in different ways. We divide depth distortion D^D in (4) into two terms: D^{D_P} and D^{D_N} for paired regions and non-paired regions, respectively.

1) Paired Regions: The constraints enforce each paired region in I_t^z and its correspondence in the other view $I_t^{z'}$ to undergo consistent width changes. Let $g_i^{L,t}$ be a grid covering paired region r and $g_k^{R,t}$ its corresponding grid in I_t^R, the constraints are given as follows:

$$D^{D_P} = \sum_t \sum_{r \in \Upsilon^t} \sum_{g_i^{L,t} \in r} \theta^{i,k,t} \cdot \|\tilde{w}(g_i^{L,t}) - \tilde{w}(g_k^{R,t})\|^2, \tag{8}$$

where Υ^t denotes the set containing all paired regions in $I^{L,t}$ and $I^{R,t}$, and $\theta^{i,k,t}$ the ratios of corresponding regions between $g_i^{L,t}$ and $g_k^{R,t}$.

2) Non-Paired Regions: We translate depth distortion in non-paired regions D^{D_N} into depth constraints that preserve the width of non-paired regions. Thus, for a non-paired region $\bar{r}$, the constraints preserve the total width of grids covering $\bar{r}$ by minimizing

$$D^{D_N} = \sum_t \sum_{g_i^{z,t} \in \bar{r}_t} \|\tilde{w}(g_i^{z,t}) - w(g_i^{z,t})\|^2, \tag{9}$$

where $\bar{r}_t$ denotes a non-paired region in $I^{L,t}$ or $I^{R,t}$.

3) Vertical Disparity: Non-zero vertical disparity would cause uncomfortable 3D viewing experience such as headache and eye fatigue. To avoid non-zero vertical disparity, we constrain the retargeted grid at the same location between left-view $I^{L,t}$ and right-view $I^{R,t}$ to be of the same height.

E. Texture-Dependent Temporal Constraints

The temporal constraints aim to constrain the temporally corresponding regions across neighboring frames to be consistently resized so as to avoid noticeable temporal incoherence artifacts. According to the user study in Sec. III, the warping function for a non-paired boundary region can differ from that of the region's temporal correspondences to some extent based on the region's textural characteristics. We hence first identify grids in non-paired boundary regions, and label them non-paired boundary grids. Then we translate the temporal incoherence distortion D^T in (4) into two sets of temporal constraints for the non-paired boundary grids and the remaining, respectively.

Before imposing temporal constraints, we first align the temporal corresponding grids across frames. We adopt the grid flow algorithm proposed in [22] to align the corresponding grids temporally, as it can effectively align grids temporally in a video involving significant object and camera motions.

Subsequently, the first set of constraints are imposed on temporally aligned non-paired boundary grids. In particular, let $g_i^{z,t}$ and $g_j^{z,t'}$ respectively denote the temporally aligned grids in frame $I^{z,t}$ and $I^{z,t'}$, where $g_i^{z,t}$ is a non-paired boundary grid. We constrain the horizontal warping inconsistency between $g_i^{z,t}$ and $g_j^{z,t'}$ based on the texture difference between the two grids as follows:

$$\begin{aligned} &\tilde{h}(g_i^{z,t}) = \tilde{h}(g_j^{z,t'}), \\ &\quad 0 < |\tilde{w}(g_i^{z,t}) - \tilde{w}(g_j^{z,t'})| < \Theta, \end{aligned} \tag{10}$$

where Θ is the threshold empirically derived from the texture difference between $g_i^{z,t}$ and $g_j^{z,t'}$.

In this work, we only set two levels of warping inconsistency for Θ. In particular, the user study results in Sec. III reveal that for most stimuli, as long as the warping inconsistency is less then $0.15 \cdot w(g_i^{z,t})$, the temporal incoherence is perceptually unnoticeable or negligible. As a result, for a region with textures sufficiently similar to that of its inconsistently resized correspondences (e.g., as depicted in Fig. 1(f) and Fig.2(f)), the degree of warping inconsistency can be up to $0.5 \cdot w(g_i^{z,t})$ or even larger. Therefore, Θ is set as

$$\Theta = \begin{cases} 0.5 \cdot w(g_i^{z,t}), & \text{if} \quad \delta(g_i^{z,t}, g_j^{z,t'}) < \varepsilon \\ 0.15 \cdot w(g_i^{z,t}), & \text{otherwise,} \end{cases} \tag{11}$$

where $\delta(g_i^{z,t}, g_j^{z,t'})$ measures the texture difference between the inconsistently resized textures of $g_i^{z,t}$ and $g_j^{z,t'}$ as follows:

$$\delta(g_i^{z,t}, g_j^{z,t'}) = (\mathbb{T}(g_i^{z,t}) - \mathbb{T}(\tilde{g}_j^{z,t'}))^2 - (\mathbb{T}(g_i^{z,t}) - \mathbb{T}(g_j^{z,t'}))^2, \tag{12}$$

where $\mathbb{T}(g_i^{z,t})$ denotes the texture descriptor for $g_i^{z,t}$.

The second set of temporal constraints applies to the remaining temporally aligned grids $g_i^{z,t}$ and their correspondences $g_j^{z,t}$ except non-paired boundary grids. We constrain $g_i^{z,t}$ and $g_j^{z,t}$ to be consistently resized in both their height and width:

$$\begin{cases} \tilde{h}(g_i^{z,t}) = \tilde{h}(g_j^{z,t'}) \\ \tilde{w}(g_i^{z,t}) = \tilde{w}(g_j^{z,t'}) \end{cases} \tag{13}$$

F. Key-Frame Based Optimization

The optimal set of retargeted grids is then obtained by minimizing the following overall distortion:

$$\alpha D^S + \beta(D^{D_P} + D^{D_N}) \tag{14}$$

subject to the derived temporal constraints. This is a quadratic programming problem, and we solve it using the active-set method [46].

To obtain optimal retargeting results, the method proposed in [12] directly performs the above optimization over all frames of a video shot.[2] However, the computational complexity of such optimization increased with the number of variables, which is $N_t \times (N_x + N_y)$, where N_t is the frame number, N_x and N_y are the numbers of grid columns and rows in a frame, respectively. Since the total number of variables increases with the number of frames, the computational cost of the in [12] is high for long video shots.

Some 2D video retargeting methods devoted efforts into reducing computation cost via reducing the number of video frames. For example, the 2D video retargeting method proposed in [22] first extracts a few key-frames which summarize the information of a video, and then performs the optimization of retargeting on these key-frames. Since the key frames well summarize information, the resizing results can be optimal for objects in all frames.

Inspired by [22], we present a key-frame based optimization method to efficiently solve the above optimization problem. In the method, only a few key-frames are resized by solving a small-scale optimization programming, where these key-frames are selected by grid flows [22]. After resizing the key-frames, we then resize the non-key frames in between every two neighboring key-frames. In existing video retargeting methods, for maintaining temporal consistency, temporal correspondences are constrained to be consistently resized across frames. Thus, given a grid in a non-key frame, it is easy to predict its retargeted size by grid interpolation from its corresponding grids in two neighboring key-frames, like [22]. However, in our method, since a non-paired boundary region and its temporal correspondences undergo inconsistent resizing. we cannot directly apply such grid interpolation. Instead, we present a new grid interpolation to address this issue. Since the time interval between two neighboring key-frames is usually short, we can assume that camera/object moves linearly between two neighboring key-frames. As such, the size of a non-paired boundary region changes linearly between two neighboring key-frames, making the sizes of these retargeted non-paired boundary regions change linearly. Thus, for non-key-frames, given a non-paired boundary grid, its retargeted size is predicted from retargeted grids on non-paired boundary regions between two neighboring key-frames, although these grids are correspondences (see Fig. 6). For the remaining grids (i.e., paired grids and non-paired non-boundary grids) in non-key-frames, as depicted in Fig. 6, their retargeted grids are estimated by linear grid interpolation.

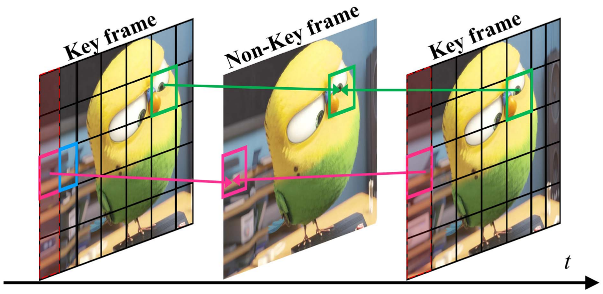


Fig. 6. Illustration of resizing a non-key frame (left-view). The dashed rectangles mark non-paired regions in key-frames. A red grid in the non-key frame is estimated from grids with the same color in the two key-frames. A red grid in the non-key frame is estimated from the corresponding non-paired grids in the two neighboring key-frames, rather than from their temporal correspondences (i.e., the blue grid).

The key-frame based scheme consumes significantly lower computation cost, since it performs optimization only on the grids in key-frames. In particular, the number of variables with the key-frame optimization is $N_k \times (N_x + N_y)$, where N_k is the number of key-frames. Typically, since $N_k \ll N_t$, the keyframe-based implementation largely reduces the computation cost, compared with the method in [12]. For example, the number of video frames in Fig. 10 is 90, while that of key-frames is only 4. In addition, we do not need to estimate disparity maps for non-key frames, which further significantly reduces computation.

V. Experimental Results

In the experiments, we first validate the effectiveness of the texture-dependent temporal constraints. Then, we compare our method with the state-of-the-art approaches by conducting qualitative and quantitative evaluations.

Dataset: We collect testing stereo videos from commercial 3D movie films for performance evaluation. Our collected stereo videos pose technical challenges on stereo video retargeting in two aspects: (1) Both foreground and background objects are of large disparity range and with significant temporal disparity changes, which exhibits vivid and impressive 3D viewing experience to viewers but requires effective depth-preserving constraints for retargeting. (2) Besides various object motions, all videos contain significant camera motions that change the locations of objects largely across frames. This makes maintaining temporal coherence challenging for stereo video retargeting.

We also test our method on the CVW dataset [29]. In most videos of the CVW dataset, foreground objects or their movements occupy a large portion of a frame. We choose challenging videos which contain multiple objects with significant motions and disparity changes.

[2] video shots are divided by shot segmentation

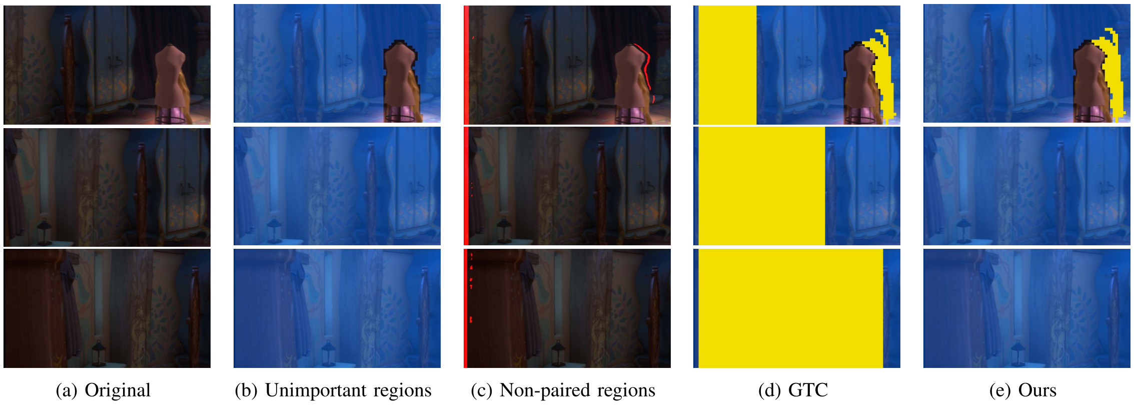


Fig. 7. Illustration of perceptually unimportant regions which are constrained by temporal constraints to be consistently resized with non-paired regions in another frames. Rows from top to bottom: frames #55, #42 and #30. In (b) unimportant regions are indicated by light blue. In (c) non-paired regions in each frame are colored red. In (d) and (e) non-salient regions which are constrained to be consistently resized with non-paired regions in another frames are colored yellow.

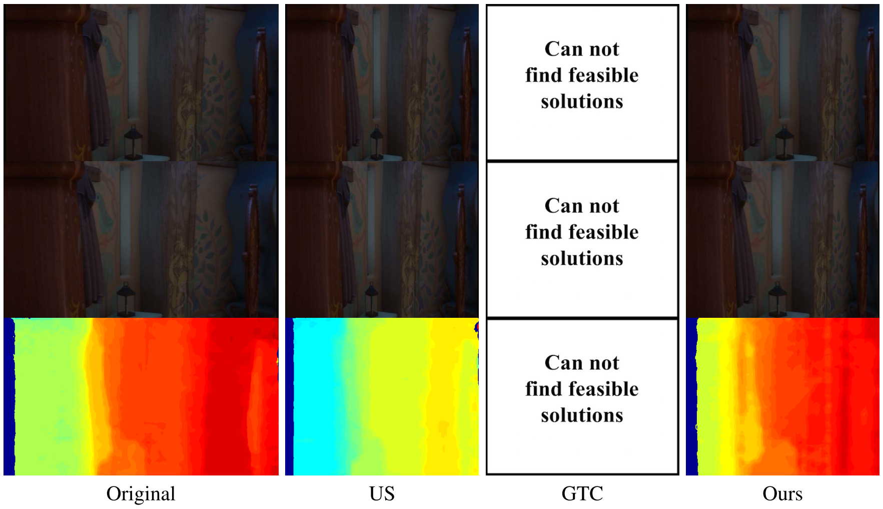

Can not find feasible solutions
Can not find feasible solutions
Can not find feasible solutions
Original
US
GTC
Ours

Fig. 8. Performance comparison for *Tangled*. Rows from top to bottom: the left-views, the right-views and the disparity maps.

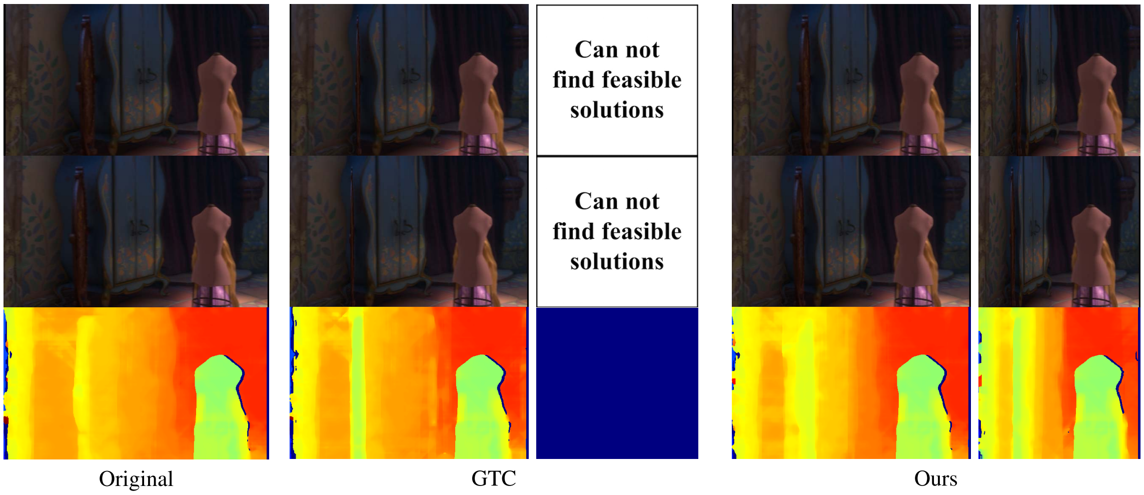

Can not find feasible solutions
Can not find feasible solutions
Original
GTC
Ours

Fig. 9. Retargerting results for *Tangled*, where the width is reduced by 10% and 40%, respectively. Rows from top to bottom: the left-views, the right-views and the disparity maps.

A. *Effectiveness of Temporal Constraints*

We evaluate the effectiveness our temporal constraints in two aspects: (1) the amount of reduction on conflicts among shape, depth, and temporal constraints and (2) the quality improvement, achieved by our temporal constraints.

Constraint conflicts occur when different constraints lead to conflicting requests for resizing a region. In particular, shape constraints preserve the shapes of salient objects by resizing perceptually unimportant regions. However, when these unimportant regions are constrained not to be resized

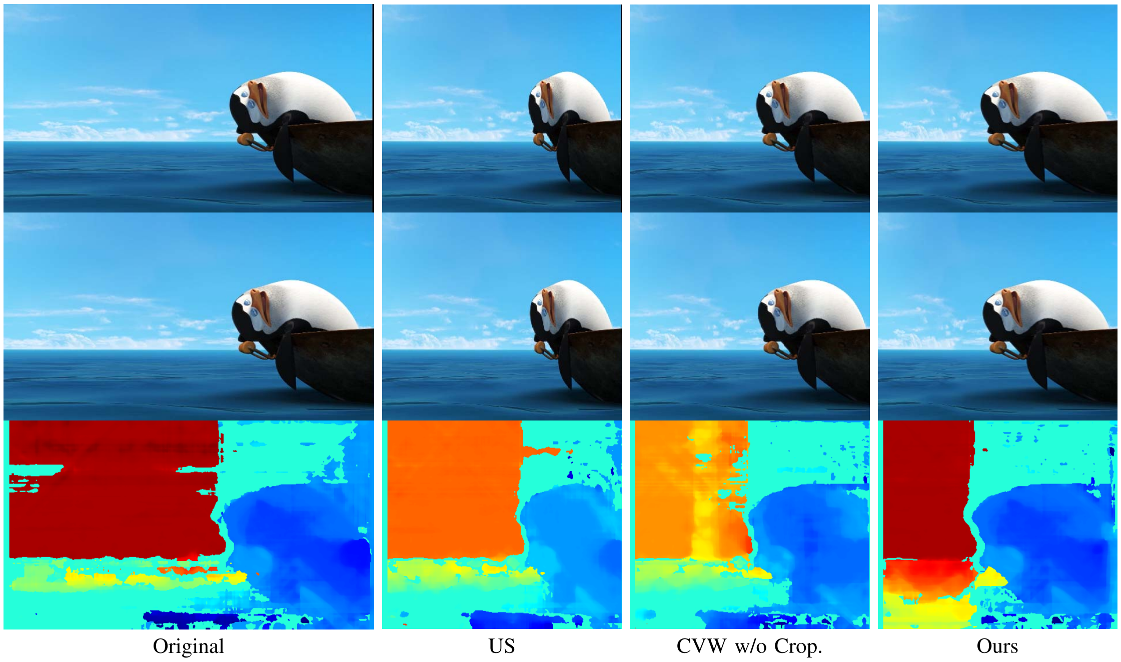


Fig. 10. Area of perceptually unimportant regions constrained not to be horizontally resized by temporal constraints.

by temporal/depth constraints, conflicts occur among these constraints. The more the constraint conflicts, the more the regions without a feasible solution, thus the severer the degradation on stereo video retargeting performance. We quantify the degree of conflicts among various constraints as the area of perceptually unimportant regions in a frame which are supposed to be resized for shape preservation, but eventually are constrained not to be resized by depth/temporal constraints.

We mainly compare our temporal constraints with global temporal constraints (GTC) proposed in the consistent volumetric warping (CVW) [29], that enforce all temporal correspondences of each region, regardless of a paired one or non-paired one, to be consistently resized across a video shot. To compare with GTC, we implement a baseline warping method (without cropping) namely GTC that employs the same shape and depth constraints as our method, but adopts the global temporal constraints proposed in CVW [29].

As illustrated in Fig. 7(d) and Fig. 10, GTC constrains a large area of unimportant regions not to be resized, which introduces severe constraint conflicts. For example, the percentage of such unimportant regions with conflicts increases to 90% in frame #30. The reason of causing such a large degree of conflicts lies in the facts that, in the test videos, many unimportant regions in a frame correspond to certain non-paired regions in neighboring frames due to camera motions. For example, both non-paired regions in frames #42 and #55 correspond to a non-paired region in frame #30. The depth constraints tend to preserve the widths of non-paired regions, whereas GTC enforces non-paired regions to be consistently resized with their temporal correspondences across neighboring frames. As a result, GTC together with depth constraints would enforce those unimportant regions that correspond to non-paired regions not be resized, thereby leading to conflicts with shape preservation which relies on resizing unimportant regions to preserve the shapes of salient objects.

In contrast, compared with GTC, our method allows much more unimportant regions to be resized by adequately relaxing the temporal constraints on non-paired boundary regions, as shown in Fig. 10, thereby largely reducing constraint conflicts.

Because GTC leads to many constraint conflicts (see Fig. 7), it fails to find feasible solutions for regions in the video shown in Fig. 8, although the frames contain lots of unimportant regions. In contrast, our method largely reduces the conflicts so as to better preserve the depth and shape information without sacrificing temporal coherence perceptually.

We further quantify the gains of our temporal constraints. We incrementally reduce the frame width by 10%, 20%, …, so as to evaluate the maximum width of a video frame that a retargeting method can be trimmed out without introducing noticeable visual artifacts. For the video shown in Fig. 8, when the trimmed width is larger than 10%, GTC fails to find a feasible solution. In contrast, the trimmable width achieved by our method is 40%, thanks to the adequate relaxation on temporal constraints that allow non-paired boundary regions to undergo inconsistent resizing temporally.

B. Qualitative Comparison

We then evaluate the performances of stereo video retargeting methods in terms of shape preservation, depth preservation, and temporal coherence. The depth preservation performance is evaluated by comparing the disparity maps of retargeted versions with their original one, where dark red and dark blue respectively indicate the highest and lowest disparity values. We compare our method with two existing methods: the uniform scaling (US) method and the CVW method [29], which is a state-of-art content-aware stereo video retargeting method and is most related to our work. Because CVW combines warping with cropping to handle difficult videos, for

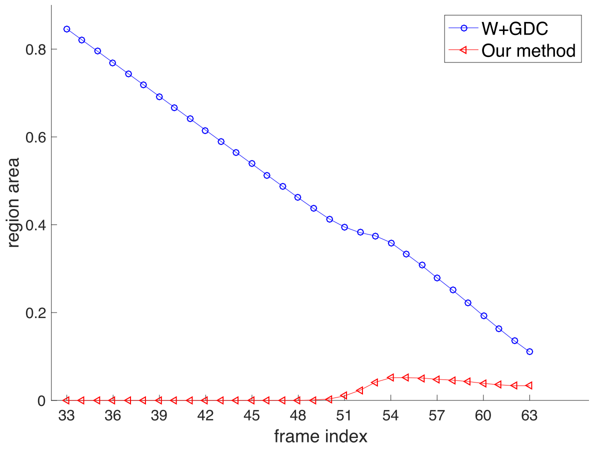


Fig. 11. Performance comparison for *Madagascar 4*. Rows from top to bottom: the left-views, the right-views and the disparity maps.

a fair comparison, we implement two versions of CVW: the full version of CVW with both warping and cropping (namely "CVW") and CVW without any cropping (namely "CVW w/o Crop.").[3]

1) Retargeting Without Cropping: Fig. 11 compares the left-views of the original and retargeted frames and their associated disparity maps for a stereo video shot containing large camera motions. It shows that, for shape preservation, our method and CVW achieve comparable performance, and both outperform US. For depth preservation, both US and CVW lead to severe depth distortions in most regions, since they do not explicitly consider depth information while performing retargeting. In contrast, since our temporal constraints trade unnoticeable temporal incoherence for depth preservation, our method achieves the best depth preservation performance. As for temporal coherence, since non-paired boundary regions usually do not contain much textural details, our temporal constraints have an enough room to inconsistently resize these non-paired regions and their temporal correspondences to a large extent to achieve shape and depth preservation without introducing noticeable temporal incoherence artifacts. Therefore, even with temporally inconsistent resizing for non-paired regions, our method still achieves comparable perceptual temporal coherence with US and CVW. Fig. 12 shows a video with more textural details in non-paired regions than the video in Fig. 11. Our temporal constraints adaptively adjust the extent of temporally inconsistent resizing of non-paired boundary regions, thereby achieving the best overall visual quality. (see the complementary materials for comparison).

2) Retargeting With Cropping: Since CVW adopts both warping and cropping, we further combine our method with cropping, namely "Ours w/ Crop", for comparison. Fig. 13 compares our method (incorporating cropping) with US and CVW for a frame in test video *IceAge4*. The test video contains multiple salient objects with significant object motions, for which camera has to be moved intensively for 'tracking the moving objects. It is therefore challenging for a stereo video retargeting method to simultaneously preserve the depth, temporal coherence and shapes of multiple objects. We can observe that CVW significantly distorts the shape of folivora and its depth, leading to poorer viewing experience. In contrast, our method simultaneously well preserves the shapes and depths of salient objects, as well as successfully maintains temporal coherence. Fig. 14 shows the right-views and disparity maps of the original and retargeted frames for *Madagascar3* that contains multiple foreground objects. Similarly, our method achieves the best performance compared with CVW and US. (see the videos in the supplementary material).

C. Subjective User Study

Since 3D viewing experience is crucial while watching a stereo video, we conduct a subjective user study to compare the performances of different retargeting methods on preserving 3D viewing experience. We invite 12 subjects with diverse ages to participate in the user study. As reported in [47], 3%–15% of the population are stereo blindness. We hence conduct vision test on all subjects, to verify whether subjects have normal stereopsis (ability of properly fusing a 3D scene) before the user study, following [48]. All subjects pass the vision test. In addition, they are not professionals in 3D image/video processing.

The subjects view stereo videos on an ASUS 3D 24-inch monitor with the resolution of 1920 × 1080, equipped with NVIDIA GeForce 3D Vision and active shuttered glasses. The width and height of the display screen are 54.6 cm and 31.4 cm, respectively. We select a moderate-size display, since stereo videos on large screen would exhibit high perceptual depth which may exceed the comfort zones of subjects. Since ITU-R BT.2021 [48] suggests that the viewing distance should be 4 times the height of the display screen. The viewing distance is set to be 125.6 cm in the user study.

Different from 2D videos, stereo videos are more complex and a human brain needs more time to fuse 3D scenes. Therefore, we allow subjects to playback, pause, and Fast forward the test videos during the subjective test.

We adopt pairwise comparison by following the setup of user studies in the literature [27], [49], [50]. We show an original video on the top and two retargeted versions which are placed in a random order to subjects each time. Then, each subject is asked to choose the retargeted version he/she prefers in terms of preserving 3D experience of the original stereo video. We do not inform the subjects of the experiment hypothesis and purpose.

We conduct subjective evaluation on 10 stereo video clips. We compare our method with the CVW and US methods. We receive 12 × 10×3= 360 pairwise comparison answers in total, where each subject evaluates 3 × 10 = 30 video pairs. Table I shows the winning frequency matrix, where the value a_{ij} in row i and column j indicates that method i receives a_{ij} preference votes to method j. As shown in Table I, our method offers the best 3D viewing experience, as 83.33% and 91.67% of subjects prefer our method to the CVW and US methods,

[3] Since CVW does not release codes, the comparisons are based on our own implementations.

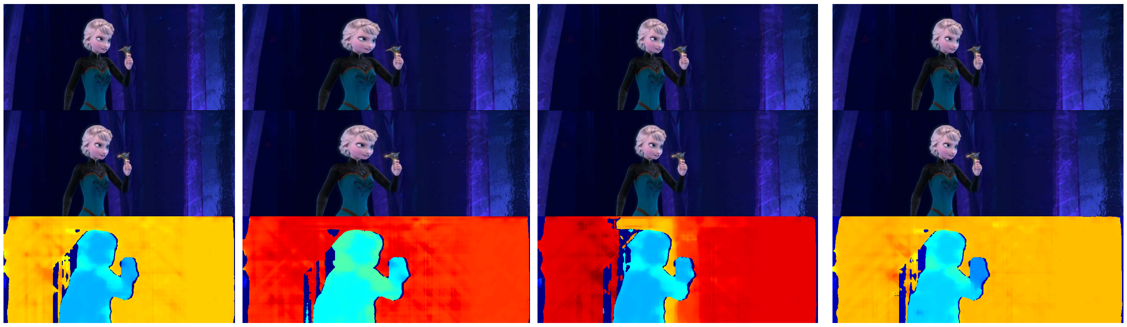

Original US CVW w/o Crop. Ours

Fig. 12. Performance comparison for *Frozen*. Rows from top to bottom: the left-views, the right-views and disparity maps.

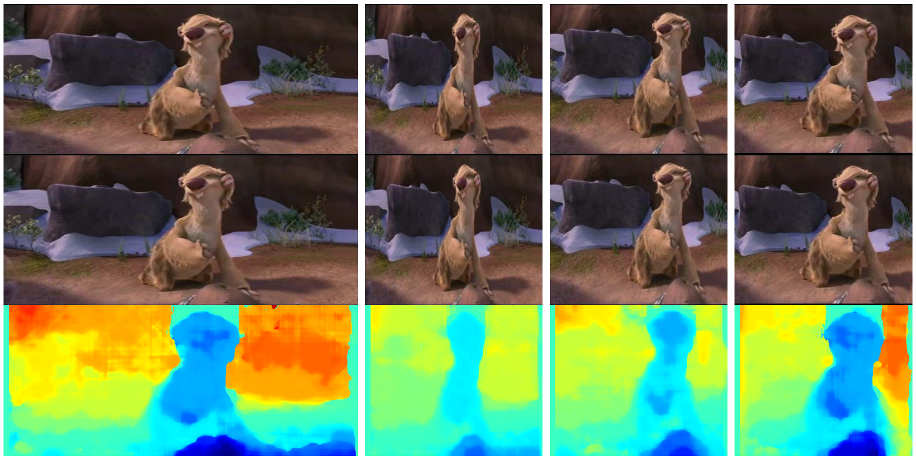

Original US CVW Ours w/ Crop.

Fig. 13. Performance comparison for *IceAge4*. Rows from top to bottom: the left-views, the right-views and disparity maps. The regions removed by individual retargeting methods are marked in the corresponding figures in the supplemental material.

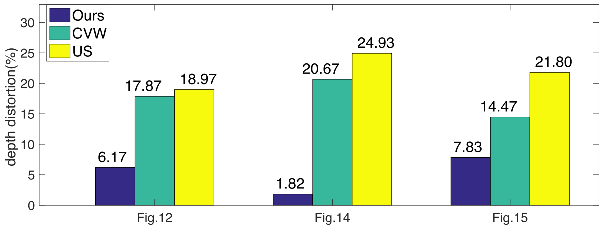


Fig. 14. Performance comparison for *Madagascar 4*. Rows from top to bottom: the left-views, the right-views and disparity maps. The regions removed by individual retargeting methods are marked in the figures in the supplemental material.

respectively. The user study results also show that the subjects do not notice significant temporal incoherence, although our method inconsistently resizes non-paired boundary regions temporally.

D. Quantitative Depth Distortion

We quantify depth distortion by the average difference between the disparity values of the retargeted video and their original values. The average disparity difference is further normalized to the range of the original disparity map as

$$\frac{1}{|d_{max}| \cdot N^{v}} \sum\nolimits_{(k,t)} |d_k^{z,t} - \tilde{d}_k^{z,t}|, \tag{15}$$

where $|d_k^{z,t} - \tilde{d}_k^{z,t}|$ is disparity difference of a point $p_k^{z,t}$, N^v is the total number of points, and $|d_{max}|$ is the range of the original disparity map.

As shown in Fig. 17, US and GTC both lead to high depth distortion due to poor depth preservation. In contrast, our method achieves the lowest depth distortion by mitigating the constraint conflicts.

E. Parameter Setting

Our stereo video retargeting method has two parameters α, β which mainly affect retargeting performance. In particular, parameter α and β are weights to control the strength of shape and depth constraints, respectively.

Following [29], we test various values of parameter α and β, to analyze the influence of these two parameters on the retargeting results as illustrated in Fig. 15 and Fig.16. As shown in Fig. 15, a small value of β corresponds to weak depth preservation constraints, which are ineffective to preserve the depth, thereby introducing severe depth distortions. In contrast, a high value of β effectively preserves the depth. Fig. 16 shows that a small to medium value of α leads to comparable performance of depth and shape preservation, but when α is too large, the shape-preserving constraints would significantly degrade the performance of depth preservation. Therefore, we empirically set $\beta = 10^5, \alpha = 1$ for all test videos in our experiments.

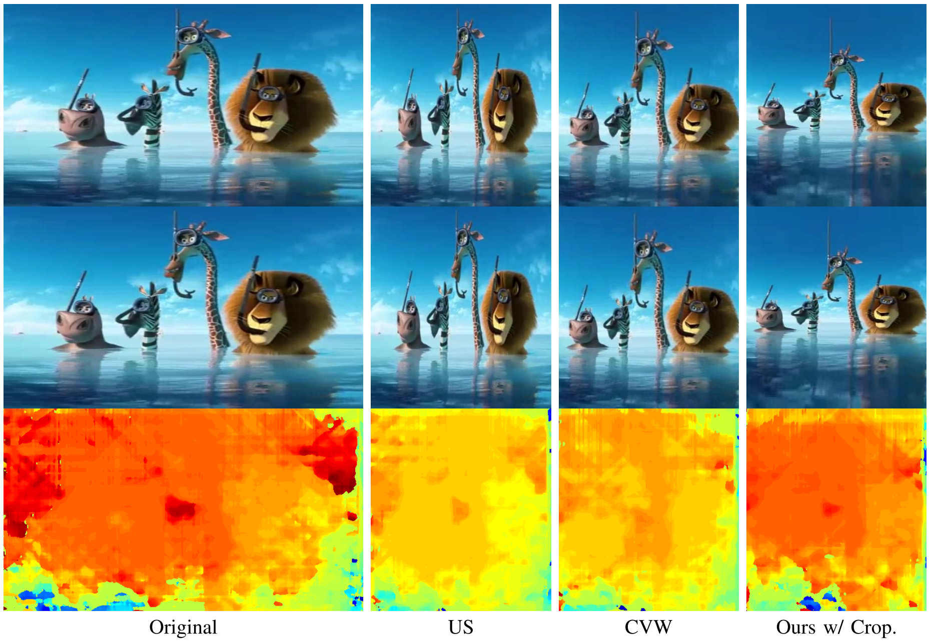


Fig. 15. Retargeting results using various values of weighting parameter β. From left to right: original, $\beta = 10^2, 10^4, 10^5$. Rows from top to bottom: the left-views, the right-views and disparity maps.

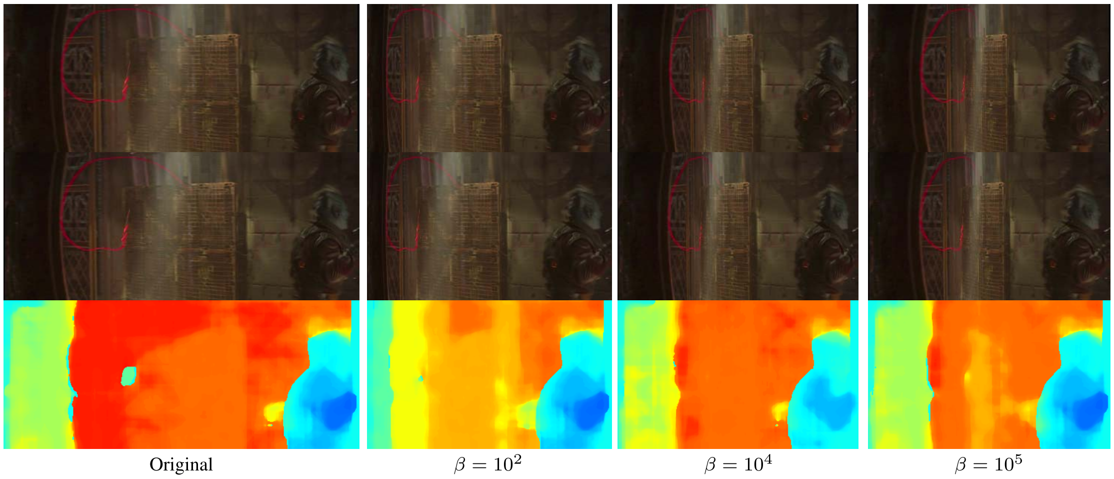


Fig. 16. Retargeting results using various values of weighting parameter α. From left to right: original $\alpha = 1, 10, 10^5$ Rows from top to bottom: the left-views, the right-views and disparity maps.

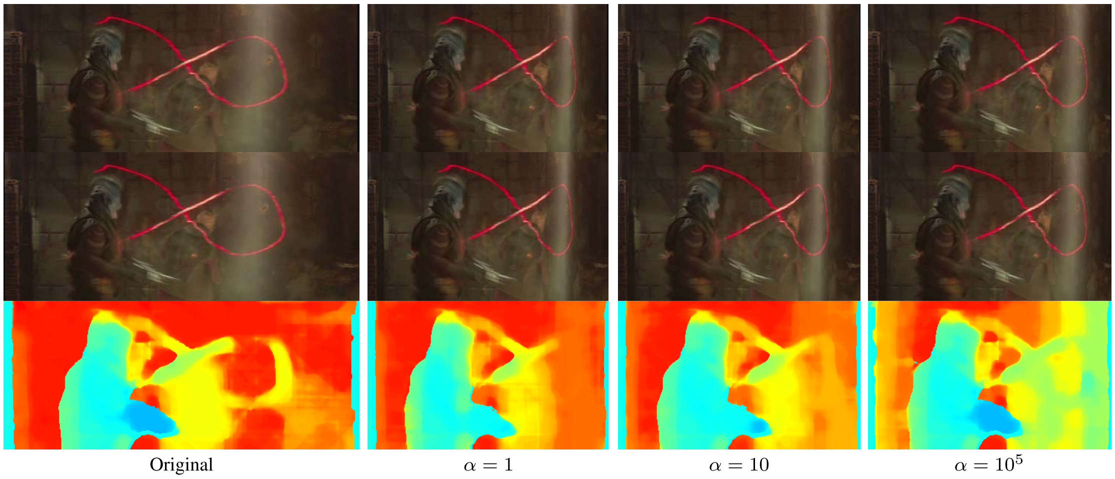


Fig. 17. Comparison of depth distortions of the proposed method, CVW, and US for three test videos.

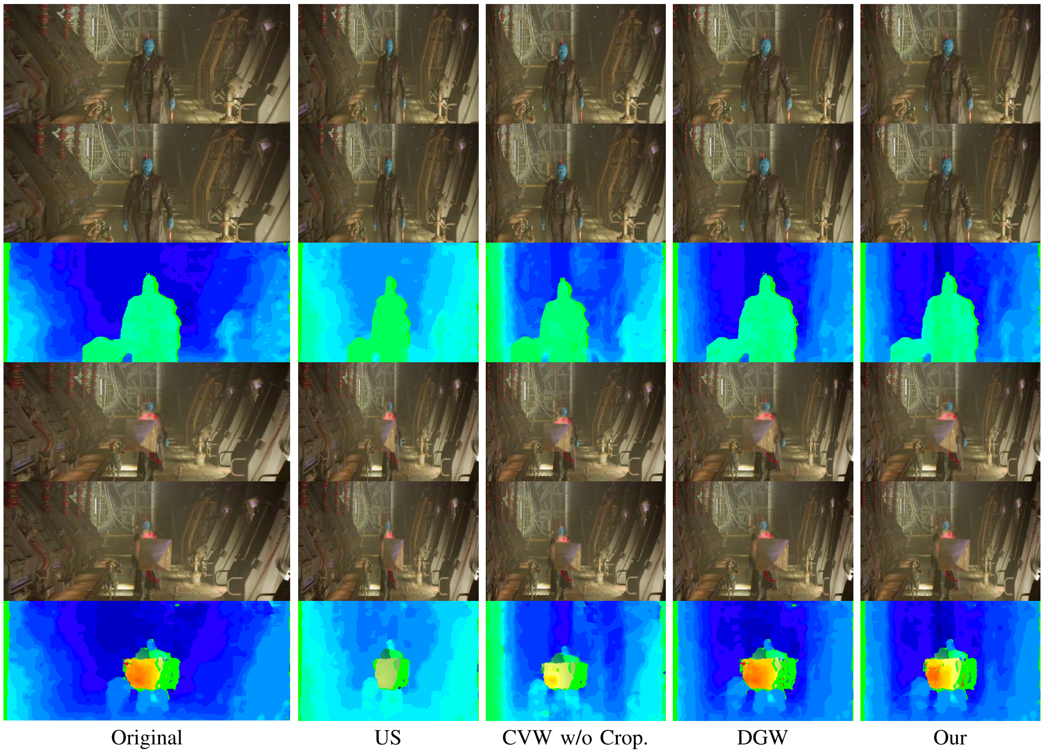


Fig. 18. Performance comparison for *Guardians of the Galaxy2*. The images in the rows 1–3 are left-views, the right-views and disparity maps of frame #17, and the images in rows 4–6 are that of frame # 41.

TABLE I
WINNING FREQUENCY MATRIX OF SUBJECTIVE PAIRWISE COMPARISON FOR TEN TEST VIDEOS

	Ours	US	CVW	Total
Ours	–	91.67%	83.33%	87.50%
US	8.33%	–	17.50%	12.92%
CVW	16.67%	82.50%	–	49.58%

TABLE II
TIME COST

Grid size	Method [12]	Ours
20×30	1.38s	0.156s

F. Time Cost

Since we proposes a key-frame optimization which achieve lower computational complexity, compared with that in [12]. We hence test the time cost of the optimization, to demonstrate the improvement on run-time complexity. As shown in Tab. II, our method is about 9 times faster than the previous version [12] when processing a 80-frame video shot, while achieving comparable visual quality.

VI. DISCUSSIONS AND LIMITATIONS

Different from DGW proposed in [28], our method does not focus on preserving the temporal depth dynamics of 3D objects. As a results, our method may not perform as well as DGW in depth preservation for stereo videos containing significant depth changes at the temporal domain. In particular, Fig. 18 shows two frames selected from a challenging shot of a live-action film. The video shot involves significant object motions along the depth direction, which leads to large depth changes temporally. Compared with US and CVW, our method achieves superior performance on shape preservation, depth preservation and temporal coherence. However, compared with DGW, although our method well preserves the depth for some frames, overall DGW achieves slightly better performance in depth preservation at both the spatial and temporal domains (e.g., temporal depth dynamics). This is because DGW explicitly considers the temporal depth dynamics of 3D objects, that is not the central focus of our paper. By contrast, however, as illustrated in Fig.9 of our paper, should a stereo video be downsized significantly or involve significant camera/object motions, DGW often leads to significantly more resource conflicts in the requirements of preserving shape, depth, and temporal coherence. In our future work, we will extend our method to effectively preserve the temporal depth dynamics of 3D objects without incurring excessive constraint conflicts.

In addition,for stereo videos where salient objects or their trajectory occupy a large portion of a frame, our method would inevitably introduce shape/depth distortions, similar to state-of-the-art warping-based methods. Although we can combine our method with cropping, the retargeted results generated by our method would be similar to that of cropping. That is, some important content/objects may be removed, since there are not enough less-important regions (i.e., "retargeting resource") for absorbing the depth/shape distortions due to grid warping.

VII. Conclusion

A novel stereo video retargeting method was proposed in this work. It can offer temporal coherence, shape preservation and depth preservation for stereo visual contents that contain various motion types, salient objects in a wide depth range and significant temporal depth changes. As compared with existing methods, our method can preserve shape and depth information better while well maintaining temporal coherence perceptually. This is achieved by adaptively relaxing temporal constraints on non-paired boundary regions to effectively mitigate conflicts among the shape, depth, and temporal constraints. Based on the proposed method, we have formulated a grid-warping-based optimization problem and proposed an efficient keyframe-based algorithm to solve it. As demonstrated by extensive experiments, our method outperforms other existing methods in general and on videos with significant camera motions in particular.

References

[1] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Trans. Graph.*, vol. 29, no. 4, p. 1, Jul. 2010.

[2] T. Yan, R. W. H. Lau, Y. Xu, and L. Huang, "Depth mapping for stereoscopic videos," *Int. J. Comput. Vis.*, vol. 102, nos. 1–3, pp. 293–307, Mar. 2013.

[3] Y. Niu, W.-C. Feng, and F. Liu, "Enabling warping on stereoscopic images," *ACM Trans. Graph.*, vol. 31, no. 6, p. 1, Nov. 2012.

[4] T. D. Basha, Y. Moses, and S. Avidan, "Stereo seam carving a geometrically consistent approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2513–2525, Oct. 2013.

[5] K.-Y. Lee, C.-D. Chung, and Y.-Y. Chuang, "Scene warping: Layer-based stereoscopic image resizing," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recongnit.*, Jun. 2012, pp. 49–56.

[6] S.-S. Lin, C.-H. Lin, S.-H. Chang, and T.-Y. Lee, "Object-coherence warping for stereoscopic image retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 759–768, May 2014.

[7] K. He, H. Chang, and J. Sun, "Content-aware rotation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 553–560.

[8] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph. (TOG)*, vol. 27, no. 3, pp. 1–9, Aug. 2008.

[9] T.-C. Yen, C.-M. Tsai, and C.-W. Lin, "Maintaining temporal coherence in video retargeting using mosaic-guided scaling," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2339–2351, Aug. 2011.

[10] Y.-S. Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Trans. Graph. (TOG)*, vol. 28, no. 5, pp. 1–10, Dec. 2009.

[11] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee, "Scalable and coherent video resizing with per-frame optimization," *ACM Trans. Graph.*, vol. 30, no. 4, p. 1, Jul. 2011.

[12] B. Li, C.-W. Lin, S. Liu, T. Huang, W. Gao, and C.-C.-J. Kuo, "Perceptual temporal incoherence aware stereo video retargeting," in *Proc. ACM Multimedia Conf. Multimedia Conf. - MM*, 2018.

[13] A. Shamir and O. Sorkine, "Visual media retargeting," in *Proc. ACM SIGGRAPH ASIA Courses SIGGRAPH ASIA*, 2009, p. 11.

[14] K. Utsugi, T. Shibahara, T. Koike, K. Takahashi, and T. Naemura, "Seam carving for stereo images," in *Proc. 3DTV-Conf., True Vis. Capture, Transmiss. Display 3D Video*, Jun. 2010, pp. 1–4.

[15] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang, "Content-aware display adaptation and interactive editing for stereoscopic images," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 589–601, Aug. 2011.

[16] J. W. Yoo, S. Yea, and I. K. Park, "Content-driven retargeting of stereoscopic images," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 519–522, May 2013.

[17] B. Li, L.-Y. Duan, C.-W. Lin, T. Huang, and W. Gao, "Depth-preserving warping for stereo image retargeting," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2811–2826, Sep. 2015.

[18] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, "Depth-preserving stereo image retargeting based on pixel fusion," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1442–1453, Jul. 2017.

[19] C.-H. Chang and Y.-Y. Chuang, "A line-structure-preserving approach to image resizing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1075–1082.

[20] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Comput. Graph. Forum*, vol. 28, no. 7, pp. 1897–1906, Oct. 2009.

[21] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graph.*, vol. 27, no. 5, p. 1, Dec. 2008.

[22] B. Li, L.-Y. Duan, J. Wang, R. Ji, C.-W. Lin, and W. Gao, "Spatiotemporal grid flow for video retargeting," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1615–1628, Apr. 2014.

[23] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph. (TOG)*, vol. 28, no. 5, pp. 1–10, Dec. 2009.

[24] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2007, pp. 1–6.

[25] Y. Zhang, S. Hu, and R. R. Martin, "Shrinkability maps for Content-Aware video resizing," *Comput. Graph. Forum*, vol. 27, no. 7, pp. 1797–1804, Oct. 2008.

[26] B. Yan, K. Sun, and L. Liu, "Matching-Area-Based seam carving for video retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 302–310, Feb. 2013.

[27] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. Graph.*, vol. 29, no. 4, p. 1, Jul. 2010.

[28] B. Li, C.-W. Lin, B. Shi, T. Huang, W. Gao, and C.-C.-J. Kuo, "Depth-aware stereo video retargeting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6517–6525.

[29] S.-S. Lin, C.-H. Lin, Y.-H. Kuo, and T.-Y. Lee, "Consistent volumetric warping using floating boundaries for stereoscopic video retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 801–813, May 2016.

[30] S. Kopf, B. Guthier, C. Hipp, J. Kiess, and W. Effelsberg, "Warping-based video retargeting for stereoscopic video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2898–2902.

[31] S. Anstis, "Picturing peripheral acuity," *Perception*, vol. 27, no. 7, pp. 817–825, Jul. 1998.

[32] A. Duchowski, *Eye Tracking Methodology: Theory Practice*, vol. 373. London, U.K.: Springer-Verlag, 2007.

[33] H. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017.

[34] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C.-C.-J. Kuo, "Measure and prediction of HEVC perceptually Lossy/Lossless boundary QP values," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 42–51.

[35] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1409–1423, Jul. 2009.

[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[37] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[38] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[39] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[40] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[41] B. Li, L.-Y. Duan, J. Wang, J. Chen, R. Ji, and W. Gao, "Grid-based retargeting with transformation consistency smoothing," in *Proc. Multimedia Modeling*, 2011, pp. 12–24.

[42] B. Li, Y. Chen, J. Wang, L.-Y. Duan, and W. Gao, "Fast retargeting with adaptive grid optimization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–4.

[43] Z. Qiu, T. Ren, Y. Liu, J. Bei, and M. Song, "Image retargeting by combining region warping and occlusion," in *Proc. Pacific-Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2013, pp. 200–210.

[44] A. Mansfield, P. Gehler, L. Van Gool, and C. Rother, "Scene carving: Scene consistent image retargeting," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 143–156.

[45] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1404–1412.

[46] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

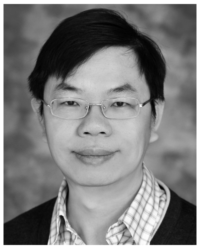
[47] B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema From Script to Screen*. Milton Park, Didcot, U.K., : Taylor & Francis, 2009.

[48] *Subjective Methods for the Assessment of Stereoscopic 3Dtv Systems*, document ITU-R BT.2021, May 2012.

[49] C.-C. Hsu, C.-W. Lin, Y. Fang, and W. Lin, "Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 377–389, Jun. 2014.

[50] Y.-H. Lin and J.-L. Wu, "Quality assessment of stereoscopic 3D image compression by binocular integration behaviors," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1527–1542, Apr. 2014.

Bing Li (Member, IEEE) received the B.S. degree in computer science from Jinan University, Guangzhou, China, in 2009, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2016. She worked as a Postdoctoral Fellow with the University of Southern California, USA, in 2016. She is currently a Postdoctoral Fellow with the King Abdullah University of Science and Technology. Her research interests include image/video processing, computer Vision, and machine learning.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Professor with the Department of Electrical Engineering, and the Institute of Communications Engineering, NTHU. His research interests include image and video processing, computer vision, and video networking. He has served as a Distinguished Lecturer for IEEE Circuits and Systems Society, from 2018 to 2019, a Steering Committee member for the IEEE Transactions on Multimedia from 2014 to 2015, and the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles received the Best Paper Award of IEEE VCIP 2015, Top 10% Paper Awards of IEEE MMSP 2013, and the Young Investigator Award of VCIP 2005. He received the Young Investigator Award presented by Ministry of Science and Technology, Taiwan, in 2006. He is also the Chair of the Steering Committee of IEEE ICME. He has been serving as the President of the Chinese Image Processing and Pattern Recognition Association, Taiwan, since 2019. He has served as a Technical Program Co-Chair for IEEE ICME 2010, and a General Co-Chair for IEEE VCIP 2018, and a Technical Program Co-Chair for IEEE ICIP 2019. He has served as an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Multimedia, the IEEE Multimedia, and the *Journal of Visual Communication and Image Representation*.

Shan Liu (Member, IEEE) received the B.Eng. degree in electronics engineering from Tsinghua University, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California. She was the formerly Director of Multimedia Technology Division, MediaTek, USA. She was also formerly with MERL, Sony, and IBM. She is currently a Tencent Distinguished Scientist and a General Manager of Tencent Media Lab. She has been actively contributing to international standards since the last decade. She has numerous proposed technologies adopted into various standards, such as HEVC, VVC, OMAF, MPEG-DASH, and PCC, and has served the Co-Editor for HEVC/H.265 v4 (a.k.a. HEVC SCC) and the emerging VVC. At the same time, technologies and products developed by Dr. Liu and her team are serving multiple millions of users daily. She holds more than 100 granted U.S. and global patents and authored more than 70 peer-reviewed technical articles. She was in the committee of Industrial Relationship of IEEE Signal Processing Society from 2014 to 2015. She has been serving as an Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology since 2018, and the Vice Chair of IEEE1857 standard WG since 2019. She also served the VP of Industrial Relations and Development of Asia-Pacific Signal and Information Processing Association from 2016 to 2017, and was named an APSIPA Industrial Distinguished Leader in 2018.

Tiejun Huang (Senior Member, IEEE) received the bachelor's and master's degrees in computer science from the Wuhan University of Technology in 1992 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong (Central China) University of Science and Technology in 1998. He is currently a Professor with the School of Electronic Engineering and Computer Science, the Chair of Department of Computer Science, and the Director of the Institute for Digital Media Technology, Peking University. His research areas include video coding and image understanding. Prof. Huang received the National Science Fund for Distinguished Young Scholars of China in 2014. He is a member of the Board of the Chinese Institute of Electronics, the Board of Director for Digital Media Project, and the Advisory Board of IEEE Computing Now.

Wen Gao (Fellow, IEEE) received the Ph.D. degree in electronics engineering from the University of Tokyo, Japan, in 1991. He was a Professor with the Harbin Institute of Technology, from 1991 to 1995, and the Institute of Computing Technology, Chinese Academy of Sciences, from 1996 to 2006. He is currently a Professor of Computer Science with Peking University. He has authored extensively, including five books and over 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, computer vision, multimedia retrieval, multimodal interface, and bioinformatics. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE International Symposium on Circuits and Systems (ISCAS), ICME, and the ACM Multimedia, and also served on the Advisory and Technical Committee of numerous professional organizations. He served or serves on the Editorial Board of several journals, such as the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transaactions on Multimedia, the IEEE Transactions on Autonomous Mental Development, the *EURASIP Journal of Image Communications, the Journal of Visual Communication, and Image Representation*.

C.-C. Jay Kuo (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively. He is currently the Director of the Multimedia Communications Laboratory and a Distinguished Professor of electrical engineering and computer science with the University of Southern California, Los Angeles, CA, USA. His research interests include multimedia computing and machine learning. Dr. Kuo is a Fellow of the American Association for the Advancement of Science (AAAS) and The International Society for Optical Engineers (SPIE).