Depth-Aware Stereo Video Retargeting

Bing Li[†], Chia-Wen Lin[‡], Boxin Shi[§], Tiejun Huang[§], Wen Gao[§], C.-C. Jay Kuo[†]
[†]University of Southern California, Los Angeles, California, USA
[‡] Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
[§]National Engineering Lab for Video Technology, School of EECS, Peking University, China

Abstract

As compared with traditional video retargeting, stereo video retargeting poses new challenges because stereo video contains the depth information of salient objects and its time dynamics. In this work, we propose a depth-aware stereo video retargeting method by imposing the depth fidelity constraint. The proposed depth-aware retargeting method reconstructs the 3D scene to obtain the depth information of salient objects. We cast it as a constrained optimization problem, where the total cost function includes the shape, temporal and depth distortions of salient objects. As a result, the solution can preserve the shape, temporal and depth fidelity of salient objects simultaneously. It is demonstrated by experimental results that the depth-aware retargeting method achieves higher retargeting quality and provides better user experience.

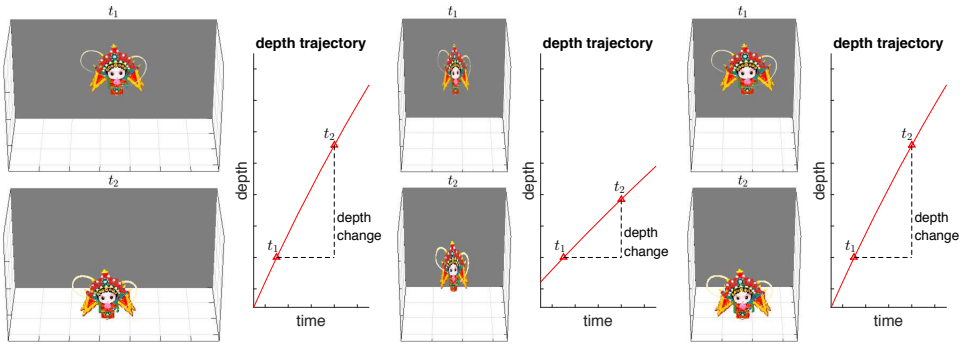
1. Introduction

3D video contents and display technologies become mature nowadays, and they offer real-world viewing experience to users. With this trend, many companies are manufacturing 3D display devices of different sizes suitable for different application environments such as theaters, TVs, and computers. Furthermore, virtual/augmented-reality devices (e.g., Google Cardboard and Oculus Rift) adopt stereo video to create immersive environments. It is however a nontrivial task to allow the same stereo content to be displayed on screens of different sizes and/or aspect ratios automatically, which is known as the resizing technique.

As compared with 2D video retargeting, stereo video retargeting poses new challenges because stereo video contains the depth information of salient objects and its time dynamics. This is particularly true when the object has a movement along the depth direction as illustrated in Fig. 1, leading to poorer 3D viewing experience. Generally speaking, there are two key factors that influence human 3D viewing experience greatly. They are the correct depth information at each time instance (i.e., a single frame) and the correct depth dynamics across multiple frames. The former determines the distance of a 3D object to the screen while the latter indicates the motion direction and speed in a 3D scene. In order to provide satisfactory 3D viewing experience, we have to design a stereo video retargeting method that takes these factors into account on top of the requirements of shape preservation and temporal coherence of traditional 2D video retargeting.

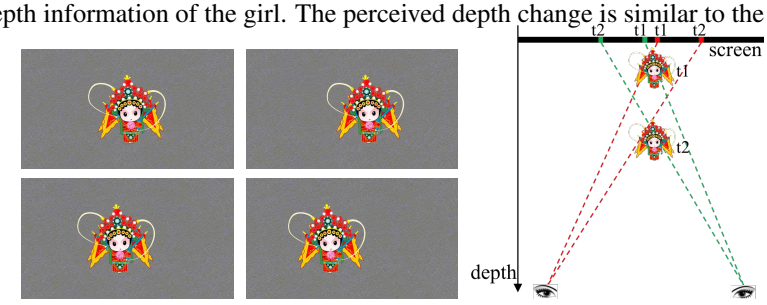
It is worth pointing out that most existing video retargeting methods do not impose the depth fidelity constraint. For example, the uniform scaling method is widely used for stereo video resizing. It up- or down-samples the left and right 2D videos of a stereo video, respectively. However, the depth of a salient object can be distorted. The comparison of the uniform scaling and the proposed depth-aware retargeting schemes is shown in Fig. 1. Uniform scaling not only shrinks the girl's size but does not capture her movement along the depth direction properly. The perceived depth change is relatively small. Clearly, the stereo video retargeting problem cannot be solved by traditional 2D video retargeting methods since they do not preserve the depth information of salient objects by analyzing the left- and right-views jointly (see Fig. 2).

In this work, we propose a depth-aware stereo video retargeting method that achieves high-quality retargeting performance by preserving the depth information of the original stereo video. This method infers the depth information and include its deviation in the total cost function for minimization. Furthermore, it adopts a grid warping scheme to facilitate the optimization framework. To the best of our knowledge, this is the first work on stereo video retargeting by considering the depth-preserving constraints. The constraints are simple and flexible to apply. Once salient objects are detected, the algorithm will preserve their depth information as faithfully as possible. It enhances users' 3D viewing experience of retargeted stereo video. Experimental results are given to demonstrate the superior performance of the proposed depth-aware retargeting method.



(a) Original (b) Uniformly Scaling (c) Depth-Aware Retargeting

Figure 1: Illustration of the advantage of the proposed depth-aware retargeting method, where the left columns in (a)-(c) show the 3D scene at frames t_1 and t_2 while the right columns in (a)-(c) show the temporal trajectories of the foreground object in the 3D scene. Uniform scaling only shrinks the girl's size but does not capture her movement along the depth direction properly. The perceived depth change is relatively small. In contrast, the proposed depth-aware retargeting method preserves both the shape and the depth information of the girl. The perceived depth change is similar to the original one.



(a) 2D frame I^{z,t_1} (b) 2D frame I^{z,t_2} (c) 3D scene

Figure 2: Illustration of a foreground object (a girl) in a 3D scene at two time instances t_1 and t_2. Their left and right views are shown in the top and bottom rows of (a) and (b) while the red and green dots on the screen of (c) indicate the x-coordinates of the girl in left and right 2D frames, respectively. Based on the disparity in the x-coordinates, we can determine the depth of the girl. The girl moves towards the camera from t_1 to t_2. Since the girl moves horizontally in each individual video as shown in the top (or bottom) row alone, traditional video retargeting algorithms cannot recover the depth change information of the stereo video.

2. Related Work

2D video retargeting. Content-aware retargeting methods can be categorized into discrete and continuous approaches [33]. Many discrete methods have been proposed for 2D videos. Cropping-based methods [11] crop a rectangular region from each 2D frame while seam carving methods [32, 33, 26] iteratively remove or insert seams. Continuous methods [14, 35, 38, 36] divide video frames into pixels or regions and warp them under the guidance of importance maps. Discrete methods tend to introduce noticeable distortion to structural objects. Continuous methods generally preserve the object shape better due to their continuous warping mechanism. According to the information used in temporal constraints, one can classify 2D video retargeting methods to local and global ones by following [20]. Local methods [14, 36] resize each frame coherently with its neighboring frames in a local time window. Global methods [20, 35, 38] exploit the temporal information of an object throughout the entire video shot for coherent resizing, thereby achieving better shape coherency along time than local methods.

Stereo image retargeting. Basha *et al.* [2] and Shen et al. [34] extended the seam carving method [32, 33, 1] to stereo image retargeting by iteratively removing a pair of seams from a stereo image pair. Several continuous methods were proposed in [4, 16, 19, 17]. They extended the warping-based 2D image retargeting methods (e.g. [6, 39]) to stereo image pairs by imposing additional depth-preserving constraints. These methods [4, 16] attempted to preserve the depth of the whole image by maintaining the depths of a set of sparse correspondences. The idea is similar to that of the depth editing methods [15, 37] while depth distortions may occur. Methods [12, 25] were proposed to remap the depth. Li *et al.* [19] imposed effective depth-preserving constraints on grid warping to achieve better depth preservation performance.

Stereo video retargeting. Research on stereo video re-

targeting is much fewer compared to stereo image retargeting, since it is more complicated due to additional requirements on temporal coherence in both shape and depth. Kopf *et al.* [13] treated a stereo video as two individual 2D videos, and formulated a 2D video retargeting method to resize the stereo video. For stereo videos (especially for those with salient objects or their moving occupying a large portion of a frame), Lin *et al.* [22] proposed to combine cropping with the grid-based 2D video retargeting. Nevertheless, since the above methods do not explicitly consider the depth information of 3D objects and its time dynamics, they usually produce severe depth artifacts in a retargeted stereo video.

3. Depth-Aware Stereo Video Retargeting

3.1. Problem Formulation

For stereo video retargeting, we adapt a stereo video sequence to a target display size and attempt to maximize users' 3D viewing experience. Given a 3D scene fused from the left- and right-views of a stereo video, 3D objects in the scene have two key attributes – shape and depth. These attributes change along time due to camera motions and object motions. In the current literature, content-aware 2D video retargeting methods [35, 38, 36] were proposed to ensure faithful preservation of the shape information spatially and temporally. That is, shapes of salient 3D objects are preserved at each frame and objects' shapes are coherently resized across frames. However, preservation of the time-varying depth attribute is often neglected. In contrast with existing methods that preserve visual contents of the left- and right-views separately, we seek for a solution that preserves the shape and the depth information of salient objects and their time dynamics in the original content as much as possible.

3D objects have different depths in a 3D scene. When humans fixate an object with two eyes (e.g., P in Fig. 3), the visual axes of eyes intersect at the object and the horopter is formed [9]. Objects behind or in front of the fixated objects are blurred. Since the human visual system (HVS) fixates a salient 3D object at a time when humans watch a stereo video, other objects are blurred and their other depth can be altered to some extent as long as there is little negative viewing experience. Furthermore, the temporal depth variation of non-salient objects and background is affected by objects' and camera's motions. It is important to scale their depths across frames coherently. Otherwise, incoherent scaling among frames often result in incorrect motion direction (e.g., confusion between moving into or out-of the screen of non-salient objects). It may also incur temporal depth discontinuity and human inability to perceive the depth due to vergence-accommodation conflicts and extended reading time [15, 24, 27]. In this work, being motivated by the above-mentioned characteristics of the HVS, we propose a depth-aware retargeting solution that not only preserves the shape and depth of salient 3D objects but also scales the whole 3D scene along time coherently.

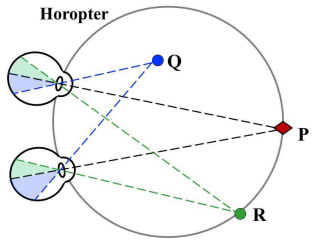


Figure 3: Objects beyond the fixation point are blurred in the HVS.

To obtain a high-quality stereo video retargeting algorithm, we formulate the following minimization problem:

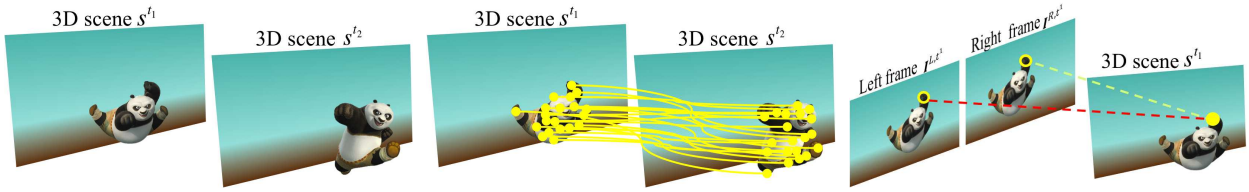
$$\min E = \min(E^S + \lambda^S \cdot E^T + \lambda^D \cdot E^D), \qquad (1)$$

where E is the total distortion, E^S, E^T and E^D denote the distortions caused by spatial shape incoherence (i.e., shape distortion), temporal shape incoherence (i.e., temporal distortion) and loss of the 3D depth information of salient objects (i.e., depth distortion), respectively, λ^S and λ^D are weighting factors. The derivation of the depth distortion, E^D, will be detailed in Sec. 3.2 while that of E^S and E^T will be given in Sec. 3.3.

Grid-based warping [15, 16, 30] has proven to be an effective means for image and video retargeting. It divides each frame into a grid mesh, and translates the problem of finding an optimal retargeted stereo video to the search of the optimal warped mesh set in all frames that minimizes the total distortion E in Eq. (1). This optimization procedure involves the search for a large number of parameters and consumes a lot of memory and time in stereo video retargeting. To lower the complexity, we adopt the axis-aligned warping scheme in [20] instead. It uses the grid width and height as parameters to control warping, and demands all retargeted grids in each column and row to be of the same width and height, respectively. As compared to those methods that use grid vertices as control parameters of a warping function, the parameter number of the grid-edge-based warping is reduced significantly. We use $w_k^{z,t}$ and $h_i^{z,t}$ to denote the width and height of a grid, denoted by $g_k^{z,t}$, in the original grid mesh, respectively. Then, getting the optimal warped mesh set is to determine $\tilde{w}_k^{z,t}$ and $\tilde{h}_k^{z,t}$ of a retargeted grid that minimize the total distortion.

3.2. Depth Distortion

We will focus on the derivation of the depth distortion, E^D, in this section, and discuss two other distortions, E^S and E^T, in Sec. 3.3. The depth distortion is used to preserve the depth information of salient objects in a stereo video.



(a) Original 3D scene (b) 3D object is represented by 3D points (c) 3D point is represented by 2D points

Figure 4: The representation of a 3D object "panda" by dense 3D points: (a) the 3D panda in s^{t_1} and s^{t_2} two scenes, (b) the 3D panda is discretized into 3D points indicated by yellow circles while the corresponding 3D points between 3D scene s^{t_1} and s^{t_2} are connected by yellow lines, and (c) a 3D point is represented by its corresponding 2D points in the left and right frames where the yellow rings indicate 2D points.

We consider both spatial depth fidelity at individual frames and temporal depth fidelity across multiple frames.

Parts of a 3D object may exhibit different movements in depth along time. One example is illustrated in Fig. 4 (a), where we show a 3D panda in s^{t_1} and s^{t_2} two scenes. It is apparently insufficient to represent the depth trajectory at the object level. Instead, we need to discretize it into representative 3D points and examine their depth trajectories as shown in Fig. 4 (b). For a 3D object decomposed into N points, E^D is measured by the weighted sum of distortions of N depth trajectories in form of

$$E^D = \sum_{i}^{N} s_i \cdot E_i^D, \tag{2}$$

where E_i^D is the distortion of the depth trajectory of the i-th point and s_i is the weight used to control the importance of the i-th depth trajectory. We calculate s_i by averaging the saliency values of 3D points in the i-th depth trajectory, where the saliency values are defined in Sec. 3.3

Let $d^{i,t}$ and $\tilde{d}^{i,t}$ be the depth values of the i-th point in the t-th frame of the original and its retargeted video, and $\mathbf{d}^i = \{d^{i,t}\}$ and $\tilde{\mathbf{d}}^i = \{\tilde{d}^{i,t}\}$ be the original and its retargeted depth trajectories of the i-th point across multiple frames of interest, respectively. Each depth trajectory is assumed to be continuous and first-order differentiable.Then, the distortion of the i-th depth trajectory is defined as

$$E_i^D = \sum_{t} \Big((\Delta d^{i,t})^2 + (\Delta \frac{\partial \mathbf{d}^i}{\partial t}|_t)^2 \Big), \tag{3}$$

where the first term

$$\Delta d^{i,t} = d^{i,t} - \tilde{d}^{i,t} \tag{4}$$

captures the spatial depth distortion while the second term

$$\Delta \frac{\partial \mathbf{d}^i}{\partial t}|_t = \frac{\partial \mathbf{d}^i}{\partial t}|_t - \frac{\partial \tilde{\mathbf{d}}^i}{\partial t}|_t \tag{5}$$

captures the temporal depth change distortion of trajectory $\mathbf{d}^i$ in the t-th frame. We use the following approximations

$$\begin{aligned} \frac{\partial \mathbf{d}^i}{\partial t}|_t &= d^{i,t} - d^{i,t+1}, \\ \frac{\partial \tilde{\mathbf{d}}^i}{\partial t}|_t &= \tilde{d}^{i,t} - \tilde{d}^{i,t+1}, \end{aligned}$$

in Eq. (5) for computation.

As mentioned in Sec. 3.1, distortion E_i^D in Eq. (3) should be expressed as a function of the grid width and height for the purpose of minimization. A 3D point is formed by its proper correspondence between the left- and right-views as shown in Fig. 4(c). It is well known that the depth of a 3D point, d_i^t, increases as its disparity along the x-axis becomes larger. Thus, one can represent E_i^D as a function of grid width; namely, $E_i^D(w^{z,t}, \tilde{w^{z,t}})$, by relating horizontal disparity to the difference of the grid width.

To obtain E^D in Eq. (2), one way is to detect all 3D points between each stereo pair, $I^{L,t}$ and $I^{R,t}$, and calculate $d^{i,t}$ using a disparity estimation algorithm. Afterwards, we align the corresponding 3D points across frames via motion estimation algorithms so as to calculate $\partial d^{i,t}/\partial t$. This approach relies on *dense* motion estimation and disparity estimation. However, estimating E^D from dense points not only consumes large amounts of memory, but also makes Eq. (1) a large-scale optimization model, leading to a high computational cost. In addition, for video frames containing large textureless regions, motion/disparity estimation may become unreliable. As a result, incorrect cost functions E_i^D can be yielded, thereby significantly degrading the visual quality of retargeted stereo video.

Instead, we aim to estimate E_i^D based on a small number of reliable 3D control points. This can be accomplished by directly representing E_i^D by sparse depth trajectories, whose performance, however, can be easily degraded by the inaccuracy of depth trajectory tracking caused by severely inaccurate motion estimation. To address this issue, we adopt a spline-interpolation-like scheme that approximates 1D curves and 2D surfaces by a few control points via grid warping. This involves two tasks. First, we use a small number of control points to build the correspondence in the spatial (left/right) and temporal domains. Second, we use the grid warping technique to approximate the depth map and its time dynamics, so as to establish the correspondence of dense 3D points and avoid depth distortions caused by tracking errors.

By combining Eqs. (2) and (3), we can express E^D as

$$E^D = \tau \cdot \big(E^C + E^W\big), \quad \tau \equiv \min_i s_i, \tag{6}$$

where

$$E^C = \sum_{i \in \mathbf{C}} \sum_t \big(\frac{s_i - \tau}{\tau} \cdot (\Delta \tilde{d}^{i,t})^2 + \frac{s_i}{\tau} \cdot (\Delta \frac{\partial \tilde{d}^i}{\partial t})^2\big)$$

is used to capture the depth distortion of a set of selected control points, denoted by $\mathbf{C}$ and

$$E^W = \sum_i \sum_t (\Delta \tilde{d}^{i,t})^2$$

is used to capture the grid warping distortion over space and time.

For E^C, there are several ways to select reliable control points from a dense set of 3D points. For example, we can remove 3D points from the textureless region and extract local features such as the SIFT and SURF features to match representative points in the two views. We adopt the SIFT features in our implementation, due to its promising performance reported in [15, 28]. Note that the number of tracked SIFT points may be too few on some frames, due to some factors such as significant motions, blurring, and occlusions. We address such problem by reinitializing new trajectories, or by skipping such frames and generating retargeting results by using interpolation like [20].

We adopt the grid warping scheme as proposed in [19] and express E^W as the spatial depth distortion of each frame by interchanging the summation order of t and i; namely,

$$E^W = \sum_i \sum_t (\Delta \tilde{d}^{i,t})^2 = \sum_t \sum_i (\Delta \tilde{d}^{i,t})^2 \cdot \Gamma_{i,t} \tag{7}$$

where $\Gamma_{i,t}$ is an indicator that indicates whether trajectory $\mathbf{d}^i$ appears in scene S^t. Then, $\sum_i (\Delta \tilde{d}^{i,t})^2 \cdot \Gamma_{i,t}$ is the total depth distortion of all 3D points from the frame pair of $I^{L,t}$ and $I^{R,t}$. By demanding

$$\sum_i (\Delta \tilde{d}^{i,t})^2 \cdot \Gamma_{i,t} = 0,$$

we get the following two warping constraints [19]:

$$\begin{cases} \sum\limits_{q_k^{z,t} \in \bar{r}_j^z} \tilde{w}_k^{z,t} = \sum\limits_{q_k^{z,t} \in \bar{r}_j^z} w_k^{z,t}, & \forall \bar{r}_j^z \in \bar{\Upsilon} \\ \tilde{w}_k^{L,t} - w_k^{L,t} = \tilde{w}_k^{R,t} - w_k^{R,t}, & \forall g_k^{z,t} \in r_j^z, \forall r_j^z \in \Upsilon \end{cases} \tag{8}$$

where $\bar{\Upsilon}$ and Υ denote the set of non-paired regions $\bar{r}_j^z$ and the set of paired regions r_j^z, respectively. Nevertheless, the conditions in (8) are too strong to be directly used in E^W since the requirement of no spatial depth distortion often conflicts with the spatial and temporal shape-preserving constraints in Sec. 3.3. Hence, we abandon the hard form as given in Eq. (8) and adopt the corresponding soft form (i.e. square differences). As a result, we have the grid warping distortion as

$$\begin{aligned} E^W = \sum_t \Big(& \sum_{q_k^z \in \bar{r}_j^z} \varpi_k (\tilde{w}_k^{z,t} - w_k^{z,t})^2 + \\ & \sum_{r_j^z \in \Upsilon} \sum_{q_k^z \in r_j^z} (\tilde{w}_k^{L,t} - w_k^{L,t} - (\tilde{w}_k^{R,t} - w_k^{R,t}))^2 \Big). \end{aligned} \tag{9}$$

3.3. Spatio-Temporal Shape Distortions

To preserve the shape of salient objects, we define the shape distortion of stereo video as the total shape distortion of all grids. Since the retargeted grids remain rectangular, the shape distortion of grid $g_k^{z,t}$ can be simply measured by the difference between the original aspect ratio and that of the retargeted version as [20][18] :

$$\begin{aligned} E^S &= \sum_z \sum_t \sum_i D(g_k^{z,t}) \\ &= \sum_z \sum_t \sum_i \|w_k^{z,t} \cdot \tilde{h}_k^{z,t} - \tilde{w}_k^{z,t} \cdot h_k^{z,t})\|^2 \cdot \delta_k^{z,t}, \end{aligned} \tag{10}$$

where $s_k^{z,t}$ is the saliency value of $g_k^{z,t}$ calculated by averaging the saliency values of all pixels in $g_k^{z,t}$. Similar to [15], we calculate pixel saliency $\delta_k^{z,t}$ using a weighted sum of the image-based saliency[8][40] and the disparity-based saliency [15].

To maintain the temporal shape coherence of 3D objects, we have to resize each corresponding object across the left and right views coherently. Given $g_k^{z,t}$ in $I^{z,t}$, we can align it with the corresponding grid $g_j^{z,t'}$ in $I^{z,t'}$ by employing the optical flow estimation [3, 23]. Then, we can sum up the temporal grid distortions at the horizontal and vertical directions, respectively, as

$$E^T = \sum_t \sum_{g_k^{z,t}} \sum_{j \in A(k)} (\|\tilde{w}_k^{z,t} - \tilde{w}_j^{z,t'}\|^2 + \|\tilde{h}_k^{z,t} - \tilde{h}_j^{z,t'}\|^2)), \tag{11}$$

where $A(k)$ is the set of aligned grids for $g_k^{z,t}$.

4. Experiments

Implementation. We minimize the objective function in Eq. (1) to find the optimal set of grid meshes for a stereo video subject to the boundary and spatial neighboring constraints [20]. This is a quadratic optimization problem, which can be solved by the active-set method [29]. Similar to existing retargeting works, our method contains a few adjustable parameters (e.g., λ^S and λ^D). They are used to weigh the shape and depth cost functions. We set $\lambda^D = 10^5$ and $\lambda^S = 10^3$ for all tested stereo videos in the experiments.

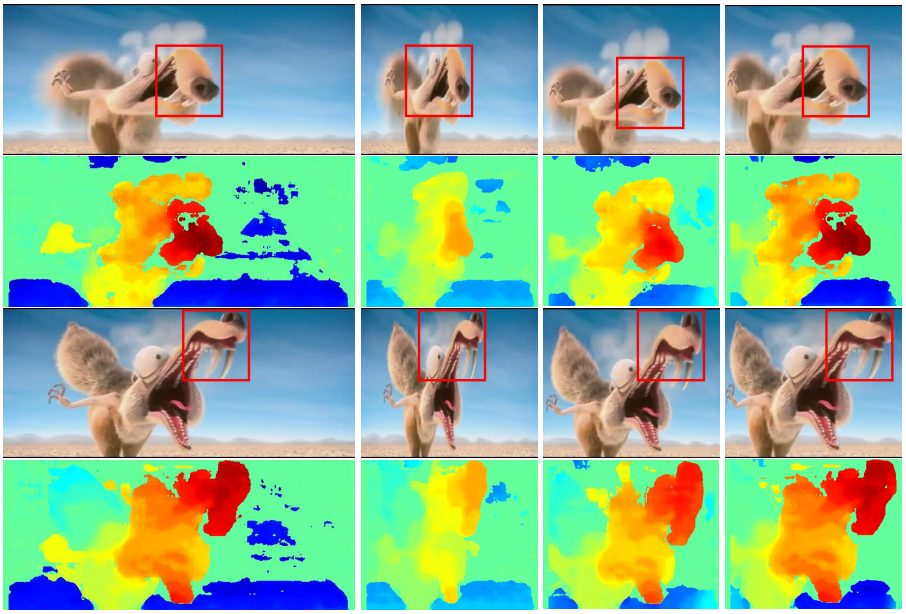


Figure 5: The left frame and the disparity map of two close-up shot frames from movie *Ice Age 4* are shown in the top and bottom rows. Regions with higher depth values are marked by red blocks in left frames.

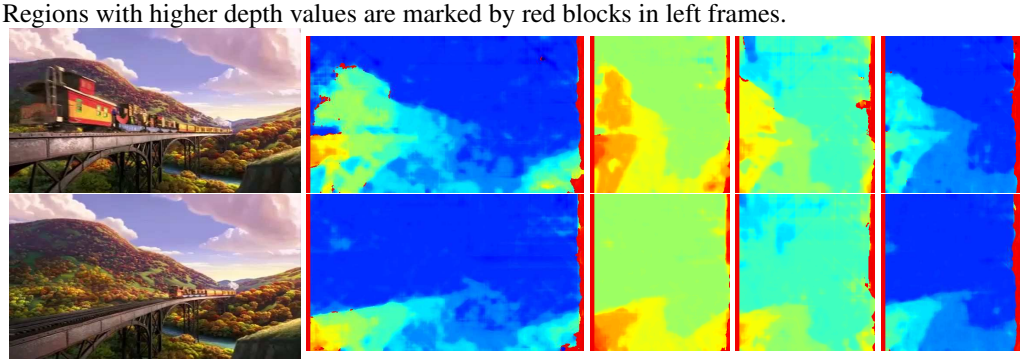


Figure 6: Comparison of disparity maps of three retargeting methods of two frames in a long shot from movie *Madagascar3*

Performance Comparison. Since there are only few existing works on stereo video retargeting, we compare the performance of our method with two existing methods: the uniform scaling (US) method and the CVW (Consistent Volumetric Warping) [22]. The CVW offers the state-of-the-art solution to warping-based stereo video retargeting, and it is most related to our work. We test the three methods on various test videos. Most of them are from the dataset used for CVW [22]. The test videos contain various types of motions and salient objects with a large depth range and significant depth variations, imposing great challenges on stereo video retargeting. We consider the same retargeting goal in [22]; namely, reducing the width of each video to 50% while preserving the height. Because the salient objects in the CVW dataset often occupy a large area (more than 50%) of a frame as shown in Fig. 5, one cannot simply crop a test video by removing unimportant side regions to achieve the goal. Instead, one has to combine cropping and warping. This strategy is adopted by both CVW and our method. A good stereo video retargeting method should preserve both the shape and depth attributes of salient 3D objects and ensure temporal coherence of shape and depth. Due to the space limit, we only show the retargeting results of two frames for each video, and provide more results in the supplemental materials[1]. Note, the performance of spatio-temporal depth preservation can be evaluated by the corresponding disparity map [10], where the red and the blue colors are used to indicate the smallest negative disparity and the largest positive disparity, respectively.

Fig. 5 shows retargeting results of a close-up shot. In this video, its background is static while the foreground object has a strong depth value in the spatio-temporal do-

[1]The effect of cropping is also analyzed in our supplemental materials.

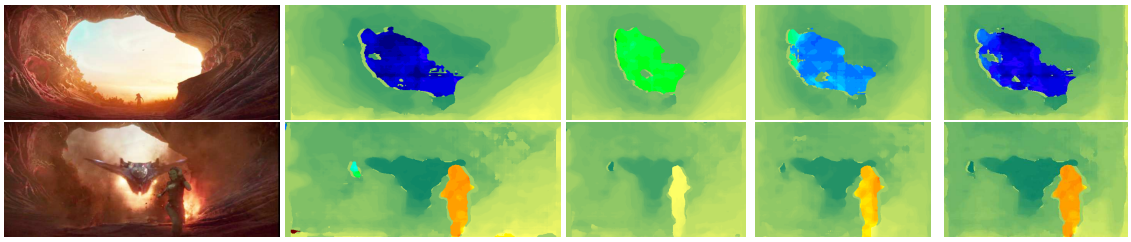


Figure 7: Comparison of disparity maps of three retargeting methods for a shot in movie *Guardians of the Galaxy2*

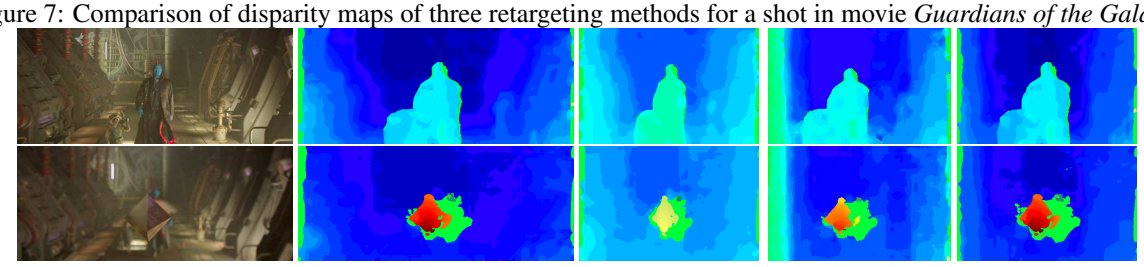


Figure 8: Comparison of disparity maps of three retargeting methods for a shot in movie *Guardians of the Galaxy2*

main. The squirrel is in front of the screen and rushes out along the depth direction. All three methods achieve high performance in terms of temporal coherence. As to spatial shape preservation, US shrinks the shape of the squirrel since it does not consider video content at all. CVW preserves the shape better than US thanks to its incorporation of the shaper cost function. However, CVW distorts the shape of squirrel's nose. Among the three methods, our method offers the highest shape fidelity. As to the spatio-temporal depth preservation, US reduces the depth of the squirrel at each frame and the depth variation across frames since it ignores the depth information. Similarly, CVW distorts the depth value at each frame and decreases the range of squirrel's motion along the depth direction. In contrast, our method preserves the spatio-temporal depth information of the squirrel well thanks to the carefully-designed depth cost function.

Fig. 6 shows retargeting results on a long shot where the camera moves slowly. For most long-shot videos, the depth of foreground objects is not as strong as that of the close-up shot. However, the depth value of the background is large so as to increase the distance of the whole 3D scene to the screen. Our method can preserve such viewing experience well by preserving the depth information. However, the US and the CVW decrease the depth values, leading to the poorer depth perception of the whole 3D scene.

Besides the CVW dataset, we also test the methods on more challenging videos, as shown in Fig. 7 and 8. These videos contain multiple foreground objects and relatively complex background, which have strong depth values in the 3D scene. Moreover, these video contain significant camera and object motions, leading to significant depth changes of objects along time. The depth changes make depth preservation very challenging to stereo video retargeting methods. Since we are unable to obtain CVW's results for these videos, we implement the constraints proposed in CVW to build a baseline called sCVW. The results show that our method can preserve the spatio-temporal depth of salient objects well, whereas US and sCVW reduce that, since they do not have effective depth-preserving constraints.

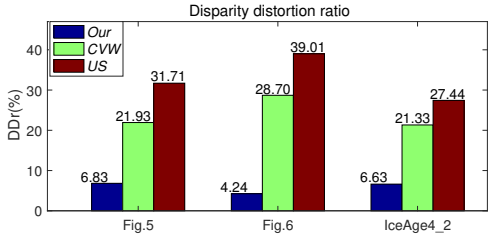


Figure 9: Estimated depth distortions of videos in Figs. 5 and 6 and IceAge4_2 respectively, where IceAge4_2 is a close-up shot from movie *Ice Age 4*.

We also compare our method with representative 2D image/video retargeting methods using images from VOC [7]. Fig. 10 shows our method can better preserve the object shape as compared to SLR [5] and SC [32], though the object size is a bit smaller.

Quantitative Analysis To the best of our knowledge, there is no widely accepted metric for evaluating spatio-temporal depth preservation for a retargeted stereo video [2]. We propose an objective metric called Disparity Distortion ratio (***DDr***). The metric calculates the ratio of the

[2] Lin *et al.* [21][22] employed *Pearson correlation coefficient* to evaluate the depth performance of "stereo image" retargeting" instead of stereo video retargeting. Moreover, such metric is not suited to evaluate depth preservation, as it evaluates the "linearity" between the depth of a stereo image pair and its counterpart of the retargeted pair.

Table 1: Winning frequencies of subjective test in comparing three stereo video retargeting methods.

	Ours	US	CVW	Total
Ours	–	96 %	76 %	86%
US	4%	–	30%	17%
CVW	24%	70%	–	47%

average disparity deviation of 3D points between the retargeted video and the original video, normalized to the disparity range of the original video as follows:

$$DDr = \frac{1}{|d_{max}| \cdot N^v} \sum_{(k,t)} |d_k^{z,t} - \tilde{d}_k^{z,t}| \qquad (12)$$

where $d_k^{z,t}$ is the depth of $g_k^{z,t}$, N^v is the total number of 3D points, and $|d_{max}|$ is the maximal magnitude of disparity in the original video. As shown in Fig. 9, our method achieves the lowest depth distortions, compared with US and CVW.

User study. We conducted the subjective evaluation on a 3D 22-inch monitor using the NVIDIA active shuttered glasses and an NVIDIA GeForce 3D Vision Solution. We invited 10 subjects to participate in the test. The subjects consisted of 7 males and 3 females. Two of the subjects were experts in the 3D perception field while the others were not. By following the same subjective test methodology in [31], we compared our method with US and CVW on 5 stereo videos. Also, by following the same setting in [28], we placed the original video in the middle and randomly put two retargeted videos at its left and right sides. Subjects were allowed to pause, forward and rewind the videos. Then, each subject was asked to answer the following question: *Q: which retargeted video is more similar to the original one ?*

In total, we receive $3 \times 5 \times 10 = 150$ answers and each method is pairwise compared by $2 \times 5 \times 10 = 100$ times. Table 1 shows the percentages of answers to the question. In total, 85% of participants were in favor of our method over CVW while 96% of participants were in favor of our method over the US method. This result clearly indicates the superiority of our method over the US and the CVW methods in preserving the fidelity of depth information. The main reason is that most subjects paid special attention to regions with stronger depth values (especially for objects moving out of the screen), making depth distortions in these regions relatively noticeable. Since our method better preserves depth information, it also leads to better 3D viewing experience generally.

Discussions Our method can be easily incorporated into exiting retargeting frameworks. For instance, we demonstrate the result of integrating our method with per-vertex optimization, which has been adopted in several 2D image/video retargeting methods (e.g.[39]). As shown in Fig. 11, per-vertex optimization often incurs shape distortions due to its high degrees of freedom for warping, whereas the

Figure 10: Comparisons with 2D retargeting methods

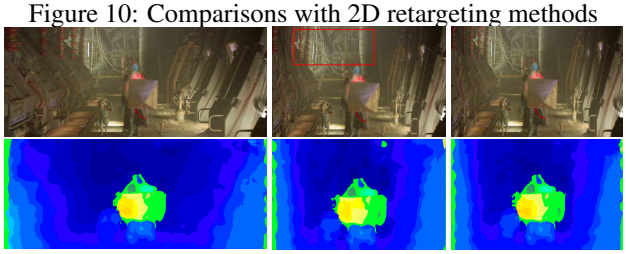


Figure 11: Examples of integrating our method with per-vertex optimization. Row from top to bottom: left frame and disparity map

result preserves depth well thanks to the depth-preserving energy. Nevertheless, the complexity of per-vertex optimization is much higher than that of axis-aligned warping. In particular, the complexity of per-vertex optimization is in form of $4 \cdot N_t \cdot (N_c \times N_r)$, where N_t is the total number of frames, N_c and N_r are the numbers of grid columns and rows in a frame. In contrast, the number of variables of axis-aligned warping is $2 \cdot N_t \cdot (N_c + N_r)$.

Besides, in practice, there is another important issue: for a stereo video, some display and viewing conditions may yield improper perceived depth, which makes viewers feel uncomfortable. To address this problem, our method can be combined with depth retargeting [15] by first constructing comfortable perceived depth maps according to a given viewing condition. We can then derive target disparity maps, and improve Eq. (3), such that the disparity of a stereo video is remapped to a desired target disparity value.

5. Conclusion

A novel approach was proposed to preserve the depth effect in stereo video retargeting. A cost function that takes the depth preservation requirement into account was derived and incorporated in the total cost function. A grid-warping-based optimization problem was formulated and solved to offer an effective stereo video retargeting solution that preserves depths of salient regions, coherently transforms other non-salient regions and achieves spatio-temporal shape coherence simultaneously. The proposed depth-aware retargeting method is flexible. It allows user interactions to emphasize the region of interest. Our method is easy to implement and can be easily incorporated in the VR or stereo video content editing toolbox with a proper interface.

References

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 2007.

[2] T. D. Basha, Y. Moses, and S. Avidan. Stereo seam carving a geometrically consistent approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35:2513–2525, 2013.

[3] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):500–513, 2011.

[4] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Trans. Multimedia*, 13:589–601, 2011.

[5] D. Cho, J. Park, T. H. Oh, Y. W. Tai, and I. S. Kweon. Weakly- and self-supervised learning for content-aware deep image retargeting. In *ICCV*, 2017.

[6] Y.-Y. Chuang and C.-H. Chang. A line-structure-preserving approach to image resizing. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recongnit.*, pages 1075–1082, 2012.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[8] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin. Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Trans. on Image Process.*, pages 3888–3901, 2012.

[9] J. Harris. *Sensation and Perception*. SAGE, 2014.

[10] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 328–341, 2008.

[11] G. Hua, C. Zhang, Z. Liu, Z. Zhang, and Y. Shan. Efficient scale-space spatiotemporal saliency tracking for distortion-free video retargeting. In *ACCV 2009*, 2009.

[12] C. Kim, A. Hornung, S. Heinzle, W. Matusik, and M. Gross. Multi-perspective stereoscopy from light fields. *ACM trans. on graph.*, 2011.

[13] S. Kopf, B. Guthier, C. Hipp, J. Kiess, and W. Effelsberg. Warping-based video retargeting for stereoscopic video. In *ICIP 2014*, 2014.

[14] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross. A system for retargeting of streaming video. *ACM Trans. Graph.*, pages 126:1–126:10, 2009.

[15] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3d. *ACM Trans. Graph.*, 29(4):75:1–75:10, 2010.

[16] 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.*, 2012.

[17] 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*, 19:1442 – 1453, 2017.

[18] B. Li, Y. Chen, J. Wang, L.-Y. Duan, and W. Gao. Fast retargeting with adaptive grid optimization. In *Proc. IEEE Int. Conf. Mutimedia Expo.*, pages 1–4, 2011.

[19] B. Li, L. Duan, C. Lin, T. Huang, and W. Gao. Depth-preserving warping for stereo image retargeting. *IEEE Trans. Image Process.*, 24(9):2811–2826, 2015.

[20] 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.*, 23(4):1615–1628, 2014.

[21] 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 Techn.*, 24(5):759–768, 2014.

[22] 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.*, 8, 2015.

[23] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[24] C.-W. Liu, T.-H. Huang, M.-H. Chang, K.-Y. Lee, C.-K. Liang, and Y.-Y. Chuang. 3d cinematography principles and their applications to stereoscopic media processing. In *ACM MM2011*, 2011.

[25] B. Masia, G. Wetzstein, C. Aliaga, R. Raskar, and D. Gutierrez. Display adaptive 3d content remapping. *Computers & Graphics*, 2013.

[26] G. Matthias, V. Kwatra, M. Han, and I. Essa. Discontinuous seam-carving for video retargeting. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recongnit.*, 2010.

[27] B. Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Taylor & Francis, 2009.

[28] Y. Niu, W.-C. Feng, and F. Liu. Enabling warping on stereoscopic images. *ACM Trans. Graph.*, 2012.

[29] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[30] D. Panozzo, O. Weber, and O.Sorkine. Robust image retargeting via axis-aligned deformation. 2012.

[31] M. Rubinstein, D. Gutierrez, A. Shamir, and A. Shamir. A comparative study of image retargeting. page 160, 2010.

[32] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 2008.

[33] A. Shamir and O. Sorkine. Visual media retargeting. In *ACM SIGGRAPH ASIA 2009 Courses*, pages 11:1–11:13, 2009.

[34] J. Shen, D. Wang, and X. Li. Depth-aware image seam carving. *IEEE Trans. on Cybernetics*, 43(5):1453–1461, 2013.

[35] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee. Scalable and coherent video resizing with per-frame optimization. *ACM Trans. Graph.*, 2011.

[36] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video-retargeting. In *Proc. IEEE Conf. Comput. Vis.*, pages 1–6, 2007.

[37] T. Yan, R. W. H. Lau, Y. Xu, and L. Huang. Depth mapping for stereoscopic videos. *Int. J. Comput. Vision*, 102:293–307, 2013.

[38] T.-C. Yen, C.-M. Tsai, and C.-W. Lin. Maintaining temporal coherence in video retargeting using mosaic-guided scaling. *IEEE Trans. Image Process.*, pages 2339–2351, 2011.

[39] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin. A shape-preserving approach to image resizing. *Comput. Graph. Forum*, pages 1897–1906, 2009.

[40] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *CVPR*, 2015.