The Visual Computer (2019) 35:1427–1446
https://doi.org/10.1007/s00371-018-01622-1

ORIGINAL ARTICLE

Disparity estimation in stereo video sequence with adaptive spatiotemporally consistent constraints

Liang Tian[1] · Jing Liu[1] · Haibin Ling[2] · Wei Guo[1]

Published online: 19 December 2018

Abstract
Numerous stereo matching algorithms have been proposed to obtain disparity estimation for a single pair of stereo images. However, simply even applying the best of them to temporal frames independently, i.e., without considering the temporal consistency between consecutive frames, may suffer from the undesirable artifacts. Here, we proposed an adaptive, spatiotemporally consistent, constraints-based systematic method that generates spatiotemporally consistent disparity maps for stereo video image sequences. Firstly, a reliable temporal neighborhood is used to enforce the "self-similarity" assumption and prevent errors caused by false optical flow matching from propagating between consecutive frames. Furthermore, we formulate the adaptive temporal predicted disparity map as prior knowledge of the current frame. It is used as a soft constraint to enhance the temporal consistency of disparities, increase the robustness to luminance variance, and restrict the range of the potential disparities for each pixel. Additionally, to further strengthen smooth variation of disparities, the adaptive temporal segment confidence is incorporated as a soft constraint to reduce ambiguities caused by under- and over-segmentation, and retain the disparity discontinuities that align with 3D object boundaries from geometrically smooth, but strong color gradient regions. Experimental evaluations demonstrate that our method significantly improves the spatiotemporal consistency both quantitatively and qualitatively compared with other state-of-the-art methods on the synthetic DCB and realistic KITTI datasets.

Keywords Spatiotemporally consistent · Adaptive temporal segment confidence · Adaptive temporal predicted disparity · Reliable temporal neighborhood

1 Introduction

As one of the most fundamental and challenging problems in visual computing, stereo matching has been an active research topic for decades, playing an important role in many advanced areas [10,28,49]. Stereo matching methods can be generally divided into two different categories: estimating disparity for the static scene and for the dynamic scene. In the static scene, many methods have been successfully proposed to recover disparity map from stereo images captured at just one time instant. They usually heavily rely on how the scene is modeled and focus on enforcing the spatial consistency between different viewpoints at one time instant [25,56].

Due to lack of considering the temporal consistency between consecutive frames, simply applying even the best of these methods to individual frames of stereo video image sequences captured from the dynamic scene yields temporally inconsistent disparity maps. It is a challenging problem for two reasons: first, the difficulties that arise from just a single pair of stereo frames, such as the presence of textureless areas, occlusions, image noise and different radiometric properties of multiple cameras. Second, the difficulties that arise from the sequence, such as fast object movement, motion blurring, etc. Both reasons may dramatically decrease the quality of disparity maps when errors (such as noise, trailing and flickering artifacts) exist and propagate in spatial and temporal domains.

✉ Jing Liu
liujing01@ict.ac.cn

[1] Key Laboratory of Augmented Reality, College of Mathematics and Information Science, Hebei Normal University, No. 20 Road East. 2nd Ring South, Yuhua District, Shijiazhuang 050024, Hebei, China

[2] Department of Computer and Information Sciences, Center for Data Analytics and Biomedical Informatics, Temple University, 382 SERC Building, 1925 North 12th St., Philadelphia, PA 19122, USA

However, because of referring to the same scene taken from only slightly different viewpoint, stereo video image sequences have both the spatial and temporal correlations to each other between consecutive frames. So if these correlations can be utilized, disparity maps will be generated by the highly related disparity between consecutive frames.

In this paper, we propose an adaptive, spatiotemporally consistent, constraints-based systematic framework that generates spatiotemporally consistent disparity maps for stereo video image sequences. It regards a scene with complex geometric characteristics as a set of segments in the disparity space, which can be viewed as a projection from a real-world 3D object. Under this assumption, the proposed framework fuses texture and segmentation information in the spatial and temporal domains to formulate the disparity estimation as a posteriori probability optimization problem with the following two objectives:

- In the temporal domain, the goal is to ensure that variations in the disparities between consecutive frames are temporally consistent and smooth;
- In the spatial domain, the goal is to encode the assumption that the disparities of each segment have a compact distribution. This strengthens the smooth variance of the disparity in each segment and retains the disparity discontinuities that align with 3D object boundaries from geometrically smooth, but strong color gradient regions.

The major contributions of this paper are summarized as follows:

(1) we propose a reliable temporal neighborhood of the current frame to prevent errors caused by noise, motion and occlusions from propagating between consecutive frames. It combines the advantages of the optical flow with an error detection framework to enforce the "self-similarity" assumption, which says that the disparity map of the current frame is only affected by its previous neighboring frames when they have similar texture distribution.
(2) we propose the adaptive temporal predicted disparity constraint to model the strength of temporal links. Because the variations in disparities between consecutive frames should be temporally consistent and smooth, we consider the adaptive temporal predicted disparity map as a prior knowledge of the current frame to enhance the temporal consistency of disparities and increase the robustness to luminance variance as well as restrict the range of the potential disparities for each pixel.
(3) we incorporate an adaptive temporal segment confidence to infer the temporal confidence that whether neighboring pixels belong to the same segment in the current frame. It strengthens smooth variations in the disparities of each segment and reduces ambiguities caused by under- and over-segmentation, as well as retains the disparity discontinuities that align with 3D object boundaries from geometrically smooth, but strong color gradient regions.

Compared to the average error rate 11.33% of the previous state-of-the-art methods, the proposed method provides an average error rate of 3.88% on the DCB datasets [39]. It is clear that our method performs almost 65% better than others in the aspect of precision. In additional to performing well on the DCB datasets, the proposed method also performs well in the realistic KITTI 2012 [7] and KITTI 2015 [33] datasets. It ranks ***tenth*** in both two dataset benchmarks [19, 20] (Abbreviated as ASTCC), respectively. This indicates that our method is generally accurate. It is worth noting that our results are comparable to the recent state-of-the-art DL- and CNN-based algorithms. Furthermore, it is the only ***one*** method without DL- or CNN-based framework in top ten of two dataset benchmarks.

The remainder of this paper is organized as follows. Section 2 gives a summary of the previous work. We present an overview of the proposed method in Sect. 3. We discuss the details of the proposed method in Sects. 4 and 5. The experimental results are presented in Sect. 6. Section 7 gives some conclusions with suggestions for future work. Note that for notation clarity, in this paper, we focus only on rectified stereo video image sequences, which involves only horizontal disparities in the row direction (considering that disparity values are inversely proportional to disparities according to the theory of epipolar geometric constraints). But the proposed method can easily be generalized to handle multi-view video image sequences.

2 Previous work

Temporally coherent disparity estimation from a stereo video sequence remains an open research topic. Several works have been developed by incorporating temporally coherent constraint into the existing model to eliminate noise, trailing and flickering artifacts in the generated result. These methods can be generally classified into four categories: motion stereo, spacetime stereo, 3D scene flow, and disparity prediction.

Motion stereo-based method obtains the disparity map by computing the ego-motion estimation between adjacent frames. Temporal consistency could be maintained over time by using spatial and temporal homographies.

Min et al. [35] achieved temporal stability by adding a coherence function to the stereo matching cost. The coherence function was computed based on the assumption that the corresponding points with motion vector between adjacent frames have similar disparity value. Zeng et al. [54]

proposed a content-adaptive temporal consistency enhancement algorithm, which classified the scene into the stationary and non-stationary regions so that the temporal consistency filtering process can be conducted in these two types of regions with an adaptive manner. However, because the stationary and non-stationary regions were detected simply based on the color difference computed at each pixel position across neighboring frames over the texture video, so their method is vulnerable to illumination variation between adjacent frames.

The Newton's Law of motion is incorporated into the temporal disparity consistency check [26]. Any disparity which significantly violates the Newton's Law of motion was modified. But this algorithm viewed the Newton's Law of motion as a hard constraint, which easily leads to error disparity in the current frame when the disparity estimation of previous frames is not correct. Zhang et al. [55] proposed a bundle optimization framework for recovering consistent depth maps from a video sequence. Their approach not only imposes the photo-consistency constraint, but also explicitly associates the geometric coherence with multiple frames in a statistical way. It thus can naturally maintain the temporal coherence of the recovered dense depth maps without oversmoothing. Furthermore, Zhang et al. [52] also proposed a novel trinocular stereo matching model, which effectively utilized the advantages of trinocular stereo images, and incorporated the visibility term with segmentation prior for robust depth estimate. They performed two motion models for handling dynamic scene. The traditional bundle optimization model and spatiotemporal optimization model were softly combined in a probabilistic way so that the depths of both static and dynamic pixels can be effectively refined.

Unlike the motion stereo work, spacetime stereo-based method does not estimate inter-frame motion, but it extends the traditional spatial support window to the temporal domain by incorporating a temporal dimension support window (named as the spacetime window) into the neighborhood where cost value aggregated [5]. The spacetime window is denoted as a rectangular 3D volume of pixels, which utilizes both spatial and temporal local appearance variation to reduce ambiguities and increase accuracy.

Richardt et al. [39,42] employed a method based on matching spatiotemporal quadric elements, which accelerated cost aggregation and allowed for weighted propagation of pixel dissimilarity metrics from previous frames to the current one. However, their method may lead to problems when large motion exists. Ramsin et al. [18] presented a two-stage algorithm for disparity estimation in stereo video sequence. They explicitly enforced the consistency of estimates in both space and time by treating the video as a spacetime volume. Initially, a frame-by-frame approach is adopted, and then a 3D optimization procedure including temporal information is applied. Their method indeed improves temporal coherence, but the quality of the generated disparity video sequence is highly dependent on the choice of the spatial and temporal penalties used in the optimization stage. Min et al. [34] proposed a weighted mode filtering method to suppress the noise and enforce the temporal consistency. The temporal consistency is improved by exploiting the spacetime support window similarity measurement based on the pixel and its temporally corresponding pixel.

The advantage of spacetime stereo-based method is the high computational efficiency. To achieve real-time performance, it only uses local "Winner-Take-All" optimization strategy in computation. But because of heavily relying on the local texture information that is easily affected by illumination variation, occlusion and texture distribution between adjacent frames, this method does not perform well in dynamic scenes that contain fast motions.

There are also approaches to obtain temporally consistent disparity map by estimating the 3D scene flow [44], which extends the traditional optical flow to 3D motion field. Because of taking stereo and motion into account simultaneously, it contains the spatial displacement of each pixel as well as the corresponding disparity variation in the temporal domain.

Cech et al. [2] took the motion of pixels from the previous frame as the correspondence seeds for the next frame. But the seed growing algorithm needs to be conducted for all frames in order to capture the objects that suddenly appear. Wedel et al. [48] studied a variant framework to consider stereo pairs from two consecutive times to compute both depth and a 3D motion vector. Hung et al. [15] proposed a unified disparity and image scene flow estimation method to preserve motion-disparity temporal consistency using robust motion trajectories. Based on those trajectories, long-range temporal constraints were advocated to correct errors caused by occasional noise or abrupt luminance variation and improved estimates temporally. But their method heavily relies on reasonable disparity initialization. If, for one region, the initial disparity estimation is consistently wrong for all frames, the following performance would not improve it much.

The main idea of disparity prediction is that the disparity maps of previous frames contain lots of information about the solution at the current frame [38]. It fully utilizes the generated disparity maps from previous frames to predict the disparity map of the current frame and hence the temporal consistency between adjacent frames is enforced.

Larsen et al. [22] enforced the temporal consistency by exploiting a temporal belief propagation function, which is composed of seven nodes, namely the current pixel, its four spatially connected pixels, and its two temporally corresponding pixels linking each frame to its previous and next frames constructed using optical flow matching. Gong et al. [9] and Yamaguchi et al. [51] extended 3D scene flow to the 3D disparity flow that models temporal consistency in depth

space between the neighboring frames, which is a 2D array of 3D vectors depicting the observation of the 3D motion in the scene from a given view. It helps to enforce the temporal consistency using the cross-validated disparity. Vretos et al. [21,47] integrated the temporal consistency in the disparity and color spaces of the video to predict disparity map. Outlier detection along the temporal dimension in the color space was used to find regions where disparities can be temporally enhanced from the previous frames. Bartczak et al. [1] proposed a method that provides dense prediction maps by reducing uncertainty due to the discrete hypotheses in [9]. Dobias et al. [6] presented a method that transfers the disparity map from previous frame to the next frame using the estimated motion of the calibrated stereo rig. But the simplification to a linear transformation leads to errors.

Although some research works have shown the improvement, the problem of how to appropriately extract information and recover consistent depths from a video remains challenging. They are typically performed using temporally consistent pixel-level cues and do not sufficiently consider the temporally consistent regional information as a cue for the disparity estimation. It is the largest distinction between previous methods and ours. On the one hand, the color-segmentation-based stereo matching algorithms [29], which lead to good performance on a single pair of stereo images, have been incorporated into the scheme of obtaining spatiotemporally consistent disparity map. But they only focus on obtaining disparity map for each single pair of stereo images and do not pay more attention to the temporally consistent constraint of segmentation, which results in that the generated disparity map suffers from the errors caused by under- and over- segmentation. On the other hand, without texture information, we cannot be sure whether the disparities from stereo matching at the current time are more confident than the temporal predicted disparities from the previous frames in texture-less and texture-repetitive regions (where traditional stereo matching usually fails and the temporal predicted disparity values may lead to a more reliable result).

3 Overview of our approach

The proposed method consists of the following phases: problem formulation, optimization and post-processing. By assuming that we have estimated the disparity maps (of the left and right viewpoints) from the previous frames, our goal is to obtain a spatiotemporally consistent disparity estimate of the current frame from the left viewpoint.

The problem formulation is the main contribution of this study. In this phase, we first compute the reliable temporal neighborhood through an optical flow-based framework (Sect. 4.1). Furthermore, constraints based on the reliable temporal neighborhood are obtained from the spatial and temporal domains:

- In the spatial domain, the initial disparity maps (D_L^t and D_R^t) of the left and right viewpoints (I_L^t and I_R^t) in the current frame t are computed using the segmentation-based stereo matching method [29]. I_L^t is also partitioned into homogenous color segments. Each pixel in I_L^t is marked as either reliable or occluded.
- In the temporal domain, the adaptive temporal predicted disparity map is obtained to model the strength of temporal links and restrict the range of the potential disparities for each pixel in the current frame (Sect. 4.2). Additionally, because the initial disparity map (D_L^t) is affected by the color-based segmentation method, there are undesired ambiguities caused by under- and over-segmentations. So the adaptive temporal segment confidence is then computed, which is used to eliminate these ambiguities and smooth the disparity variations (Sect. 4.3).

In addition, the spatial and temporal constraints are viewed as soft constraints to formulate the disparity estimation as a Markov random field posteriori probability optimization problem. More details will be discussed in Sect 4.4.

In the optimization and post-processing phases, by using the α-$expansion$ approach, the Markov random field-based energy function is iteratively optimized to obtain the spatiotemporally consistent disparity map of the current frame. And, the fitted plane-based filling occlusion procedure is employed to fill the occlusion and refine the estimated disparity map (Sect. 5).

4 Problem formulation

4.1 Reliable temporal neighborhood

The definition of the temporal neighborhood is less straightforward because adjacent frames can contain ambiguities, such as variations in illumination and texture, motion, and occlusions, resulting in temporal correspondences that are non-contiguous and may span large displacements.

Optical flow matching is used to compute a dense flow field between pairs of adjacent images. Because of preserving large displacements, we use the approach proposed by [43] to construct the temporal neighborhood. Although the temporal neighborhood is defined by considering all optical flow matching correspondences between adjacent frames, this may produce a large number of false correspondences because of known limitations of optical flow matching. So the reliable temporal neighborhood is denoted by applying the error cross-checking to test the validity of each temporal correspondence and avoid undesired false matches.

Let $f_p^{(i,t)} = (\Delta u_p^{(i,i)}, \Delta v_p^{(i,t)})$ be the optical flow displacement vector at pixel p from the previous frame i to current frame t. Meanwhile, $f_q^{(t,i)} = (\Delta u_q^{(t,i)}, \Delta v_q^{(t,i)})$ denotes the optical flow displacement vector at pixel q from frame t to frame i. Then, we calculate the ratio of reliable matches between p and q in frame i and t as follows:

- According to $f^{(t,i)}$, the optical flow matching pixel of $q = (u_q, v_q)$ in current frame t is p in previous frame i, where $p = (u_q + \Delta u_q^{(t,i)}, v_q + \Delta v_q^{(t,i)})$.
- According to $f^{(i,t)}$, the optical flow matching pixel of p in previous frame i is $q' = (u_q + \Delta u_q^{(t,i)} + \Delta u_p^{(i,t)}, v_q + \Delta v_q^{(t,i)} + \Delta v_p^{(i,t)})$ in current frame t.
- If $|q' - q| \leq T_{op}$, then the optical flow between pixel q in frame t and pixel p in frame i can be viewed as a reliable match.
- Iterating steps (1) to (3) for all pixels.

This process is repeated between the current frame and each of its previous frames (e.g., 20 frames). The reliable temporal neighborhood of the current frame is denoted as the set of five neighboring previous frames with the highest ratio of reliable matches. Each frame in the reliable temporal neighborhood is noted as the reliable temporal frame.

The reliable temporal neighborhood enforces the "self similarity" assumption that the disparity of the current frame is only affected by its previous neighboring frames, if and only if they have similar texture distributions. Furthermore, when the disparities of a single reliable temporal frame fail to the local optimal solution in certain areas, we expect that the disparities in other reliable temporal frames may be correct for the same region, with spatiotemporal consistency. Additionally, this prevents errors caused by false optical flow matches from propagating between consecutive frames.

4.2 Adaptive temporal predicted disparity map

It is very important to restrict the scope of the disparity variance for each pixel. On the one hand, given an inappropriate disparity variance range, conventional stereo matching methods for single pairs of images are often prone to finding local minima or incorrect estimates caused by luminance variations and noise. On the other hand, the lack of scope of the disparity variance often also implies the need to search over a wider range of candidate disparities, which requires more computation and memory, especially for stereo image sequences.

However, we know that the variations in disparities between consecutive frames are temporally consistent and smooth. Then, the disparities from the previous frames can be used as a useful guide for the disparities of the current frame. Meanwhile, the reliable temporal neighborhood has been proven to enforce the "self-similarity" assumption described in Sect. 4.1. So the adaptive temporal predicted disparity map, which is based on the disparity and texture information of each reliable temporal frame in the reliable temporal neighborhood, is used as prior knowledge of the disparities of the current frame. This restricts the range of the potential disparities for each pixel to remove errors caused by variations in luminance and texture. Furthermore, it models the strength of temporal links between adjacent frames.

First, we use the reliable temporal frame i in the reliable temporal neighborhood as an example to illustrate how we compute the temporal predicted disparity map ($R_L^{(i,t)}$) between frame i and the current frame t of the left viewpoint. Suppose that I_L^i and I_R^i (I_L^t and I_R^t) are the texture images in frame i (frame t) and that D_L^i and D_R^i (D_L^t and D_R^t) are the disparity maps in frame i (frame t). Given each pixel's optical flow between frames i and t, each pixel $p \in R_L^{(i,t)}$ is defined as a 3D motion vector $(\Delta u_p, \Delta v_p, \Delta d_p)$, where $(\Delta u_p, \Delta v_p)$ are image coordinate displacements between p in frame i and its corresponding optical flow matching pixel q in frame t. Δd_p is the predicted disparity variation between optical flow matching corresponding pixels (p and q) from the reliable temporal frame i to the current frame t. Our goal is to estimate Δd_p through a global Markov random field optimization function (E):

$$E = E_{\mathrm{d}} + E_{\mathrm{s}} \tag{1}$$

It consists of the data term (E_{d}) and the smoothness term (E_{s}). According to the hypothesis that all corresponding pixels in the spatial and temporal domains are the projection from the same 3D object at different time instances and viewpoints, the data term (E_{d}) penalizes dissimilarities of corresponding pixels whose color or intensity values should be constant or similar. Meanwhile, the smoothness term (E_{s}) penalizes the local variations in the optical flow fields.

As shown in Eqs. 2 and 3, the data term (E_{d}) consists of three terms related to the left and right optical flows, and the stereo between frames i and t.

$$E_{\mathrm{d}} = \sum_{p \in I_L^i} O^{(i,t)}(p) \cdot (E_{fl} + E_{lr} + E_{fr}) \tag{2}$$

$$O^{(i,t)}(p) = O_{fl}^{(i,t)} \cdot O_{lr}^{(t)} \cdot O_{fr}^{(i,t)} \tag{3}$$

Here, $O_{lr}^{(t)}$ are the non-occluded pixels for the stereo image pair in the current frame t. $O_{fl}^{(i,t)}$ and $O_{fr}^{(i,t)}$ are the non-occluded pixels for the left and right optical flows between frames i and t.

Given each pixel $p(u_p, v_p)$ in the reliable temporal frame i from the left viewpoint (u_p and v_p are the corresponding image coordinates). The stereo matching pixels of p in frame i from the right viewpoint are denoted by $p'(u_p +$

$D_L^i(p), v_p)$. The optical flow pixel that corresponds to p in frame t can be defined as $q(u_p + \Delta u_p, v_p + \Delta v_p)$ according to the optical flow between frames i and t. The predicted disparity of q from p is denoted as $D_L^i(p) + \Delta d_p$. Then, the stereo matching pixel of q from the right viewpoint in frame t is defined as $q'(u_p + \Delta u_p + D_L^i(p) + \Delta d_p, v_p + \Delta v_p)$.

$$E_{fl} = \sum_{c\in R,G,B} \sum_{q\in I_L^t, p\in I_L^i} \Psi\left\{q\left(u_p + \Delta u_p, v_p + \Delta v_p, c\right), p\left(u_p, v_p, c\right)\right\} \quad (4)$$

$$E_{lr} = \sum_{c\in R,G,B} \sum_{q\in I_L^t, q'\in I_R^t} \Psi\left\{q'\left(u_p + \Delta u_p + D_L^i(p) + \Delta d_p, v_p + \Delta v_p, c\right), q\left(u_p + \Delta u_p, v_p + \Delta v_p, c\right)\right\} \quad (5)$$

$$E_{fr} = \sum_{c\in R,G,B} \sum_{q'\in I_R^t, p'\in I_R^i} \Psi\left\{q'\left(u_p + \Delta u_p + D_L^i(p) + \Delta d_p, v_p + \Delta v_p, c\right), p'\left(u_p + D_L^i(p), v_p, c\right)\right\} \quad (6)$$

According to the above definitions, E_{fl} (Eq. 4) is the color difference of the optical flow matching pixels (i.e., p and q) between frames i and t from the left viewpoint (the blue line in Fig. 1). E_{lr} (Eq. 5) is the color difference of the stereo matching pixels (i.e., q and q') from the left and right viewpoints in frame t (the orange line in Fig. 1). E_{fr} (Eq. 6) is the color difference of the optical flow matching pixels (i.e., p' and q') between frames t and i from the right viewpoint (the green line in Fig. 1). c denotes one of three color channels. $\Psi\{x, y\}$ is a robust function as $\sqrt{(x-y)^2 + 0.0001}$.

The basic assumption behind the data term is that if Δd_p is correct, then all corresponding pixels (p, p', q, and q') can be viewed as projections from the same 3D object and should have similar colors. It helps to reduce the influence of occlusion, noise, and illumination variations.

The smoothness term is defined to penalize the local variations in the optical flow fields.

$$E_s = \sum_{\substack{p\in I_L^i \\ p_i\in N_p}} \min\{|\Delta d_p - \Delta d_{p_i}|, T_s\} \cdot T[S_L^i(p), S_L^i(p_i)] \quad (7)$$

$$T[S_L^i(p), S_L^i(p_i)] = \begin{cases} 1 & S_L^i(p) = S_L^i(p_i) \\ 0 & S_L^i(p) \neq S_L^i(p_i) \end{cases} \quad (8)$$

N_p is the spatial four-neighborhood of pixel p. The constant threshold value T_s is equal to 13. We have already obtained the disparity map of previous frame, i, so S_L^i is defined as the disparity-based segmentation result using the algorithm described in [4]. According to the disparity information, each disparity-based segment in frame i can be viewed as a projection from a real-world 3D object. So $T[S_L^i(p), S_L^i(p_i)]$ enhances the assumption that disparity discontinuities have been aligned with the disparity-based segment boundary. It means that disparity variations inside each disparity-based segment are smooth. For each neighboring pixel (p and p_i):

- If $S_L^i(p) = S_L^i(p_i)$ (i.e., p and p_i belong to the same disparity-based segment), the disparity variation between them should be smooth.
- If $S_L^i(p) \neq S_L^i(p_i)$ (i.e., p and p_i do not belong to the disparity-based same segment), there may be sharp disparity variations between them.

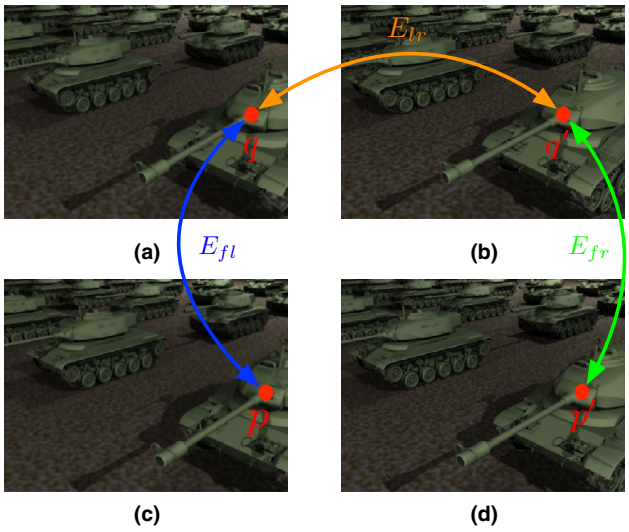

Fig. 1 The data cost combines three terms in the spatial domain (orange line) and the temporal domain (blue and green lines) between frame i and frame t. **a** and **b** Are the texture images of left and right viewpoints in frame t, and **c** and **d** are the texture images of left and right viewpoints in frame i. Frame i is one of the reliable temporal frames in the reliable temporal neighborhood of frame t

Optimization of the energy function defined in Eq. 1 is a NP-hard. However, an approximate solution with strong optimality properties can be obtained using the $\alpha - expansion$ algorithm based on graph-cuts [29]. We generate a random sequence consisting of one proposal for each allowed disparity variation, in the range of one to the allowed maximum value.

After optimization, the generated disparity map is viewed as the temporal predicted disparity map ($R_L^{(i,t)}$) from reliable temporal frame i to the current frame, t. However, because the optical flow is vulnerable to illumination variations, changes in texture, motion, and occlusion, there may be some false matches in the optical flow that reduce the credibility of a single temporal predicted disparity map. In order to enforce

the "self similarity" assumption and prevent errors caused by false optical flow matches, we iteratively obtain all the predicted disparity maps between each frame in the reliable temporal neighborhood and the current frame t. Then, we assign the adaptive temporal weight (w_d) to each predicted disparity map and aggregate them to achieve the adaptive temporal predicted disparity map of the current frame t as Eq. 9:

$$\bar{d}_q^t = \frac{\sum_{p\in\Omega} w_d \cdot R_L^{(i,t)}(p,q)}{\sum_{p\in\Omega} w_d} \tag{9}$$

$\bar{d}_q^t$ is the adaptive predicted disparity value of pixel q in the current frame t. Ω is the reliable temporal neighborhood. $R_L^{(i,t)}(p,q)$ is the predicted disparity value of q in the current frame t based on the information of its optical flow matching p in the reliable temporal frame i ($i \in \Omega$). We can see that the adaptive temporal predicted disparity ($\bar{d}_q^t$) is a weighted average of its temporal corresponding optical flow matching pixels in the reliable temporal frames.

$$w_d = w_o(p,q) \cdot w_u(p,q) \cdot w_c(p,q) \cdot w_f(i,t) \tag{10}$$

The adaptive temporal weight (w_d in Eq. 10) consists of four types of weights: the temporal occlusion weight ($w_o(p,q)$), the spatial closeness weight ($w_u(p,q)$), the temporal texture similarity weight ($w_c(p,q)$), and the temporal closeness weight ($w_f(i,t)$).

Firstly, the temporal occlusion weight ($w_o(p,q)$) is defined as a bool value. If optical flow matching pixel p of q is an optical flow occluded pixel between adjacent frames i and t, $w_o(p,q)$ is 0; otherwise, $w_o(p,q)$ is 1.

Optical flow matching often fails in texture-less and texture-repetitive regions, because there is not enough visual information to obtain a correspondence. So texture variances and gradients are used as cues to reliably estimate the optical flow by computing the similarity of the local texture structure of optical matching pixels p and q in the texture image. We define a neighborhood patch N_p (N_q) (i.e., with a radius of 15) centered at p (q) (see Fig. 2). It is evenly divided into four annular subregions because the annular spatial histogram is translation and rotation invariant.

We compute the normalized intensity eight-bin gray histogram $\Phi_p = \{\phi_p^{(k,j)}, k = 0,1,2,3, j = 0\ldots7\}$ of each subregion N_p^i, and $\Phi_q = \{\phi_q^{(k,j)}, k = 0,1,2,3, j = 0\ldots7\}$ of each subregion N_q^i to represent the annular distribution density of N_p and N_q as a 32-dimensional feature vector:

$$C^{(i,t)}(p,q) = \sum_{k=0}^{3}\sum_{j=0}^{7} \Phi(\phi_p^{(k,j)}, \phi_q^{(k,j)}) \tag{11}$$

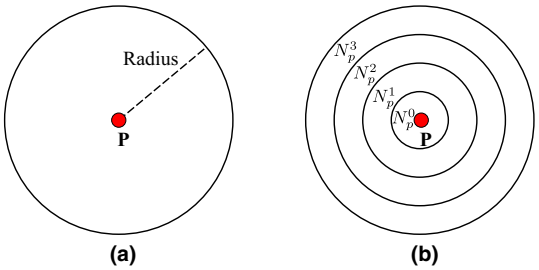

Fig. 2 An example for the surrounding neighborhood patch, N_p, for: **a** pixel p; and **b** its corresponding subregions

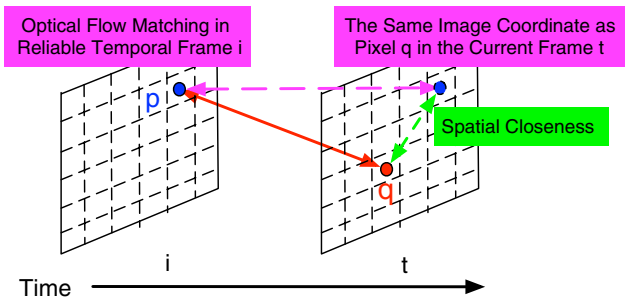

Fig. 3 Conceptual diagram of the spatial closeness weight

Equation 12 is the Hamming distances [12] between the annular distribution densities of p and its neighboring pixel q.

$$\Phi(\phi_p^{(k,j)}, \phi_q^{(k,j)}) = \begin{cases} 1 & |\phi_p^{(k,j)} - \phi_q^{(k,j)}| \geq 0.1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Temporal texture similarity weight ($w_c(p,q)$) is used to measure the closeness with respect to the texture distribution histogram (Eq. 11) of the Hamming distance (Eq. 12) between optical matching pixels p and q as:

$$w_c(p,q) = e^{-\frac{C^{(i,t)}(p,q)}{r_c}} \tag{13}$$

$w_u(p,q)$ is the spatial closeness weight that determines the reliability of the estimated optical flow. On the one hand, the optical flow algorithm is easily affected by large motions and noise. On the other hand, we assume that the disparity variations between adjacent frames are smooth and stable over time. So as shown in Fig. 3, we regard an estimated optical flow match as erroneous if the absolute Euclidean difference between two correspondences is out of range:

$$w_u(p,q) = e^{-\frac{U^{(i,t)}(p,q)}{r_u}} \tag{14}$$

$$U^{(i,t)}(p,q) = \sqrt{(u_p^i - u_q^t)^2 + (v_p^i - v_q^t)^2 + 0.0001} \tag{15}$$

where (u_p^i, v_p^i) and (u_q^t, v_q^t) are the image coordinates of pixel p in the reliable temporal frame i, and its corresponding optical flow matching pixel q in the current frame t.

Luminance variation between frames is another important factor that may lead to incorrect matching results. For stereo image sequences, a larger distance between frames will result in more significant luminance changes. We define the temporal closeness weight as Eq. 16, which measures the reliability of the optical flow match. It is clear that this reliability decreases with an increase in the distance between frames.

$$w_f(i,t) = e^{-\frac{|i-t|}{r_f}} \tag{16}$$

4.3 Adaptive temporal segment confidence

Color-segmentation-based stereo matching methods have obtained a great development and become the mainstream of stereo matching algorithms [50,51]. These methods can generate spatially consistent disparity maps and have been shown to be the most successful stereo matching techniques for static scenes. They are based on the following assumptions:

- The variance of disparity values in each color segment is smooth;
- The color segment boundaries are forced to coincide with object boundaries. That is to say, (1) object boundaries discontinuities in the three dimensions are forced to align with disparity discontinuities between color segments.

That is to say, object boundaries discontinuities in the three dimensions are forced to align with disparity discontinuities between color segments; and neighboring color segments showing similar color are more likely to originate from the same real-world surface than neighboring segments of completely different color.

Unfortunately, the accuracy of the color-segmentation-based algorithms is easily affected by initial color segmentation. On the one hand, colors around object boundaries discontinuities are often similar, a direct consequence of that is the under-segmentation which groups pixels from different objects but with similar colors into one color segment. This leads to blend the boundary between different objects. On the other hand, neighboring color segments with total different colors distribution may have similar disparities. It leads to that pixels with different colors, but on the same object are over-segmented into different color segments. This causes computationally inefficiency and ambiguities on color segment boundaries.

To reduce the ambiguities caused by under- and over-segmentation, we apply an adaptive temporal segment confidence to strengthen smooth variations of disparities in the spatial domain and retain disparity discontinuities that align with object boundaries from geometrically smooth, but strong color gradient regions. We suppose that the disparity maps of previous reliable temporal frames are known, and incorporate them as prior knowledge to infer the probability that two neighboring pixels in frame t belong to the same segment according to the disparity-based segmentation of their optical flow matching pixels in the previous reliable temporal frames. This probability is referred to as the temporal segment confidence of neighboring pixels at frame t.

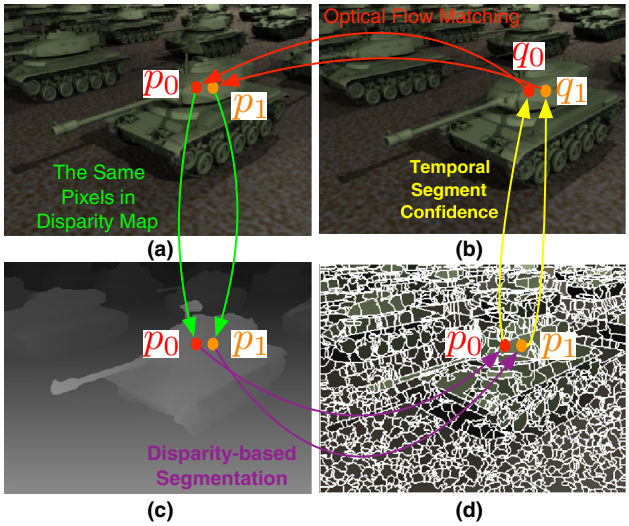

Fig. 4 Conceptual diagram of the temporal segment confidence between p_0 and p_1. **a** and **b** Are the 40th and 41th frames of the left "Tanks" image sequence. **c** The correspondent disparity map of **a**. **d** The disparity-based segmentation result of **c**. White lines are the boundaries of disparity-based segments

In the following part, the current frame (t) and its reliable temporal frame (i) are taken as an example to illustrate the entire process of computing the adaptive temporal segment confidence. Firstly, as shown in Fig. 4c, d, we apply the disparity-based segmentation algorithm [4] to divide the texture image frame (i) into different disparity-based segments.

Given q_0 and q_1 are neighboring non-occlusion pixels in the current frame t, p_0 and p_1 are their corresponding optical flow matching non-occlusion pixels in the reliable temporal frame i (represented by red lines in Fig. 4). According to the disparity-based segmentation result of frame i (represented by purple lines in Fig. 4), the temporal segment confidence between q_0 and q_1 (represented by yellow lines in Fig. 4) is:

$$S(q_0, q_1) = \begin{cases} 1 & S_L^i(p_0) = S_L^i(p_1) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Because the disparity variations between adjacent frames are smooth and stable, if p_0 and p_1 belong to the same disparity-based segment in the reliable temporal frame i

$(S_L^i(p_0) = S_L^i(p_1))$, we can assume that q_0 and q_1 most likely belong to the same segment in the current frame t $(S(q_0, q_1) = 1)$; otherwise, $S(q_0, q_1) = 0$.

Due to the low credibility of the optical flow, we iteratively obtain all temporal segment confidences of neighboring pixels (q_0 and q_1) based on the disparity-based segmentation information of their optical matching pixels (p_0 and p_1) in each reliable temporal frames. Then, we assign the adaptive temporal weight (w_s in Eq. 19) to each generated temporal segment confidence and aggregate them to obtain the adaptive temporal segment confidence of arbitrary neighboring pixels (q_0 and q_1) in the current frame, t.

Let $\bar{S}^t(q_0, q_1)$ be the adaptive temporal segment confidence between neighboring pixels q_0 and q_1 in the current frame t:

$$\bar{S}^t(q_0, q_1) = \frac{\sum_{(p_0,p_1)\in\xi} w_s \cdot S(q_0, q_1)}{\sum_{(p_0,p_1)\in\xi} w_s} \tag{18}$$

$$w_s = w_c(p_0, q_0) \cdot w_c(p_1, q_1) \cdot w_u(q_0, q_1) \cdot w_f(i, t) \tag{19}$$

where p_0 and p_1 are the optical flow matching pixels of q_0 and q_1 from the reliable temporal frame i. $w_c(p_0, q_0)$ and $w_c(p_1, q_1)$ are the temporal texture similarity weights (Eq. 13). $w_u(q_0, q_1)$ is the image coordinate distance weight (Eq. 14). $w_f(i, t)$ is the temporal closeness weight (Eq. 16). If q_0 and q_1 are close to each other and their disparity variations are smooth and stable, the distance between their optical flow matching pixels (p_0 and p_1) should be small. Otherwise, the optical flow match is unreliable.

According to the above discussion, we can obtain all adaptive temporal segment confidences between arbitrary neighboring pixels in the current frame t. The adaptive temporal segment confidence between arbitrary neighboring pixels is viewed as a soft constraint to strengthen smooth variations in disparities of each segment. It also retains the disparity discontinuities that align with object boundaries from geometrically smooth, but strong color gradient regions. If the adaptive temporal segment confidence is small, the arbitrary neighboring pixels in the frames t may belong to different segments, which indicate that a large disparity variation is allowed and the disparity discontinuities may be located between them. Otherwise, if the adaptive temporal segment confidence is large, the arbitrary neighboring pixels in the frames t may belong to the same segment, which indicate that disparity variation between them is smooth and the disparity discontinuities may be not located between them.

4.4 Energy function

When the adaptive temporal predicted disparity map and adaptive temporal segment confidence have been obtained, the process for computing the disparity map of the current frame t is cast as an iterative energy optimization to enforce temporal consistency and smooth variations in the temporal domain as well as the spatial domain. This energy function is denoted as:

$$E^t = E_d^t + E_s^t \tag{20}$$

4.4.1 Adaptive temporal predicted disparity constraint

Conventional methods consider stereo matching pixels originating from the same three-dimensional point should have a similar appearance. They assume that the surface of each object in three-dimensional space is Lambertian with perfectly diffuse appearance. It reflects the same luminance regardless of the viewing angle. So the luminance consistency hypothesis is often used to penalize the appearance dissimilarity of matching pixels between I_L^t and I_R^t in the conventional stereo matching methods. But the Lambertian assumption is usually violated by specular reflections, whose position and colors change substantially depending on the viewpoint in practice. Furthermore, varying colors for the same scene point can as well be the consequence of different camera device' sensor characteristics. So the accuracy of the luminance consistency hypothesis relies heavily on the illumination condition, and its confidence usually is low in texture-less and texture-repetitive areas.

In contrast, because the variations in disparities between consecutive frames are temporally consistent and smooth, the adaptive temporal predicted disparity map can be referred to as a prior knowledge of the current frame.

After exploring the complementary characteristics of the luminance consistency hypothesis and the adaptive temporal predicted disparity map, the adaptive temporal predicted disparity constraint (E_d^t) is incorporated into our framework to restrict the range of each pixel's potential disparities for enhancing the strength of temporal links between consecutive frames. Furthermore, because of reflecting the reliability of the luminance consistency and temporally consistent, it also reduces the problem caused by the luminance variance:

$$E_d^t = \sum_{q\in I_L^t} \lambda_d \cdot (1 - O^t(q)) \cdot \Gamma(q, q') + O^t(q) \cdot \lambda_o \tag{21}$$

$$\Gamma(q, q') = w_q^H \cdot C^t(q, q') + w_q^T \cdot A^t(q) \tag{22}$$

where λ_d is a positive constant value. q' is the matching pixel of q in the other viewpoint. $O(q)$ is the occlusion mask, and λ_o is a positive penalty used to avoid maximizing the number of occluded pixels. $C^t(q, q')$ is similar to the pixel-wise cost function (Eq. 11) to measure the appearance dissimilarity between stereo matching pixels (q and q') in frame t.

$A(q)$ is the components from the adaptive temporal predicted disparity map, which are defined as:

$$A(q) = \min\left\{|D_L^t(q) - \bar{d}_q|, 7\right\} \tag{23}$$

where $D_L^t(q)$ is the assigned disparity of non-occlusion pixel q of the left viewpoint in current frame t. $\bar{d}_q$ is the adaptive temporal predicted disparity value of q.

w_q^H and w_q^T are the pixel-wise confidence weight that are related to the confidence probability of disparities from the luminance consistency hypothesis and the adaptive temporal predicted disparity, respectively:

$$w_q^H = 1 - \frac{\eta_q^{1\text{st}}}{\eta_q^{2\text{nd}}} \qquad w_q^T = 1 - w_q^H \tag{24}$$

w_q^H quantifies how distinctive the best and the second best matching costs (defined as $\eta_q^{1\text{st}}$ and $\eta_q^{2\text{nd}}$ respectively) of pixel q. When pixel q locals at the texture-less or texture-repetitive regions and illumination variation areas, $\eta_q^{1\text{st}}$ is close to $\eta_q^{2\text{nd}}$ and w_q^T is larger than w_q^H. So $A(q)$ can restrict the range of the potential disparities around the adaptive temporal predicted disparity ($D_L^t(q)$) to reduce the matching ambiguities.

4.4.2 Adaptive temporal segment confidence constraint

Conventional color-segmentation-based stereo matching algorithms can tend to under-segmentation when pixels with similar colors but on different objects are grouped into one segment, and over-segmentation when pixels with different colors but on the same object are partitioned into different segments. As a direct consequence of under-segmentation, foreground and background boundaries are blended if they have similar colors at disparity discontinuities, whereas over-segmentation will cause ambiguities between segment boundaries.

To avoid that, E_s^t uses the adaptive temporal segment confidence as a soft guide to reduce ambiguities caused by over- and under-segmentation and strengthen smooth variation of disparities, while retaining the disparity discontinuities that align with object boundaries from geometrically smooth, but strong color gradient regions:

$$E_s^t = \sum_{\substack{q_0 \in I_L^t \\ q_1 \in N_{q_0}}} \bar{S}^t(q_0, q_1) \cdot \lambda_{\bar{s}} \cdot \min\left\{\left|D_L^t(q_0) - D_L^t(q_1)\right|, 5\right\} \tag{25}$$

- Case I: q_0 and q_1 belong to the same color segments and $\bar{S}^t(q_0, q_1)$ is close to 0. It means that q_0 and q_1 may belong to the different objects in three dimensions but with similar color distribution. The initial color segmentation result may lead to the under-segmentation. So the disparity discontinuity ($D_L^t(q_0) \neq D_L^t(q_1)$) is allowed to avoid blending the boundary between different objects.
- Case II: q_0 and q_1 belong to the same color segments and $\bar{S}^t(q_0, q_1)$ is close to 1. It means that q_0 and q_1 may belong to the same objects in three dimension with similar colors, the disparity discontinuity ($D_L^t(q_0) \neq D_L^t(q_1)$) is not allowed.
- Case III: q_0 and q_1 belong to different color segments, and $\bar{S}^t(q_0, q_1)$ is close to 0. It means that q_0 and q_1 may belong to different objects in three dimension with different colors, and the disparity discontinuity ($D_L^t(q_0) \neq D_L^t(q_1)$) is allowed.
- Case IV: q_0 and q_1 belong to different color segments, and $\bar{S}^t(q_0, q_1)$ is close to 1. It means that q_0 and q_1 may belong to the same objects in three dimension with different colors. The initial color segmentation result may lead to the over-segmentation. The disparity discontinuity ($D_L^t(q_0) \neq D_L^t(q_1)$) is not allowed to eliminate the ambiguities on color segment boundaries.

5 Optimization and post-processing

Optimization of the energy function (Eq. 20) is realized by using the algorithm in [29]. E_p^t is expressed as a unary terms, and E_s^t is the pairwise terms. The choice of the proposal disparity map is another crucial factor for optimization. We generate a random sequence consisting of one proposal for each allowed disparity value, in the range of minimum disparity to maximum disparity. During each optimization, the result of the current iteration is used as the initial disparity map of the next iteration.

The post-processing consists of two steps: first of all, the RANSAC-based plane fitting procedure [29] is used to estimate disparities of occluded pixels. Furthermore, the weighted joint bilateral filter with the slope disparity compensation filter [31] is applied to refine the disparity map.

6 Experimental results

In this section, a series of experiments were performed to verify the effectiveness of the proposed method. As listed in Table 1, all parameters are fixed throughout our experiments and kept constant for all image pairs.

Table 1 Parameter settings for all experiments

r_u	r_c	r_f	$\lambda_{\bar{s}}$	λ_d	λ_o
9	15	21	10	45	600

We first conducted evaluations on the synthetic DCB datasets [39] to compare the performance with other state-of-the-art stereo video image sequence disparity estimate methods. Each frame is 400 × 300 pixels in size with a disparity range of 64 pixels. The "Book" sequence contains 41 frames, while each other sequence contains 100 frames. Furthermore, we also evaluated the robustness of the proposed method using the realistic KITTI 2012 [7] and KITTI 2015 [33] stereo multi-view extension datasets. These datasets comprises 394 scenes in training datasets with associated semi-dense ground truth from a laser scanner and 395 scenes in testing ones without ground truth. Each scene contains 20 frames. Obtained from an autonomous moving platform driving around the metropolitan area of Karlsruhe, the KITTI datasets have rich scene features such as non-Lambertian surfaces (e.g., reflectance, transparency), fast motions (e.g., high speed), a large variety of materials (e.g., matte vs. shiny), and variable lighting conditions (e.g., sunny vs. cloudy). A rotating laser scanner mounted behind the left camera recorded ground truth depth.

Note that for notation clarity, in the following experiments, we focused only on recovering the disparity map of the left camera in each dataset. The percentage of error pixels (ρ), where the true disparity ($G_L^t(p)$) and the estimated disparity ($D_L^t(p)$) differ by more than a error threshold (T_{eval}) averaged over all images(T_{all}), is used as the evaluation metrics:

$$\rho = \frac{1}{T_{\text{all}}} \sum_{p \in I_L^t} \left| D_L^t(p) - G_L^t(p) \right| \leq T_{\text{eval}} \tag{26}$$

6.1 Evaluation using the synthetic DCB datasets

To confirm the accuracy of the proposed method, we first compared its performance with those of other state-of-the-art stereo video image sequence disparity estimate methods, that is, the 3D scene flow stereo-based algorithm (Liu [27]), the space-time stereo-based algorithm (Hosni [14], Pham [37]) and the motion stereo-based algorithm (Jiang [16]).

From the qualitative comparison in Fig. 5, we could find out that the proposed algorithm can obtain more smooth disparities on the surface of 3D objects (see the "Book" and the "Street" sequences). Furthermore, the noise is eliminated well both in texture-less (see the background of "Temple" and the "Tanks" sequences) and texture-repetitive regions (see the wall of "Tunnel" sequences) by considering the adaptive temporal predicted disparity map as a prior knowledge of the current frame to restrict the range of the potential disparities for each pixel. Additionally, it reduces ambiguities caused by under- and over-segmentation with the adaptive temporal segment confidence. In the under-segmentation regions (red regions in Fig. 5), the proposed method avoid blending of the foreground and background where the RGB color between the foreground and background is similar at disparity discontinuities. In the over-segmentation regions (yellow regions in Fig. 5), the proposed method retains the disparity discontinuities aligning with 3D object boundaries from geometrically smooth, but strong color gradient regions. As shown the 7th, 8th rows in Fig. 5, the disparity maps generated by the proposed method are temporally coherent and exhibit less artifacts than those generated by our previous method in a frame-by-frame manner. Results for the full stereo video sequences are shown in the supplementary material.

To further verify the effect and robustness of removing spatial-temporal artifacts, we show some consecutive frames of the "Book" sequence in Fig. 6. From this figure, we can notice that the frames contain temporally inconsistent disparities in color regions. The experimental results indicate that the proposed method significantly improves the quality of the estimated disparity map compared with other methods in temporal domain. Furthermore, it enhances the spatiotemporal consistency between consecutive frames and removes undesired flickering and trailing artifacts (see the color regions in each sequence). Our result is qualitatively very similar to the ground truth.

The quantitative evaluation results with means and variances of the error rates are listed in Table 2. We can confirm that Liu et al. [27] estimated the 3D scene flow in an interactive manner, combining a state-of-the-art stereo algorithm with the scene flow concept to capture temporal correspondences. However, they only used information between two adjacent frames, which makes it vulnerable to illumination variations and texture distributions. Furthermore, a significant amount of computation time is required to ensure a very accurate result. Compared with their results, our average error rate was 49% smaller, over 441 frames of all DCB sequences.

Pham et al. [37] incorporated the information permeability algorithm into the space-time stereo scheme for obtaining spatiotemporal disparity estimates. Their core idea was to first aggregate the matching costs between adjacent frames in the space domain and then in the time domain, using the color similarities as weights in the aggregation step. However, the spatial windows related to moving pixels may not significantly overlap over time, compromising the temporal coherence. Compared with their results, our average error rate was nearly 74% smaller, over 441 frames of all DCB sequences.

The space-time stereo-based algorithm proposed by Hosni et al. [14] applies a 2D fast edge-preserving filter to the 3D spatiotemporal domain for efficiently achieving temporally consistent disparity maps. But it has the disadvantage that the intrinsic quality of the filter results in unclear object boundaries. Compared with their results, our average error rate was 65% smaller, over 441 frames of all DCB sequences.

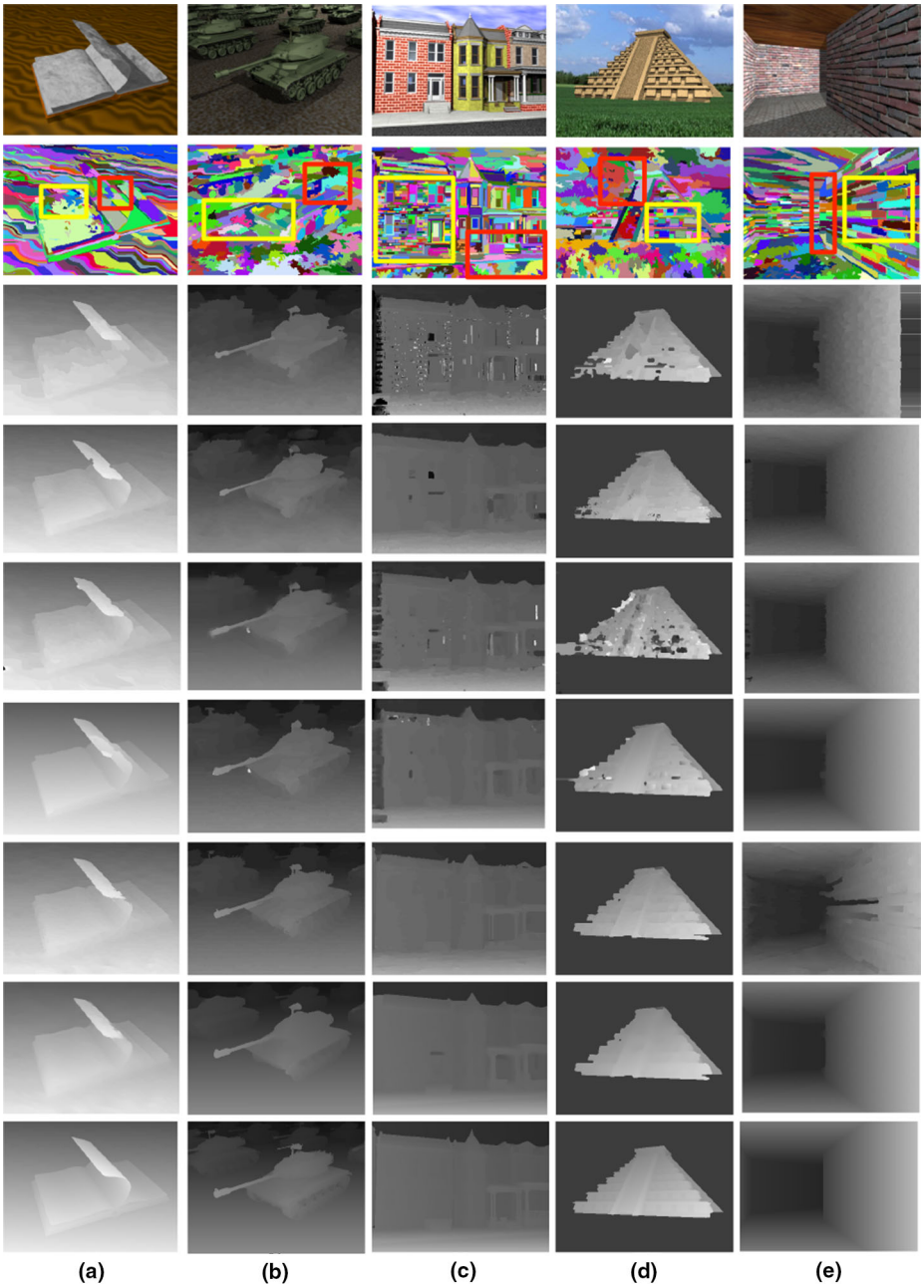

Fig. 5 The qualitative comparison between the proposed method and other state-of-the-art methods. **a** The 20th frame of Book sequence, **b** the 40th frame of Tanks sequence, **c** the 93th frame of Street sequence, **d** the 59th frame of Temple sequence, **e** the 84th of Tunnel sequence. The first and second rows in each column are the left texture image and its initial color-based segmentation. From the third row to the bottom in each column are the results obtained from: Liu [27], Pham [37], Hosni [14], Jiang [16], Initial Results [29], ours, the ground truth

Jiang et al. [16] applied 3D registration to estimate the motion of a stereo rig using feature pixels and transferred the previous disparity values to the current frame based on an ego-motion transformation between adjacent disparity maps. The estimated ego-motion is based on global spatial and temporal homography between adjacent frames. However, there are few feature pixels compared with the total number of pixels in the image, so the estimated global ego-motion model cannot maintain the temporally consistent motion of all pixels over time. Furthermore, illumination variations, occlusions, and noise often cause significant disparity errors, which results in ambiguities when modeling the homography. Compared with their results, our error rate was 67% smaller.

Additionally, we also compared the proposed method with the frame-by-frame manner using our previous work [29] that is viewed as the initial disparity maps without considering the spatiotemporally consistent constraints (referred to as "Initial results"). Compared with initial results, our error rate was 49% smaller.

Based on the above discussion, our method is superior to that of other state-of-art methods with respect to the mean error rate, which implies that our results are quantitatively closer to the ground truth. Additionally, the lower variance of our method implies that it is temporally stable for a video input.

Fig. 6 Disparity maps of consecutive frames in the "Book" sequences. For each column, from top to bottom is the result with Liu [27], Pham [37], Hosni [14], Jiang [16], Initial Results [29], ours and the ground truth. For each row, from left to right is the results of consecutive frames from 20th to 24th

Table 2 Error rate of our method with the results using other state-of-art algorithms, over all frames for each DCB sequence

	Error rate (%)											
	Book		Tanks		Street		Temple		Tunnel		Average	
	Mean	Variance	Mean	Variance	Mean	variance	Mean	Variance	Mean	Variance	Mean	Variance
Liu [27]	5.62	3.28	7.21	1.21	9.22	1.05	7.32	2.06	7.50	4.26	7.59	2.44
Pham [37]	19.39	1.37	13.79	1.01	16.43	1.20	10.77	1.35	13.99	6.30	14.87	2.24
Hosni [14]	10.10	0.75	12.20	1.83	8.70	1.04	6.60	**1.16**	17.70	**3.10**	11.06	1.57
Jiang [16]	14.28	0.76	11.52	1.50	12.10	0.98	13.90	2.51	7.24	4.54	11.80	2.06
Initial results [29]	4.33	0.50	5.69	1.31	6.73	**0.74**	4.30	1.30	7.91	5.48	5.79	1.86
Proposed method	**3.27**	**0.80**	**4.85**	**0.54**	**4.74**	1.46	**3.84**	1.24	**2.71**	3.52	**3.88**	**1.51**

Minimum value in each column is bold

6.2 Evaluation using the realistic KITTI stereo multi-view extension datasets

It is important to evaluate the proposed method on different datasets to test its adaptability and accuracy. So we conducted evaluations on the challenging real-world scenes (KITTI 2012 and KITTI 2015 stereo multi-view extension datasets [7,33]) to assess the performance of the proposed method. For the task of obtaining stereo, two datasets are nearly identical, but the newer one contains more luminance variance, texture-less regions and fast motions that are more difficult to stereo matching task. The KITTI 2012 dataset contains 194 training and 195 testing scenes, while the KITTI 2015 dataset contains 200 training and 200 testing scenes. Each

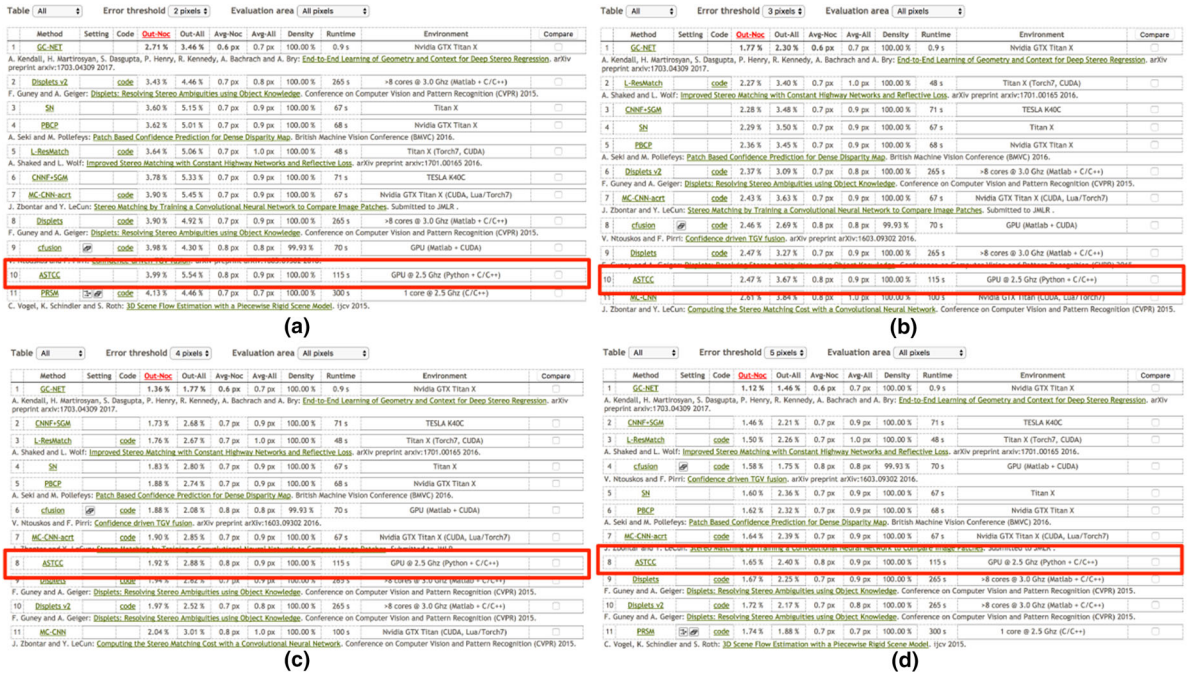

Fig. 7 Quantitative comparisons on the KITTI 2012 stereo multi-view extension scenes in the benchmark [19]. This error rate is computed as the average number of error pixels of *tenth* frame over all testing scenes, for which the estimated disparity differs from the ground truth by less than a fixed error threshold (from **a** to **d** the error thresholds are 2,3,4,5). Out-Noc: percentage of erroneous pixels in non-occluded areas; Out-All: percentage of erroneous pixels in entire image; Avg-Noc: average disparity error in non-occluded pixels; Avg-All: average disparity error in all pixels. The "Runtime" column measures the time, in seconds, required to process one scene pair images, and the "Environment" illustrates the computer configuration. Our results (ASTCC) have been marked using red rectangles. We have achieved comparable results for the KITTI 2012 dataset

scene in the above two datasets consists of 20 consecutive frames. The ground truth disparity maps for testing scenes are withheld and two online benchmarks [19,20] are provided where researchers can evaluate their methods on these testing scenes. Error rate is measured as the average percentage of error pixels of **tenth** frame over all testing scenes. Submissions are allowed once per hour and three times per month.

Figure 7 presents the evaluation results on 195 testing scenes of the KITTI 2012 between the proposed method and other state-of-the-art algorithms. Among approximately 92 methods listed on the benchmark [19] (at the time of July 2018), in terms of the error threshold 2 and 3, the proposed method yields 3.99% and 2.47% error rates, ranking the ***tenth*** place. On the other hand, in terms of the error threshold 4 and 5, the proposed method ranks the ***eighth*** with the error ratios of 1.92% and 1.65%.

Figure 8 shows the evaluation results on the 200 testing scenes of the KITTI 2015 dataset. The error rates of the proposed methods are 2.94%, ranking the ***tenth*** place among approximately 66 methods listed on the benchmark [20] (at the time of July 2018).

Some pairs of disparity maps generated by our method are presented in Figs. 9 and 10. We can see that our method obtains piecewise smooth and visually plausible results. It dose not only preserve geometry details near depth discontinuities (white arrows), but also performs well on challenging regions such as luminance variance (orange arrows) and texture-less regions (red arrows). But the proposed method also suffers from the transparent regions, such as automobile window glass and fences (black rectangles).

It is worth noting that the proposed method follows the conventional global stereo matching framework [13,29] without incorporating deep learning (DL) or convolutional neural network (CNN) knowledge. As listed in Tables 3 and 4, we compare the performances between our method and other state-of-the-art DL- or CNN-based methods whose results are available in two benchmarks [19,20]. In these tables, the methods in benchmarks without published are not listed. We can see that our results (ASTCC) are comparable to the state-of-the-art DL- and CNN-based algorithms. This indicates that our method is generally accurate. Furthermore, the proposed method is the only ***one*** without DL- or CNN-based framework in top ten of two benchmarks [19,20].

On the contrary, we compared our results to those of conventional stereo matching methods (without DL- or CNN-based framework) in Tables 5 and 6. The methods in two benchmarks without published are also not listed. For

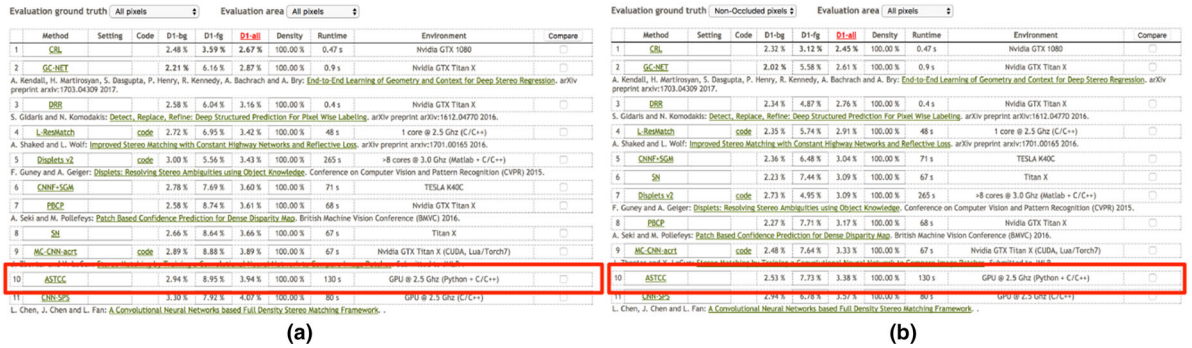

Fig. 8 Quantitative comparison of different methods on the KITTI 2015 stereo multi-view extension scenes. **a** The evaluation result on all image pixels. **b** The evaluation result on non-occlusion image pixels. Error rate is computed as the average number of error pixels of $tenth$ frame over all testing scenes, for which the estimated disparity differs from the ground truth by $\leq$ 3px or $\leq$ 5% error. The "Runtime" and "Environment" columns have the same meaning as in Fig. 7. D1: Percentage of stereo disparity outliers in first frame. bg/fg/all: Percentage of outliers averaged only over background/foreground/all regions. Our results (ASTCC) have been marked using red rectangles. We can obtain a satisfy results for the KITTI 2015 stereo dataset

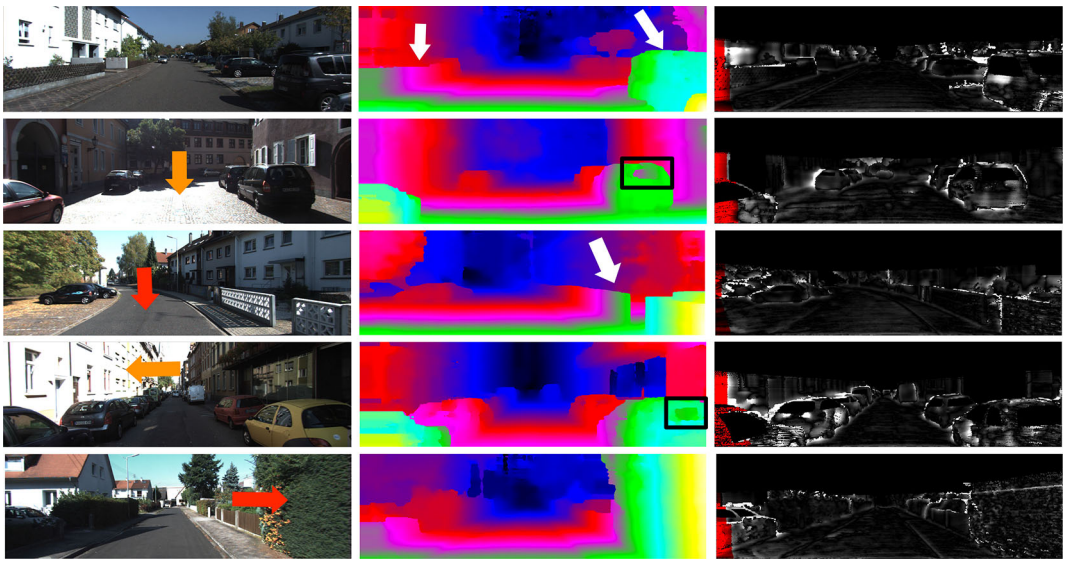

Fig. 9 Qualitative results generated by the proposed method using the KITTI 2012 stereo multi-view extension scenes: From up to bottom, the scenes are the 0, 3, 5, 9 and 10. For each scene, from left to right we show the rectified left images of the $tenth$ frame in each scene, the estimated disparity maps and the corresponding disparity error maps. The error map scales linearly between 0 (black) and $\geq$5 (white) pixels error. Red denotes all occluded pixels. The false color map is scaled to the largest ground truth disparity

example, as listed in Table 5 which is obtained from the KITTI 2012 benchmark, the proposed method significantly outperforms the conventional stereo matching methods on the non-occluded regions. It achieves a 3 pixel error threshold of 2.47%, while the second best performing method provides 2.78%. The accuracy increased almost 12%.

As listed in Table 6 obtained from the KITTI 2015 benchmark, the proposed method significantly outperforms the conventional stereo matching methods in all categories, which shows the effectiveness of the adaptive, spatiotemporally consistent, constraints-based framework. It further proves the better accuracy of global method for stereo matching. For example, our method yields 8.95% error ratio in the "D1-fg" column, while the second best performing method increases shapely to 10.52%. The accuracy increased almost 18%. The corresponding visual comparisons are illustrated in Fig. 11. It can be demonstrated that:

– Because of incorporating the adaptive temporal predicted disparity constraint, we reduce the ambiguities and noises caused by luminance variance and less texture (such as the ground, tree, shadow and sky in orange arrows and rectangles in Fig. 11), where often appear in conventional stereo matching methods.

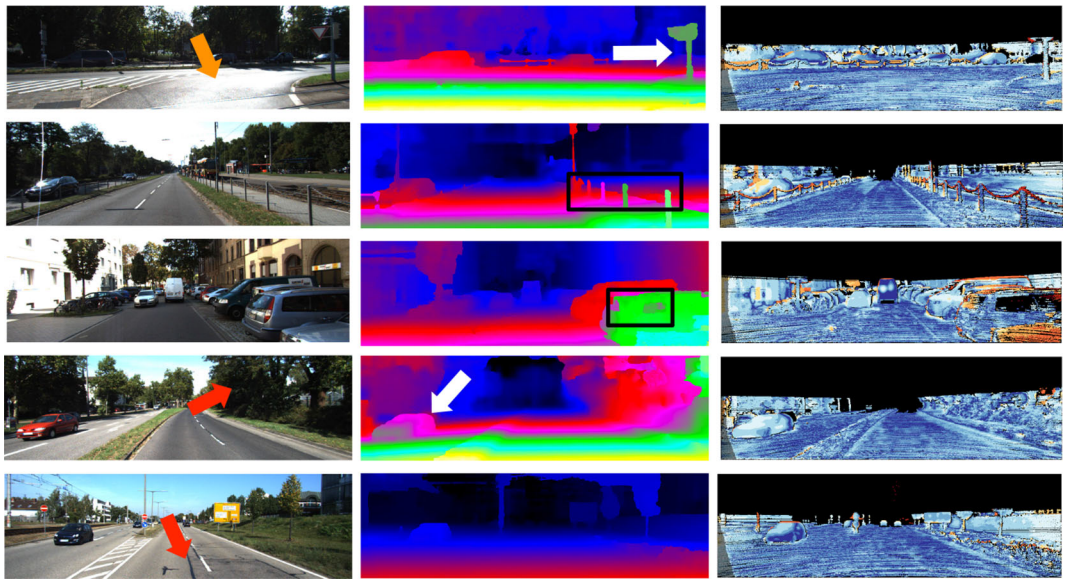

Fig. 10 Qualitative results of the KITTI 2015 stereo multi-view extension dataset: From up to bottom, the scenes are the 0, 3, 6, 13 and 17. For each scene, from left to right we show the rectified left images of the *tenth* frame, the estimated disparity maps and the corresponding disparity error maps. The error map uses the log-color scale described in [33], depicting correct estimates ($\leq$ 3px or $\leq$ 5% error) in blue and wrong estimates in red color tones. Dark regions in the error images denote the occluded pixels which fall outside the image boundaries. The false color maps of the results are scaled to the largest ground truth disparity

Table 3 Comparison to the state-of-the-art DL- or CNN-based stereo matching methods in the KITTI 2012 benchmark. "Out-Noc," "Out-All," "Avg-Noc," "Avg-All," and "Runtime" columns have the same meaning as in Fig. 7

	Error ratio (%)								
	Rank in Benchmark [19]	> 2 pixels		> 3 pixels		> 4 pixels		> 5 pixels	
		Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All
GC-NET [17]	1	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46
L-ResMatch [41]	2	3.64	5.06	2.27	3.40	1.76	2.67	1.50	2.26
Displets v2 [11]	6	3.43	4.46	2.37	3.09	1.97	2.52	1.72	2.17
MC-CNN-acrt [53]	7	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39
cfusion [36]	8	3.98	4.30	2.46	2.69	1.88	2.08	1.58	1.75
Ours (ASTCC)	10	3.99	5.54	2.47	3.67	1.92	2.88	1.65	2.40
Content-CNN [30]	16	4.98	6.51	3.07	4.29	2.39	3.36	2.03	2.82
Deep Embed [3]	18	5.05	6.47	3.10	4.24	2.32	3.25	1.92	2.68

The evaluation table ranks all methods according to the number of non-occluded erroneous pixels when the error threshold is 3

– Because of considering the adaptive temporal segment confidence as a soft guide, we can avoid to mixture the foreground and background at object boundaries as well as preserve the edges better with less error pixels (white rectangles in Fig. 11).

6.3 Quantitative evaluation for each constraint term using the KITTI stereo multi-view extension datasets

In addition, we conducted experiments to investigate how the individual constraint term in Eq. 20 affects the results by using the KITTI 2012 and KITTI 2015 stereo multi-view extension training datasets. Because only providing the ground truth of the **tenth** frame, we only calculated the error ratios of non-occluded erroneous pixel of the **tenth** frame in each training scenes. In each experiment, we omitted one part of our method and retained the remaining parts. The error ratios are listed in Table 7.

Firstly, we omitted the adaptive temporal predicted disparity constraint ($ATPDC$) term by setting $w_q^H := 1$ and $w_q^T := 0$, meaning the scope of disparity variance was no longer restricted by the strength of temporal links. Error occurred because of the inclusion of some pixels that are eas-

Table 4 Comparison to the state-of-the-art DL- or CNN-based stereo matching methods in the KITTI 2015 benchmark

	Error ratio (%)			
	Rank	D1-bg	D1-fg	D1-all
GC-NET [17]	2	2.21	6.16	2.87
DRR [8]	3	2.58	6.04	3.16
L-ResMatch [41]	4	2.72	6.95	3.42
Displets v2 [11]	5	3.00	5.56	3.43
PBCP [40]	7	2.58	8.74	3.61
MC-CNN-acrt [53]	9	2.89	8.88	3.89
Ours (ASTCC)	10	2.94	8.95	3.94
DispNetC [32]	14	4.32	4.41	4.34
Content-CNN [30]	18	3.73	8.58	4.54

"D1-bg," "D1-fg," and "D1-all" columns have the same meaning as in Fig. 8. "Rank" is the rank index in the benchmark [20]

ily affected by the luminance variation as well as the texture distribution. The temporally consistent and smooth distribution of disparities between adjacent frames were violated. The error rate of KITTI 2012 and KITTI 2015 datasets in non-occlusion regions sharply increased to 3.23% and 3.99%, respectively.

Secondly, the adaptive temporal segment confidence constraint ($ATSCC$) term was turned off by setting $\bar{S}^t(q_0, q_1) := 1$. Then, the smoothness term became the traditional first-order smoothness one that typically leads to a frontal-parallel disparity map and contains ambiguities caused by the over- and under-segmentation. The error rate in non-occlusion regions sharply increased to 2.85% and 3.64%, respectively.

We can see that the proposed method generates higher-quality results and is robust to different scenes when all terms were applied.

Table 6 Comparison to state-of-the-art conventional stereo matching methods in the KITTI 2015 benchmark

	Error ratio (%)			
	Rank	D1-bg	D1-fg	D1-all
PRSM [46]	12	3.02	10.52	4.27
3DMST [23]	21	3.36	13.03	14.97
SPS-St [51]	26	3.84	12.67	5.31
Ours (ASTCC)	**10**	**2.94**	**8.95**	**3.94**
MDP [32]	28	4.19	11.25	5.36

"D1-bg," "D1-fg" and "D1-all" columns have the same meaning as in Fig. 8. "Rank" is the rank index in the benchmark [20]. Minimum value in each column is bold

6.4 Discussion

Evaluations have demonstrated that the proposed method significantly improves the spatiotemporal consistency both quantitatively and qualitatively. However, similar to most disparity estimation methods, our method suffers from the transparent regions, such as automobile window glass (black rectangles in Figs. 9 and 10); disparity estimate in these regions may lead to errors.

Furthermore, another limitation of our method is that if the object is extremely thin and long, its color is similar to the background that leads to incorrect segmentation. As shown in the black rectangle in the second row of Fig. 10, the iron chain is extremely thin and long. It was segmented into the background. So if there is no extra prior knowledge, obtaining true disparity of the iron chain is very difficult.

Additionally, the proposed method was implemented on a PC with Core i5-2500 3.30 GHZ CPU and 8 GB RAM. It is obvious that the computational time is proportional to the image size. For example, it took approximately 15 s to obtain results on DCB data and 110 s on the real-world scene. Currently, all steps were implemented offline. Next, we aim to implement our algorithm on a GPU to achieve a good balance between accuracy and efficiency.

Table 5 Comparison to the state-of-the-art conventional stereo matching methods in the KITTI 2012 benchmark

	Error ratio (%)								
	Rank in Benchmark [19]	> 2 pixels		> 3 pixels		> 4 pixels		> 5 pixels	
		Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All
PRSM [46]	12	4.13	**4.46**	2.78	**3.00**	2.15	**2.31**	1.74	**1.88**
SPS-St [51]	13	4.30	5.39	2.83	3.64	2.24	2.89	1.90	2.46
VC-SF [45]	15	4.56	4.93	3.05	3.31	2.35	2.55	1.92	2.09
JSOSM [24]	19	4.76	5.82	3.15	3.94	2.45	3.06	2.05	2.55
Ours (ASTCC)	**10**	**3.99**	5.54	**2.47**	3.67	**1.92**	2.88	**1.65**	2.40

"Out-Noc," "Out-All," "Avg-Noc," "Avg-All," and "Runtime" columns have the same meaning as in Fig. 7. The evaluation table ranks all methods according to the number of non-occluded erroneous pixels when the error threshold is 3. Minimum value in each column is bold

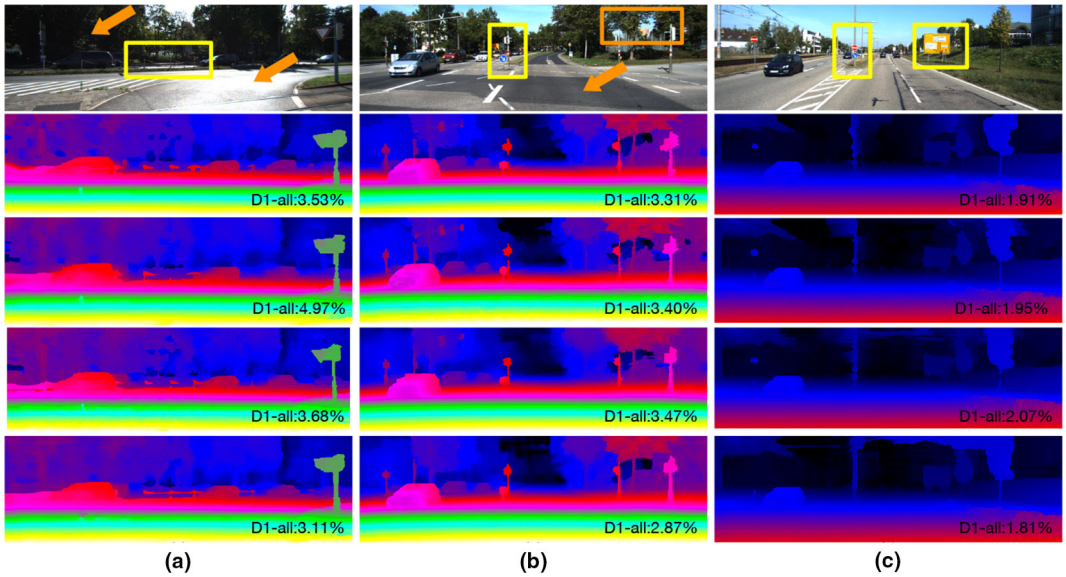

Fig. 11 Comparison results between the proposed method and conventional stereo matching methods using the KITTI 2015 scenes. **a** Testing 0 scene. **b** Testing 10 scene. **c** Testing 19 scene. For each column, disparity maps arranged from top to bottom are generated by PRSM [46], 3DMST [23], SPS-St [51], MDP [32] and ours. Black numbers are the percentage of error pixels averaged over all ground truth

Table 7 Error ratio in the KITTI 2012 and 2015 multi-view extension training datasets with different constraint terms turned off

	Error ratio (%)			
	KITTI 2012		KITTI 2015	
	Mean	Variance	Mean	Variance
Initial	3.43	2.79	4.40	3.64
ATPDC off	3.23	2.44	3.99	4.06
ATSCC off	2.85	2.33	3.64	3.19
All terms on	**2.47**	**1.88**	**2.94**	**2.20**

The error ratios are denoted as the number of non-occluded erroneous pixels when the error threshold is 3. "ATPDC": Adaptive Temporal Predicted Disparity Constraint. "ATSCC": Adaptive Temporal Segment Confidence Constraint. Minimum value in each column is bold

7 Conclusion

In this paper, we proposed an adaptive, spatiotemporally consistent, constraints-based systematic method that generates spatiotemporally consistent disparity maps for stereo video image sequences. The major contributions are the reliable temporal neighborhood, the adaptive temporal predicted disparity map and the adaptive temporal segment confidence. The evaluations indicate that the proposed method performs almost 65% better than others in the aspect of precision on the DCB datasets [39]. In addition, our method ranks ***tenth*** in both realistic KITTI 2012 and KITTI 2015 datasets benchmarks [19,20], respectively. It is worth noting that our results is comparable to the recent state-of-the-art DL- and CNN- based algorithms. In the future, we will intend to focus on obtaining accurate disparity estimate in the transparent regions and transforming our method to a parallel GPU implementation.

Funding This study was funded by the National Natural Science Foundation of China (Grant No.: 61802109), the Natural Science Foundation of Hebei province (Grant No.: F2017205066), the Science Foundation of Hebei Normal University (Grant No.: L2017B06, L2018K02).

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Bartczak, B., Jung, D., Koch, R.: Real-Time Neighborhood Based Disparity Estimation Incorporating Temporal Evidence, pp. 153–162. Springer, Berlin (2008)
2. Čech, J., Sanchez-Riera, J., Horaud, R.: Scene flow estimation by growing correspondence seeds. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3129–3136. IEEE (2011)
3. Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C.: A deep visual correspondence embedding model for stereo matching costs. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 972–980 (2015)
4. Dahan, M.J., Chen, N., Shamir, A., Cohen-Or, D.: Combining color and depth for enhanced image segmentation and retargeting. Vis. Comput. **28**(12), 1181–1193 (2012)
5. Davis, J., Ramamoorthi, R., Rusinkiewicz, S.: Spacetime stereo: a unifying framework for depth from triangulation. In: Proceedings.

2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, vol. 2, pp. II–359. IEEE (2003)
6. Dobias, M., Sara, R.: Real-time global prediction for temporally stable stereo. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 704–707 (2011)
7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the Kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361. IEEE (2012)
8. Gidaris, S., Komodakis, N.: Detect, replace, refine: deep structured prediction for pixel wise labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5248–5257 (2017)
9. Gong, M.: Real-time joint disparity and disparity flow estimation on programmable graphics hardware. Comput. Vis. Image Underst. **113**(1), 90–100 (2009)
10. Guerrero, P., Winnemöller, H., Li, W., Mitra, N.J.: Depthcut: improved depth edge estimation using multiple unreliable channels. Vis. Comput. **34**(9), 1165–1176 (2017)
11. Guney, F., Geiger, A.: Displets: resolving stereo ambiguities using object knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4165–4175 (2015)
12. Hamming distance. https://en.wikipedia.org/wiki/Hamming_distance
13. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)
14. Hosni, A., Rhemann, C., Bleyer, M., Gelautz, M.: Temporally Consistent Disparity and Optical Flow via Efficient Spatio-Temporal Filtering, pp. 165–177. Springer, Berlin (2012)
15. Hung, C.H., Xu, L., Jia, J.: Consistent binocular depth and scene flow with chained temporal profiles. Int. J. Comput. Vis. **102**(1–3), 271–292 (2013)
16. Jiang, J., Cheng, J., Chen, B., Wu, X.: Disparity prediction between adjacent frames for dynamic scenes. Neurocomputing **142**, 335–342 (2014)
17. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression (2017). arXiv preprint arxiv:1703.04309
18. Khoshabeh, R., Chan, S.H., Nguyen, T.Q.: Spatio-temporal consistency in video disparity estimation. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 885–888. IEEE (2011)
19. Kitti 2012 stereo benchmark. http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo
20. Kitti 2015 stereo benchmark. http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo
21. Kordelas, G.A., Alexiadis, D.S., Daras, P., Izquierdo, E.: Revisiting guided image filter based stereo matching and scanline optimization for improved disparity estimation. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 3803–3807. IEEE (2014)
22. Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
23. Li, L., Yu, X., Zhang, S., Zhao, X., Zhang, L.: 3d cost aggregation with multiple minimum spanning trees for stereo matching. Appl. Opt. **56**(12), 3411–3420 (2017)
24. Li, X., Liu, J.: Efficient stereo matching using segment optimization. In: ICIP (2016)
25. Li, Y., Zhang, J., Zhong, Y., Wang, M.: An efficient stereo matching based on fragment matching. Vis. Comput. 1–13 (2018). https://doi.org/10.1007/s00371-018-1491-0
26. Lin, S.H., Chung, P.C.: Temporal consistency enhancement of depth video sequence. In: 2014 International Conference on Information Science, Electronics and Electrical Engineering (ISEEE), vol. 3, pp. 1897–1900. IEEE (2014)
27. Liu, F., Philomin, V.: Disparity estimation in stereo sequences using scene flow. In: Proceedings of the British Machine Vision Conference, pp. 55.1–55.11. BMVA Press (2009)
28. Liu, J., Li, C., Fan, X., Wang, Z., Shi, M., Yang, J.: View synthesis with 3d object segmentation-based asynchronous blending and boundary misalignment rectification. Vis. Comput. **32**(6), 989–999 (2016)
29. Liu, J., Li, C., Mei, F., Wang, Z.: 3d entity-based stereo matching with ground control points and joint second-order smoothness prior. Vis. Comput. **31**(9), 1253–1269 (2015)
30. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5695–5703 (2016)
31. Matsuo, T., Fukushima, N., Ishibashi, Y.: Weighted joint bilateral filter with slope depth compensation filter for depth map refinement. VISAPP **2**, 300–309 (2013)
32. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4040–4048 (2016)
33. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3061–3070 (2015)
34. Min, D., Lu, J., Do, M.N.: Depth video enhancement based on weighted mode filtering. IEEE Trans. Image Process. **21**(3), 1176–1190 (2012)
35. Min, D., Yea, S., Vetro, A.: Temporally consistent stereo matching using coherence function. In: 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010, pp. 1–4. IEEE (2010)
36. Ntouskos, V., Pirri, F.: Confidence driven tgv fusion (2016). arXiv preprint arXiv:1603.09302
37. Pham, C.C., Nguyen, V.D., Jeon, J.W.: Efficient spatio-temporal local stereo matching using information permeability filtering. In: 2012 19th IEEE International Conference on Image Processing, pp. 2965–2968 (2012)
38. Qi, F., Zhao, D., Liu, S., Fan, X.: 3d visual saliency detection model with generated disparity map. Multimed. Tools Appl. **76**(2), 3087–3103 (2017)
39. Richardt, C., Orr, D., Davies, I., Criminisi, A., Dodgson, N.A.: Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In: European Conference on Computer Vision, pp. 510–523. Springer (2010)
40. Seki, A., Pollefeys, M.: Patch based confidence prediction for dense disparity map. In: BMVC, vol. 2, p. 4 (2016)
41. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective loss (2016). arXiv preprint arxiv:1701.00165
42. Sizintsev, M., Wildes, R.P.: Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 493–500. IEEE (2009)
43. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439 (2010)
44. Taniai, T., Sinha, S.N., Sato, Y.: Fast multi-frame stereo scene flow with motion segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6891–6900. IEEE (2017)
45. Vogel, C., Roth, S., Schindler, K.: View-consistent 3d scene flow estimation over multiple frames. In: European Conference on Computer Vision, pp. 263–278. Springer (2014)

46. Vogel, C., Schindler, K., Roth, S.: 3d scene flow estimation with a piecewise rigid scene model. Int. J. Comput. Vis. **115**(1), 1–28 (2015)
47. Vretos, N., Daras, P.: Temporal and color consistent disparity estimation in stereo videos. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 3798–3802. IEEE (2014)
48. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3d motion understanding. Int. J. Comput. Vis. **95**(1), 29–51 (2011)
49. Xing, G., Liu, Y., Zhang, W., Ling, H.: Light mixture intrinsic image decomposition based on a single rgb-d image. Vis. Comput. **32**(6–8), 1013–1023 (2016)
50. Xu, S., Zhang, F., He, X., Shen, X., Zhang, X.: Pm-pm: patchmatch with potts model for object segmentation and stereo matching. IEEE Trans. Image Process. **24**(7), 2182–2196 (2015)
51. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: European Conference on Computer Vision, pp. 756–771. Springer (2014)
52. Yang, W., Zhang, G., Bao, H., Kim, J., Lee, H.Y.: Consistent depth maps recovery from a trinocular video sequence. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1466–1473. IEEE (2012)
53. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. **17**(1–32), 2 (2016)
54. Zeng, H., Ma, K.K.: Content-adaptive temporal consistency enhancement for depth video. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 3017–3020. IEEE (2012)
55. Zhang, G., Jia, J., Wong, T.T., Bao, H.: Consistent depth maps recovery from a video sequence. IEEE Trans. Pattern Anal. Mach. Intell. **31**(6), 974–988 (2009)
56. Zhu, S., Yan, L.: Local stereo matching algorithm with efficient matching cost and adaptive guided image filter. Vis. Comput. **33**(9), 1087–1102 (2017)

Liang Tian was born in 1981 in Hebei Province, China. He is a Ph.D. candidate at Hebei Key Laboratory of Computational Mathematics and Application, College of mathematics and information science, Hebei Normal University. His research interests include computer vision, image processing, deep learning and augmented reality.

Jing Liu was born in 1985 in Hebei Province, China. He received his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2016. He is currently a lecturer at the Hebei Normal University. His main research interests are augmented reality, medical image analysis and computer vision.

Haibin Ling received the B.S. degree in mathematics and the MS degree in computer science from Peking University, China, in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in Computer Science in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. Since fall 2008, he has been with Temple University where he is now an Associate Professor. Dr. Ling's research interests include computer vision, augmented reality, medical image analysis, and human–computer interaction. He serves as associate editors for IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition (PR), and Computer Vision and Image Understanding (CVIU), and had also served or will serve as Area Chairs for CVPR 2014, CVPR 2016, and CVPR 2019.

Wei Guo is a professor and Ph.D. supervisor at Hebei Key Laboratory of Computational Mathematics and Application, College of mathematics and information science, Hebei normal university. Her research interests include wavelet analysis, image processing and augmented reality.