IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 November 2023, accepted 21 November 2023, date of publication 24 November 2023,
date of current version 30 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336700

RESEARCH ARTICLE

Seam Generation Matrix Based on a Guided Energy-Depth Map for Image and Video Stitching

SEONGBAE RHEE[1], GWANG HOON PARK[2], (Senior Member, IEEE), AND KYUHEON KIM[3], (Member, IEEE)
[1]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Seoul 17104, South Korea
[2]Department of Computer Science and Engineering, Kyung Hee University, Seoul 17104, South Korea
[3]Department of Electronic Engineering, Kyung Hee University, Seoul 17104, South Korea

Corresponding author: Kyuheon Kim (kyuheonkim@khu.ac.kr)

This work was supported in part by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) Grant funded by the Korea Government through Defense Acquisition Program Administration (DAPA) (Space-Layer Intelligent Communication Network Laboratory, 2022) under Grant 21-106-A00-007; and in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2021-0-02046) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT An image captured by a single camera has a smaller viewing angle than that of the human eye. One method to expand this viewing angle is a technique known as image stitching, which generates a wider view from images captured with multiple cameras. Although this technique has found uses in multiple industries, it is vulnerable to parallax distortion, wherein objects disappear from or repeatedly appear in stitched images when the parallax between cameras differs significantly. To minimize parallax distortion, seam-based and multi-homography-based methods have been proposed. In particular, the seam-based method enables faster image stitching owing to its intuitive procedure; however, the seam generation matrix may still incur parallax distortion under certain restrictive circumstances, and a longer stitching time is required when this method is applied to video sequences. This motivated us to develop the Guided Energy–Depth Map, which uses the energy function, depth information, and guidance map to minimize parallax distortion from a human visual perspective and reduce the time required to apply the stitching process to video sequences. Based on Average Seam Error (ASE) evaluation, the proposed method produces better seams than energy functions in 25 out of 32 experimental datasets, and the improvement rate of ASE evaluation is 15.58%. Also, the Frame Selection module for video stitching proposed in this paper takes only 7.27% of the time to find a specific frame for seam regeneration compared to the instance segmentation-based frame selection method.

INDEX TERMS Depth, energy map, image stitching, seam, moving search.

I. INTRODUCTION

Immersive media content, represented by panoramic and 360-degree videos, has recently been implemented for various applications – such as games, sports, and shopping – and is expected to spread to specialized fields including education, military, medical care, and manufacturing [1]. This content is characterized by ultra-high-definition video with a wider field of view than that of humans; however, current image capture technologies are limited to single conventional cameras with a narrow field of view. To obtain a wider field of view, images can be either captured using a camera with a wide-angle lens, or generated using image stitching, a technique that combines images captured by multiple conventional cameras [2]. However, images captured with a wide-angle lens generally exhibit radial distortion owing to the lens curvature. Because the accurate appearance of objects is important in immersive media content, such content is mainly generated via image stitching.

Image stitching includes the processes of keypoint extraction and matching, homography estimation overlap region calculation, warping, and synthesis [2]. Although image stitching mitigates the issue of radial distortion, the existence of a large parallax between cameras may incur parallax

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

distortion, wherein objects disappear or overlap in the stitched image [3].

The multi-homography- and seam-based image stitching methods are representative approaches designed to minimize the parallax distortion in stitched images [4]. The multi-homography-based method reduces the parallax distortion by dividing an image into multiple patches, estimating the homography for each segmented patch, and minimizing the differences between patches by warping [5]. However, this approach is vulnerable to local distortion, which requires a highly complex correction procedure [4].

In contrast, the Seam-based Image Stitching (SIS) method constructs a Seam Generation Matrix (SGM) using a visual cognitive energy function in the overlapping regions of two input images, and synthesizes along the seam with minimal accumulation of SGM [6]. Although SIS is highly dependent on accurate homography, the use of seams in image stitching has the advantage of a fast stitching because it is a simple structure that synthesizes two warped images along the seam. In addition, because a large weight could be assigned to the object region of the SGM, a seam is generated to avoid an object, which prevents parallax distortion from occurring in the object region. However, in image stitching based on seam generation matrix, correct seam generation may be limited depending on the position of the seam that is initially generated, performance of the object detector, and placement of objects, which may cause parallax distortion [3].

As described above, image-stitching methods based on both multi-homography and seam may distort the stitched image. Although methods based on multi-homography are difficult to implement in practice due to local distortion, the limitations of the method based on seam can be overcome by defining the proper SGM. This encouraged us to propose an image stitching method that can overcome parallax distortion by constructing a seam generation matrix, which combines depth information, predicted using a deep learning network, and visual cognitive energy functions. In addition, we propose a method to apply video stitching by exploiting the efficiency of seam-based image.

Section II examines the limitations of seam-based image stitching, and Section III describes the proposed method. We present an analysis of our experimental results in Section IV, and provide concluding remarks in Section V.

II. BACKGROUND

In general, Seam-based Image Stitching (SIS) comprises the following modules, as illustrated in Figure 1: Keypoint Extraction, Keypoint Matching, Homography, Warping, Composition, Object Weight, and Blending [7]. The Keypoint Extraction module extracts robust keypoints regardless of the image size, rotation, and brightness of the various input images. The Keypoint Matching module removes outliers from the extracted keypoints using RANSAC [8] and matches keypoints corresponding to the same location to identify overlapped regions. In the Warping module, the overlapping regions are warped to the same plane through homography, which is estimated using the matched keypoints. The Composition module of the SIS method synthesizes images using a seam, which is generated along points, producing a minimum difference between overlap regions to generate a naturally stitched image. However, when a seam is generated in the object region, distortions such as the duplication or disappearance of objects may occur. Thus, the Object Weight module detects the object area and assigns a large weight value to that area to prevent seam generation. In addition, the difference in the brightness or luminance between the images to be stitched is compensated for in the Blending module. Moreover, Seam-based Video Stitching (SVS) is realized by applying SIS to every frame of a video sequence.

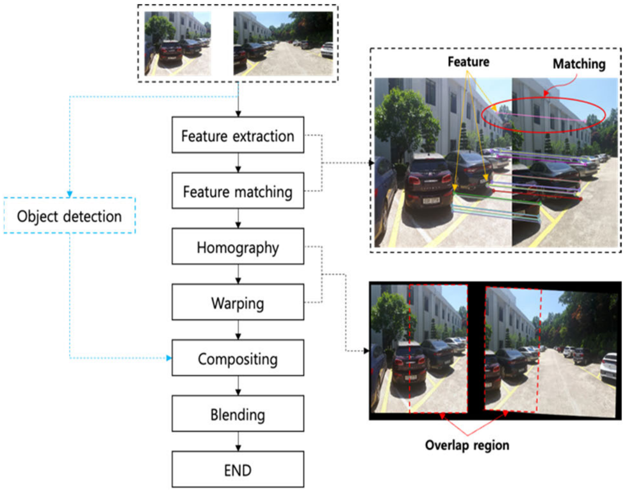


FIGURE 1. Pipeline of seam-based image stitching.

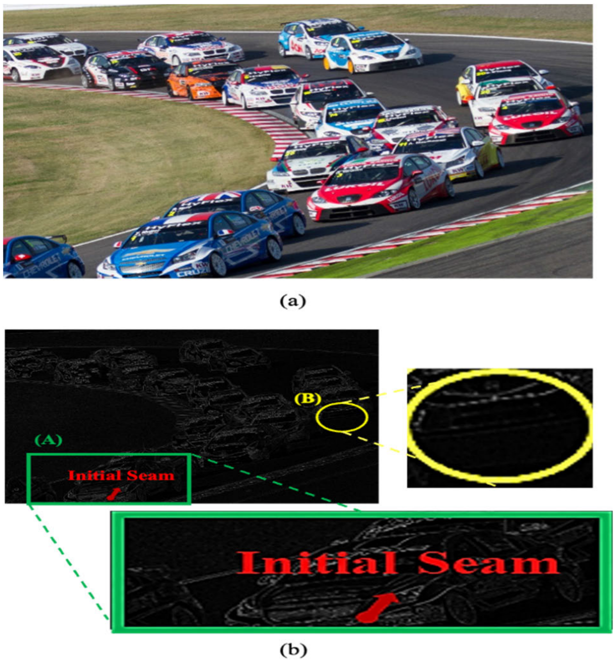
The Seam Generation Matrix (SGM) of SIS could be composed of a pixel intensity [9], a visual perception energy function [10], [11], and an object weight [3], [7], [12], which is used to prevent parallax distortion caused by the seam generated in the object region. For example, larger weights can be assigned to the corresponding edges of the background and object in the SGM, as well as to the inner regions of the object in the overlapped regions. This prevents a seam from being generated in these regions because it is generated along the path where the cumulative sum of the SGM is minimal [13].

However, conventional SIS has difficulties not only generating accurate edges of the background and objects, including the inner regions of the objects, but also assigning appropriately large weights in the SGM. Additionally, it does not consider the case in which the minimum path for a seam is not generated owing to the characteristics of the input images, such as when an object extends horizontally across an image. The following sections provide more details on the limitations of conventional SIS.

A. LIMITATIONS OF CONVENTIONAL SEAM GENERATION MATRIX IN SEAM-BASED IMAGE STITCHING

Section II-A presents an analysis of the limitations of the conventional SGM used in the SIS. As previously explained, an SGM could be composed of a visual perception energy function [10], where the visual cognitive energy function in (1) is defined as the sum of the magnitudes of the change of the horizontal and vertical pixels. Thus, the visual perception energy function in an image is set to be larger in the outline because of the larger pixel change, and smaller in the inner region of the object because of the smaller pixel change. For example, an input image and its visual cognitive energy function are shown in Figure 2(a) and (b), respectively. As shown in area (A) of Figure 2(b), the inner region of the object is dark because the weight value for this region is smaller, which means that a seam could be generated in the inner region of the object, resulting in parallax distortion.

$$\mathrm{E}\left(\mathrm{x},\mathrm{y}\right)=\left|\frac{\partial}{\partial X}I\left(x,y\right)\right|+\left|\frac{\partial}{\partial Y}I\left(x,y\right)\right| \tag{1}$$


(a)

(b)

FIGURE 2. **Visual cognitive energy function.**

As the visual cognitive energy function is simply the magnitude of the pixel change, it may be difficult to set an appropriate weight value for the region where the seam should not be generated. For example, the yellow area (B) in Figure 2(b) is an object region; therefore, a higher weight value should be set, and a white line should appear. However, a smaller weight value was set and the area appears dark, as the pixel values are similar between the background and object. Consequently, it is difficult to mitigate the parallax distortion that occurs in SGM-based image stitching using only the energy function based on visual perception.

To overcome this limitation, other SGM construction methods have been proposed for detecting objects and setting higher weight values in the inner regions of objects [3], [7]. Various object detection methods – including those that define areas with dense edges as object regions, as well as deep-learning-based object detection and instance segmentation methods – have also been proposed [14], [15]. For example, the result of detecting the objects in Figure 2(a) using the density of the edge-based object detection method and setting the weight value in the inner regions of the objects is shown in Figure 3.

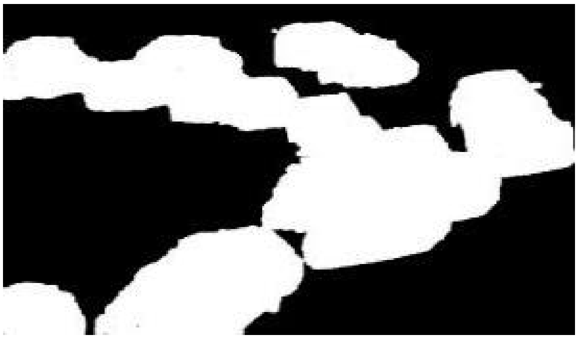


FIGURE 3. **Object segmentation.**

The weighted regions shown in Figure 3 are similar to the object regions in Figure 2(a) but are not accurate. These incorrectly detected object regions can cause parallax distortion in the stitched image because of the seams generated in the incorrect object regions. This problem can be overcome by using object detection [16], [17] or instance segmentation [18], [19] based on deep learning with high detection accuracy.

In addition, to overcome parallax distortion, it is necessary to define the magnitude of the weight to be set after accurately

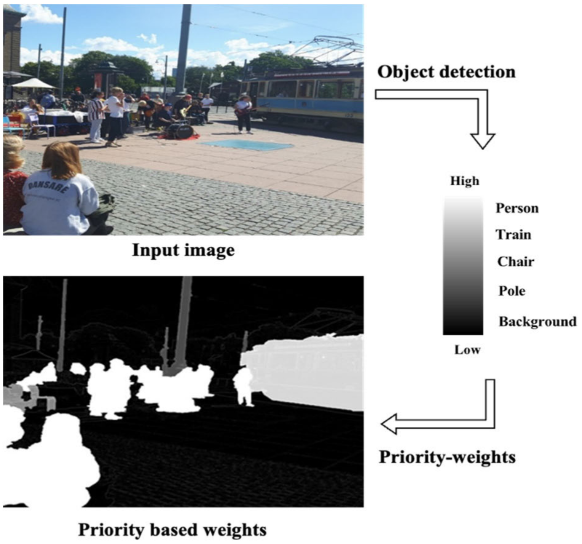


FIGURE 4. **Object-priority-based weights.**

detecting the inner region of the object. It is possible to simply assign the same weight value in all the inner regions of objects, which can cause parallax distortion in any object region, regardless of human consumption. In other words, it would be preferable not to cause parallax distortion in object regions that are more sensitive to human consumption than in any other object region. This can be obtained by setting different weight magnitudes for the object regions depending on human consumption sensitivity. For example, as shown in Figure 4, the priority of the detected objects through instance segmentation is set in advance according to the object type, and the weight value can be set differently according to this priority [3]. This method can induce parallax distortion in less sensitive objects. However, this method is limited because undefined objects in the deep learning network are not detected. Therefore, it is necessary to set individual weight values for all the objects to reduce the parallax distortion that occurs in a stitched image.

B. LIMITATION IN SEAM FINDER OF SEAM-BASED IMAGE STITCHING

Section II-B explains the limitations of the Seam Finder in SIS. As described above, Seam Finder generates seams along a path with the minimum cumulative sum in an SGM. As shown in Figure 5, Seam Finder is generally classified into three different models: three-pixel, five-pixel, and seven-pixel models, according to the strategy it uses to search the minimum path. For example, when a seam is generated in the vertical direction, the three-pixel model considers only the three elements shown as green pixels in Figure 5 (a) at the top of the current seam node as the next seam node. In addition, the five-pixel model considers not only the three elements at the top of the current node, but also the elements to the left and right of the current node as the next seam node, as shown in Figure 5 (b). Finally, the seven-pixel model considers all elements except the previous node as the next seam node, as shown in Figure 5 (c).

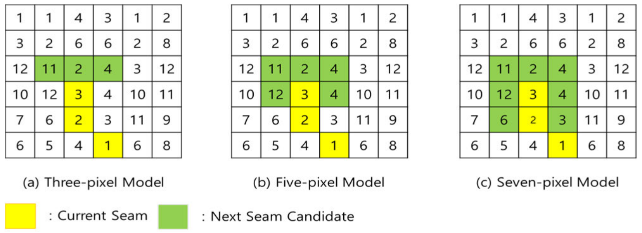


FIGURE 5. Seam finder: (a) three-pixel model (b) five-pixel model (c) seven-pixel model.

Thus, the larger the number of elements considered as the next seam node in Seam Finder, the less likely a parallax distortion is to occur. As an example of a visualization of SGM shown in Figure 6 (a), where the area indicated in blue represents the region of an object with large weight values and the line indicated in red represents the seam generated thus far, it is impossible to generate a seam that avoids the object region with a three-pixel model; thus, parallax distortion may occur. On the other hand, A five- or seven-pixel model for generating seams that do not pass through the object region may be more suitable for finding the next seam node.

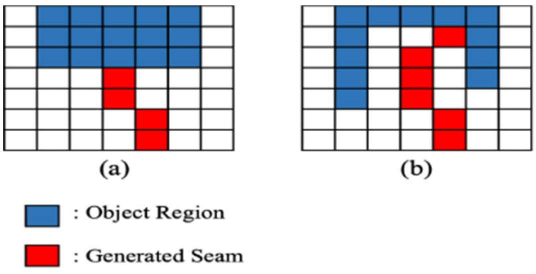


FIGURE 6. Examples of seam finder limitations.

In the situation shown in Figure 6 (b), a seam that avoids the object region could only be generated by the seven-pixel model because the next seam node should be selected in the element at the bottom of the current seam node. As the number of elements considered as the next seam node increases, a seam is generated to avoid the object region and the probability of parallax distortion is reduced. Thus, the most appropriate Seam Finder model may be the seven-pixel model.

However, the seven-pixel model has a higher seam generation complexity than the three- and five-pixel models because it requires exception handling to avoid infinite processing and repetitions of the seam generation trial and a regression process at the current node to determine the minimum cumulative sum path in all possible paths. This increased complexity gets rid of the efficiency of SIS.

Furthermore, parallax distortion may occur even in the seven-pixel model. For example, as shown in Figure 7, when the generated seam is surrounded by the object region, it inevitably passes through that region. At this time, it may be more appropriate to induce parallax distortion to occur at the point where the parallax distortion is minimal in terms of the human visual perspective, rather than to find a seam with higher-complexity Dynamic Programming, such as the seven-pixel model [11].

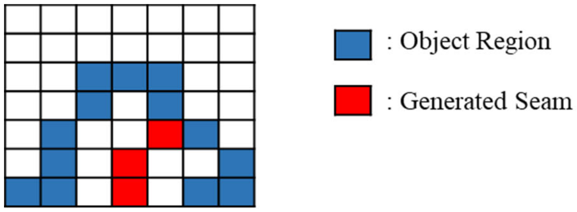


FIGURE 7. Limitation of seam generation with seven-pixel model.

Owing to the limitations described in Section II-A and II-B, it was found that the conventional SIS method can cause parallax distortion. In order to mitigate this distortion when constructing the SGM, all internal regions of objects must be assigned different weights, and their outline

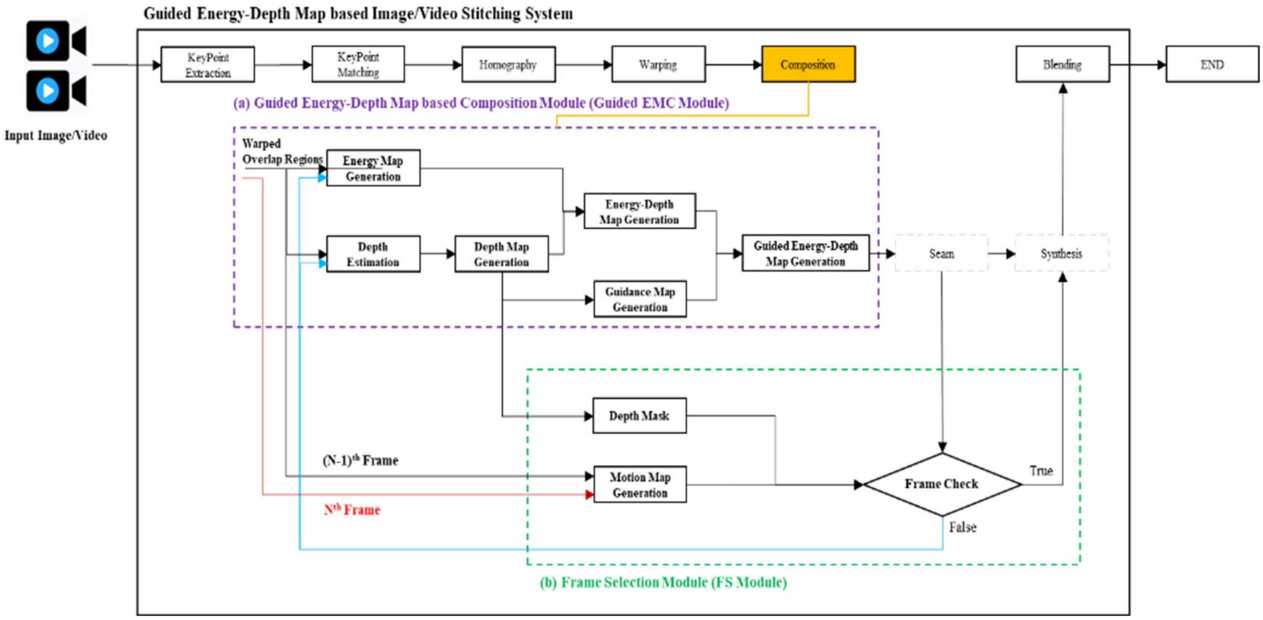


FIGURE 8. Structure of image/video stitching system based on guided energy-depth map.

information must also be accurately considered. In addition, it is necessary to define a specific rule for inducing parallax distortion at a point where parallax distortion is minimal from a human visual perspective when parallax distortion is inevitable. This led us to propose a seam generation method based on a Guided Energy-Depth (GED) map that enhances our previous method based on an energy-depth map [20], enables weights to be set for both the interior and contour of an object, and induces parallax distortion at a point where the parallax distortion is minimum from the human perspective. Additionally, the proposed Seam-based Video Stitching (SVS) method based on a GED map was applied to video sequences to improve the video stitching procedures described in Section III.

III. PROPOSED SEAM GENERATION METHOD BASED ON GUIDED ENERGY-DEPTH MAP

As described in the previous section, conventional Seam-based Image Stitching (SIS) consists of Keypoint Extraction, Keypoint Matching, Homography, Warping, Composition, and Blending modules. In the Composition module, the Seam Generation Matrix (SGM) is constructed with visual cognitive energy functions in warped overlapping regions, and images are synthesized along the minimal cumulative path of the SGM. However, it was verified in Section II that parallax distortion may occur in stitched images because conventional SGM does not properly set weights for the outline or inner region of objects, and the magnitude of the weights is set without considering parallax. To address these issues, we developed an SGM configuration method that assigns proper weights for the outline and inner regions of objects by considering parallax distortion. Furthermore, we enhanced the efficiency of seam-based video stitching by selectively determining whether the Composition module should be executed for each frame.

As shown in Figure 8, the proposed method introduces a Guided Energy-Depth (GED) map to the Composition module, and includes a new Frame Selection (FS) module in addition to the six conventional modules. The Guided Energy–depth Map-based Composition (Guided EMC) module proposed in this study encompasses the Energy Map Generation, Depth Estimation, Depth Map Generation, Energy-Depth Map Generation, Guidance Map Generation, and Guided Energy-Depth Generation submodules, as shown in Figure 8 (a). The FS module, which supports efficient video stitching, is composed of the Depth Mask, Motion Map Generation, and Frame Check submodules, as shown in Figure 8 (b). Detailed descriptions of all submodules are provided in Sections III-A and III-B.

A. GUIDED ENERGY-DEPTH MAP COMPOSITION MODULE

The Guided EMC module constructs an SGM to assign weight values to the outline and inner regions of objects. When parallax distortion is inevitable, the model confines this distortion to a point that minimizes it from a human perspective. The submodules comprising the Guided EMC module are depicted in Figure 9.

The Energy Map Generation, Depth Estimation, Depth Map Generation, and Energy−Depth Map Generation submodules use visual cognitive energy functions and depth information to set weights for the outline and inner regions of objects to generate a seam that minimizes parallax distortion. First, the Energy Map Generation submodule generates an energy map using (1) in overlapping regions. This visual

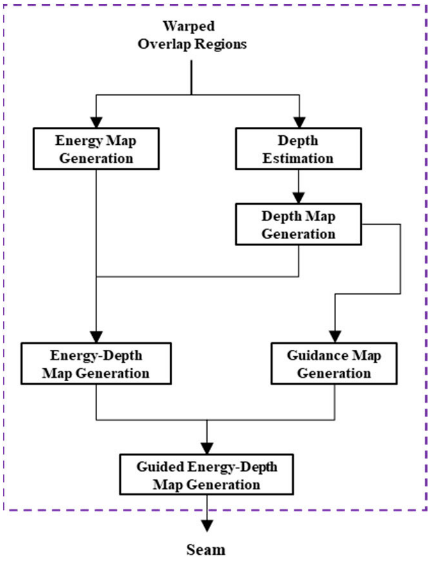


FIGURE 9. Structure of guided energy-depth map composition module.

cognitive energy function is defined as the summation of the change in the magnitude of the horizontal and vertical pixels in the image, providing the outline information of objects. Weights for the inner regions of objects are set according depth information by the Depth Estimation, Depth Map Generation, and Energy–Depth Map Generation submodules.

The Depth Estimation submodule is a procedure for generating depth information that uses the distance difference between the same points from at least two images [21], as shown in Figure 10, or estimates depth information using a deep learning-based depth information estimator [22], [23]. The use of depth information for image stitching has also been proposed in [29] and [30]. To reduce parallax distortion, [29] estimated optimal homography, whereas [30] constructed an SGM using depth information. However, since [29] and [30] used the depth-difference between input images, the depth information could be set to a small in an area where the depth is similar. Thus, we used the Depth Estimation submodule, which uses the sum of depth information, rather than its difference.

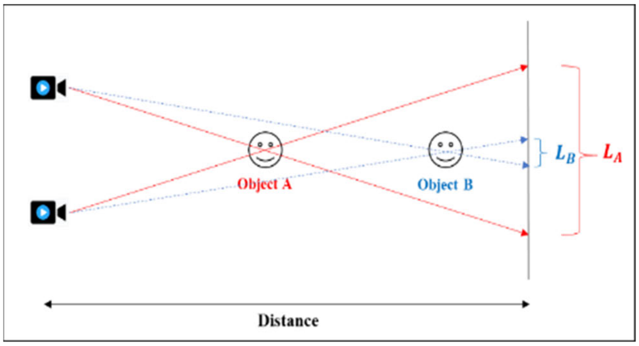


FIGURE 10. Parallax according to distance.

Before calculating the sum of depth information, the depth information is N-bit quantized in the subsequent Depth Map Generation submodule, where a larger quantization value is assigned to a region with a closer distance, and a smaller one is assigned to a region with a farther distance. For example, 255 was set in the closest region and 0 in the farthest region when the depth information was quantized to 8-bits. This quantized depth information is divided into k levels as in (2) and thus, the depth map ($\boldsymbol{D_k}$) is composed of values from 0 to k−1. For example, when quantized depth information (Q) with 8-bits are divided into five levels, the depth map ($\boldsymbol{D_k}$) is set such that the range 0−51 is set to 0, 52−102 to 1, 103−153 to 2, 154−204 to 3, and 205−255 to 4.

$$D_k = floor(k * \frac{Q}{2^N}) \tag{2}$$

Based on information from the Energy Map Generation and Depth Map Generation submodules, the Energy–Depth Map Generation submodule constructs the energy–depth map for generating a seam in the far region to minimize parallax distortion by using the characteristic that the farther distance from the camera to the object decreases parallax, as shown in Figure 10. This energy–depth map is constructed via (3) and (4), where $\boldsymbol{E}$ and $\boldsymbol{E_{max}}$ represent the energy map and its maximum value, respectively, and k and $\boldsymbol{D_k}$ in (4) represent the number of divisions of the depth map and level values assigned to the depth map, respectively. As expressed in (4), an energy−depth map ($\boldsymbol{E_D}$) is defined as the summation of the visual cognitive energy function and the maximum energy value proportional to the depth levels. As an example, the energy, depth, and energy–depth maps generated from Figure 11(a) as input images are shown in (b).

$$E_{max} = Max(E) \tag{3}$$

$$E_D = E + E_{max} * (D_k + 1) \tag{4}$$

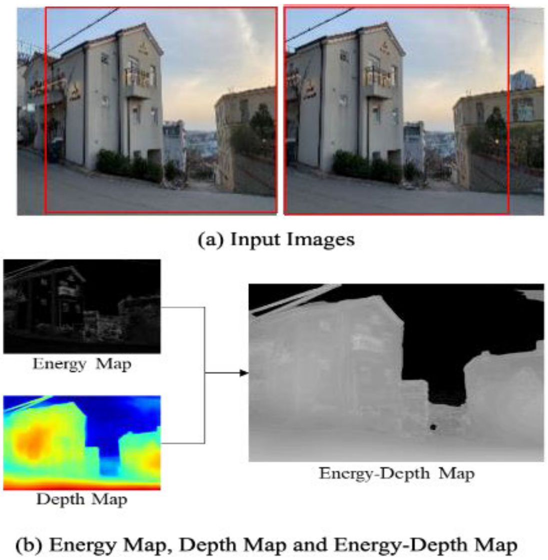
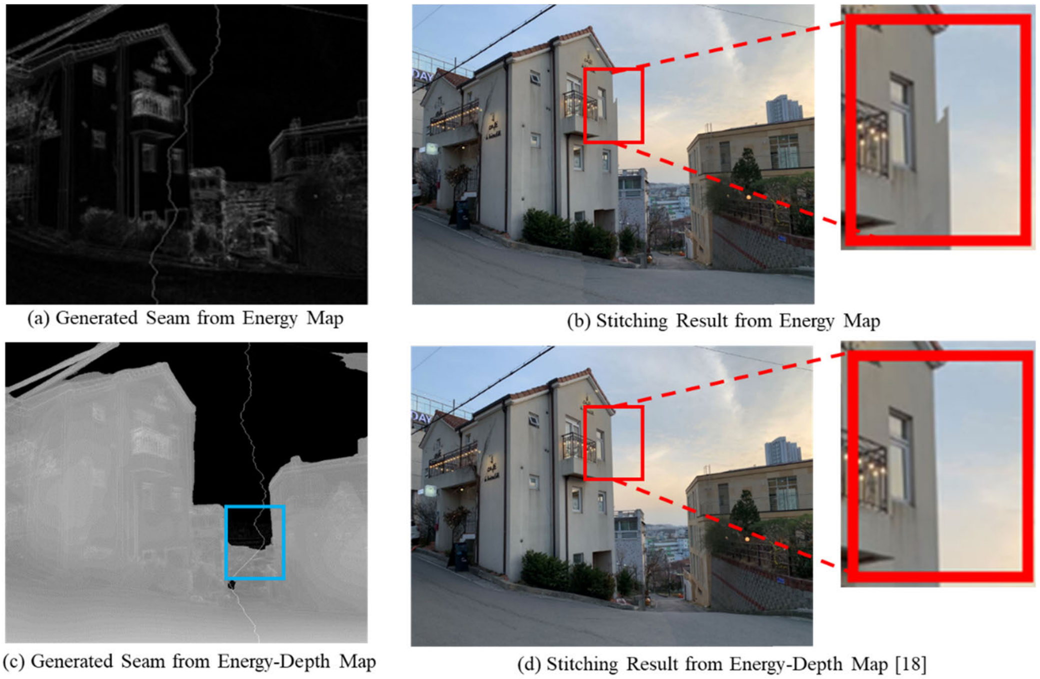


FIGURE 11. Examples of energy, depth, and energy-depth maps: (a) input images (b) energy map, depth map and energy-depth map.



(a) Generated Seam from Energy Map

(b) Stitching Result from Energy Map

(c) Generated Seam from Energy-Depth Map

(d) Stitching Result from Energy-Depth Map [18]

FIGURE 12. **Comparison of image stitching results using the energy and energy-depth maps: (a) seam generated from energy map (b) stitching result from energy map (c) seam generated from energy-depth map (d) stitching result from energy-depth map.**

Examples of image stitching results obtained using the energy map [10] and previous energy−depth map [20] are shown in Figure 12 along with their corresponding seams. As shown in the region demarcated by the red rectangle in Figure 12(b), a parallax distortion was incurred by a seam generated in the close region. In contrast, the seam generated by the energy−depth map, along with its image stitching result, appear to be more natural, as the seam was generated in an area with minimal parallax distortion.

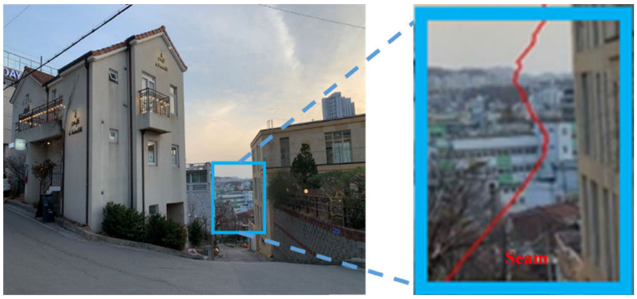
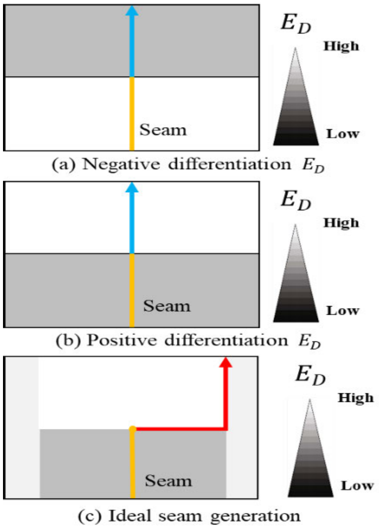


FIGURE 13. **Results of seam generation and image stitching from energy-depth map: (a) generated seam from energy-depth map (b) generated seam and stitching result from energy-depth map.**

However, the seam generated using the energy−depth map, indicated by the blue box in Figure 12(c), is located in the middle of the far object region displayed as a black region rather than a boundary between the far and near object regions. When the stitching process is conducted, the seam passes through the building, as shown in the blue rectangular region of Figure 13, leading to parallax distortion. When the stitching process is conducted based on this generated seam, it passes through the buildings, as shown in the blue rectangular region Figure 13, which could cause parallax distortion. This is a limitation of applying depth information to the energy depth map E_D in (4) when it varies significantly at a boundary.



(a) Negative differentiation E_D

(b) Positive differentiation E_D

(c) Ideal seam generation

FIGURE 14. **Limitations of energy-depth map.**

For example, as shown in Figure 14(a) and (b), the value of E_D can be changed from lower to higher or vice versa when the energy map (E) in (4) is set to zero. Regardless of the positive or negative differentiation of E_D it can generate incorrect seams because the Seam Finder explained in Section II-A considers two pixels in the vertical direction. Even positive differentiation E_D, such as in Figure 14(b), can produce more noticeable stitching errors because the wrong seams are located near the object region. Thus, it is necessary to generate a seam along the boundary, as shown by the red arrow in Figure 14(c), to overcome the parallax distortion that may occur at the boundary, where the E_D of the energy–depth map varies greatly.

To overcome this limitation that arises at the boundary where $\boldsymbol{E_D}$ varies, a guidance map generated by the Guidance Map Generation submodule in Figure 9 is proposed for extracting the location representing the boundary of the depth map to generate a seam along the boundary of the energy–depth map, as shown in Figure 14(c). As indicated in (5), the guidance map ($\boldsymbol{G_m}$ $(\boldsymbol{x}, \boldsymbol{y})$) is a binary matrix of the $\boldsymbol{D_k}$value of the depth map, which varies along the horizontal and vertical directions. As an example, the guidance map of Figure 15 takes the value 0 in the same energy–depth map region and one value in the region where the energy–depth map is changed, which is displayed as a red line in Figure 15.

$$G_m(x, y) = 1 - \delta(|D_k(x, y) - D_k(x+1, y)| + |D_k(x, y) - D_k(x, y+1)|) \quad (5)$$

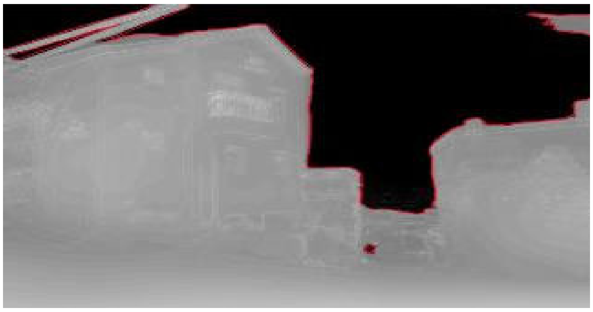
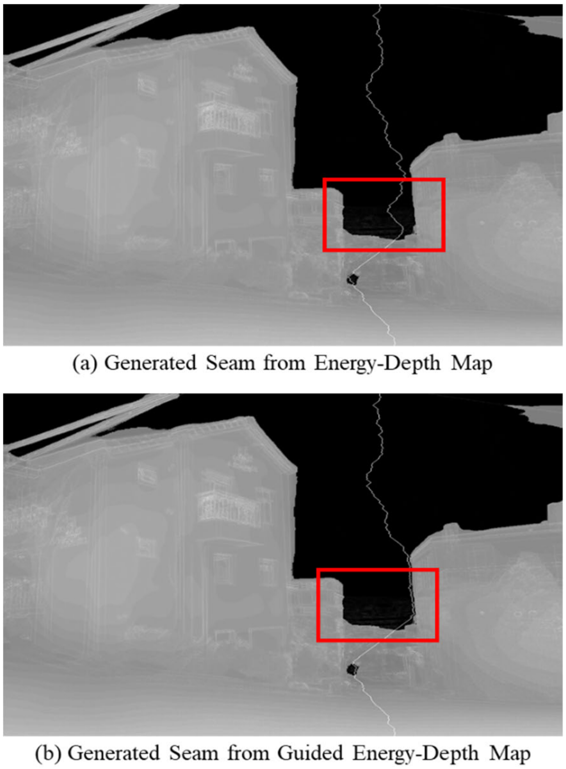


FIGURE 15. Guidance map.

Based on the energy depth and guidance map, a Guide Energy–Depth (GED) map is proposed for more precise seam generation. The GED map uses the energy–depth map ($E_D(x, y)$) to set weights for both the outlier and inner regions of objects and the background, respectively, and generates a seam at points where parallax distortion is minimal from a human perspective. Because the use of only the energy-depth map ($E_D(x, y)$) could cause parallax distortion at a boundary with varying E_D as described above, the GED map is defined in (6) to reduce the maximum energy (E_{max}) in the energy–depth map, where the value of the guidance map (G_m (x, y)) is 1.

$$GED(x, y) = E_D(x, y) - E_{max} * G_m(x, y) \quad (6)$$

The seams generated by the energy–depth map and GED map are depicted in Figure 16(a) and (b), respectively, where the seam generated by the GED map appears to be more precise in the region enclosed within the red rectangle. Thus, the Guided EMC module generates proper seams by using the GED map.



(a) Generated Seam from Energy-Depth Map

(b) Generated Seam from Guided Energy-Depth Map

FIGURE 16. Comparison of generated seams generated using the (a) energy-depth map (b) guided-energy-depth map.

B. FRAME SELECTION MODULE FOR VIDEO STITCHING

For Seam-based Video Stitching (SVS), two methods may potentially be used to apply a Seam-based Image Stitching (SIS) algorithm to video sequences: using the respective seams generated from every frame of a video sequence [24], [25], or using a single seam generated from the first frame for an entire sequence [26]. Although the first method can produce more precise stitching results, it is highly time-intensive. In contrast, the second method reduces the processing burden. However, since the seam location is fixed, parallax distortion may occur when objects are moving in the remaining frames of the video sequence. As a compromise between accuracy and efficiency, methods have been designed to select certain video frames for seam regeneration [27], [28].

For example, [27] proposed the selection of new video frames for seam generation by comparing the preserved seam with the locations of moving objects. However, this approach has a limitation wherein background information must be obtained in advance for the detection of moving objects. Furthermore, because this method does not update the predefined background information in subsequent frames, the actual background may be misclassified as a moving object. For example, when a stopped vehicle is defined as a background and subsequently moves, the region that previously encapsulated the vehicle might be classified as a moving object.

On the other hand, the method proposed in [28] selects video frames that require new seam generation by detecting moving objects through deep learning-based instance segmentation such as Mask R-CNN [18]. Although deep learning-based object detection or instance segmentation could reduce processing time and guarantee higher detection accuracy, they cannot detect objects that were not defined during model training. Furthermore, it may be time-consuming in that deep learning-based object detection or instance segmentation algorithm has to be applied to all video frames.

To mitigate these limitations, this study proposes the Frame Selection (FS) module, which selects video frames for new seam generation through an efficient and generalizable process. As shown in Figure 17, the proposed FS module consists of the Depth Mask, Motion Map Generation, and Frame-Check submodules.

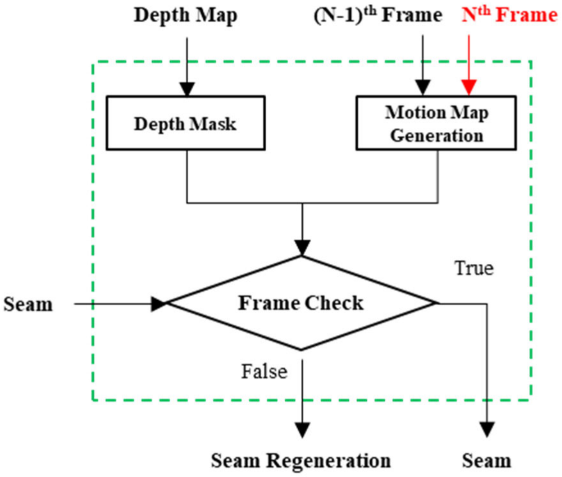


FIGURE 17. Overview of frame selection module.

The FS module uses depth information based on parallax characteristics to select specific video frames for seam regeneration, similar to the Guided EMC module. The Depth Mask submodule initializes and updates depth information as a depth mask (M) to determine whether a new seam should be generated from the current frame. As shown in (7), the initial depth mask (M_0) is initialized as the first depth map ($D_{\mathrm{k}0}$) generated in the first video frame. In addition, the depth mask is updated differently depending on whether a new seam must be generated. This selective procedure is expressed by (8), where the depth mask (M_i) is maintained as the previous depth mask (M_{i-1}) when $Seam_i = Seam_{i-1}$, and updated with the newly generated depth map ($D_{k\,i}$) when $Seam_i \neq Seam_{i-1}$.

$$M_0 = D_{k_0} \tag{7}$$

$$M_i = \begin{cases} D_{k\,i}, & \text{if } Seam_i \neq Seam_{i-1} \\ M_{i-1}, & Otherwise \end{cases} \tag{8}$$

where, $Seam_i$ represents a seam generated from the i^{th} frame.

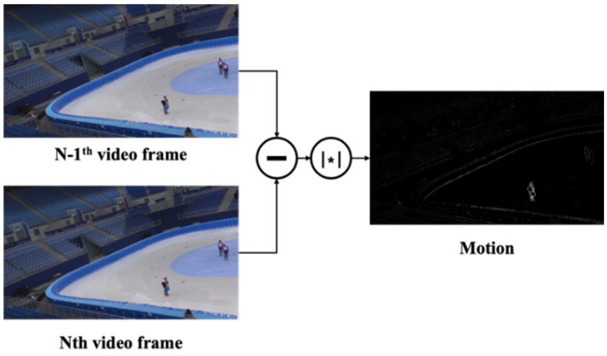
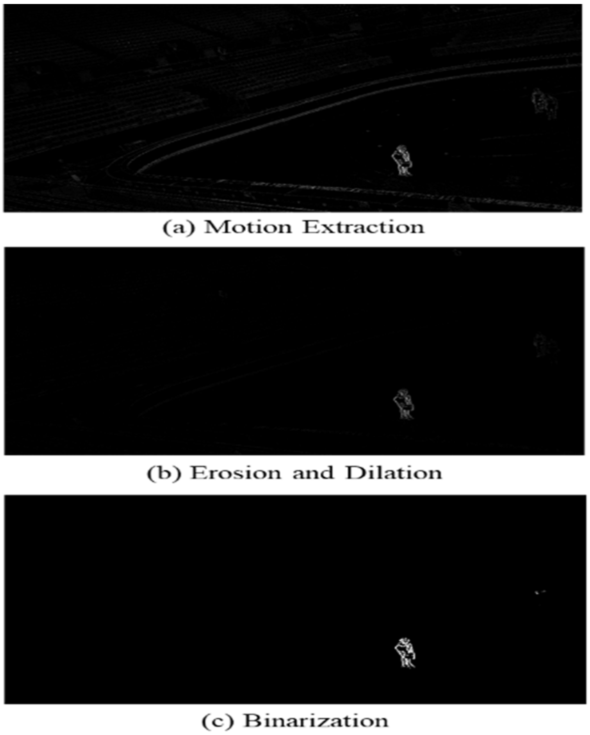


FIGURE 18. Motion extraction from input video frames.

The FS module accelerates video stitching by using motion map (MOT_B) in place of object detection to select frames for new seam generation. As shown as Figure 18, the Motion Map Generation submodule extracts motion (MOT_N) from the previous (f_{N-1}) and current (f_N) frames according to (9).

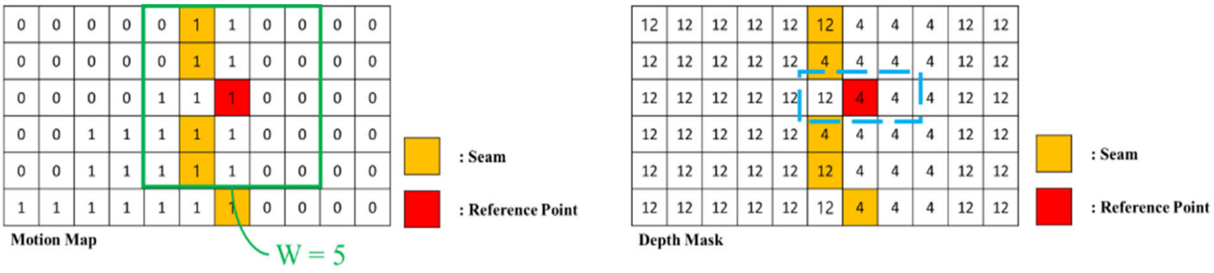

(a) Motion Extraction

(b) Erosion and Dilation

(c) Binarization

FIGURE 19. Motion map generation process: (a) motion extraction (b) erosion and dilation (c) binarization.

The MOT_N shown in Figure 19(a) indicates the distance the object has traveled as well as noise caused by camera shaking and changes in illumination of the capturing location. To remove this noise in the motion, it is converted to the refined motion (MOT_R) through erosion and dilation processes, as shown in Figure 19(b). Finally, the MOT_R is converted to the motion map (MOT_B) as shown in Figure 19(c) through binarization based on the threshold (θ_{th})



0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	1	1	1	0	0	0	0
0	0	1	1	1	1	1	0	0	0	0
0	0	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	0	0	0	0

(a) Motion Map Window Search

12	12	12	12	12	12	4	4	4	12	12
12	12	12	12	12	4	4	4	4	12	12
12	12	12	12	12	12	4	4	4	12	12
12	12	12	12	12	4	4	4	4	12	12
12	12	12	12	12	12	4	4	4	12	12
12	12	12	12	12	12	4	4	4	12	12

(b) Depth Mask Side Search

FIGURE 20. Motion map window search (a) and depth mask side search (b).

using (9).

$$MOT_N(x,y) = |f_N(x,y) - f_{N-1}(x,y)| \tag{9}$$

$$MOT_B(x,y) = \begin{cases} 0, if \quad MOT_R(x,y) < \theta_{th} \\ 1, Otherwise \end{cases} \tag{10}$$

To determine whether a new seam should be generated in the current video frame, the previous seam, motion map, and depth mask are used as the reference point, error, and error threshold, respectively, for a window search in the Frame Check submodule. The error (err) is defined as the sum of the MOT_B divided by the square of the window size (W) within the window block ($\boldsymbol{W_B}$) at the reference point, as shown in (11). In accordance with (12), the depth-based error threshold (θ_{depth}) is defined as one minus the sum of the values on the left (MOT_{BL}) and right (MOT_{BR}) divided by $2k$ at the reference point in the depth mask, where k corresponds to the k-th level in the depth map and depth mask. As an example, the orange points in Figure 20 indicate the previous seam, while the red point indicates the reference point for the window search, where err is set to 0.52 when the summation of the motion map is 13 within a window of size 5. Furthermore, θ_{depth} is set to 0.385 when the values of MOT_{BL}, MOT_{BR}, and k are 12, 4, and 13, respectively, as shown in Figure 20(b). Finally, to determine whether a new seam should be generated from the current video frame, the magnitude of err is compared with $\alpha \cdot \theta_{depth}$ as expressed in (13). At this stage, α is an insensitivity value of the depth-based error threshold. As an example, a new seam is generated in Figure 20 because the err value of 0.52 exceeds that of θ_{depth} when α is 1.

$$err(x,y) = \frac{1}{W^2}\sum_{u,v} MOT_B(u,v)\{x,y \in \boldsymbol{Seam}, u,v \in \boldsymbol{W_B}\} \tag{11}$$

$$\theta_{depth}(x,y) = 1 - \frac{M_L(x,y) + M_R(x,y)}{2k}\{x,y|x,y \in \boldsymbol{Seam}\} \tag{12}$$

$$Decision = \begin{cases} 1, & if \ err(x,y) > \alpha \cdot \theta_{depth}(x,y) \\ 0, & Otherwise \end{cases} \tag{13}$$

In the following Section IV, we assess the proposed method in terms of efficiency and performance.

IV. EXPERIMENTS

All experiments were conducted on a PC platform running Microsoft Windows 10 and Python 3.7 with OpenCV 4.7.0, an Intel i9-12900K processor with 64 GB memory, and an NVIDIA GeForce RTX 4090 GPU. The following subsections present experimental results obtained for the Guided Energy–depth Map-based Composition (Guided EMC) and Frame Selection (FS) modules.

A. RESULTS OF GUIDED ENERGY–DEPTH MAP BASED COMPOSITION MODULE

In Section III-A, we propose a Guided EMC module to minimize parallax distortion in human recognition systems by inducing seam generation in smaller parallax regions using depth information. In this Section IV-A, the stitched images produced from the proposed method were compared with those obtained from an energy function. To ensure an accurate comparison, the same seam generation process – consisting of keypoint extraction, matching, homography, warping, and the seam finder – was applied for both the energy function and Guided EMC module. Also, to compare performance with minimal intervention of the seam finder, all seams were generated by the three-pixel seam finder [31]. Furthermore, the bending process was not applied, allowing the parallax distortion in the stitched images to be closely distinguished. The Guided Energy-Depth (GED) map was generated using the deep learning-based depth prediction model MiDaS [23]. In addition, two types of test images representing smaller and larger parallax effects were used for the comparative experiments, as shown in Figure 21(a) and (b). To improve visualization, we enhanced brightness in the overlapping regions, as shown in Figure 21.

To quantitatively evaluate the accuracy of seams generated through each method, we adopted the seam evaluation method used for seam optimization-based image stitching [4]. As an evaluation metric, the Average Seam Error (ASE) is obtained by accumulating the pixel differences of two input

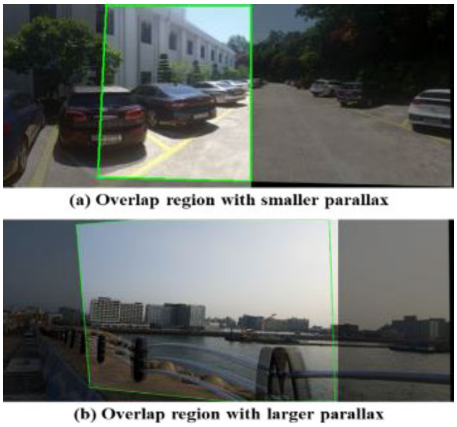

(a) Overlap region with smaller parallax
(b) Overlap region with larger parallax

FIGURE 21. Overlapping regions with different parallax (a) overlap. region with smaller parallax (b) overlap region with larger parallax.

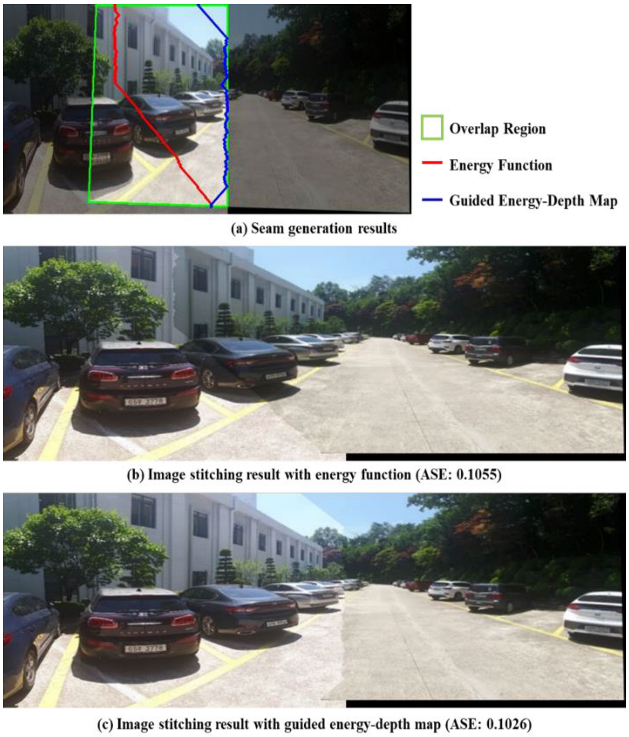

Overlap Region
Energy Function
Guided Energy-Depth Map
(a) Seam generation results
(b) Image stitching result with energy function (ASE: 0.1055)
(c) Image stitching result with guided energy-depth map (ASE: 0.1026)

FIGURE 22. Seam generation and image stitching results of energy function and GED map with little parallax distortion: (a) seam generation results (b) image stitching result with energy function, ASE: 0.1055 (c) image stitching result with GED map, ASE: 0.1026.

images along each seam as expressed in (14).

$$ASE = \frac{1}{N}\sum \left| I_L\left(x, y\right) - I_R\left(x, y\right) \right| \{x, y|x, y \in \boldsymbol{Seam}\} \tag{14}$$

The seams generated using the energy function and GED map in an image with a smaller parallax error in Figure 21(a) are shown in Figure 22. The red seam generated by the energy function passed through a vehicle region. In contrast, the blue seam generated by the GED map passed through a vehicle at a relatively greater distance from the camera. And, the ASE measures of the seams generated by the energy function and GED map were 0.1055 and 0.1026, respectively, indicating a lower degree of error in the latter.

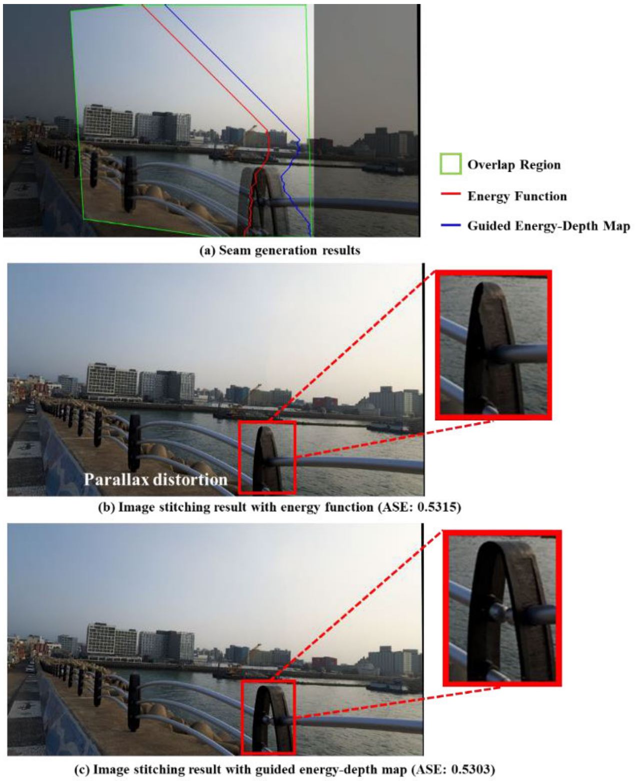
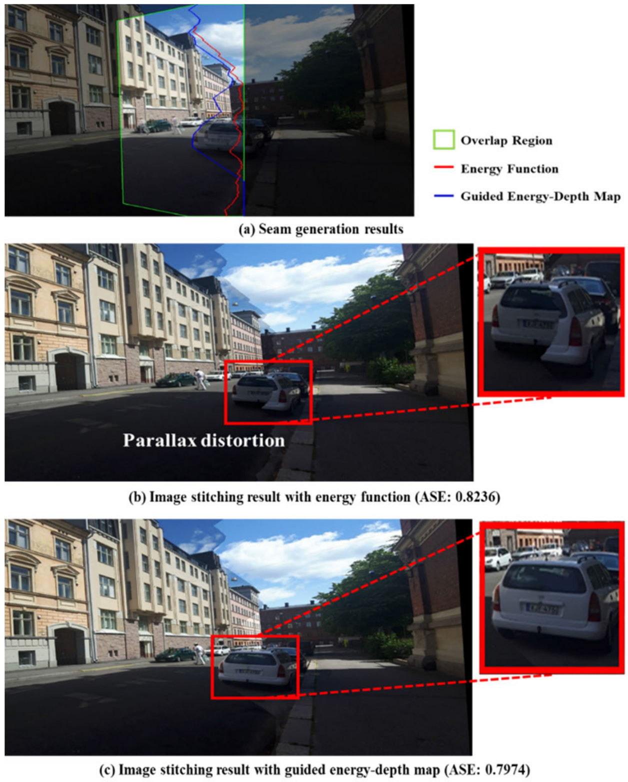

Overlap Region
Energy Function
Guided Energy-Depth Map
(a) Seam generation results
Parallax distortion
(b) Image stitching result with energy function (ASE: 0.5315)
(c) Image stitching result with guided energy-depth map (ASE: 0.5303)

FIGURE 23. Seam generation and image stitching results of energy function and GED map for greater parallax distortion: (a) seam generation results (b) image stitching result with energy function, ASE: 0.5315 (c) image stitching result with GED map, ASE: 0.5303.

A situation with greater parallax distortion is presented in Figure 23. The seam generated by the energy function can be seen to incur greater visible distortion in Figure 23(b), whereas that generated with the GED map produced undistorted object in Figure 23(c). The ASE measures of the seams generated by the energy function and GED map were 0.5315 and 0.5303, respectively, demonstrating the superior accuracy of the latter.

To precisely verify the performance of the GED map, an additional image stitching test was performed on an image with a greater parallax, as shown in Figure 24. As seen in the region enclosed in the red box, the seam obtained from the energy function resulted in parallax distortion, whereas that obtained from the proposed GED map did not. In the quantitative evaluation, the ASE values measured for the seams generated by the energy function and GED map were 0.8236 and 0.7974, respectively. Therefore, the seam generated through the proposed method was more accurate than that generated using the energy function.

Another experiment was conducted on test images that exhibited even more extensive distortion. Unlike images used in the prior experiments, which were taken in parallel to the left and right sides of the device, these images were taken

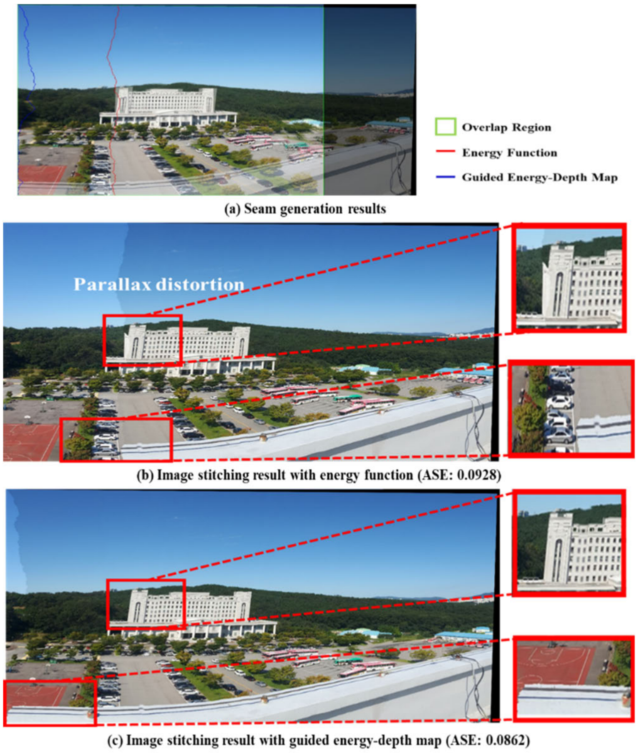


(a) Seam generation results

(b) Image stitching result with energy function (ASE: 0.8236)

(c) Image stitching result with guided energy-depth map (ASE: 0.7974)

FIGURE 24. Seam generation and image stitching results of energy function and GED map for greater parallax distortion: (a) seam generation results (b) image stitching result with energy function, ASE: 0.8236 (c) image stitching result with GED map, ASE: 0.7974.



(a) Seam generation results

(b) Image stitching result with energy function (ASE: 0.0928)

(c) Image stitching result with guided energy-depth map (ASE: 0.0862)

FIGURE 25. Seam generation and image stitching results of energy function and guided energy-depth map for extreme parallax distortion: (a) seam generation results (b) image stitching result with energy function, ASE: 0.0928 (c) image stitching result with guided energy-depth map, ASE: 0.0862.

with cameras located at the front and rear of the device. The images used for this experiment are presented in Figure 25. For example, one of the input images depicts the presence of a wall at the bottom, which does not appear in the other input image. This large difference between the two images is therefore highly problematic when attempting to minimize parallax distortion in the stitching process. The results in the Figure 25 indicate that the image obtained with the seam generated by the energy function exhibited visible parallax distortion in regions corresponding to the lower wall and building, whereas that obtained with the seam generated using the GED map did not exhibit clear parallax distortion. The ASE of the seam generated by the energy function was 0.5315, whereas that of the seam generated by the GED map was 0.5303, again demonstrating the higher accuracy of the latter.

To verify the general performance of the Guided EMC module, additional experiments were conducted using the SEAGULL public dataset [32]. Figures 26-29 present experimental results for images corresponding to ID 25, 31, 67, and 71, respectively, in the SEAGULL dataset. As seen in Figure 26, the stitched images generated by the energy function exhibited parallax distortion in the vehicle, whereas those generated by the GED map exhibited minimal parallax distortion. Similar results were obtained for the other three images, as shown in Figures 27-29.

Table 1 presents the results of a quantitative evaluation in terms of ASE on 28 images from the SEAGUL dataset, demonstrating that the proposed GED map achieved superior accuracy in 75% of the cases.

As demonstrated throughout the experiments, the Guided EMC module produces more precise stitching results by minimizing parallax distortion with respect to human recognition. Through the ASE evaluation, we confirmed that the proposed method produced more accurate seams than the energy function in 25 of the 32 experimental cases, with an average improvement of 15.58%.

B. RESULTS OF THE FRAME SELECTION MODULE FOR VIDEO STITCHING

We conducted additional experiments to assess the efficiency of the Fame Selection (FS) module, which is deployed when applying the Guided EMC Module to video sequences. As described previously, simple implementations of image stitching of video sequences can involve stitching the entire video sequence based on a single seam generated from the first frame, or generating seams for each individual frame. These two methods were applied to the test video sequence **Woman**, shown in Figure 30, where a woman walks from left to right and from right to left 14 times, with the configuration in Table 2.

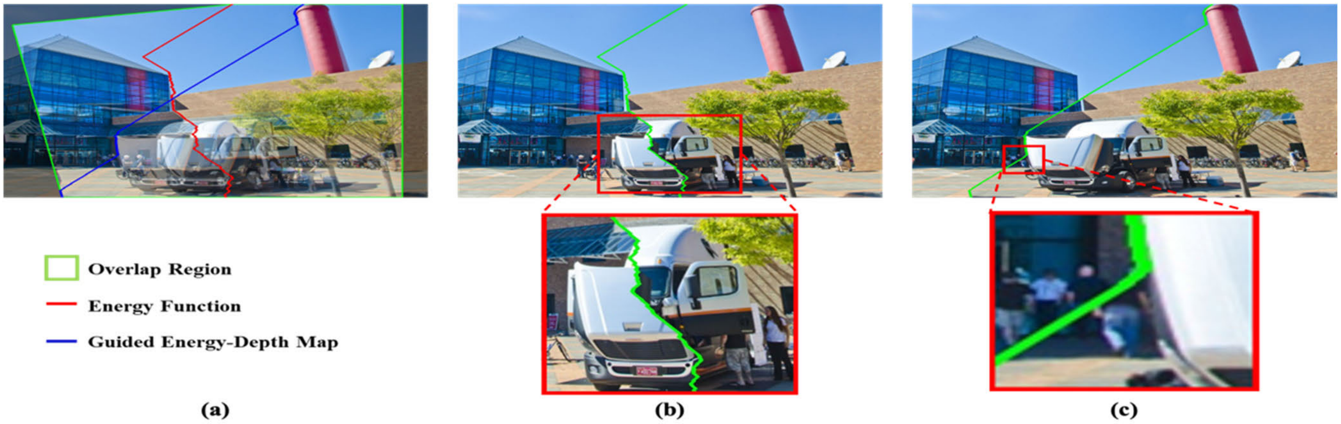


FIGURE 26. Seam generation and image stitching results of energy function and GED map for ID 25 of SEAGULL dataset: (a) seam generation results (b) image stitching result with energy function (c) image stitching result with GED map.

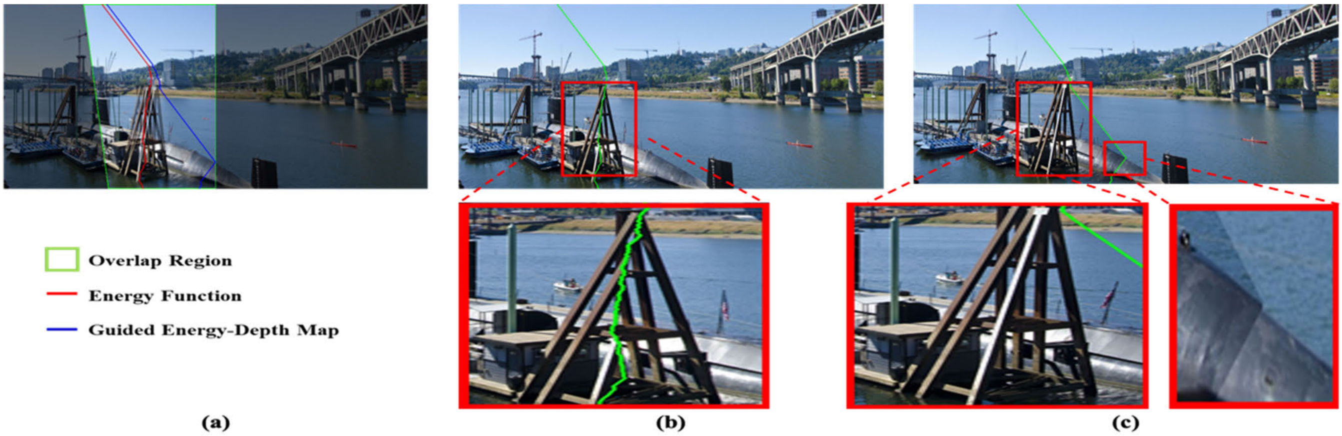


FIGURE 27. Seam generation and image stitching results of energy function and GED map for ID 31 of SEAGULL dataset: (a) seam generation results (b) image stitching result with energy function (c) image stitching result with GED map.

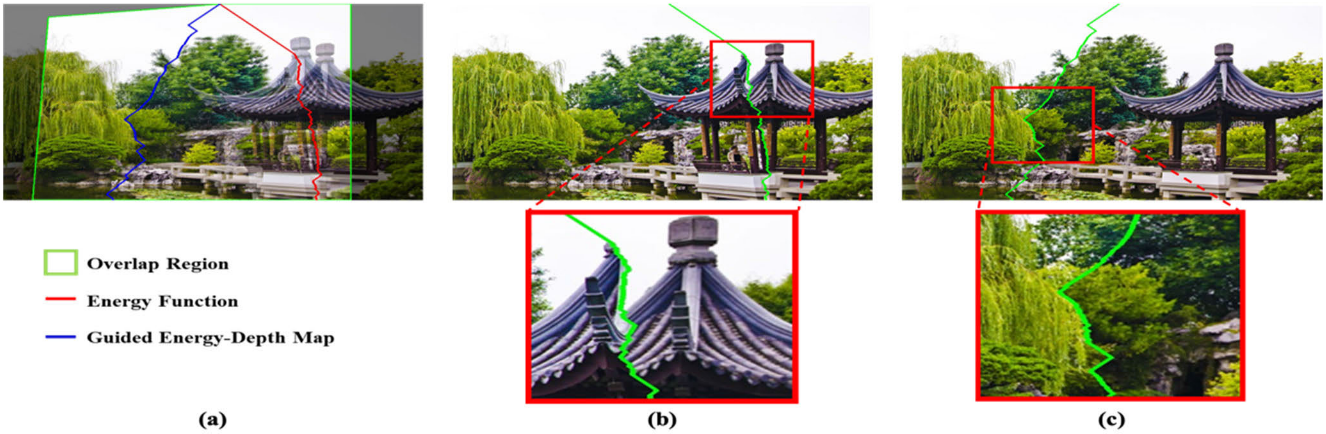


FIGURE 28. Seam generation and image stitching results of energy function and GED map for ID 67 of SEAGULL dataset: (a) seam generation results (b) image stitching result with energy function (c) image stitching result with GED map.

First, a single seam generated in the first video frame was used to stitch all subsequent video frames. The woman passed through the seam 14 times in the video, resulting in a parallax distortion of 13.02 seconds over 780 frames, as shown in Figure 31. Thus, stitching on the basis of a single seam was confirmed to result in a parallax distortion.

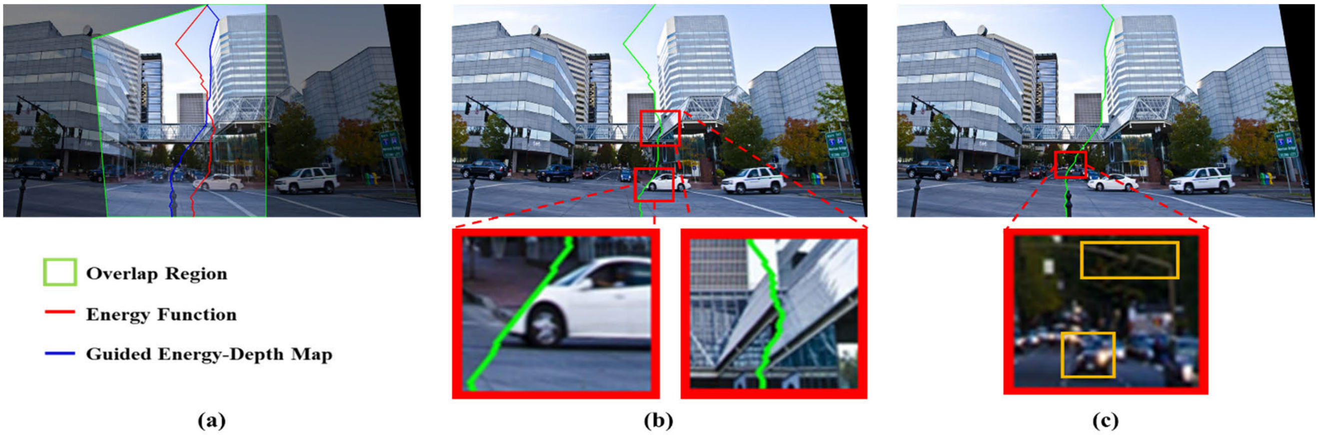


FIGURE 29. Seam generation and image stitching results of energy function and GED map for ID 71 in SEAGULL dataset: (a) seam generation results (b) image stitching result with energy function (c) image stitching result with GED map.

TABLE 1. Results of average seam error evaluation on seagull dataset.

Scene ID	001	005	007	009	011	013	019
Energy function	0.3556	0.2945	0.7522	0.7470	0.7462	**0.2967**	0.4965
Proposed	**0.2136**	**0.1286**	**0.4208**	**0.4753**	**0.7071**	0.7222	**0.4454**
Scene ID	025	029	031	033	035	037	039
Energy function	0.7495	0.6138	0.3209	0.6888	0.5684	0.7891	0.7567
Proposed	**0.6165**	**0.5364**	**0.3038**	**0.5815**	**0.5570**	**0.5539**	**0.4808**
Scene ID	041	043	045	047	051	053	057
Energy function	0.2505	0.3209	0.3868	0.7580	0.4751	**0.2247**	0.7418
Proposed	**0.2498**	**0.3038**	**0.2374**	**0.5299**	**0.2633**	0.3907	**0.4309**
Scene ID	059	063	067	069	071	073	075
Energy function	**0.1598**	**0.2622**	0.6108	**0.5677**	0.8074	**0.2790**	**0.5970**
Proposed	0.1954	0.3600	**0.3334**	0.6475	**0.7423**	0.3362	0.7262

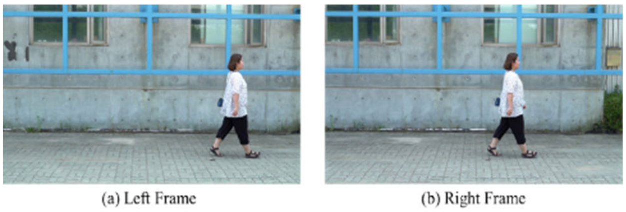


FIGURE 30. Nth frames in video sample (a) left frame (b) right frame.

TABLE 2. Configuration of video dataset woman.

Video Configuration	
Duration	227 (sec)
Frames per second (FPS)	59.94
Total frames	10,028
Height, Width	1080, 1920

Although the other aforementioned approach, wherein a new seam is generated for every video frame, might eliminate the parallax distortion seen in Figure 31, it incurs a higher computational cost along with a new type of parallax distortion. As shown in Figure 32, the Regions of Interest (ROIs) of the N_{th} and $N+1_{th}$ stitched frames abruptly changed owing to the lack of seam consistency, resulting in a unique type of parallax distortion that occurs between frames.

To maintain seam consistency, new seams must be generated only in specific frames selected from the video sequence. As explained previously, Herrmann et al. [28] selected specific frames using an instance segmentation method; however, this approach has the limitation of considerably extending computational time. In contrast, the proposed FS module achieved consistent seam generation while maintaining computational efficiency. A comparative experiment between the method proposed in [28] and the FS module was conducted on the test video sequences denoted as **Woman**, **People**, and **Skating**, with details presented in Table 2, Table 3, and Figure 33.

The results of an experiment to measure the computational time consumed by both methods with the **Woman** sequence are listed in Table 4. The sequence in question comprised

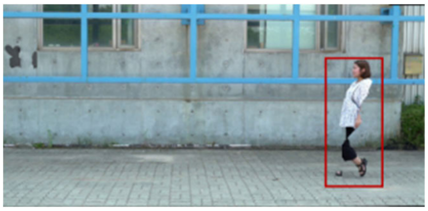


FIGURE 31. Parallax distortion during video stitching based on a single seam generated from the first frame.

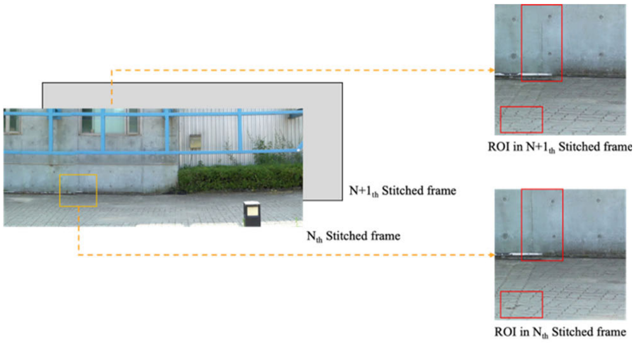
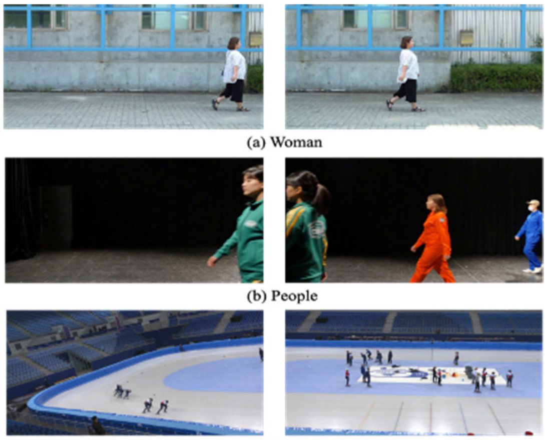


FIGURE 32. Parallax distortion between frames when using a unique seam for each frame.



(a) Woman

(b) People

(c) Skate

FIGURE 33. Nth frames of the video datasets (a) Woman (b) People (c) Skating.

10,028 frames. To process this sequence, the method proposed in [28] required 0.2215s per frame for searching with Yolact [19], resulting in a total search time of 2,221.6394s. In contrast, the FS Module of the proposed method required 0.0092s per frame; thus, the total search time was 92.2269s. However, the proposed method also generates a new depth mask with MiDaS [23] for every selected frame, which takes 0.1332s per frame. Nonetheless, the total time required for video content stitching by the proposed method was still only 139.3928s. Because the number of selected frames depends on the position of the newly generated seam, it does not represent the detection accuracy of each method.

TABLE 3. Configuration of video datasets people and skating.

Video Configuration		
Dataset	**People**	**Skating**
Frames per second (FPS)	29.97	59.94
Total frames	4,680	2,216
Height, Width	1080, 1920	2160, 3840

TABLE 4. Results of computational time experiment on woman sequence.

	Herrmann et al. [28]	Proposed method
Average search time per frame (sec)	0.2215 (with Yolact [19])	0.0092
Number of search frames	10,028	10,028
Total time for search (sec)	2,221.6394	92.2269
Number of selected frames	298	354
Additional time per selected frame (sec)	-	0.1332 (with MiDaS [23])
Additional time for selected frames (sec)	-	47.1659
Total Time (sec)	2,221.6394	139.3928

Table 5 lists results obtained in the computational experiment on the **People** sequence. Although this sequence comprises less frames than the **Woman** sequence, it includes scenes showing many moving people, which require more frequent seam regeneration. The method proposed in [28] required 0.1827s per frame for searching, for a total time of 854.9568s. In contrast, the proposed method required 0091s per frame for searching and an additional 0.1352s per selected frame for depth mask generation, for a total time of 99.5403s. These results further confirm the superior efficiency of the proposed method in terms of computational time.

The results of the experiment with the **Skating** video sequence are listed in Table 6. The method proposed in [28] required 0.6916s per frame for searching, for a total time of 1,532.7112s. Because Yolact [19] detected a variety of objects for this sequence, only the 'person' class was set to be detected in the experiment. However, some objects were not detected owing to the low detection confidence [19]. In contrast, the proposed method required 0.0334s per frame for searching, and an additional 0.2266s per selected frame for depth mask generation, for a total time of 179.1887s. Thus, the proposed method achieved more accurate and faster video stitching results irrespective of video content, as it uses a motion map to select frames for new seam generation.

To summarize, the experimental results presented in Section IV-A confirm that the proposed image-stitching method reduces parallax distortion using the GED map,

TABLE 5. Results of computational time experiment on people sequence.

	Herrmann et al. [28]	Proposed method
Average search time per frame (sec)	0.1827 (with Yolact [19])	0.0091
Number of search frames	4,680	4,680
Total time for search (sec)	854.9568	42.4787
Number of selected frames	366	422
Additional time per selected frame (sec)	-	0.1352 (with MiDaS [23])
Additional time for selected frames (sec)	-	57.0616
Total Time (sec)	854.9568	99.5403

TABLE 6. Results of computational time experiment on skating sequence.

	Herrmann et al. [28]	Proposed method
Average search time per frame (sec)	0.6916 (with Yolact [27])	0.0334
Number of search frames	2,216	2,216
Total time for search (sec)	1,532.7112	74.0587
Number of selected frames	527	464
Additional time per selected frame (sec)	-	0.2266 (with MiDaS [23])
Additional time for selected frames (sec)	-	105.1300
Total Time (sec)	4,578.4554	179.1887

whereas the results presented in Section IV-B confirm that the proposed method yields faster and more accurate performance when applied to video data.

V. CONCLUSION

In this study, we developed the Guided Energy—Depth (GED) map-based image and video stitching system and obtained experimental results to verify its superior performance and efficiency. The system incorporates a Guided Energy—depth Map Composition (Guided EMC) module and a Frame Selection (FS) module for image and video stitching, respectively.

To perform image stitching, the Guided EMC module constructs an energy—depth map using a visual cognitive energy function and depth information to ensure minimal parallax distortion. This is achieved by adding a guidance map to the energy—depth map to generate a more appropriate seam, enabling the Guided EMC Module to set weights proportional to the depth information for the inner and borderline regions of objects. Consequently, parallax distortion is not only subjectively minimized in terms of the human visual perspective, but also quantitatively minimized as demonstrated in our experiments.

To extend the proposed model for video stitching, we introduced the FS module that uses a depth map to efficiently stitch video frames. To select a specific frame for seam regeneration, the FS module defines a motion map using residual information from the current and previous video frames, and then sets different decision thresholds according to the depth information. The FS module does not apply a deep learning-based algorithm to every frame, instead only selecting frames that do not meet the condition in the Frame Check submodule. Consequently, the proposed method required only 7.2769% of the time burden incurred by the conventional method [28] to stitch video sequences. Thus, the proposed image and video stitching system is expected to provide more accurate and efficient stitching results for various panoramic image and video services.

However, it may be desirable to obtain more accurate depth information to further minimize parallax distortion. This information may be represented by the absolute or relative depth values for distinguishing multiple objects in images, which may represent a subsequent technical challenge for future research.

REFERENCES

[1] S. Rhee, J. Kang, and K. Kim, "Parallax distortion detection and correction method for video stitching by using LDPM image assessment," *J. Broadcast Eng.*, vol. 25, no. 5, pp. 685–697, Sep. 2020.

[2] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2007.

[3] S. Rhee, J. Kang, and K. Kim, "Image stitching focused on priority object using deep learning based object detection," *J. Broadcast Eng.*, vol. 25, no. 6, pp. 882–897, Nov. 2020.

[4] W. Lyu, Z. Zhou, L. Chen, and Y. Zhou, "A survey on image and video stitching," *Virtual Reality Intell. Hardw.*, vol. 1, no. 1, pp. 55–83, Feb. 2019.

[5] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2339–2346.

[6] J. Gao, Y. Li, T.-J. Chin, and M. S. Brown, "Seam-driven image stitching," in *Proc. EUROGRAPHICS*, 2013, pp. 45–48.

[7] J. Kang, J. Kim, I. Lee, and K. Kim, "Minimum error seam-based efficient panorama video stitching method robust to parallax," *IEEE Access*, vol. 7, pp. 167127–167140, 2019.

[8] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] A. Mills and G. Dudek, "Image stitching with dynamic elements," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1593–1602, Sep. 2009.

[10] N. Li, T. Liao, and C. Wang, "Perception-based energy functions in seam-cutting," 2017, *arXiv:1701.06141*.

[11] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *Proc. ACM SIGGRAPH papers*, Jul. 2007, p. 10.

[12] Y. Tang and J. Shin, "De-ghosting for image stitching with automatic content-awareness," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2210–2213.

[13] H. Wen and J. Zhou, "An improved algorithm for image mosaic," in *Proc. Int. Symp. Inf. Sci. Eng.*, Dec. 2008, pp. 497–500.

[14] D. Lee, J. Yoon, and S. Lim, "Image stitching using multiple homographies estimated by segmented regions for different parallaxes," in *Proc. Int. Conf. Vis., Image Signal Process. (ICVISP)*, Sep. 2017, pp. 71–75.

[15] X. Gu, P. Song, Y. Rao, Y. G. Soo, C. F. Yeong, J. T. C. Tan, H. Asama, and F. Duan, "Dynamic image stitching for moving object," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2016, pp. 1770–1775.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[19] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[20] K. Kim et al., "Depth-based image stitching using MegaDepth network," in *Proc. Korean Soc. Broadcast Eng. Conf.*, The Korean Institute of Broadcast and Media Engineers, 2021, pp. 275–278.

[21] K. Mühlmann, D. Maier, J. Hesser, and R. Männer, "Calculating dense disparity maps from color stereo images, an efficient implementation," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 79–88, 2002.

[22] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from Internet photos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2041–2050.

[23] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[24] J. Hu, D.-Q. Zhang, H. Yu, and C. W. Chen, "Discontinuous seam cutting for enhanced video stitching," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[25] M. Tennøe, E. Helgedagsrud, M. Næss, H. K. Alstad, H. K. Stensland, V. R. Gaddam, D. Johansen, C. Griwodz, and P. Halvorsen, "Efficient implementation and processing of a real-time panorama video pipeline," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2013, pp. 76–83.

[26] H. Liu, C. Tang, S. Wu, and H. Wang, "Real-time video surveillance for large scenes," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2011, pp. 1–4.

[27] B. He and S. Yu, "Parallax-robust surveillance video stitching," *Sensors*, vol. 16, no. 1, p. 7, Dec. 2015.

[28] C. Herrmann, C. Wang, R. S. Bowen, E. Keyder, and R. Zabih, "Object-centered image stitching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 846–861.

[29] M. Lin, T. Liu, Y. Li, X. Miao, and C. He, "Image stitching by disparity-guided multi-plane alignment," *Signal Process.*, vol. 197, Aug. 2022, Art. no. 108534.

[30] X. Chen, M. Yu, and Y. Song, "Optimized seam-driven image stitching method based on scene depth information," *Electronics*, vol. 11, no. 12, p. 1876, Jun. 2022.

[31] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[32] K. Lin et al., "SEAGULL: Seam-guided local alignment for parallax-tolerant image stitching," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 370–385.

SEONGBAE RHEE received the B.S. degree in electronics engineering and the M.S. degree in electronics and information convergence engineering from Kyung Hee University, Yongin, South Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electronics and information convergence engineering. His research interests include signal processing, multimedia systems, pattern recognition, and computational intelligence.

GWANG HOON PARK (Senior Member, IEEE) received the B.S. and first M.S. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1985 and 1987, respectively, and the second M.S. and Ph.D. degrees in electrical engineering and applied physics from Case Western Reserve University, OH, USA, in 1991 and 1995, respectively. He was a Principal Research Engineer with the Information and Telecommunication Research and Development Center, Hyundai Electronics Industries, Icheon, South Korea, from 1995 to 1997; and an Associate Professor with the Department of Computer Science, Yonsei University, Wonju, South Korea, from 1997 to 2001. Since 2001, he has been a Professor with the Department of Computer Engineering, Kyung Hee University, South Korea. He has been a co-inventor of more than 700 patents registered worldwide. His research interests include video signal processing, multimedia systems, pattern recognition, and computational intelligence. He received the King Sejong Award; the Grand Prize for Patent Technology by the Head of the Korean Intellectual Property Office, in 2004, in recognition of the excellence of the patent he invented; and the Service Merit Medal by the President of the Republic of Korea on World Standards Day, in 2014, in recognition of his contribution to video coding standardization activities.

KYUHEON KIM (Member, IEEE) received the B.S. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 1989, and the M.Phil. and Ph.D. degrees in electrical and electronic engineering from the University of Newcastle Upon Tyne, U.K., in 1996. From 1996 to 1997, he was with Sheffield University, U.K., as a Research Fellow. From 1997 to 2006, he was with the Electronics and Telecommunications Research Institute, South Korea, as the Head of Interactive Media Research Team, where he standardized and developed T-DMB specification, and conducted a Head of Korean delegates for MPEG standard body, from 2001 to 2005. Since 2006, he has been conducting research with Kyung Hee University, Seoul. He has published numerous technical papers. His current research interests include interactive media processing, digital signal processing, and digital broadcasting technologies. He was a recipient of the Ministry Award from the Ministry of Information and Communication, in 2003, and the Prime Minister Award, in 2005.