Composition Transformation of a Single Landscape Image using Segmentation and 3D Reconstruction

Yuya Suganuma and Norihiko Kawai

Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata, Osaka, Japan

ABSTRACT

This study proposes a composition transformation method to generate visually more appealing images from a single landscape image. The method reconstructs a 3D scene from a single image, virtually changes the camera position to achieve desired compositions, and performs inpainting on the missing regions while considering region-specific characteristics. In the experiments, we generated images using centered and rule-of-thirds compositions. The effectiveness of the proposed method was demonstrated through comparisons with conventional approaches.

Keywords: Composition transformation, Inpainting, 3D reconstruction

1. INTRODUCTION

With the widespread use of smartphones and social networking services, opportunities to capture and share photographs have increased significantly. As a result, users often seek visually attractive images that follow well-established photographic composition rules, such as centered composition or the rule of thirds.

Existing approaches for enhancing photo aesthetics mainly rely on two-dimensional image processing, including color correction and tone adjustment. While these methods can improve the visual appearance of an image, they cannot alter fundamental compositional factors determined at the time of capture. To address this limitation, several studies have explored composition transformation techniques based on cropping and outpainting.[1–3] However, aggressive cropping often reduces image resolution or produces unnatural aspect ratios. In addition, depending on the target composition, the region requiring extrapolation may become large, resulting in images that differ significantly from the original scene.

As an alternative approach, three-dimensional scene representations derived from a single image have been investigated to enable viewpoint changes. Methods such as Tour Into the Picture[4] and learning-based novel view synthesis[5,6] allow virtual camera movement. Nevertheless, existing approaches either restrict camera motion to avoid visual artifacts or rely on generative models whose performance degrades when large extrapolated regions are required.

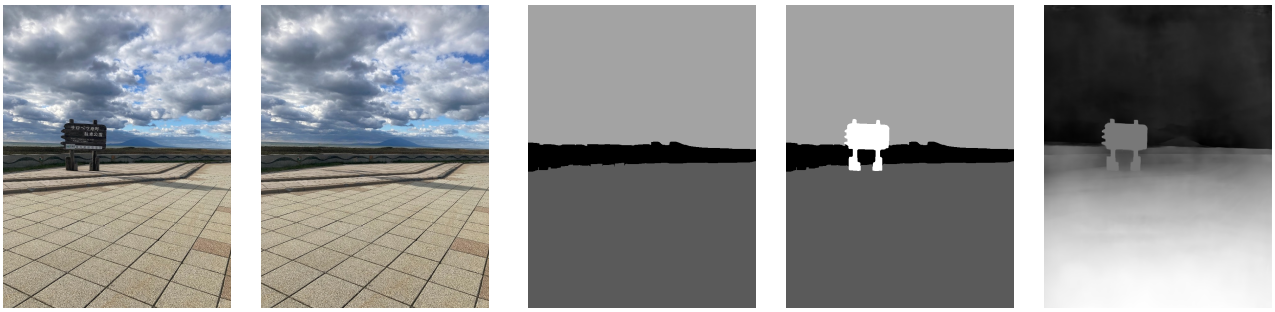
In this study, we propose a region-based composition transformation method for single landscape images using 3D reconstruction. The method generates a textured mesh from a single image and transforms the composition by virtually changing the camera position. Missing regions caused by viewpoint changes are then filled by inpainting while considering region-specific characteristics. This approach enables visually consistent composition transformation for single outdoor landscape images, even when the extrapolated regions are relatively large.

2. PROPOSED METHOD

2.1 Overview

The proposed method first takes a landscape image that includes a primary object, the sky, and the ground, and generates four additional images from it: a background image with objects removed, a region label image of the input image, a region label image of the background image, and a depth image of the input image. Using the input image and the four generated images, textured 3D meshes are then generated. Next, the optimal composition is determined by moving a virtual camera. Gaps between meshes due to camera movement in the composition-transformed image are filled by inpainting. Each step is described in detail below.

Further author information:
Yuya Suganuma: E-mail: m1m24a21@oit.ac.jp
Norihiko Kawai: E-mail: norihiko.kawai@oit.ac.jp



(a) Example input image (b) Background image (BG) (c) Region label image of BG (d) Region label image of input (e) Depth image of input

Figure 1. Preparation of intermediate images.

2.2 Preparation of Intermediate Images

First, IOPaint,[7] an inpainting method, is applied to the input image, as shown in Fig. 1(a), by manually selecting the object to generate a background image, as shown in Fig. 1(b). Next, semantic segmentation[8] is applied to the background image to generate region label images representing sky, ground, and other regions, as shown in Fig. 1(c). SAM,[9] a segmentation method, is then applied to the input image to generate a region label image of the object. We obtain a region label image corresponding to the input image by integrating the results as shown in Fig. 1(d). We also generate a depth image corresponding to the input image using DepthMap, a function in Stable Diffusion Web UI,[10] as shown in Fig. 1(e).

2.3 Generation of Textured 3D Mesh

The five images in Fig. 1 are used to independently generate the 3D mesh for each of the object, ground, and other regions. First, the 3D coordinates of the depth image are calculated using the standard projection model. Next, for the ground and other regions, a plane is fitted to the 3D coordinates of each region using RANSAC[11] followed by the least-squares method to suppress unnatural texture distortion during the camera-motion-based composition transformation described below. For the object region, the depth value of the ground pixel located directly beneath the lowest part of the object region is used as the uniform value of the entire object. This maintains consistency between the object and its shadow position on the ground when the camera is moved. Finally, texture mapping is performed on the generated meshes by projecting the background image.

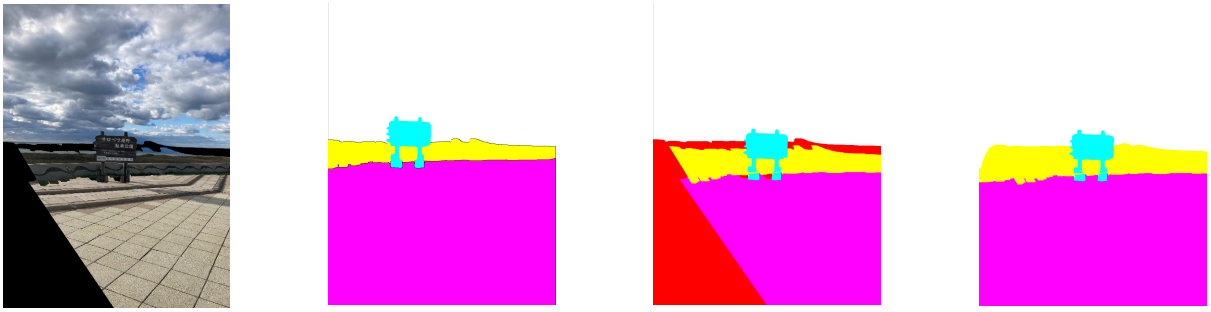
2.4 Composition Transformation

In the composition transformation, the virtual camera is manually translated in a direction perpendicular to the optical axis of the image so that the primary object aligns with the desired compositional layout (e.g., centered or rule-of-thirds placement), as shown in Fig. 2(a). Here, the sky region is assumed to be at infinity; therefore, its position remains fixed in the rendering process even when the camera is moved.

Next, the same composition transformation is applied to the region label image colorized from Fig. 1(d), as shown in Fig. 2(c). In this transformation, the red region indicates missing areas for which no information exists in the original input image. Subsequently, patch-based inpainting[12] is applied to obtain a region label image as shown in Fig. 2(d). These labels are later used during inpainting of the composition-transformed image to suppress the generation of unnatural textures.

2.5 Inpainting

Region-specific inpainting is performed. First, for the ground region, a homography transformation is applied to transform the ground region of the input image into a frontal view, as shown in Fig. 3(a). This process compensates for projective distortions. Next, patch-based inpainting[12] is applied to the frontal view image to obtain a ground image with no missing areas, as shown in Fig. 3(b). Finally, by mapping it onto the ground plane, we obtain a texture-filled image of the ground region after the composition transformation. Next, the



(a) After composition transformation

(b) Colorized region label image

(c) Composition transformation from (b)

(d) Inpainted image from (c)

Figure 2. Composition transformation and region label images.

(a) Ground texture before inpainting

(b) Ground texture after inpainting

Figure 3. Ground texture.

missing regions in the other region are filled. Using only the textures in the other regions shown in Fig. 2(d) as references, we apply inpainting[12] to suppress the generation of unnatural textures. Finally, inpainting[12] is applied to the remaining missing regions to obtain the final output image with all the missing regions filled in.

3. EXPERIMENTS

3.1 Overview

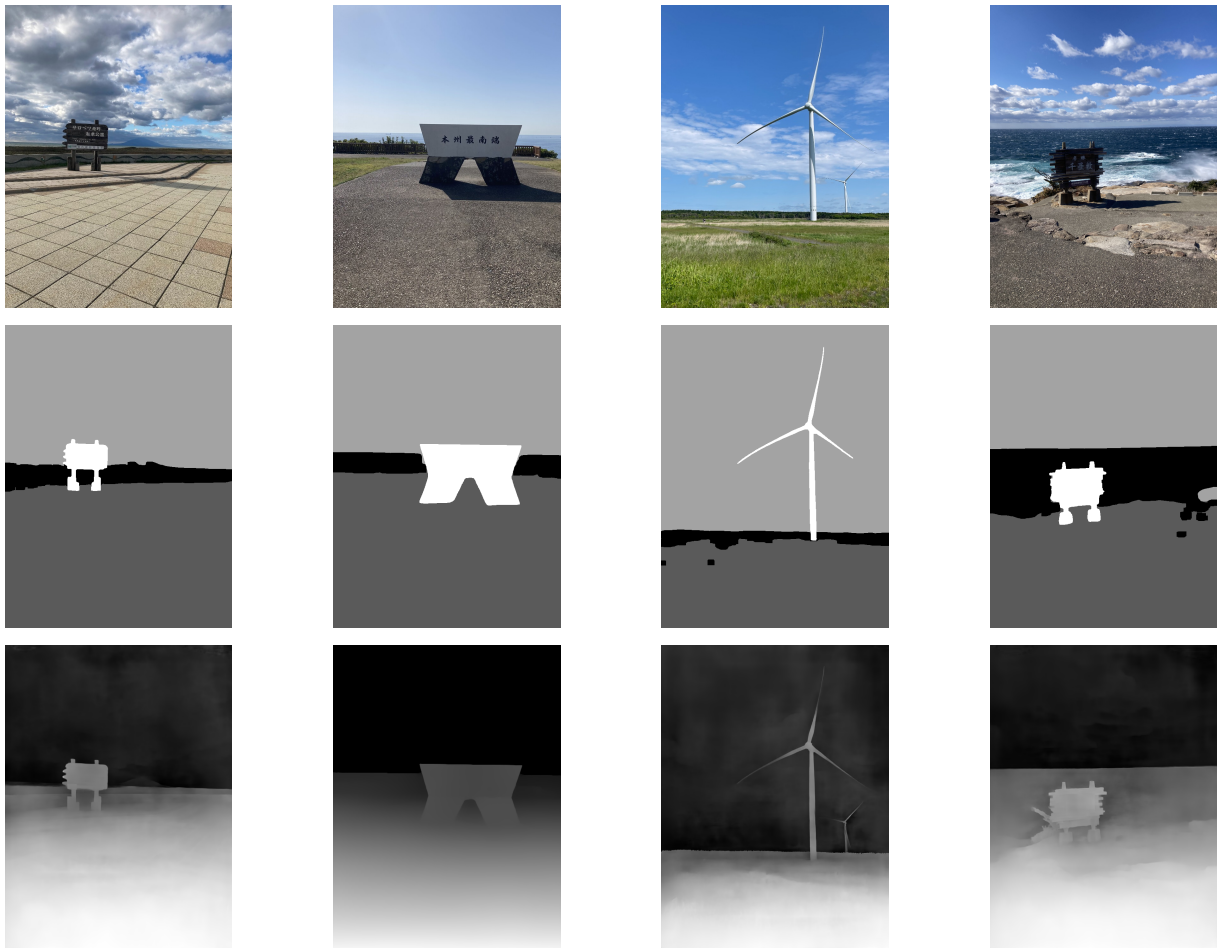
We conducted experiments to perform composition transformations to two common photographic styles: centered and rule-of-thirds compositions. The composition transformations applied to the input images shown at the top of Figs. 4(a)-(d), whose resolution is 720 × 960 pixels, are referred to as Experiments 1 through 4, respectively.

In the experiments, we compared the results of the proposed method with those obtained using a combination of cropping and outpainting, an approach adopted by Zhong et al.,[1] as a conventional baseline. Although Zhong et al. determine cropping and outpainting regions based on evaluations of image quality and compositional aesthetics, in our experiments, we manually set the cropping and outpainting regions so that the size and position of the primary object match the transformed compositions produced by the proposed method.

3.2 Experiment Results

The middle and bottom of Fig. 4 show the generated region label and depth images. Figure 5 shows the images after transformation to centered and rule-of-thirds compositions, respectively. Figure 6 presents the experimental results for Experiments 1 through 4, respectively. In each figure, (a) and (b) show the results for the centered and the rule-of-thirds compositions. Panels (c) and (d) show comparison images obtained by first applying outpainting using Stable Diffusion Web UI,[10] followed by cropping to a 720 × 960 resolution to match the centered and rule-of-thirds compositions.

From these experimental results, the comparison images generated by the conventional method, shown in Fig. 6(c) and (d), contain large interpolated regions, which often lead to unnatural textures. In contrast, the results of the proposed method, shown in Fig. 6(a) and (b), demonstrate that composition transformation can be achieved without generating visually implausible objects.



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3 (d) Experiment 4

Figure 4. From top to bottom: input images, region label images, and depth images.

In composition transformations where the primary objects are shifted substantially from their original positions, the size of the missing regions increases, as observed in Fig. 6. In these cases, most of the missing regions correspond to ground areas. Therefore, the use of region-specific inpainting combined with fronto-parallel rectification of ground textures contributes to preserving the natural appearance of the generated images.

On the other hand, several limitations can also be observed in Fig. 6. In the rule-of-thirds result of Experiment 1, small gaps appear between the ground and other regions, causing slight leakage of sky textures. In the rule-of-thirds result of Experiment 2, camera movement results in the ground and sky regions becoming adjacent, making their boundaries more noticeable. In the rule-of-thirds result of Experiment 3, slight positional shifts at the base of the primary objects can be observed due to camera motion. In addition, in the centered result of Experiment 1 and the rule-of-thirds result of Experiment 4, blurring occurs along the horizontal boundaries between the sky and other regions.

Overall, the proposed method achieves composition transformations involving significant object movement while reducing the size of extrapolated regions. In particular, for ground regions, removing perspective distortion prior to inpainting allows natural textures to be synthesized. However, gaps between regions caused by camera movement are likely due to inaccuracies in the segmentation used for generating region label images, indicating that further refinement of the segmentation step is necessary. Moreover, the blurring observed at region boundaries suggests that boundary-aware inpainting should be considered in future work.

(a) Experiment 1 (b) Experiment 2 (c) Experiment 3 (d) Experiment 4

Figure 5. From top to bottom: images after transformation to centered composition and rule-of-thirds composition.

4. CONCLUSION

This study proposed a method for performing image composition transformation from a single landscape image using region-based segmentation and 3D reconstruction. The experimental results demonstrated that the proposed method can achieve composition transformations without introducing significant deviations from the original scene or generating unnatural objects. However, the current method sometimes produces artifacts. Future work will focus on improving segmentation accuracy and incorporating boundary-aware inpainting.

REFERENCES

[1] Zhong, L., Li, F.-H., Huang, H.-Z., Zhang, Y., Lu, S.-P., and Wang, J., "Aesthetic-guided outward image cropping," *ACM Transactions on Graphics* **40**(6), 1–13 (2021).

[2] Hong, C., Du, S., Xian, K., Lu, H., Cao, Z., and Zhong, W., "Composing photos like a photographer," in [*Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7053–7062 (2021).

[3] Pan, Z., Cao, Z., Wang, K., Lu, H., and Zhong, W., "Transview: Inside, outside, and across the cropping view boundaries," in [*Proc. IEEE/CVF International Conference on Computer Vision*], 4198–4207 (2021).

[4] Horry, Y., Anjyo, K., and Arai, K., "Tour into the picture: using a spidery mesh interface to make animation from a single image," in [*Proc. SIGGRAPH*], 225–232 (1997).

[5] Wiles, O., Gkioxari, G., Szeliski, R., and Johnson, J., "Synsin: End-to-end view synthesis from a single image," in [*Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7465–7475 (2020).

[6] Li, J., Feng, Z., She, Q., Ding, H., Wang, C., and Lee, G. H., "Mine: Towards continuous depth mpi with nerf for novel view synthesis," in [*Proc. IEEE/CVF International Conference on Computer Vision*], 12558–12568 (2021).

[7] Sanster, "IOPaint." https://github.com/Sanster/IOPaint. (Accessed: 18 November 2025).

[8] sithu31296, "semantic-segmentation." https://github.com/sithu31296/semantic-segmentation. (Accessed: 18 November 2025).

[9] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R., "Segment anything," in [*Proc. IEEE/CVF International Conference on Computer Vision*], 4015–4026 (2023).

(a) Proposed: centered (b) Proposed: thirds (c) Conventional: centered (d) Conventional: thirds

Figure 6. Composition transformation results for four scenes. From left to right: proposed (centered), proposed (rule of thirds), conventional (centered), and conventional (rule of thirds). Rows correspond to Experiments 1–4 from top to bottom.

[10] AUTOMATIC1111, "Stable Diffusion web UI." https://github.com/AUTOMATIC1111/stable-diffusion-webui. (Accessed: 18 November 2025).

[11] Fischler, M. A. and Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM* **24**(6), 381–395 (1981).

[12] Kawai, N. and Yokoya, N., "Image inpainting considering symmetric patterns," in [*Proc. International Conference on Pattern Recognition*], 2744–2747 (2012).