2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3503-4839-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VRW58643.2023.00285

Scene Transformer: Automatic Transformation from Real Scene to Virtual Scene

Runze Fan[1] Lili Wang[1,2] * Chan-Tong Lam[3] Wei Ke[3]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Peng Cheng Laboratory, Shengzhen, China
[3]Macao Polytechnic University, Macao, China.

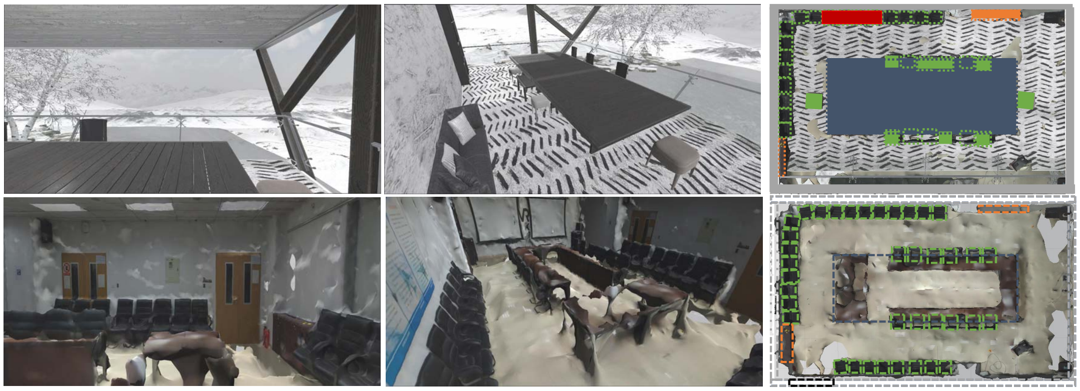

Figure 1: Indoor Scene Transformation problem is defined as transforming an input virtual scene into a new virtual scene constrained by the input real scene. The first column shows the mixed scene and the original real scene from the same user's perspective. The second column shows the original virtual scene (top) and the real scene (down). The third column is the transformed virtual scene and the original real scene, where the rectangle represents the the virtual furniture and the dotted box represents the real furniture.

Abstract

Given a real scene and a virtual scene, the indoor scene transformation problem is defined as transforming the layout of the input virtual scene. The transformed layout preserves as much as possible the relationship between the furniture in the input virtual scene, and the input real scene provides the user with as much passive haptic as possible when exploring the virtual scene. We propose a real-scene-constrained deep scene transformer to solve this problem. First, we introduce the deep scene matching network to predict the matching relationship between real furniture and virtual furniture. Then we introduce a layout refinement algorithm based on the refinement parameter network to arrange the matched virtual furniture into the new virtual scene. At last, we introduce a deep scene generating network to arrange the unmatched virtual furniture into the new virtual scene.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality;

1 Introduction

Affected by the epidemic of COVID-19, people are restricted in their mobility, VR environment exploration provides a solution to reduce the impact, for example, VR navigation in the 3D model scene, but the interactivity and fidelity are limited. Mixed reality provides a more creative way for users to explore the scene with a high level of interactivity and fidelity. As shown in Fig. 1, in a real conference room, users can wear an augmented reality HMD, and the sightseeing room is mixed rendered in the real conference room. Users will have the illusion that they are sitting on the sofa enjoying the scenery outside the window when the truth is that they are sitting on a chair staring at a blank wall. In order to give the MR solution, there is one challenge to be addressed. Given a real scene and a virtual scene, how to arrange the virtual furniture in the virtual scene to the real scene so that the real scene looks similar to the virtual scene and is interactive. We call this challenge the indoor scene transformation problem.

There are 2 problems related to the scene transformation problem. The first one is the scene retargeting problem [2,4] and the second one is the scene synthesis problem [1,6]. While, the former allows furniture in the virtual scene to be discarded, and the latter does not consider the relationship between the furniture in the virtual scene.

In this paper, we propose a real-scene-constrained deep scene transformer to transform the layout of the input virtual scene. The transformed layout preserves as much as possible the relationship between the furniture in the input virtual scene, and the input real scene provides the user with as much passive haptic as possible when exploring the virtual scene. We focus on the main furniture in the scene without considering some tiny decorations. There are 3 main steps in our scene transformation process: scene matching, layout generating of matched virtual furniture, and layout generating of unmatched virtual furniture. First, we introduce the deep scene matching network to predict the matching relationship between real furniture and virtual furniture. Then we introduce a layout refinement algorithm based on the refinement parameter network to arrange the matched virtual furniture into the new virtual scene,i.e., predicting suitable layouts of them to match the real scene. At last, we introduce a deep scene generating network to arrange the unmatched virtual furniture in the new virtual scene.

*Lili Wang is corresponding author: wanglily@buaa.edu.cn

979-8-3503-4839-2/23/$31.00 ©2023 IEEE
DOI 10.1109/VRW58643.2023.00285

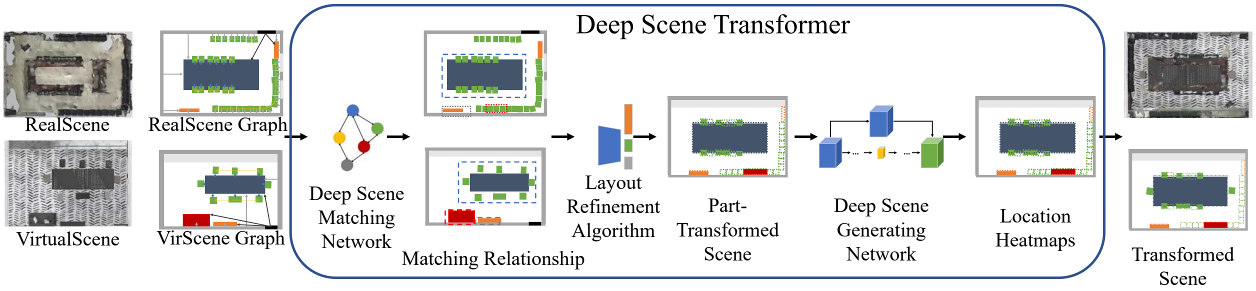

Figure 2: The pipeline of the deep scene transformer. Given the real scene and the virtual scene, we first introduce a deep scene matching network to predict the matching relation between real furniture and virtual furniture. Then, for the matched virtual furniture, a layout refinement algorithm is introduced to arrange them. Next, we introduce a deep scene generating network to arrange the unmatched virtual furniture.

2 Deep Scene Transformer

2.1 Pipeline

Our goal is to transform the input virtual scene S_v into a new virtual scene $S_v{}^{'}$ while maintaining the category and quantity of virtual furniture. The transforming process is implemented by changing the layout of the virtual furniture. The interactivity and fidelity are implemented by matching the virtual furniture F_v in S_v with the real furniture F_r in S_r as much as possible while maintaining the real furniture relationship and virtual furniture relationship.

As shown in Fig. 2, we propose a 3-step pipeline according to the goal. 1) We introduce a deep scene matching network to predict the matching matrix M between F_v and F_r; 2) We introduce a layout refinement algorithm based on the refinement parameter network to predict the layout of matched virtual furniture $F_v^{m'}$; 3) We introduce a deep scene generating network to predict the layout of unmatched virtual furniture $F_v^{um'}$.

2.2 Deep Scene Matching Network

We first represent the real and virtual scenes as scene graphs, furniture is encoded as nodes, and relationships between furniture are encoded as edges. Then we apply two GNNs with the same architecture to propagate and aggregate messages encoded by nodes and edges. With aggregated features, we apply an affinity layer to calculate the affinity between furniture in the two scenes. We adopt fully-connected layers as the affinity layer. Next, we use a sinkhorn layer to predict the mathcing matrix [3].

2.3 Layout Refinement Algorithm

With the predicted matching matrix, virtual furniture can be divided into matched virtual furniture and unmatched virtual furniture. In this part, we transform the matched virtual furniture. We start with a multi-channel top-down projection representation to represent the real scene and the virtual furniture to be transformed. The first 3 channels of this representation are the orthographic top-down view image. The next channel is an orthographic top-down depth image. The final channel is an orthographic top-down category image. Then we use a refinement parameter network to predict 3 parameters to control the layout refinement process, which are S, K, and O. S determines which side of the real and virtual furniture aligns. K determines whether virtual furniture can be scaled without maintaining the aspect ratio. O determines whether to allow virtual furniture beyond the area where real furniture is located. This network takes the projection representation of the real scene and the virtual furniture as input, uses CNN to encode the information, and uses 3 classifiers to predict the parameters respectively. Next, based on the predicted parameters, we first move and then scale the virtual furniture according to the layout of the corresponding matched real furniture.

2.4 Deep Scene Generating Network

In this part, we transform the unmatched virtual furniture. We first represent the part-transformed scene with the multi-channel top-down projection representation. Then we apply the hourglass network [5] to predict the location probability distribution map of the unmatched virtual furniture. Besides, the hourglass network takes an additional input,i.e., the original size of the unmatched virtual furniture. This additional input is injected into the network via Featurewise Linear Modulation. Given the predicted map, we take the pixel with the maximum probability as the exact location. For a properly arranged scene, the front direction of the furniture should point to the center of the scene. Based on this principle, we determine the orientation of the unmatched virtual furniture. For the size, we try to keep the original size. In the case of insufficient space, we appropriately shrink the transformed virtual furniture and arrange it in the empty space.

3 Conclusion

We introduce the indoor scene transformation problem and propose a real-scene-constrained deep scene transformer to solve this problem which contains 3 steps. First, a deep scene matching network is introduced to predict the matching relationship between real furniture and virtual furniture. Then a layout refinement algorithm is introduced to transform the matched virtual furniture. Third, a deep scene generating network is introduced to transform the unmatched virtual furniture.

Acknowledgments

This work was supported by National Key R&D plan 2019YFC1521102, by the National Natural Science Foundation of China through Projects 61932003, by Beijing Science and Technology Plan Project Z221100007722004.

References

[1] M. Fisher, M. Savva, Y. Li, P. Hanrahan, and M. Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Trans. Graph.*, 34(6), oct 2015.

[2] C.-K. Huang, Y.-L. Chen, I.-C. Shen, and B.-Y. Chen. Retargeting 3d objects and scenes with a general framework. *Comput. Graph. Forum*, 35(7):33–42, oct 2016.

[3] Y. Kushinsky, H. Maron, N. Dym, and Y. Lipman. Sinkhorn algorithm for lifted assignment problems. *SIAM J. Img. Sci.*, 12(2):716–735, jan 2019.

[4] J. Lin, D. Cohen-Or, H. Zhang, C. Liang, A. Sharf, O. Deussen, and B. Chen. Structure-preserving retargeting of irregular 3d architecture. *ACM Trans. Graph.*, 30(6):1–10, dec 2011.

[5] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[6] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Trans. Graph.*, 38(4), jul 2019.