# SceneDirector: Supplementary Materials

Shao-Kui Zhang, Hou Tam, Yi-Ke Li, Ke-Xin Ren, Hongbo Fu, Song-Hai Zhang, *Member, IEEE*

## Appendix A Overview

The supplementary materials contain the following contents that strengthen our contributions:

- **Section B**: The details of the proposed dataset, including styles, statistics and examples.
- **Section C**: The details of the group transitioning from $G^{t-1}$ to $G^t$ corresponding to Section 3.1 of the main text.
- **Section D**: Further illustrations on the attribute smoothness in Section 3.2 of the main text.
- **Section E**: The technical details of the application "SceneForce".
- **Section F**: Engineering details of the timers for the user study in Section 5.1 of the main text.
- **Section G**: The user study platform for measuring the aesthetic and plausibility in Section 5.3 of the main text.
- A demonstration video is attached separately through the online submission system, to demonstrate the proposed method.
- Samples of our dataset are also attached in the supplementary files [1].

## Appendix B The Proposed Dataset

In sum, our dataset contains ten styles elaborated by professional interior designers, including grey modern, gentle light, modern contrast, western pastoral, luxury, new Chinese, minimalism, Wabi-Sabi, American country style and Mediterranean style. Since this dataset is initially proposed to support the coherent group set (CGS) in the main text, so we name it **CGS-0**. Figure 1 shows the ten styles of the dataset. Compared with 3D-Front [1] where the averaged occurrence of objects is $24.865$, ours is $28.515$. It seems to be a minor improvement, but the standard deviation of occurrence of objects is $98.505$ for 3D-Front and is $31.286$ for ours. This is because, in 3D-Front, most objects are only used once or twice among all the layouts, which refers to the imbalance problem for 3D scene datasets [2]. In contrast, an object is thoroughly used in at least one entire CGS. Besides, since this dataset is created for functionally and stylistically coherent groups of objects, we classify them with leader objects representing functions and styles. This dataset will be publicly available. Designing 3D content is expensive. Due to our budget, the dataset contains $1719$ layouts before the submission of this paper. We consider creating this dataset a long-term task. Though the dataset in this paper is sufficient for our method, more layouts will be created in the future. We will also make the entire annotation system open-source. As far as we investigated, we are the first to release an annotation tool for research on 3D scenes.

We first invite three professional interior designers to select a list of furniture categorized in the ten styles as shown in figure 1, with an average of 200 objects in each style. All those common and essential furniture in the living room, dining room and bedroom are provided, such as sofa, bed, nightstand, TV stand, etc. We also ensure that there are multiple choices of objects for each category in each style, e.g., the dinning table category of the gentle light style should have multiple instances (CAD models) to support it. After that, We invited more than 20 designers with interior design and art-related backgrounds to create the CGS-0.

The designers first choose a style from the provided ten styles. Then they select a topic such as living room, dining room or bedroom. Next, one dominant object expressing the topic to the greatest extent is selected with several subordinate objects from the chosen style, in order to design 10-20 different coherent group layouts. The dominant object appears exactly once in each layout while the subordinate objects can be placed repeatedly according to the needs of designers and layout strategies. The 10-20 layouts show how the scene changes gradually and variously w.r.t the contexts and constraints. The designers use the same platform as in our user study to craft the layouts. They can translate, rotate and re-scale furniture objects. After they finish designing a CGS, 10-20 layouts consisting of the same object pool with a consistent style are constructed. More evolutionary chains of CGS are shown in the supplementary files.

- *Shao-Kui Zhang and Hou Tam are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China.*
*E-mail: zhangsk18@mails.tsinghua.edu.cn, th21@mails.tsinghua.edu.cn*
- *Yi-Ke Li and Ke-Xin Ren is with the Academy of Arts & Design, Tsinghua University, Beijing, China.*
*E-mail: lyk20@mails.tsinghua.edu.cn, rkx20@mails.tsinghua.edu.cn*
- *Prof.Hongbo Fu is with the School of Creative Media, City University of Hong Kong, Hong Kong.*
*E-mail: hongbofu@cityu.edu.hk*
- *Song-Hai Zhang is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China and Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China.*
*E-mail: shz@tsinghua.edu.cn*

*Manuscript received - -, 2022; revised - -, -.*

[1] Due to the size limit, the demonstration video and the samples are compressed.

## Appendix C
## Group Transitioning

This section formulates how we re-organize object from $G^{t-1}$ to $G^t$. Assuming an set containing all objects is $\tilde{O}$ as shown in equation 1, which is also the set that containing the actual objects being placed to 3D scenes.

$$\tilde{O} = \{o_i = (x_i, y_i, z_i, u_i, \gamma_i)\} \tag{1}$$

The tuple $(x_i, y_i, z_i)$ stands for the current position of $o_i$ relative to the dominant object in 3D space, and variable $u_i$ stands for whether $o_i$ is already used in the scene. $\gamma_i$ refers to a specific instance, e.g., a pink chair demanded by $G^t$. Assuming the target coherent group generated by a layout concept is $G^t = \{o'_j = (x'_j, y'_j, z'_j, \gamma_j)\}$, where each target object $o'_j$ will be selected from $\tilde{O}$. The tuple $(x'_j, y'_j, z'_j)$ stands for the target position of $o'_j$ in the next transient movement. Thus, the problem is how to select a subset $\hat{O} \subset \tilde{O}$, so that the re-organization from $G^{t-1}$ to $G^t$ is as tidy as possible. The re-organization process is formulated in algorithm 1. For each transient time, this algorithm is executed once, so positions of objects in $\tilde{O}$ constantly change.

**ALGORITHM 1:** Finding the appropriate objects set for the next generated layout $G^t$.

```
Input: Entire object set $\tilde{O}$ and target layout $G^t$.
Output: An optimal subset $\hat{O} \subset \tilde{O}$ with the overall
        minimal moving distances of objects.
for Each object $o_i \in \tilde{O}$ do
    $u_i = False;$
end
$\hat{O} = \emptyset$ ;
for Each $o'_j \in G^t$ do
    $D = \infty$ ;
    for Each $o_i \in \tilde{O}$ do
        if not $u_i$ and $\gamma_i == \gamma_j$ then
            $d = ||(x_i, y_i, z_i) - (x'_j, y'_j, z'_j)||_2$;
            if $d \leq D$ then
                $I = o_i$;
                $D = d$;
            end
        end
    end
    $\hat{O} = \hat{O} \cup \{I\}$ ;
    $o_i = (x'_j, y'_j, z'_j, True)$ ;
end
```

Algorithm 1 follows a greedy approach. The outer loop iterates through the entire $G^t$ to find a best matched $o_i \in \tilde{O}$. Each $o_i$ can be used up to once. If $o_i$ is already occupied by a $o'_j \in G^t$, the $u_i$ of $o_i$ is marked as "true". The inner loop iterates through the $\tilde{O}$. The basic idea is to greedily find a nearest $o_i \in \tilde{O}$ based on the Euclidean distance. After that the position $(x_i, y_i, z_i)$ of $o_i$ will be set to the target position $(x'_j, y'_j, z'_j)$ suggested by $G^t$. Subsequently, $G^{t+1}$ would run algorithm 1 based on the positions set by $G^t$. Note that the proposed coherent group set guarantee that all target objects $o'_j \in G^t$ can always find an object from the object set $\tilde{O}$.

TABLE 1: Smoothness on the scenes in figure 2.

| | 2a | 2b | 2c | 2d |
|---|---|---|---|---|
| 2a | 0 | 0.274 | 0.698 | 0.979 |
| 2b | 0.274 | 0 | 0.833 | 0.735 |
| 2c | 0.698 | 0.833 | 0 | 0.655 |
| 2d | 0.978 | 0.735 | 0.655 | 0 |

## Appendix D
## Smoothness

We exploit the relative direction between the dominant object and subordinate objects. Edges are defined between each subordinate object $o_i$ and the dominant object as unit vectors of $o_i$'s relative direction with respect to the dominant object. The kernel between two edges is the inner product of their unit vectors and is clamped to 0 if less than 0. This metric shows how consistent the directions of two subordinate objects are with the dominant object. The distance is excluded since it might devalue the similarity metric as the room area goes larger or smaller. Table 1 is an example of smoothness on the four scenes in figure 2, where values are normalized in $[0, 1]$. First of all, a scene has no difference from itself, so values in the diagonal are all zero.

First the difference between the scene in figure 2a and the scene in figure 2b is very small, because we simply add several subordinate objects in figure 2b. However, compared with the scene in figure 2c w.r.t 2a, the layout strategy changes, so the value increases from $0.272$ to $0.698$. Compared with the scene in figure 2d w.r.t 2a, the layout strategies further diverge and the value is close to $1.0$. From the scene in figure 2b to the scene in figure 2c, objects are not only added but also removed a bit, where we removed the three small corner-side tables. Thus, the value between them is even higher, though the two scenes in figure 2a and figure 2b are simialr. In contrast, two scenes in figure 2b and figure 2d share the similar groups of objects, so the value between them results in $0.735$. Finally, the two scenes in figure 2c and 2d have a medium difference on both layout strategies and object set, so their value remains $0.655$. Note that each result value of smoothness is shared by groups $G_i \in \mathcal{S}'$ derived from the same coherent group $\mathcal{S}$, to alleviate the precomputing cost.

## Appendix E
## Scene Force

There have been plenty of works researching 3D hand gestures in the context of smart environments. [3] designs semaphoric gestures to trigger discrete commands in smart environments. [4] proposes 3D gestures in indoor synthesis. But gestures are only used to change the position, orientation and scale of individual objects. [5] proposes hand gestures that generate commands to directly control objects in a game. [6] builds a 3D modelling system that could manipulate and deform the meshes with hand gestures. Essentially, these works propose one-to-one gesture-command mapping without much intelligent response by environments. Additionally, only a single object is involved in each interaction. In contrast, our application modifies the entire environment intelligently according to users' interactions, i.e., groups of objects are involved according to the proposed method. As shown in figure 10 in the main text, with users moving their hand from left to right, the positions of groups of objects change and the layout are established accordingly.

We use an HTC Vive Pro Eye HMD for eye-tracking. We use one HTC Vive controller and two HTC Vive trackers for hand-tracking. The application runs at the minimum of the HMD's framerate (90 Hz). We use equation 2 to calculate $\vec{x}$ in the virtual environment depending on hand movements in each frame and the scale ratio. The scale ratio is a positive constant to amplify the movement of objects in a scene. $\Delta s$ denotes scene syntheses movement per frame. $\Delta h$ denotes hand movement per frame, and $a$ denotes axis in x and z.

$$\Delta s_a = r \times \Delta h_a \tag{2}$$

# Appendix F
# Timers

This section illustrates how we add timers to correctly count time in the evaluation of the main text. First, the navigating time includes users using the mouse to rotate the perspective camera by holding the left click, users using the mouse to translate the perspective camera by holding the right click, users using UP/DOWN/LEFT/RIGHT to translate the camera and users using the wheel to zoom in or zoom out. For the former two, the timer records time from starting to hold the trigger until releasing the trigger. For the latter two, since the keyboard and wheel events are discrete, the timer first caches a series of events with intervals less than $t'$. Subsequently, the timer records the time from the start of the series until the end of it. In this paper, we empirically set $t' = 0.2s$. Second, the research time includes users entering keywords such as "nightstand" or "pink fairy tale, users drawing sketches. The time records from their start interacting with the search panel until they click a result item. For adding objects, the time records from that a user clicks a searched item until the object is inserted. If an operation of searching or adding is interrupted, the time is recorded to "others".

As illustrated in the main text, transformations, including translation, rotation, and scale, are adjustable through three panels. The timer starts to record the time when an object is clicked. If a user uses the blue panel, the timer simply ends recording the time when the object is released. If a user uses the local coordinate system, the timer also keeps recording time, but it ends recording time if the user releases the coordinate system for more than $t'$. For example, a user uses the arrows representing the axes to translate objects by holding and dragging the axes. If she/he releases the axes for a while, we consider it is the end of a single interaction. If a user uses the right-top panel, the timer ends recording time after the user inputs the values. Note that if a user selects an object and does nothing, the time is recorded to "others". For removing objects, the timer records from an object being clicked until the object are removed. Our method of group editing is used through the blue panel, so the timer records the time similar to recording transformation time.

# Appendix G
# Aesthetic and Plausibility

Figure 3 shows how we conduct a user study in Section 5.3 in the main text, where users could select a favored scene or select no preference, and they can jump to other questions. Subjects can also zoom in on each scene for better perceptions during the study.

# References

[1] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, and H. Zhang, "3d-front: 3d furnished rooms with layouts and semantics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 933–10 942.

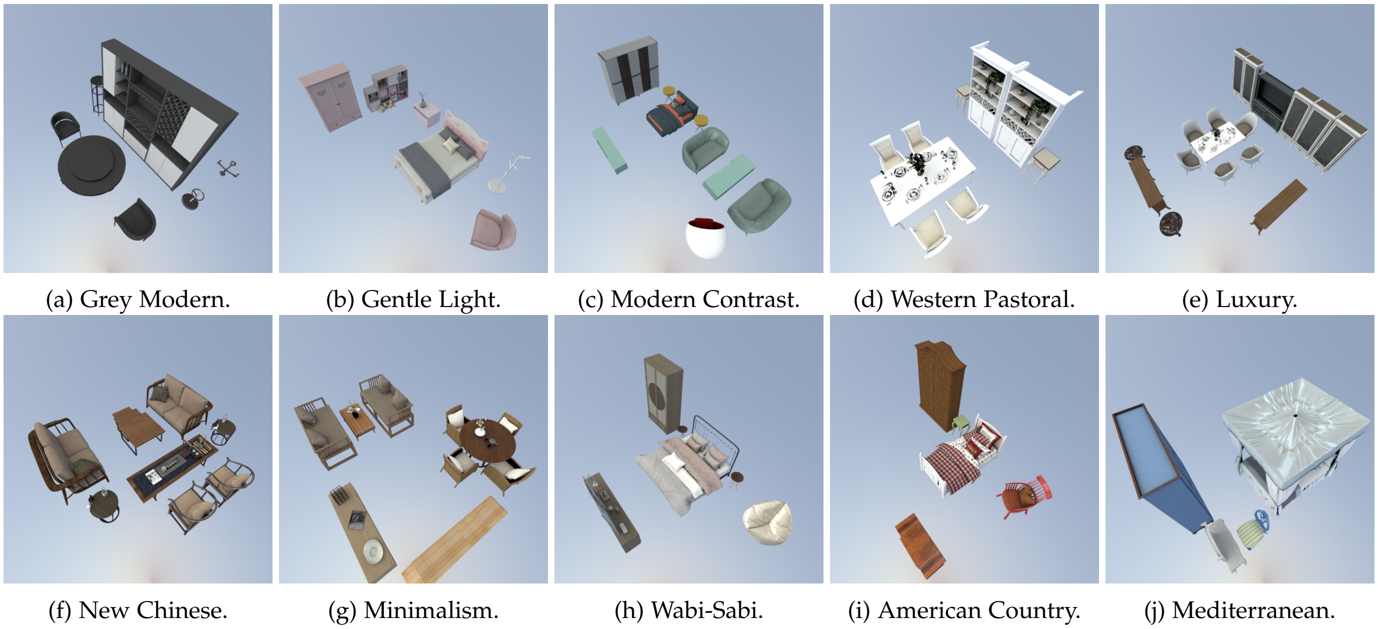
[2] S.-H. Zhang, S.-K. Zhang, W.-Y. Xie, C.-Y. Luo, Y.-L. Yang, and H. Fu, "Fast 3d indoor scene synthesis by learning spatial relation priors of objects," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[3] J. Kela, P. Korpipää, J. Mäntyjärvi, S. Kallio, G. Savino, L. Jozzo, and S. D. Marca, "Accelerometer-based gesture control for a design environment," *Personal and Ubiquitous Computing*, vol. 10, no. 5, pp. 285–299, 2006.

[4] Y. Li, X. Wang, Z. Wu, G. Li, S. Liu, and M. Zhou, "Flexible indoor scene synthesis based on multi-object particle swarm intelligence optimization and user intentions with 3d gesture," *Computers & Graphics*, vol. 93, pp. 1–12, 2020.

[5] N. H. Dardas and M. Alhaj, "Hand gesture interaction with a 3d virtual environment," *The Research Bulletin of Jordan ACM*, vol. 2, no. 3, pp. 86–94, 2011.

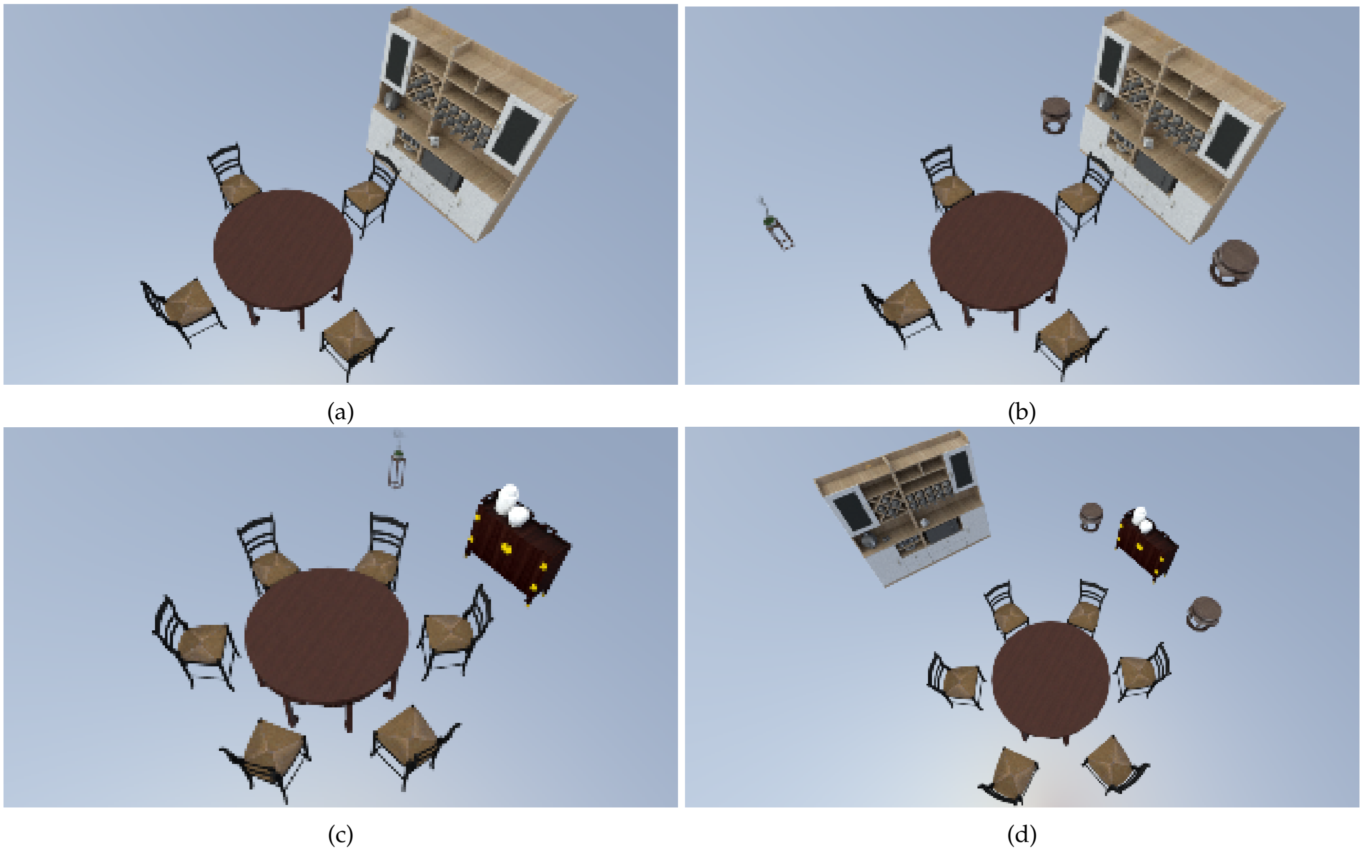
[6] H. Kim, G. Albuquerque, S. Havemann, and D. W. Fellner, "Tangible 3d: Hand gesture interaction for immersive 3d modeling." *IPT/EGVE*, vol. 2005, pp. 191–9, 2005.



(a) Grey Modern. (b) Gentle Light. (c) Modern Contrast. (d) Western Pastoral. (e) Luxury.

(f) New Chinese. (g) Minimalism. (h) Wabi-Sabi. (i) American Country. (j) Mediterranean.

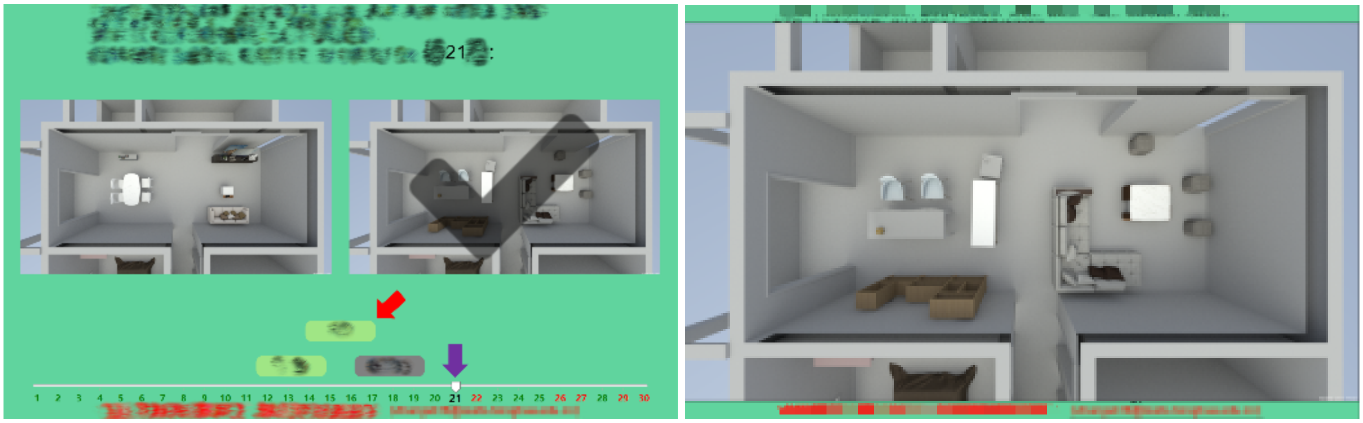
Fig. 1: Examples of the 10 styles in our datasets.



(a) (b)

(c) (d)

Fig. 2: Four example scenes for illustrating differences between scenes.



(a) (b)

Fig. 3: The user study platform for measuring the aesthetic and plausibility. 3a: users could select a favored scene or select no preference (see the red arrow), and they can also jump to other questions (see the purple arrow) if some questions require a further consideration. 3b: zooming in each scene for better perceptions.