Appendices

A. Detailed Analysis

Figure A1 is a sensitivity analysis of the text guidance scale. s_T is a hyper-parameter defined in Eq. 5, which determines the degree of fidelity to the text instruction. With larger s_T, the source scene can be converted to a 3D scene that is more faithful to the text instruction.

Figure A2 shows the source and converted scenes by each method and corresponding depth maps. With Instruct 3D-to-3D, we converted the source 3D scene with the text instruction of "make it into pixel art". With DreamFusion and CLIP-NeRF, we converted with the target text of "A pixel art of a trex fossil". As we noted in section 4.2, DreamFusion tends to generate 3D scenes that ignore the source 3D scene and its geometry. As described in Figure A2, the depth map of the 3D scene converted by DreamFusion does not correspond to the rendered image, and the converted 3D scene has a broken 3D structure. On the other hand, the depth map of a 3D scene converted with our method corresponds to its appearance, and the 3D structure is correctly obtained.

Figure A3 shows how the 3D scene changes during the 3D conversion process. We can see that the 3D scene gradually makes a major structural change and then refines the details.

B. 3D scenes used in the user study

Figure A4 and Figure A5 shows the source and converted 3D scenes used in the user study. As described in section 4.4, we showed these 3D scenes as videos, and the videos are in the submitted supplementary material.

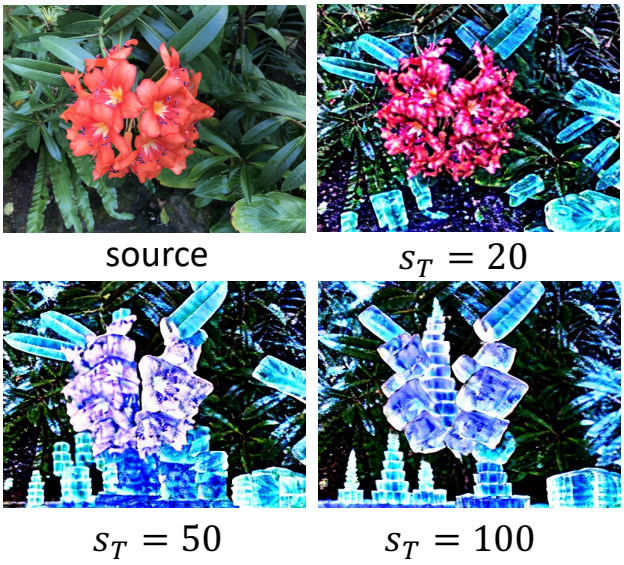

Figure A1: Effects of s_T in the converted 3D scenes. These 3D scenes are converted from the source 3D scene with the text instruction of "make it an ice statue". We can generate a 3D scene that is faithful to the text instruction with larger s_T.

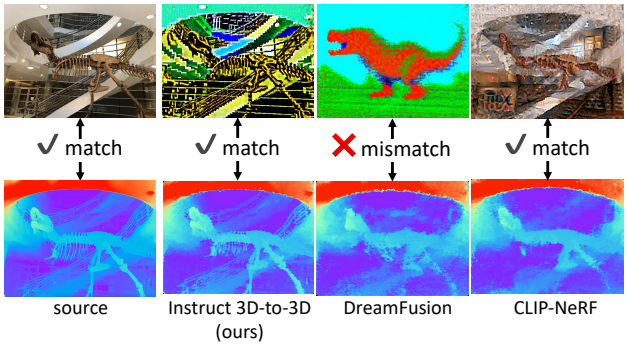

Figure A2: The source and converted scenes and their depth maps.

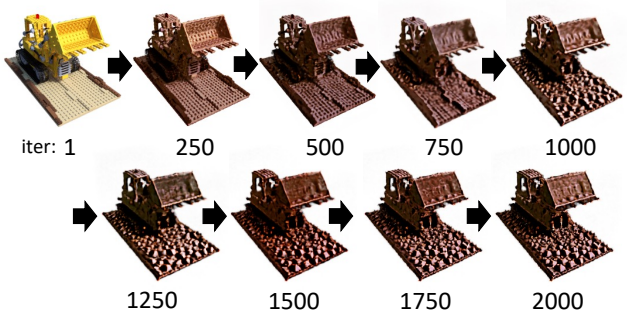

Figure A3: Gradual change of the 3D scene during the 3D conversion process.

Figure A4: Converted 3D scenes from NeRF synthetic dataset used in the user study

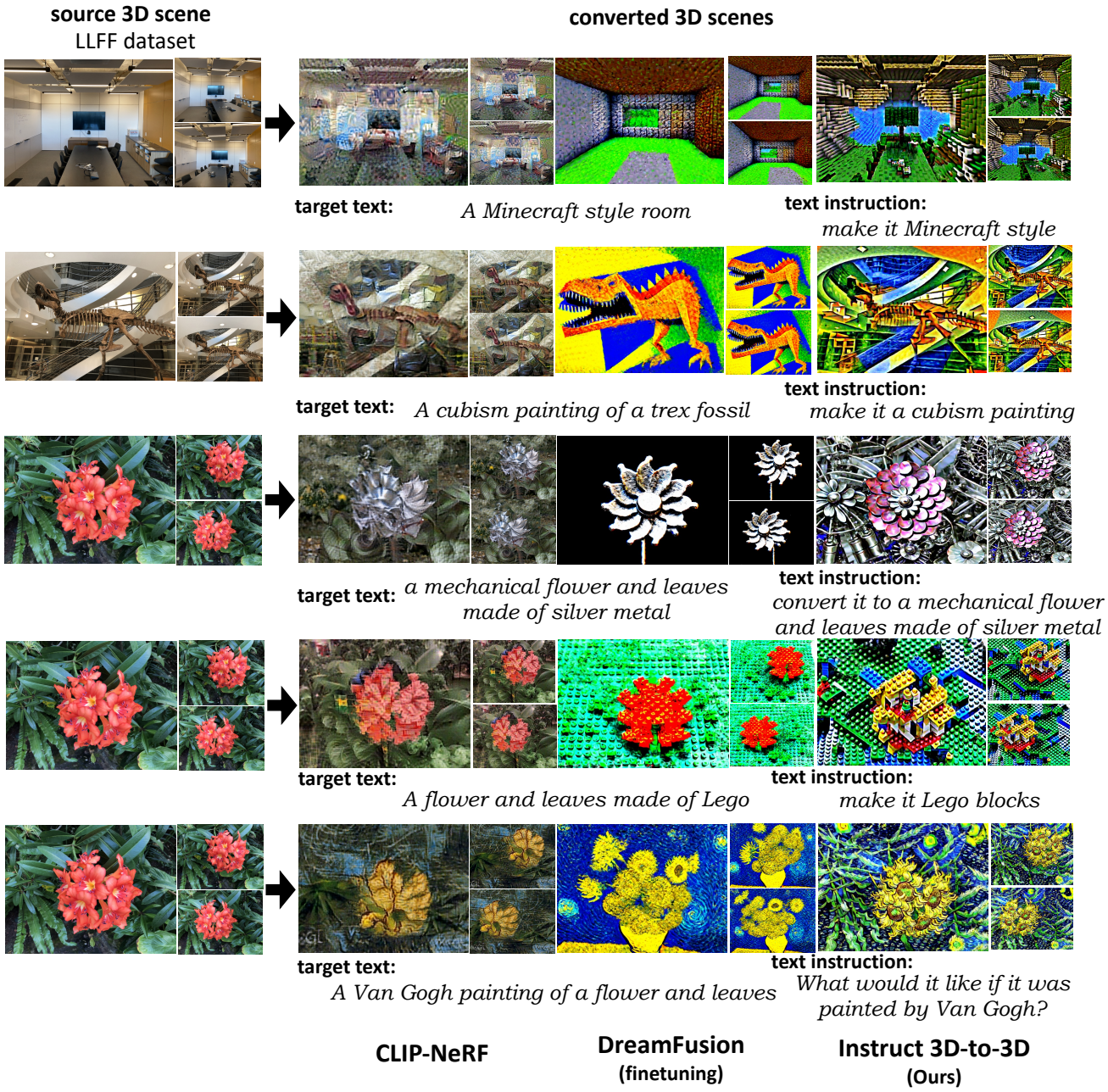

Figure A5: Converted 3D scenes from LLFF dataset used in the user study

C. Examples of Converted 3D scenes of Instruct 3D-to-3D

The followings are examples of 3D scenes converted with Instruct 3D-to-3D. Our Instruct 3D-to-3D is able to generate high-quality converted 3D scenes for a variety of text instructions.

Figure A6: Converted 3D scenes from the 3D scene of lego.

Figure A7: Converted 3D scenes from the 3D scene of hotdog.

Figure A8: Converted 3D scenes from the 3D scene of chair.

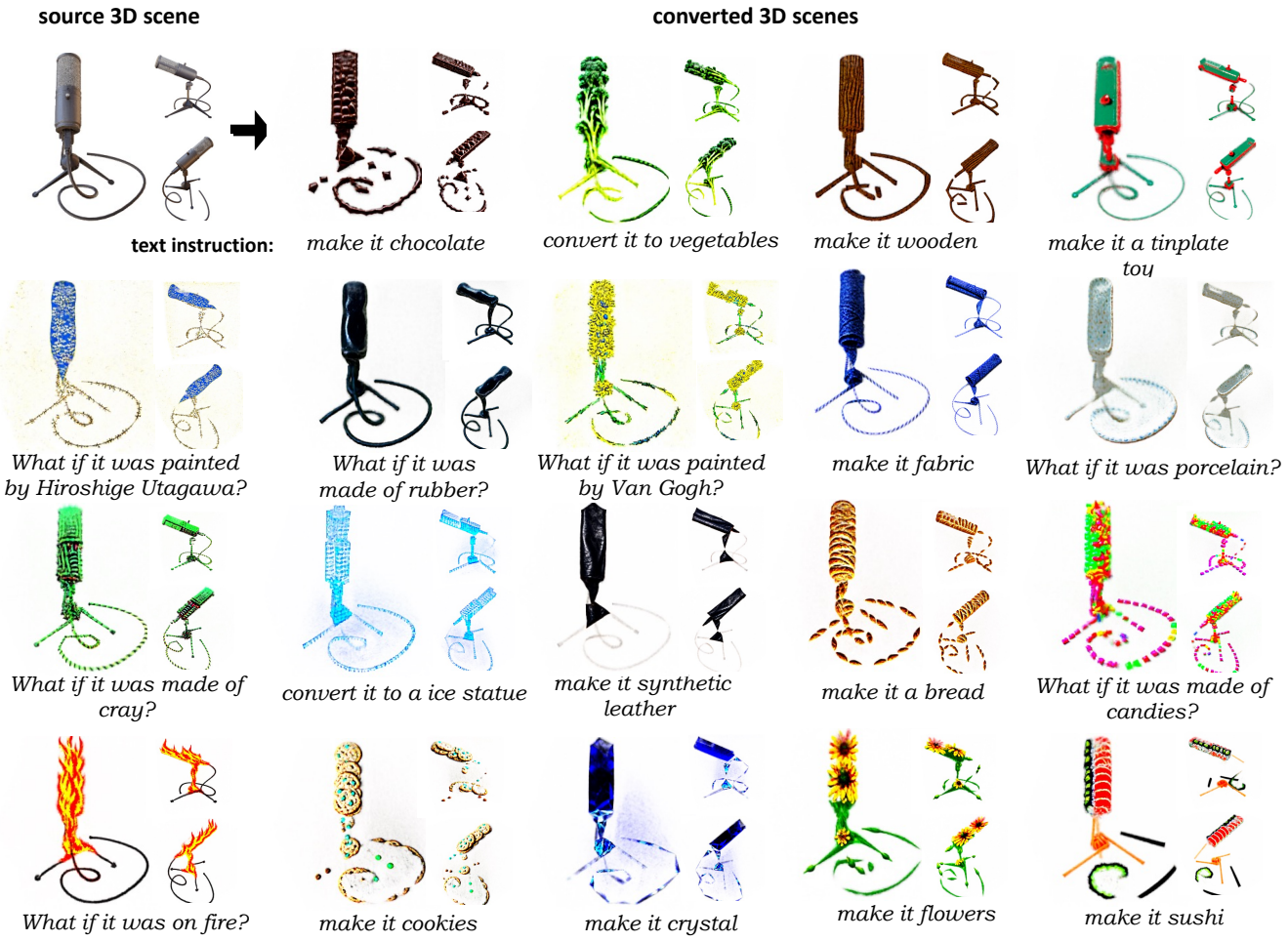

Figure A9: Converted 3D scenes from the 3D scene of mic.

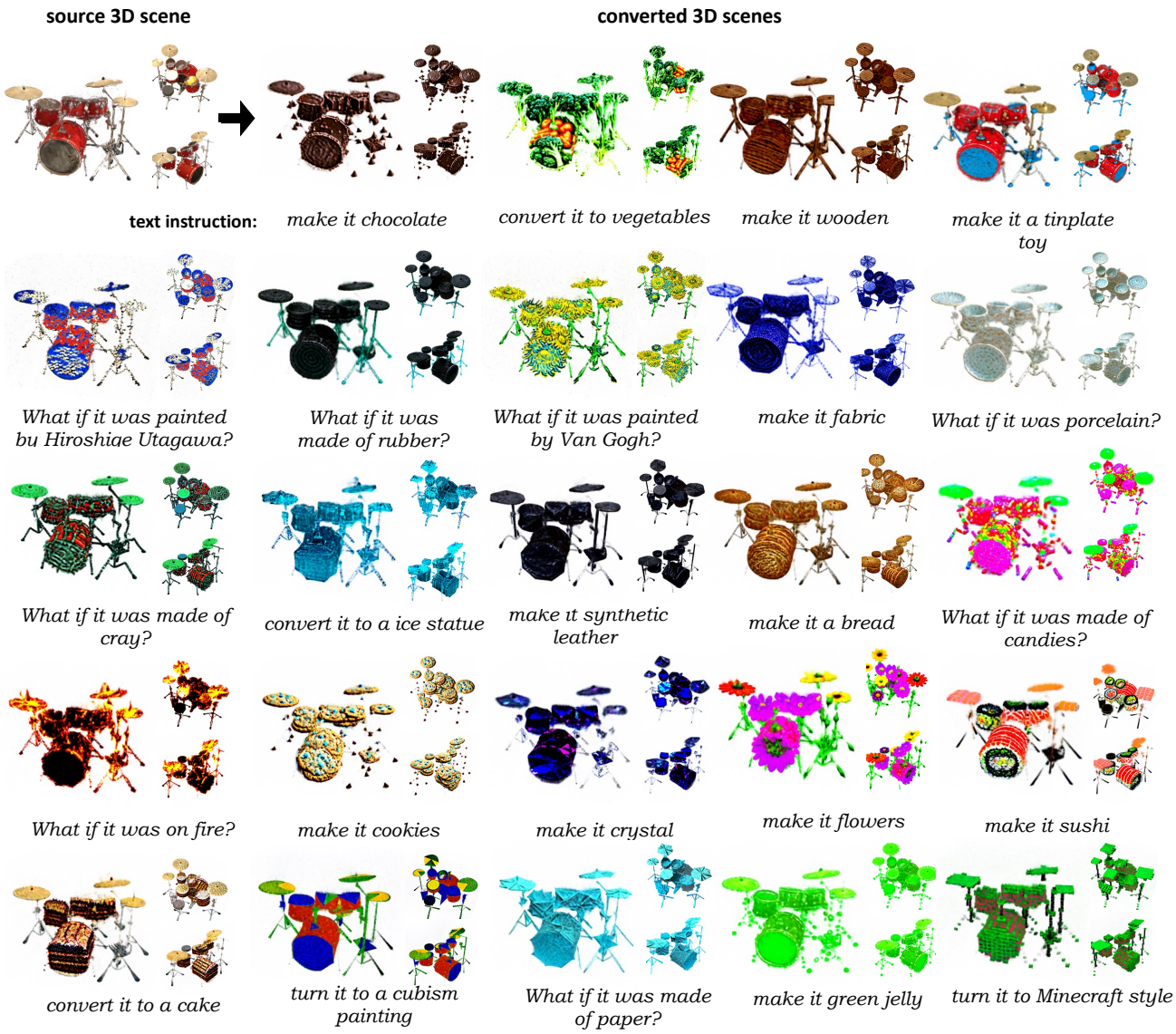

Figure A10: Converted 3D scenes from the 3D scene of drums.

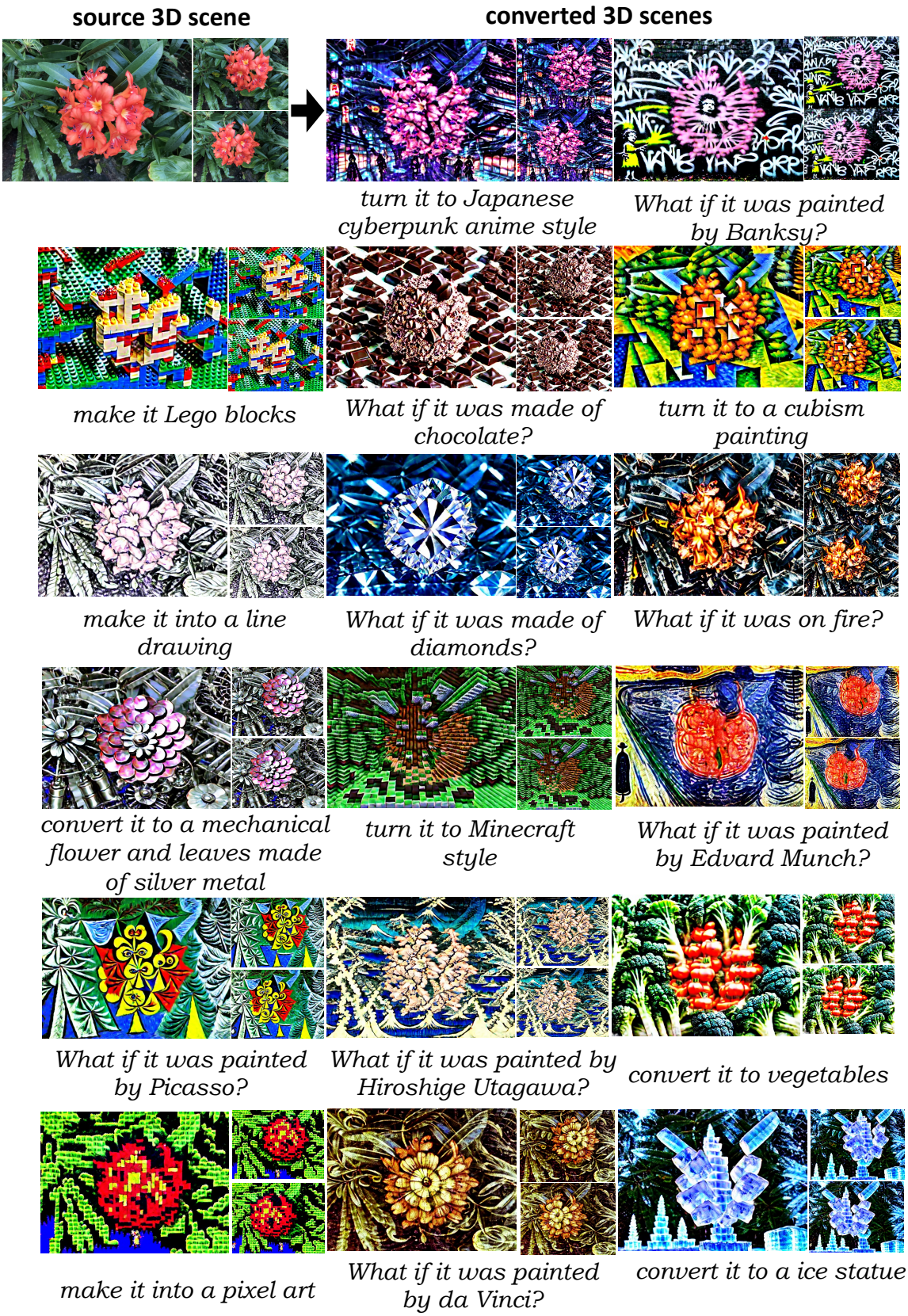

Figure A11: Converted 3D scenes from the 3D scene of flower.

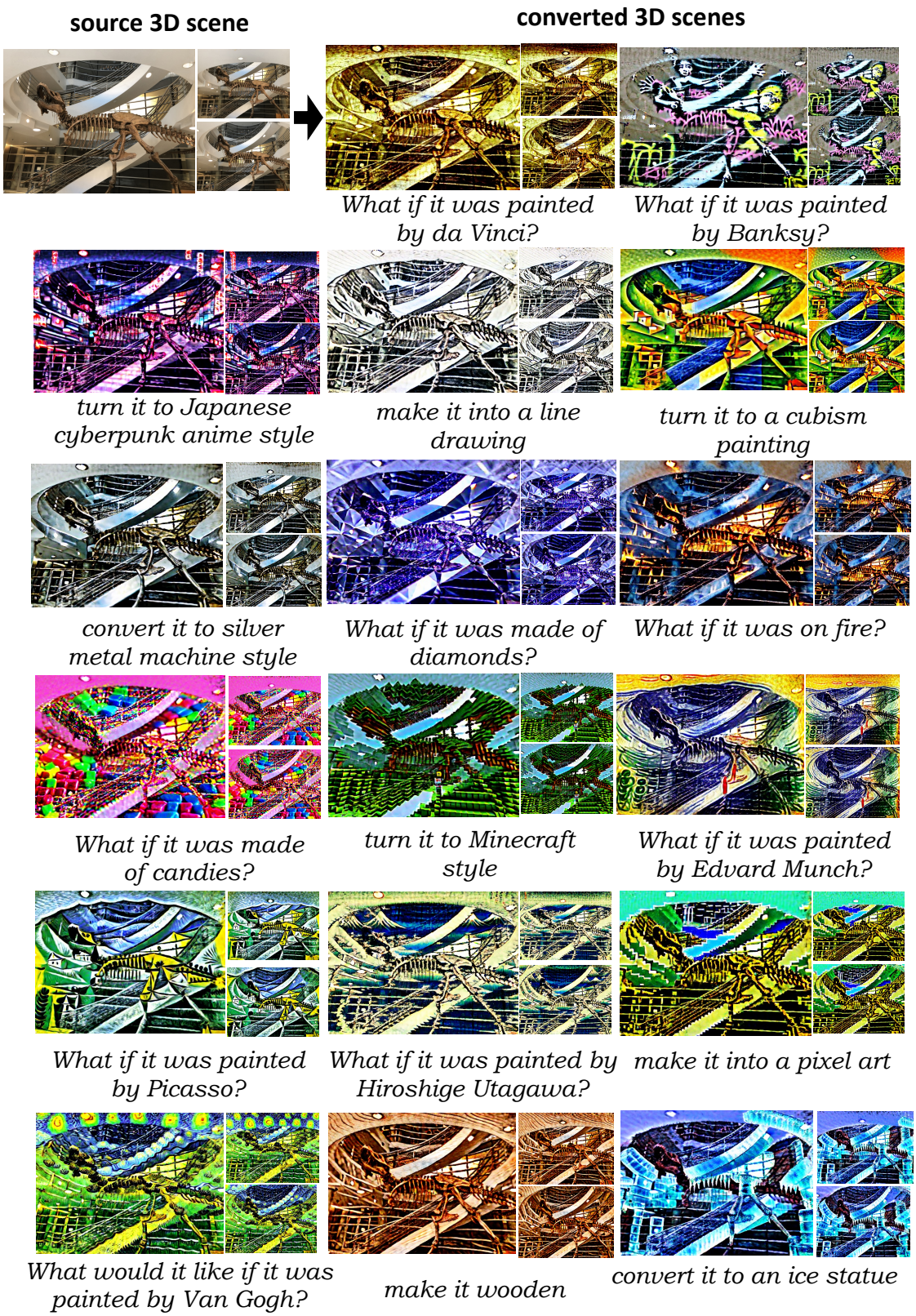

Figure A12: Converted 3D scenes from the 3D scene of trex.

Figure A13: Converted 3D scenes from the 3D scene of fern.

Figure A14: Converted 3D scenes from the 3D scene of horns.