Supplementary Material for Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis

Nikhil Singh
MIT
Media Lab
nsingh1@mit.edu

Jeff Mentch
Harvard University
SHBT
jsmentch@g.harvard.edu

Jerry Ng
MIT
Mechanical Engineering
jerryng@mit.edu

Matthew Beveridge
MIT
EECS
mattbev@mit.edu

Iddo Drori
MIT
EECS
idrori@mit.edu

As supplementary material, we present and review a number of input/output examples across several categories with distinct properties[1]. A summary of these results is shown in Table 1. We additionally present a more detailed diagram of our architecture, shown in Fig. 11.

Finally, to gain a qualitative view of intra-scene and adjacent-scene consistency, we plot our test set input images according to the corresponding output audio characteristics by a visualization shown in Figure 12. We produce multiband T_{60} estimations from all output IRs, and then used t-SNE [3] to reduce the data dimensionality to two dimensions. We then solve a linear assignment problem to transform this into a grid representation. Several instances of within-scene clusters are visible, as well as closeness of related scenes. This suggests that while our method does make errors (outliers are also visible), it learns to treat similar scenes similarly while capturing variation.

Topic	Figure #	Images
Famous and iconic places	1	6
Musical environments	2	6
Artistic renderings	3	6
DALL•E-generated spaces	4	6
Limitations (i.e. challenging examples)	5	4
Animated scenes	6	6
Virtual backgrounds	7	6
Historical places	8	5
Video games	9	4
Common and identifiable scenes	10	6
Total		55

Table 1. Additional Results.

[1]Link to audiovisual examples page: https://web.media.mit.edu/ nsingh1/image2reverb/

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Scott Gray. Dall·e: Creating images from text. *OpenAI Blog*, 2021.

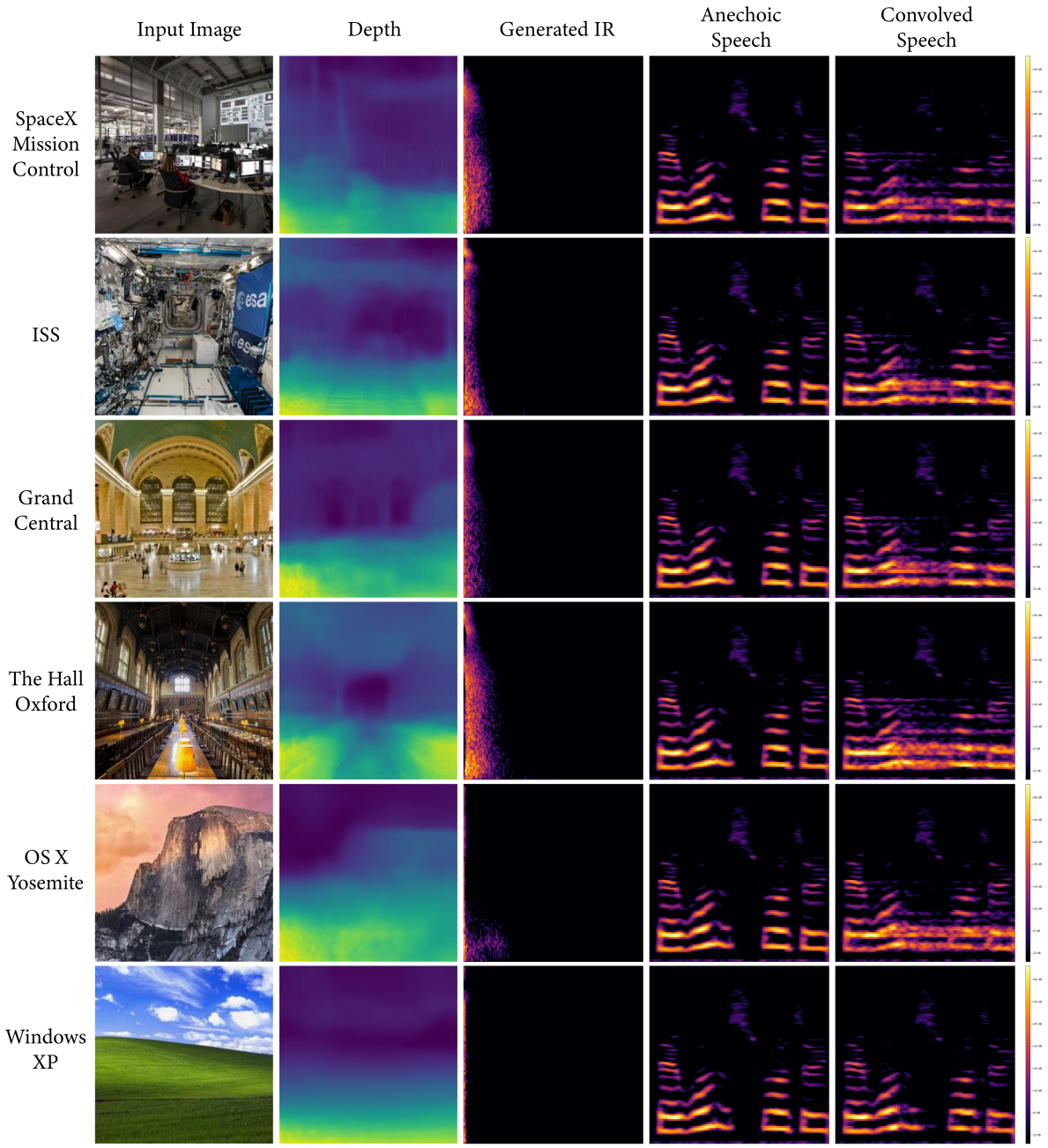

Figure 1. Famous and iconic spaces. Columns show input images, depth maps, generated IRs, and a dry anechoic speech signal before and after the generated IR was applied to the signal via convolution. The input images come from spaces that may be impractical or impossible to record in. The indoor spaces here show longer impulse responses compared to the outdoor scenes which is typically observed and expected in real-world settings. Larger indoor spaces also tend to exhibit greater T_{60} times with longer impulse responses which we see here, though the ISS image has a longer impulse response than we expect.

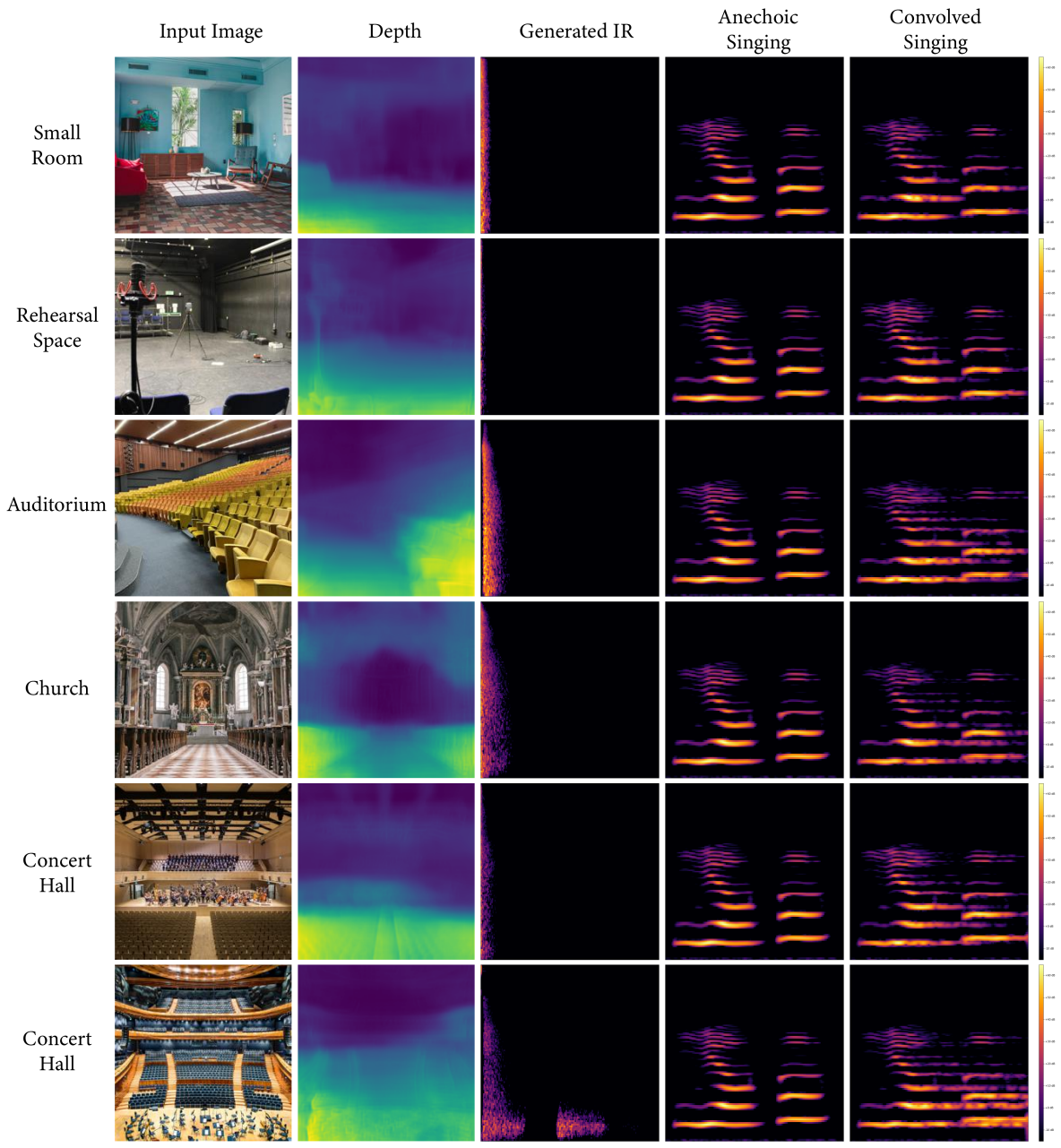

Figure 2. Music. Columns show input images, depth maps, generated IRs, and an anechoic vocal singing signal before and after the generated IR was applied to the signal via convolution. The input images come from spaces relevant to music including a typical small room, an acoustically treated rehearsal space, an auditorium, a church, and 2 large concert halls. Generally, larger spaces tend to exhibit longer decay times in the output, however some examples such as the concert halls with visible acoustic treatment appear to have a shorter decay than more reverberant spaces like the church or auditorium with more reflective surfaces. The final concert hall shows an atypical impulse response with a visible discontinuity in the IR tail. This is not commonly observed among our model outputs, but illustrates the nature of artifacts which can occasionally occur.

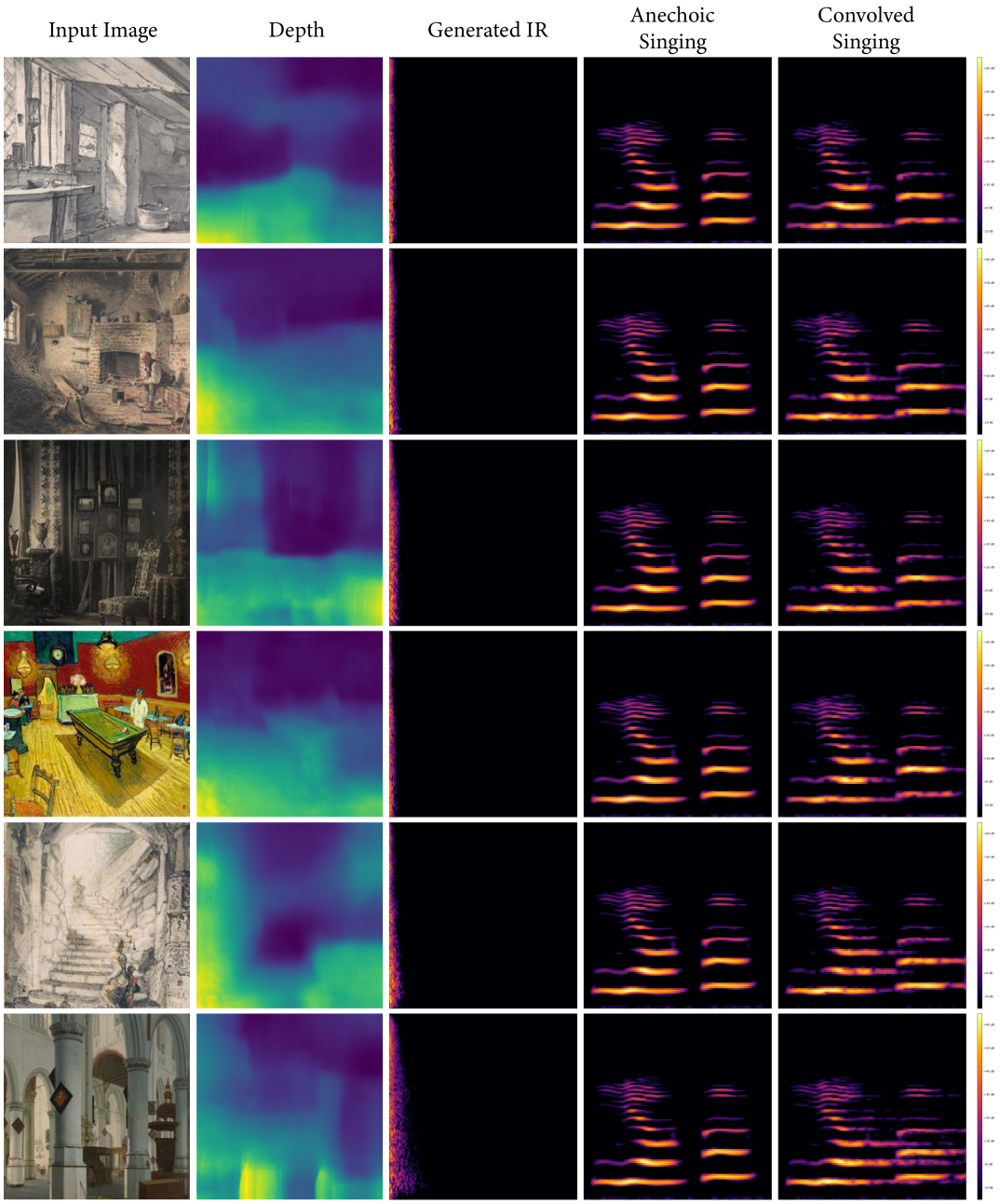

Figure 3. Art. Columns show input images, depth maps, generated IRs, and an anechoic operatic singing signal before and after the generated IR was applied to the signal via convolution. Images here are drawings, paintings and a vintage art photograph ca. 1850. Artistic depictions of spaces were not included in our training dataset. In many cases, plausible impulse responses are generated from such input images. In general, larger depicted spaces, like the church in the bottom row, exhibit longer decay times as is observed with standard 2D photographs.

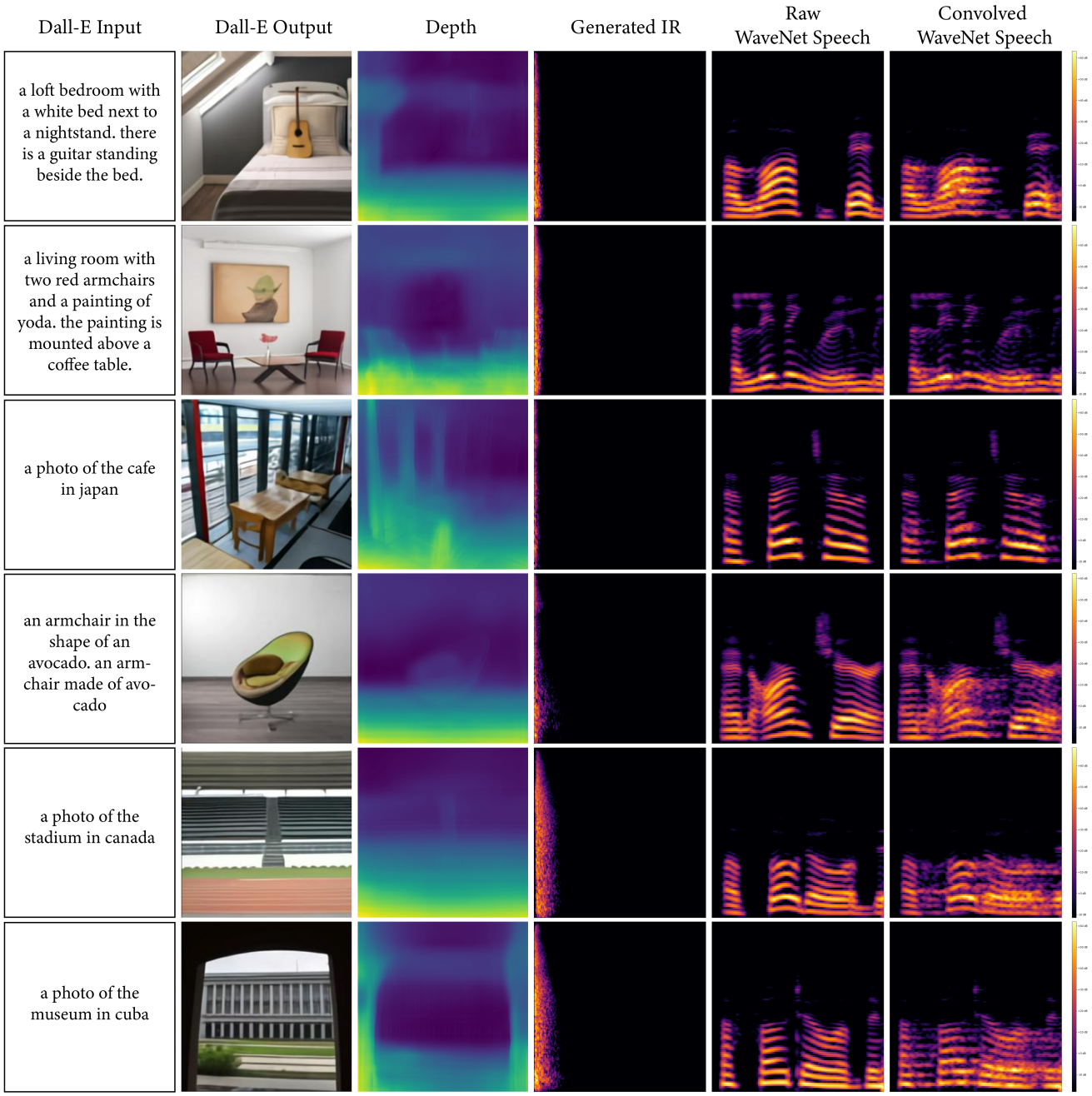

Figure 4. DALL·E. Images generated from text by DALL·E [4] used here as input images. The same corresponding input text was synthesized via text-to-speech as our signal of interest and convolved with the generated IR. This reflects synthetic speech in a synthetic environment, indicating a path for synthesizing realistic IRs from text. It also shows how our model might work with other state-of-the-art generative media models to produce more consistent and realistic results in different domains.

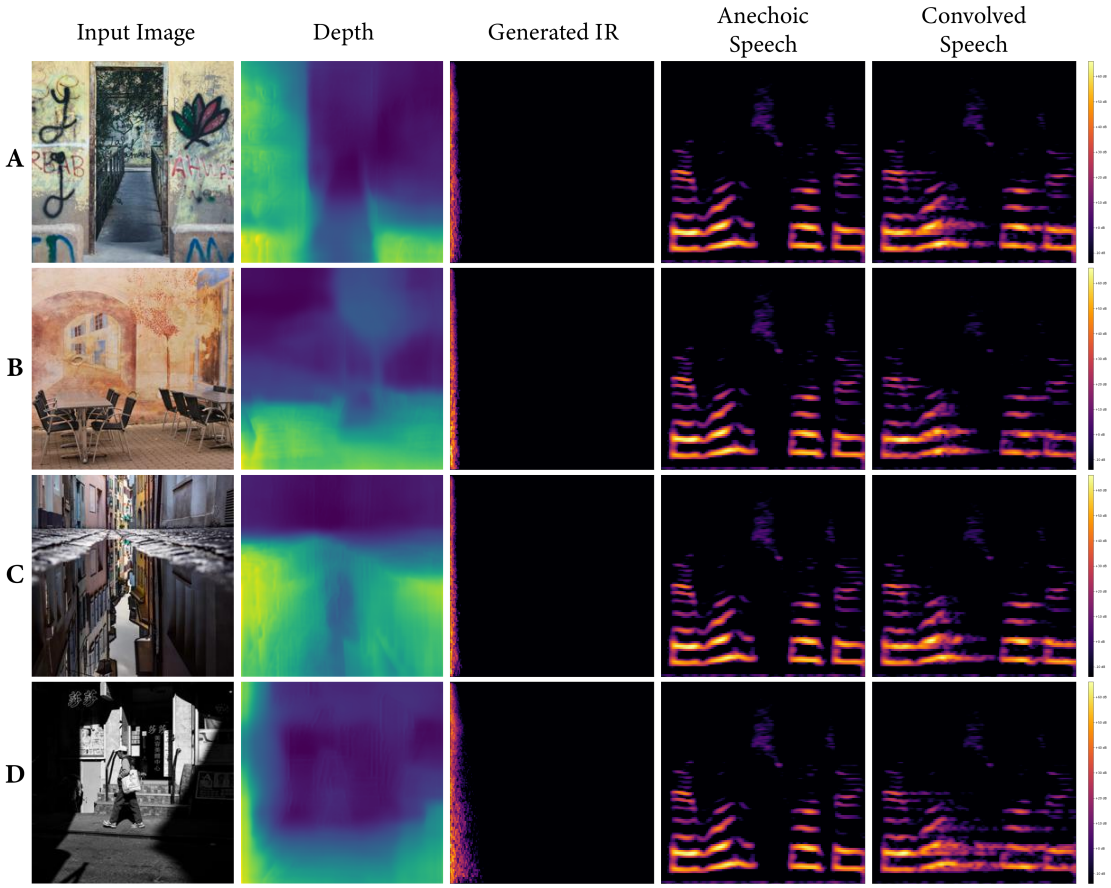

Figure 5. Challenging images. Input images containing street murals, reflections, and shadows demonstrating cases where depth is inaccurately estimated. (A) A painted doorway giving the illusion of depth. (B) A wall with a mural of a street and tree where the depth of the wall is inaccurately estimated. (C) A low-angle photo of a reflective puddle. (D) An outdoor street image with strong shadows which results in a depth map and generated IR more similar to a room than an outdoor space. These more extreme scenarios are chosen to clearly illustrate the limitations of our approach.

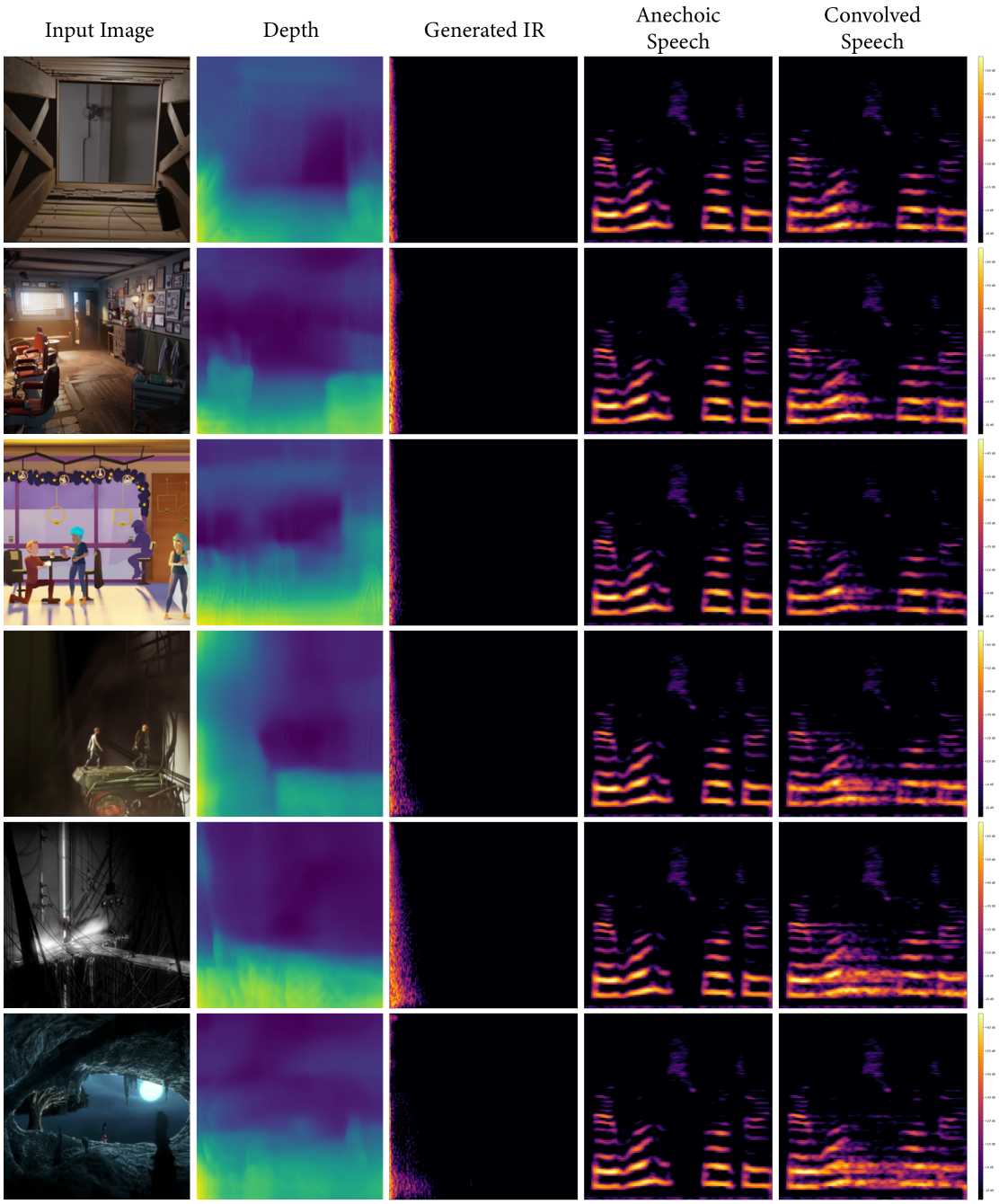

Figure 6. Animated films. Scenes from Blender open animation films used as input images (speech convolved with generated IRs). Columns show input image, calculated depth map, spectrogram of generated IR, an anechoic passage reading sample, and the same passage with the generated IR applied via convolution. In general, we find that our model plausibly estimates the reverberant characteristics of these spaces. For example, the wooden small space is very brief. The barbershop appears longer due to some artefacts, but the broadband decay is relatively quick as can be heard in the audio. Seemingly larger spaces again correspond to longer IRs. This is a case of Real2Sim transfer, where we can approximate IRs directly that sound as measured IRs, but in virtual environments where this measurement is not possible.

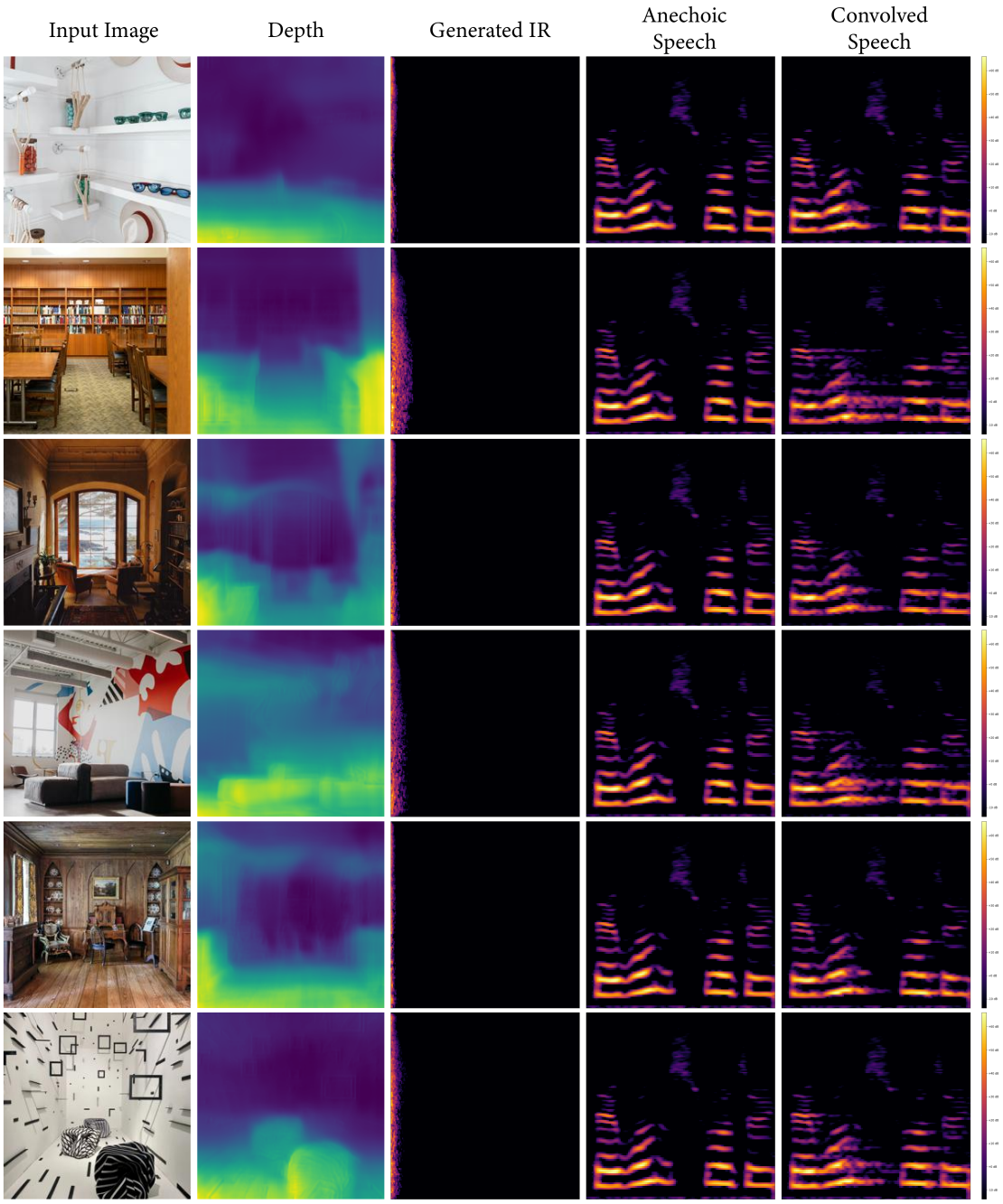

Figure 7. Virtual backgrounds. Images which may serve as virtual backgrounds used as input images to our model. These reflect spaces that may be used for videoconferencing or other online meetings. Realistic IRs may be generated and used in these contexts to increase the sense of being in a shared space with others.

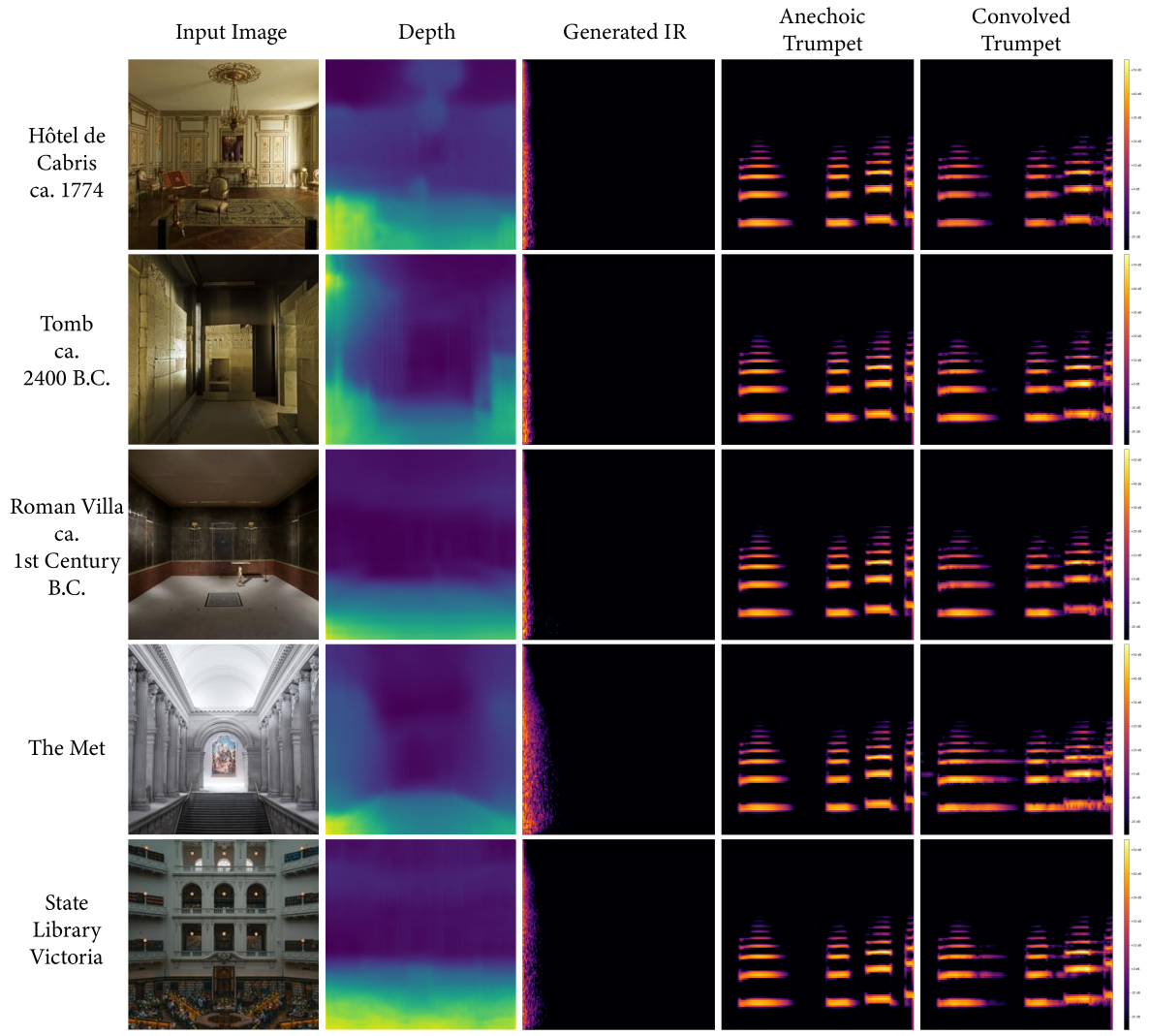

Figure 8. Historical and notable places. Additional examples of unusual and historical spaces which may be difficult or impossible to obtain IRs from.

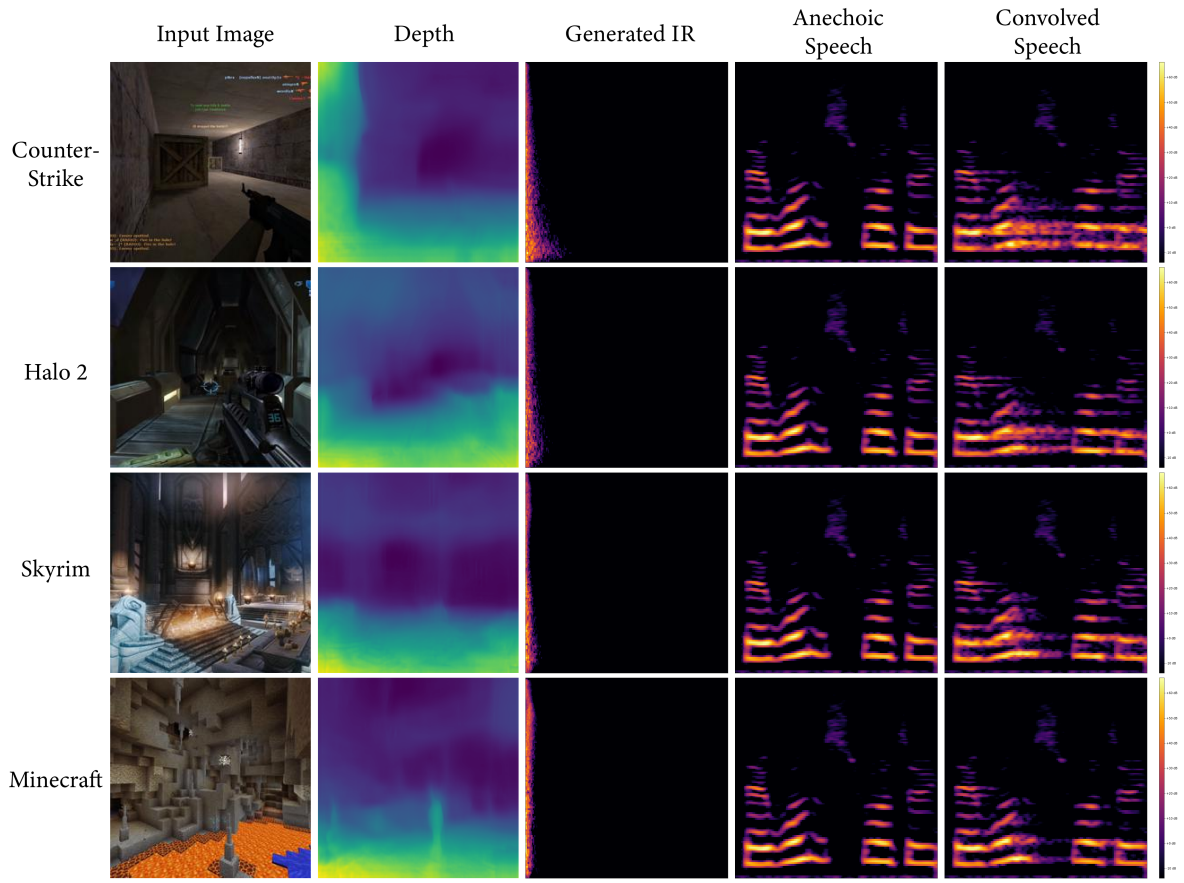

Figure 9. Video games. Impulse responses generated and applied via convolution from screenshots of four 3D video games. Video games are one example of a virtual space that might benefit from easily generated impulse responses. While the medium sized room from Counter-Strike and the large hallway from Halo 2 may be plausible IRs, the large hall shown in the Skyrim screenshot and the cavern in the Minecraft example do not have correspondingly long reverberant tails as would be expected showing possible examples of where the scale of the space was not accurately estimated. 3D rendered images were not included in our dataset but are a ripe area of future work which might greatly increase the performance of our model on both real scenes and virtual scenes such as these video game examples.

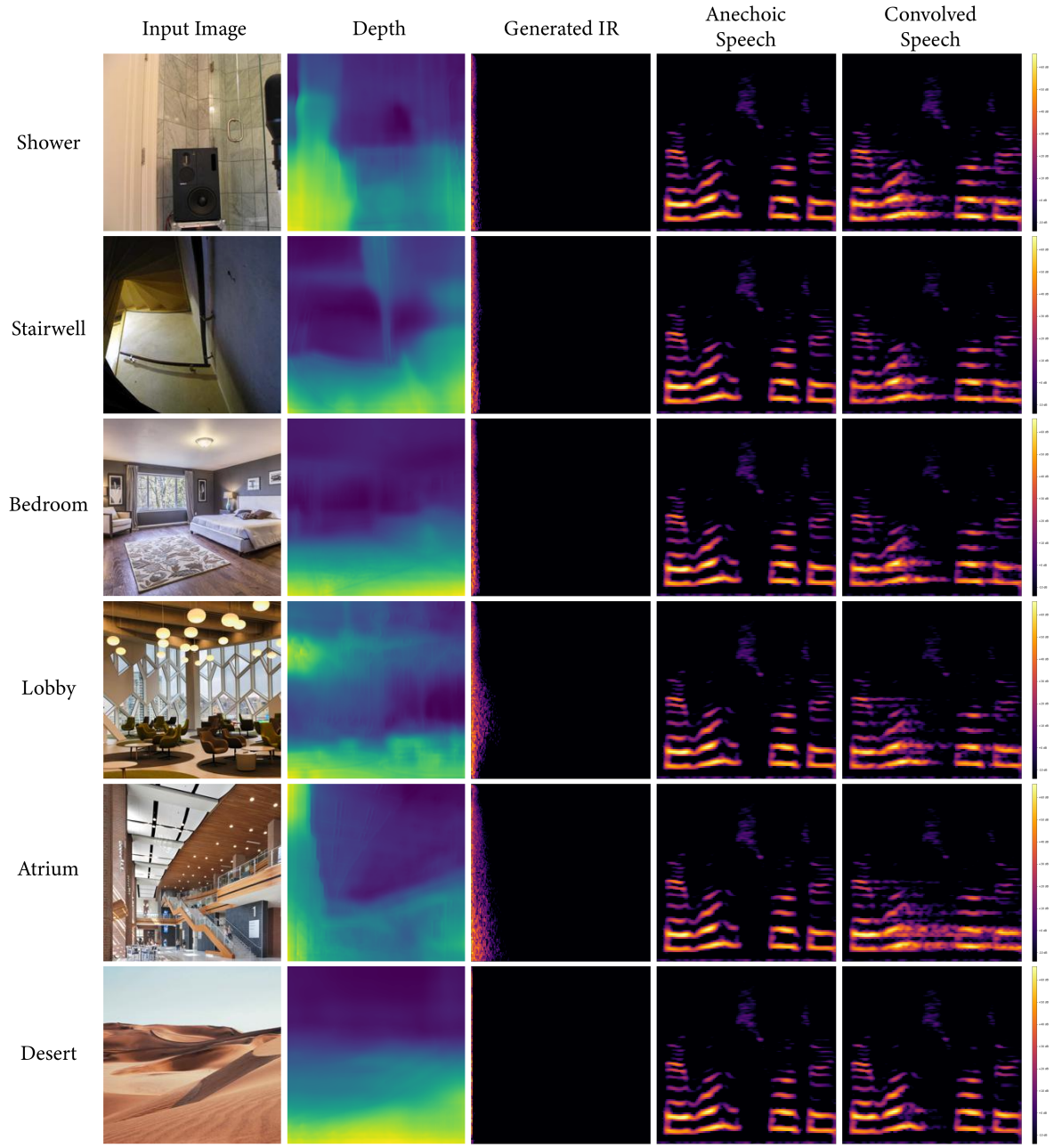

Figure 10. Common and identifiable scenes. Input images and the resulting IRs are shown and convolved with an anechoic speech signal. Input images here reflect spaces that are regularly encountered in everyday life yet may not often be recorded in. These types of scenes are useful for audio post-production as they may be commonly found in movies and television shows. Small and outdoor scenes are observed to have very brief IRs while in comparison, the larger building interior has a much longer output IR as expected.

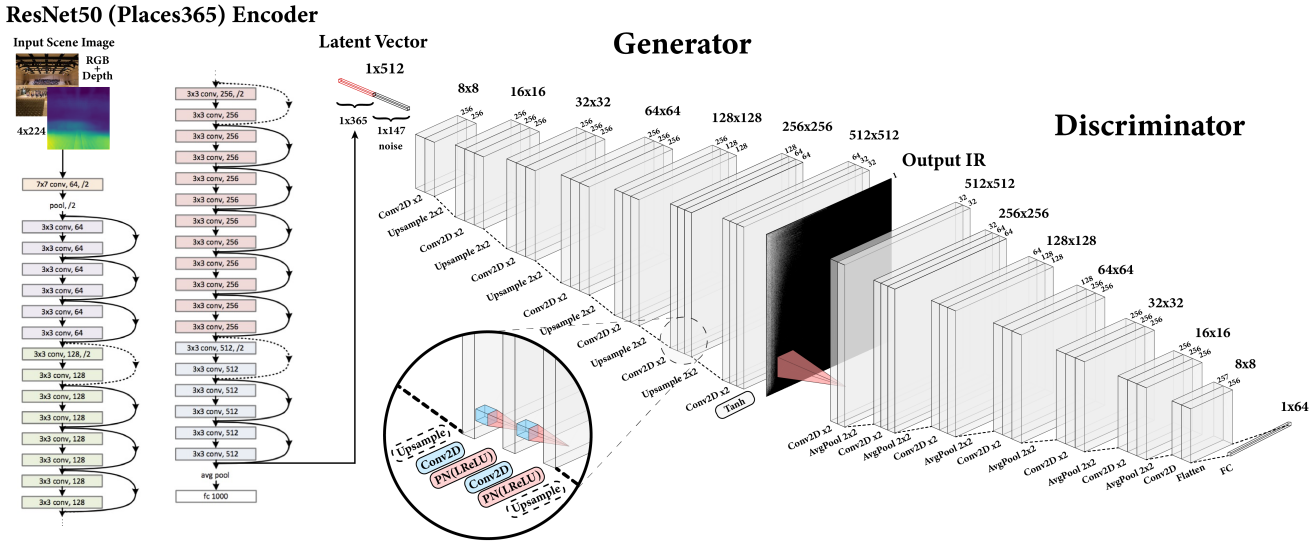

Figure 11. Detailed overview of Image2Reverb model architecture. Left: the ResNet50 encoder pre-trained on Places365 (figure at left adapted from [1]). Right: the generator and discriminator. The output of the encoder consists of 365 features, to which we concatenate noise to produce a 512d latent vector. The generator and discriminator contain upsampling and downsampling convolutions respectively. A leaky rectified linear unit (LReLU), with $\alpha = 0.2$, is used after each convolutional layer in the model in both the discriminator and the generator with the final layer of the generator using a $\tanh$ activation. PN denotes pixelwise normalization, which we use in the generator. The composition of blocks is based on ProGAN [2]. The final step in the discriminator is a fully connected layer with a linear activation (scalar output).

Figure 12. Manifold-based visualization of our test set. We compute multi-band T_{60} estimates for output audio IRs for each image, and then perform nonlinear dimensionality reduction with t-SNE to obtain two-dimensional feature vectors for each example. We produce a grid by solving a linear assignment problem, as is commonly done to visualize large image datasets. Our visualization shows local clusters of same and similar scenes in many cases, but also some variation within scenes. In some outdoor settings, this variation grows considerably large, resulting in increased scattering. In other cases, we observe closeness between different views of the same scene and similar scenes.