Replay:
Multi-modal Multi-view Acted Videos for Casual Holography
Supplementary material

Roman Shapovalov* Yanir Kleiman* Ignacio Rocco* David Novotny
Andrea Vedaldi Changan Chen† Filippos Kokkinos Ben Graham Natalia Neverova
Meta UT Austin† *equal contribution
{romansh, yanirk, irocco, dnovotny, vedaldi, fkokkinos, benjamingraham, nneverova}@meta.com

A. Scene Variety

The Replay dataset contains a large variety of scenes in terms of actions, number of participants, environments, and props. The scenes are acted out by a diverse cast of actors of different age, gender, and ethnicity. In the overview video and Fig. 1, we show the representative sample of the different scenes one can find in the dataset. In Fig. 2, we show an example of a moment in time that was captured by the twelve visual sensors that we provide for each scene. All sensors are temporally synchronized: we provide frame timestamps in the time frame that is shared across sensors of each scene.

B. Overview video and dynamic results

We attach a video demonstrating:

- the diversity of the collected scenes (cf. Figure 1),
- results of the color calibration (cf. Figure 3),
- visualisation of the estimated camera poses and static scene points,
- the diversity of viewpoints (cf. Figure 2) and foreground segmentation available for benchmark scenes,
- the results of the novel-view synthesis when moving in space or time.

We show the novel-view synthesis rendered in three settings, specified in the slide titles. Below is the description of those.

Fly-around – Fixed Timestamp. In this setting, we train a model on a fly-around stage where the camera used for training goes around the scene. To render these videos, we generated camera poses by fitting a circle to camera centres from the training data. The camera wearer went along an elliptical trajectory during the fly-around, so approximating it with a circle forces the method to extrapolate the views in some parts of the trajectory. For the models taking time as input, we fix the timestamp to the middle of the sequence.

Acting – Fixed Timestamp. In this setting, we train the model using captures from the frontal cameras during the acting stage that contain substantial actors' motions. Here we again fit a circle to the centres of cameras used for training and fix the timestamp to the middle of the stage, hence visualisations are static. This setup shows how well the models are able to decouple viewpoints from motions in time.

Acting – Fixed Novel Viewpoint. We additionally render the acting stage in fixed-viewpoint mode. We fix the camera pose to the one used for evaluation (*i.e.*, static DSLR-1, held out from training) and use the sub-sampled range of training timestamps. This setting shows how well the methods can model the geometry and appearence changing in time. In particular, time-agnostic methods (NeRF and TensoRF) are bound to produce static renders in this mode. Please note that this is a viewpoint extrapolation scenario, since all training cameras were located at a significant distance from DLSR-1.

Figure 1. **Scene diversity.** Representative frames from **28 different scenes** in the Replay dataset. Scenes are individually synchronised and calibrated.

Figure 2. **Viewpoint diversity.** The **same moment in time** as viewed by the twelve different sensors we provide for each scene. Top two rows: DSLR cameras. Bottom row: Head mounted GoPro cameras, and a 360°ceiling camera.

Figure 3. **Color calibration.** For 8 different scenes (rows), we show a pair of frames from a DSLR and a GoPro sensor before color equalisation (left) and after equalisation (right). Please note the discrepancy in colors on the left due to hardware differences. We equalise images by transforming them into a common sRGB color space and matching a color-checker target. Equalised colors enable combining sensors for training new-view synthesis methods and using a sensor of different type for evaluation.