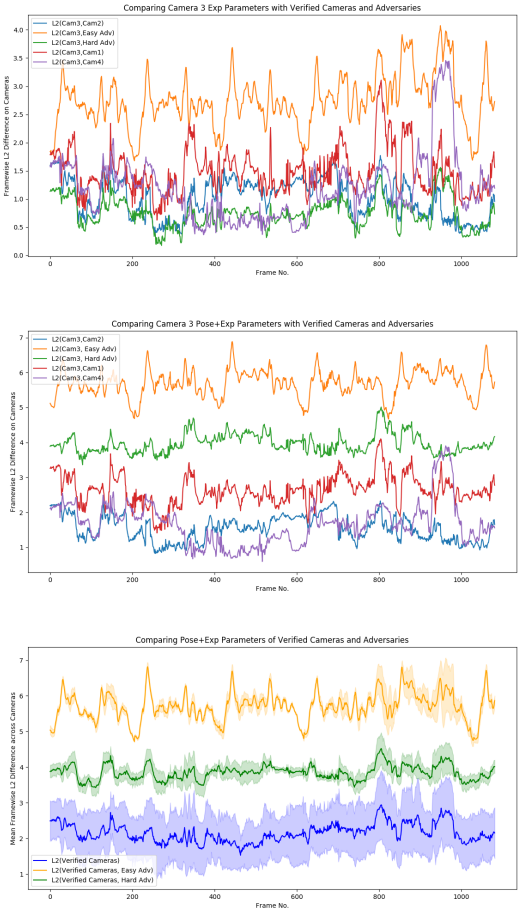

Figure 8: *Top:* We compare expression parameters of camera three against all other inputs. While the easy adversarial case (orange) is separable from most of the real data, the hard adversarial case (green) is not. *Middle:* We compare the full parameter set of camera three against all other inputs. Both easy and hard fake cases are separable from the real video. *Bottom:* We look at the mean frame-wise l_2 difference across the real cameras only (blue), the real cameras and the easy fake case (yellow), and the real cameras and the hard fake case (green). We plot one standard deviation from the mean in all cases. All fake cases are separable from the real cases.

A. Appendix

A.1. Synthetic Data Collection

We collect synchronized video of a seated speaker from four different views. The speaker is prompted to make a wide variety of facial expressions, and to turn their head dramatically in the space. FLAME is then fit to each frame using RingNet. For each frame, we save the shape, expression, and jaw pose parameters of the FLAME model.

Next, we create synthetic adversarial examples in the FLAME model space using VOCA. Given a neutral FLAME model, and an audio sample, VOCA animates the model to synchronously mouth the audio. We create two fake videos: In the first, we keep the expression neutralized, and maintain the shape, or identity-related, parameters. In the second, we match the average expression across the real videos in addition to matching the shape parameters (Figure 9).

Table 2: To detect a fake, the fake must be outside of the range of normal fitting error among the real cameras. Shape is most consistent across cameras, and jaw pose, which controls how open the mouth is, has the most variation.

Parameters	Mean frame-wise standard deviation across real cameras
Shape	± 0.1037
Expression	± 0.2025
Jaw pose	± 1.1690

A.2. Synthetic Data Experiments

We aim to confirm that in an idealized scenario, where we are perfectly able to capture face geometry from an image to create a detailed model, that we are able to detect small changes in mouth pose. First, we need to determine whether the FLAME models fit across different views look the same. Qualitatively, we see in Figure 9 that the shape, expression, and jaw pose remain invariant across views.

In Table 2, we see that a fake video can most easily vary in jaw pose while remaining undetected. The fit of RingNet is much more sensitive to the shape and expression parameter space across views. To determine if our fakes are detectable, given this sensitivity to the model fitting, we employ a one versus all approach. For each frame, we look at the l_2 difference between the parameter vectors for every possible pair of cameras. We use both the expression and jaw pose parameters alone, and the full set of shape, expression, and jaw pose parameters.

In Figure 8, we see that if we use expression parameters, we are only able to isolate the easy fake case from the rest of the pack. However, if we look at the full set of parameters, we are able to isolate both the easy and hard fake cases from our real video, except for a small section at frame 900. When we plot the mean l_2 difference between the set of real cameras, and sets including different fakes, we see that we can differentiate between the real and fake cases.

We find that the parameter space is fairly invariant across our real views. We can also detect fake videos that only differ from the real videos in mouth pose and expression. These results demonstrate a proof of concept for using geometric cues with social verification to detect video manipulation.

Figure 9: *Top:* Four input frames from four different views, synchronized in time. *Middle:* We show the RingNet fitting output, which stays fairly consistent across views, though it is not able to capture the furrowed brow, or much interesting face shape. *Bottom:* On the left, the easy faked case has a neutral expression, and a different viseme from the original input. On the right, the hard faked case has the averaged expression of the input models, and the new viseme from the easy case.