Supplementary Materials for "Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion"

	Dubbing	Wav2Lip	Impersonator	FaceSwap	itw
LipForensic	0.50	0.85	0.34	0.67	0.79
PWL	0.50	0.58	**0.86**	**0.79**	0.53
Ours	**0.89**	**0.91**	0.84	**0.79**	**0.94**

Table 1. The performance in terms of AUCs on 10-second video clips after compression of real test videos.

1. Overview

We first evaluate the robustness of our system against unseen video perturbations. Next, we provide some quantitative and qualitative results to support the analysis made in the main paper. In the main paper we compared with the related works using the average AUCs across all individuals. In order to give a better insight into the comparison, we first present per-individual results for each of the related works. We then present a more detailed version of qualitative results using both training and testing datasets.

1.1. Robustness Test

For this experiment, we re-saved the real test videos of each individual using the ffmpeg compression quality of 40. Shown in Table 1 are the results when compared to the best performing related techniques. In our case the average performance is reduced by 5% from 0.92 to 0.87, whereas in case of LipForensics and PWL the reduction is 14% and 6%. Even though our performance is reduced, our approach still performs better than previous techniques.

1.2. Comparison with State-Of-The-Art

Shown in Table 2 are the per-individual results for all the related methods that were presented in the main paper.

1.3. Videos for Qualitative Analysis

Here we provide the videos used for qualitative analysis of the words presented in the Figure 1 and Figure 6 of the main paper. For Obama, Trump, and Oliver we provide occurrences of the word "hi", "tremendous", and "billion" in the real and fake videos. Therefore, there are a total of six videos for this section:

- Obama_hi_real.mp4,
- Obama_hi_fake.mp4,
- Trump_tremendous_real.mp4,
- Trump_tremendous_fake.mp4,
- Oliver_billion_real.mp4, and
- Oliver_billion_fake.mp4.

In each video, the output probability of the word-specific classifier is shown in red on the top left corner (a value of 1

XceptionNet					
	Audio Dubbing	Wav2Lip	Impersonator	FaceSwap	in-the-wild
Obama	0.50	0.94	0.74	0.96	0.47
Trump	0.50	0.84	0.70	0.82	0.54
Biden	0.50	0.49	0.69	0.67	0.45
Harris	0.50	0.80	0.48	0.24	-
O' Brien	0.50	0.69	0.44	0.11	-
Oliver	0.50	0.93	0.26	0.15	-
PWL					
	Audio Dubbing	Wav2Lip	Impersonator	FaceSwap	in-the-wild
Obama	0.5	0.56	0.96	0.96	0.83
Trump	0.5	0.51	0.95	0.94	0.41
Biden	0.5	0.53	0.65	0.66	0.55
Harris	0.5	0.45	0.94	0.94	-
O' Brien	0.5	0.84	0.69	0.67	-
Oliver	0.5	0.88	0.99	0.93	-
LipForensics					
	Audio Dubbing	Wav2Lip	Impersonator	FaceSwap	in-the-wild
Obama	0.50	1.00	0.83	1.00	0.98
Trump	0.50	1.00	0.68	0.98	0.97
Biden	0.50	0.93	0.15	0.30	0.91
Harris	0.50	1.00	0.08	0.71	-
O' Brien	0.50	0.96	0.48	0.90	-
Oliver	0.50	0.97	0.39	0.98	-
ID-Reveal					
	Audio Dubbing	Wav2Lip	Impersonator	FaceSwap	in-the-wild
Obama	0.50	0.77	0.81	0.71	0.59
Trump	0.50	0.66	0.92	0.88	0.77
Biden	0.50	0.47	0.75	0.59	0.47
Harris	0.50	0.73	0.98	0.98	-
O' Brien	0.50	0.66	0.63	0.56	-
Oliver	0.50	0.69	0.98	0.93	-

Table 2. Accuracy in terms of AUC on 10-second video clips for the six individuals and five different video falsification scenarios. The average AUC across all individuals in given in the last row. From top-bottom are the AUCs for XceptionNet, PWL, LipForensics and ID-Reveal.

is for real and 0 is fake). The occurrences of the words are selected from the training dataset. This is done to demonstrate the facial gestures associated with specific words during training.

In each case, it can be observed that a specific facial gesture is present in real videos which is missing in the fake videos. For example, the occurrences of the word "hi" is associated with an upward head movement which is missing in the fake examples. Similarly, in case of the word "tremendous", notice the presence of lip rounding and chin raise action in multiple occurrences of the word in real videos, whereas these actions are missing in the fake videos.

1.4. Word Analysis for in-the-wild videos

Here we show how the results of our method can be interpreted during the evaluation of a test video. For this we provide four example videos, a real and a fake video of Obama and Trump. The real videos are from test-split of real dataset and fake videos are from in-the-wild dataset. The videos are named as:

- Obama_itw_test.mp4,
- Obama_real_test.mp4,
- Trump_itw_test.mp4, and
- Trump_real_test.mp4.

Given a test video of 10-second length, we show the output of word-specific classifier for each word. Shown on the x-axis of the plot is time and on the y-axis is the probability that the word occurrence is real. Shown in orange is the probability of the word in the test video and shown in the blue is the average real probability of the word in real dataset during training. The region in blue indicates the standard deviation of training probability. The gaps in the plot indicate that the word-specific classifier was missing. The current time is indicated by the red dot on the plot and the current word is displayed on the top of the video.

These word-level probabilities, can be used to isolate the words which obtain low probability of being real. For example, in
Obama_itw_test.mp4 many words have a low probability of being real with a minimum probability of zero for the word "coverage". Similarly in Trump_itw_test.mp4 video, the word "protected" has the zero probability of being real. Whereas in the videos Obama_real_test.mp4 and Trump_real_test.mp4, the real probability for each of the words is close to training real dataset (average of 0.8).

Shown in Figure 1 are the distributions of the 25 facial-gesture features for the word "coverage" for Obama. In each panel, shown in blue is the distribution of one facial-gesture feature in real training videos of Obama. Shown with red line is the value of facial-gesture feature in the current test video of Obama which in this case is the fake video shown in Obama_itw_test.mp4. The word "coverage" in this example fake video have an out-of-distribution value for AU26 i.e. jaw drop. The out-of-distribution value can also be observed for lip-ver motion where the value in the fake is lower than any of the value seen during training.

Similarly, shown in Figure 2 are the distributions of the 25 facial-gesture features for the word "protect" for Trump. The red line in each panel is the value of facial-gesture feature in the fake test video of Trump shown in Trump_itw_test.mp4. For the word "protect" the value for AU17 (chin raise) and AU23 (lip tightner) in the fake is lower than any of the value seen during training.

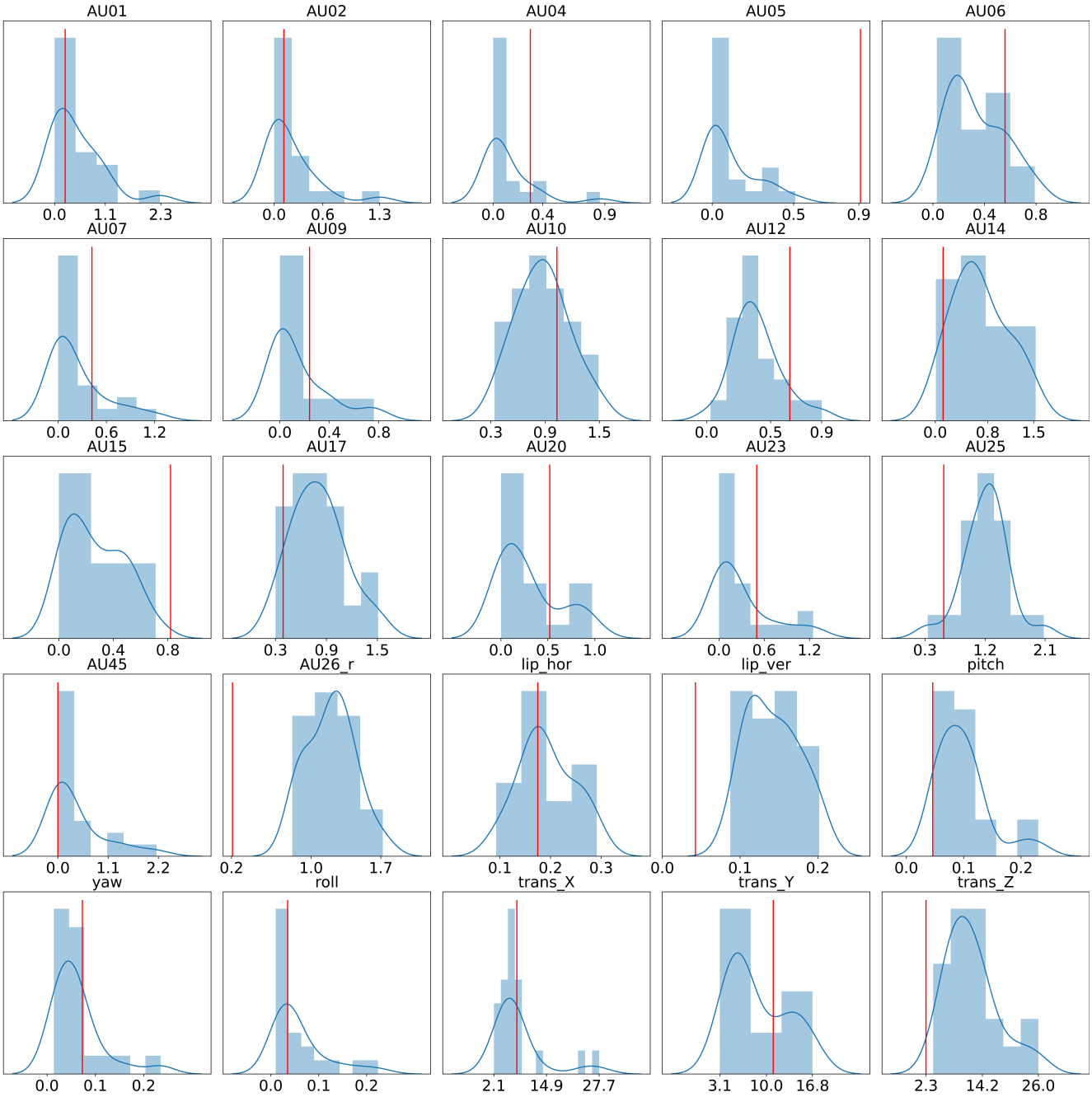

Figure 1. In each panel, shown in blue is the distribution of one facial-gesture feature in real training videos of Obama for the word "coverage". The name of the facial-gesture feature is given on top of the panel. Shown with red line is the value of the facial feature in the current test video of Obama which in this case is the fake video shown in Obama_itw_test.mp4.

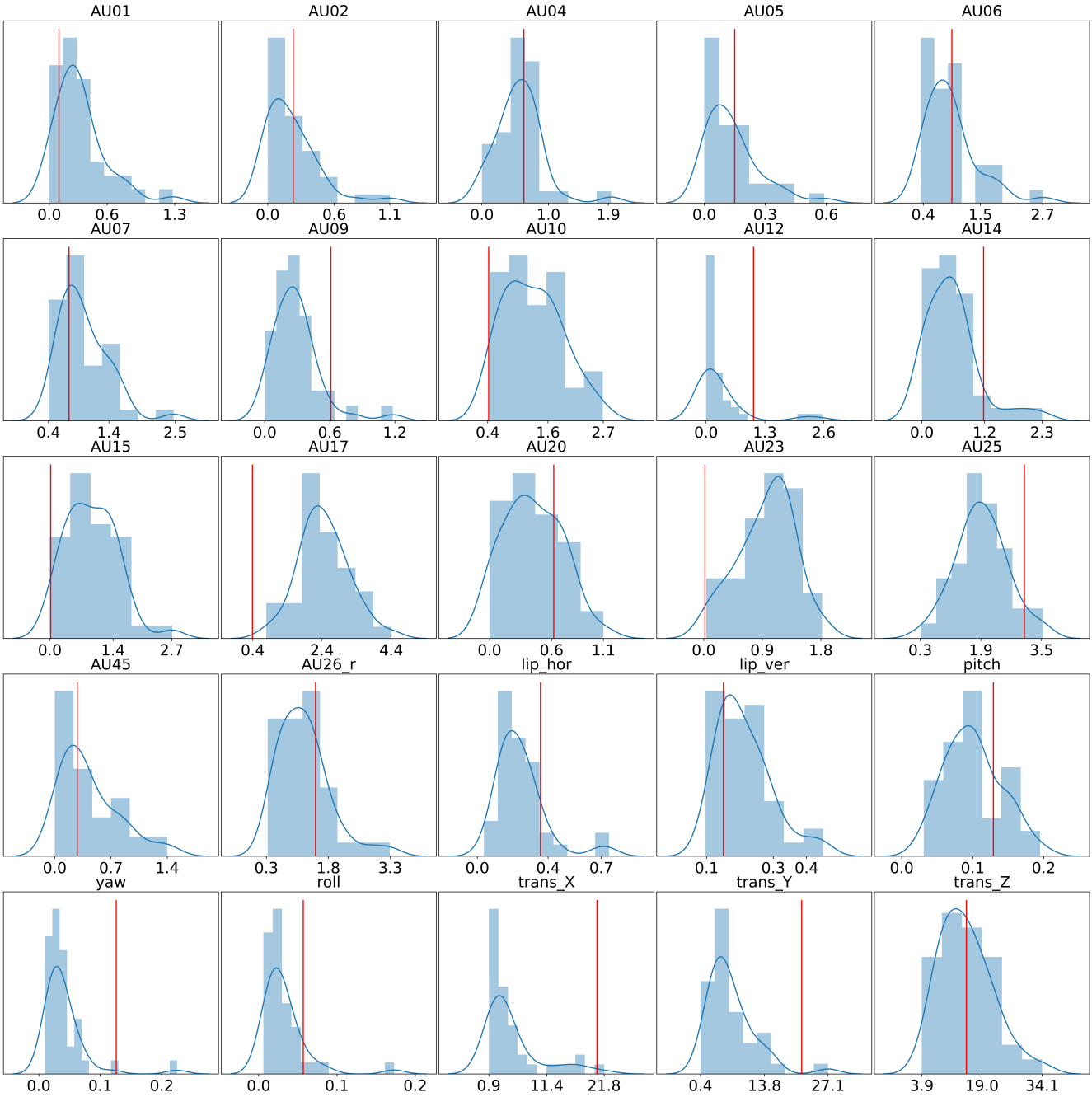

Figure 2. In each panel, shown in blue is the distribution of one facial-gesture feature in real training videos of Trump for the word "protect". The name of the facial-gesture feature is given on top of the panel. Shown with red line is the value of the facial feature in the current test video of Trump which in this case is the fake video shown in Trump_itw_test.mp4.