A. Models architectures

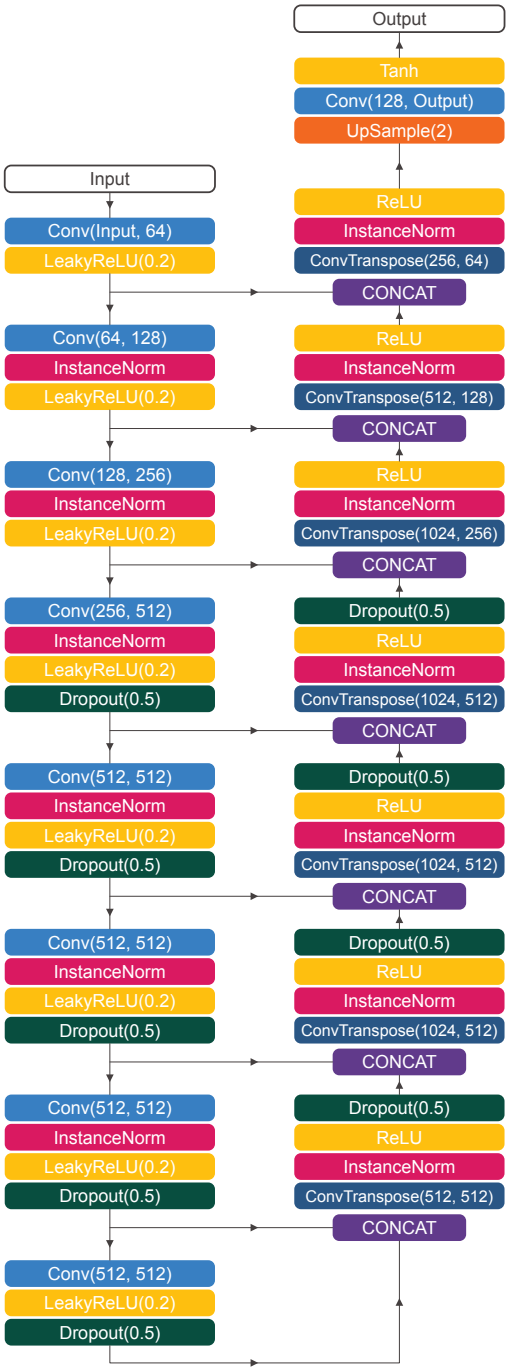

Figure 8: Generator's architecture. It is a UNet in which Conv and ConvTranspose correspond respectively to 2D convolution and 2D transpose convolution with a kernel size of 4, a stride of 2, and a padding of 1. We set the slope coefficient of the leaky ReLU to 0.2 and the probability of dropouts to 50%.

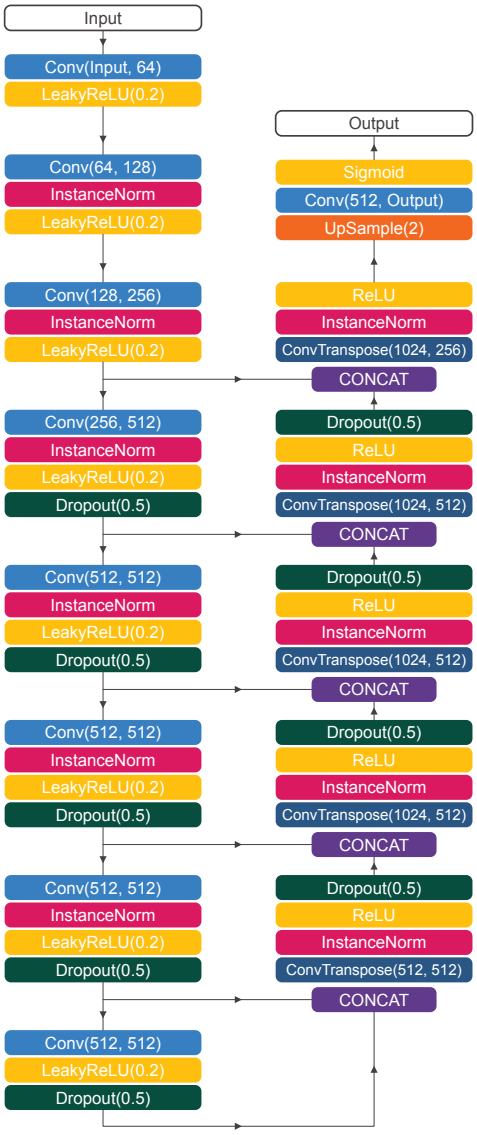

Figure 9: Detector's architecture. It is a UNet in which Conv and ConvTranspose correspond respectively to 2D convolution and 2D transpose convolution with a kernel size of 4, a stride of 2, and a padding of 1. We set the slope coefficient of the leaky ReLU to 0.2 and the probability of dropouts to 50%.

B. Additional results on models trained using MSE

This section provides more results from models trained using MSE as reconstruction loss.

B.1. Watermarks Detection

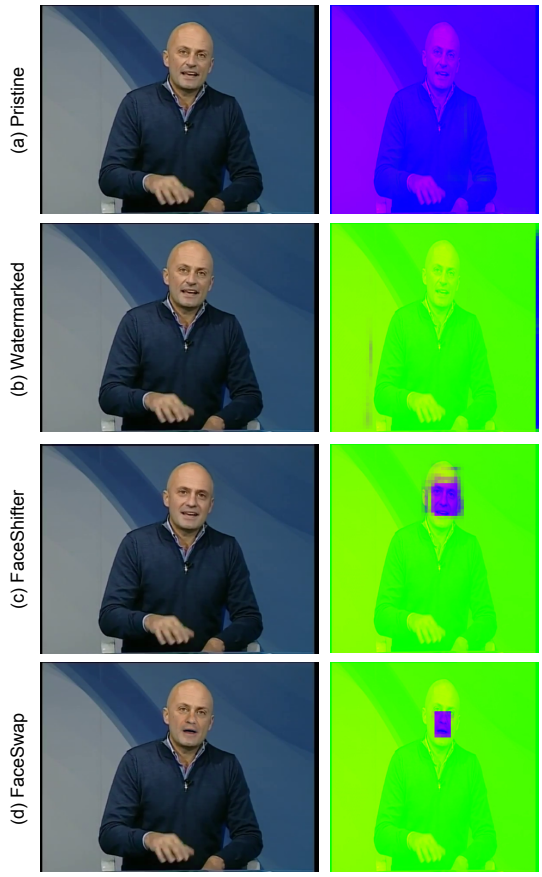

Figure 10: Example of input/output pairs from the detector. The model was trained using $\alpha = 0.002$ and MSE as reconstruction loss. The output is green in regions where the watermark was detected and blue where it was not.

In the paper, models trained with the MSE were found, on average, less accurate on deepfake detection than that with SSIM. This claim is supported by the raw detections in Figure 10, in which the watermarked image (b) exhibits more undetected areas (i.e., not green) compared to that from a model trained with SSIM (Figure 5 in the paper). Although the undetected areas are not around the face in this case, there may be other samples in which they are, causing a miss-classification. Additionally, the detection of modified regions in the deepfake images (c and d) are smaller compared to SSIM models, making them harder to classify for our automatic detection method.

B.2. Robustness to Compression

In addition to raw detections, we analyze the robustness against compression from models trained with MSE. The results in Figure 11 are, for FS, very similar to that of the paper (i.e., on models trained with SSIM). More precisely, when using the compression module with high α values (α = 0.005 and 0.007), the models are effectively made robust to compression, reaching accuracies above 75% on high compression qualities (> 75). Additionally, with a low α value ($\alpha = 0.002$), the model can detect the watermark, but not modified regions, lowering the accuracy to around 50%. On the other hand, the results on FSh are very different compared to that using SSIM. Indeed, the accuracy for each model is close to 50%, even for high α values. It again confirms our conclusions in the paper that SSIM trains more accurate models for our task than MSE.

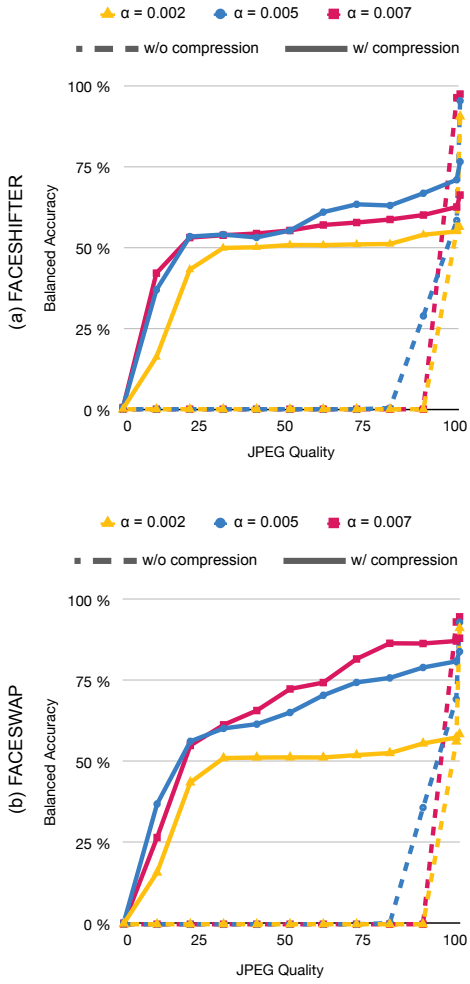

Figure 11: Balanced accuracy of Deepfake detection versus the JPEG quality of our solution using SSIM as reconstruction loss with different values of α, and Faceshifter (a) or Faceswap (b) as Deepfake generation models.

C. Additional watermarked images

Figure 12: Images watermarked with a model trained using SSIM as reconstruction loss and different values of α. The first row images are pristine ones, followed by 3 rows of watermarked images from a model trained without compression, then 3 others from a model trained with compression.

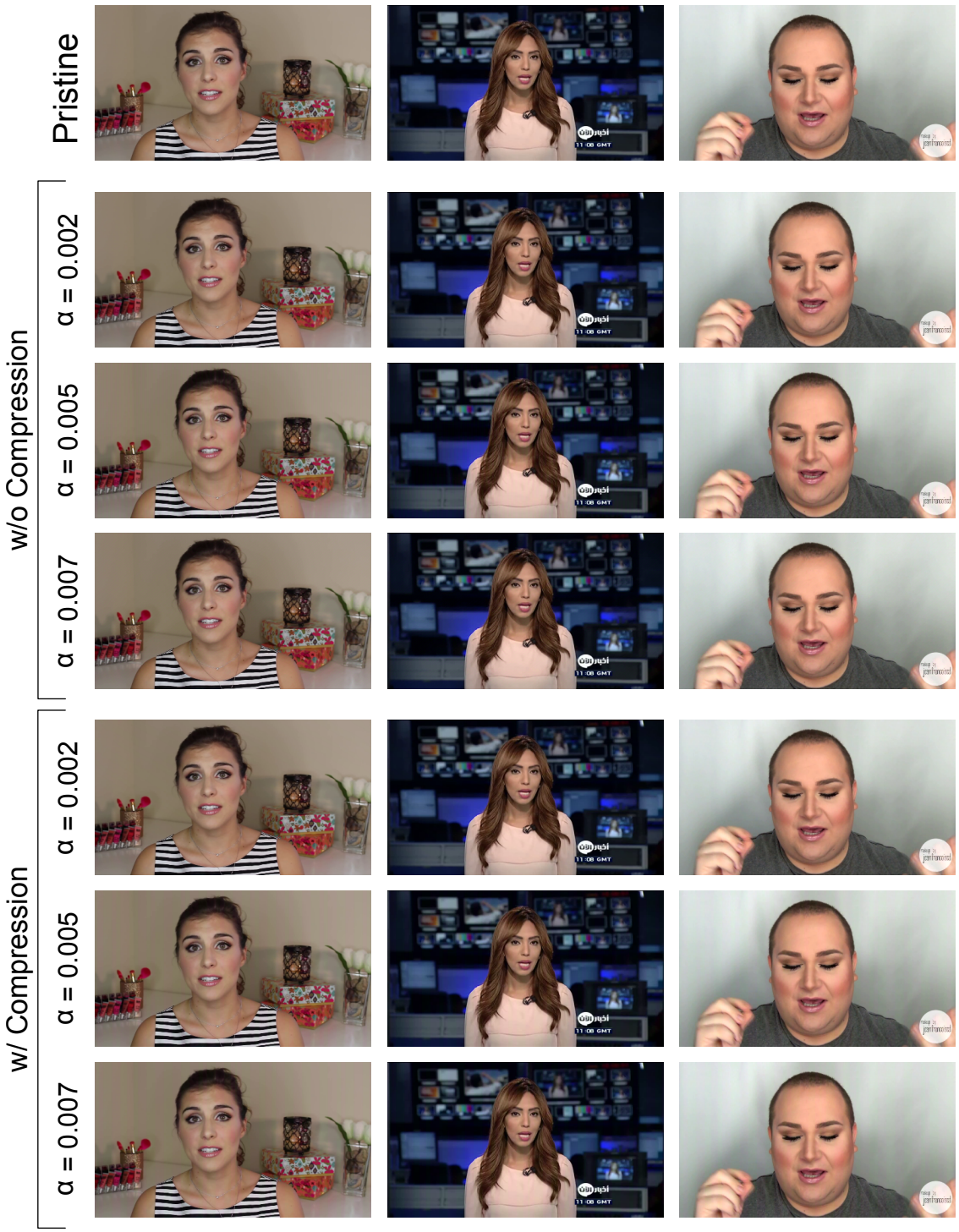

Figure 13: Images watermarked with a model trained using MSE as reconstruction loss and different values of α. The first row images are pristine ones, followed by 3 rows of watermarked images from a model trained without compression, then 3 others from a model trained with compression.

D. Additional watermark raw detecions

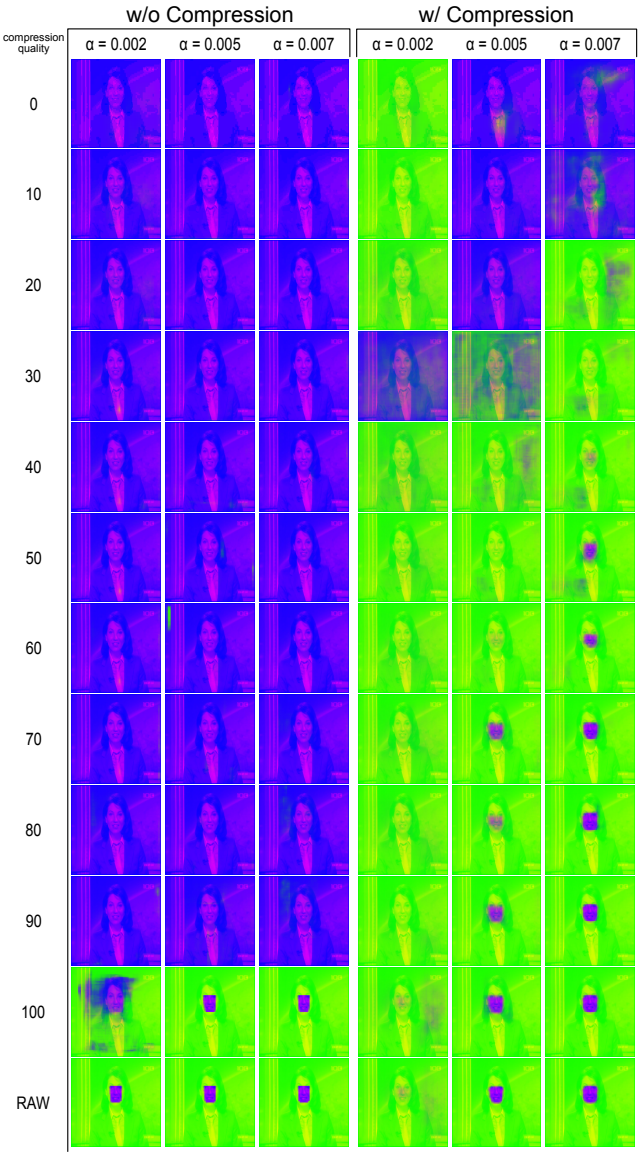

Figure 14: Raw detections (heat-map) of faceswap images from models trained using SSIM as reconstruction loss and different α values (0.002, 0.005, and 0.007). The first 3 columns are from models trained without compression, followed by 3 columns from models trained with compression. Each row correspond different encoding qualities during testing, ranging from 0 to 100, with a last row of uncompressed (RAW) images.

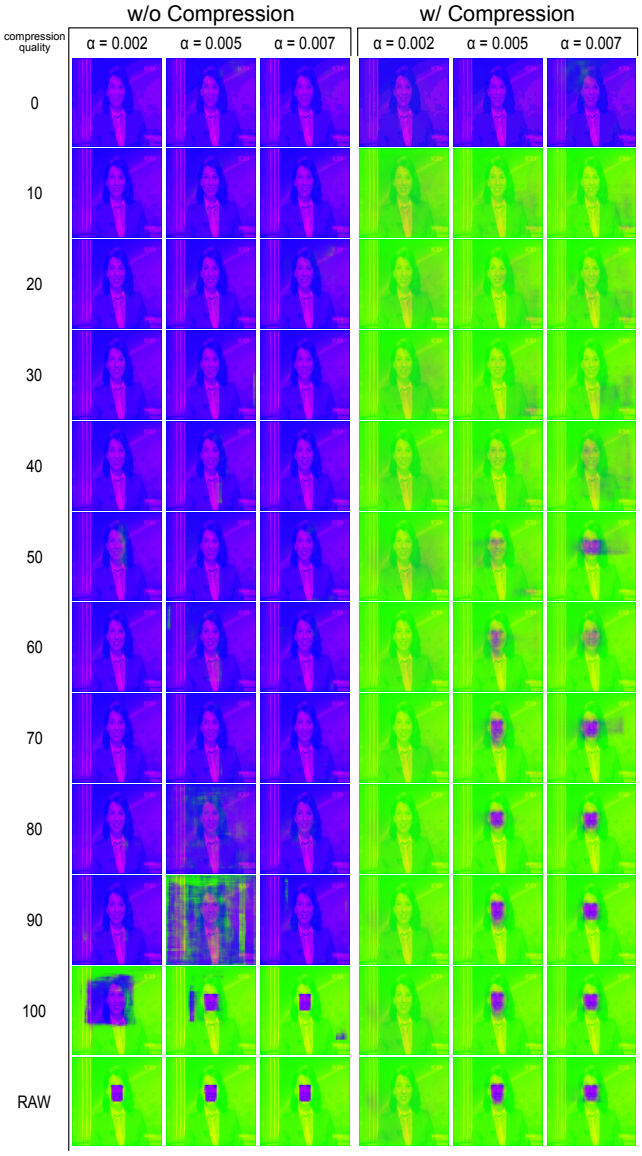

Figure 15: Raw detections (heat-map) of faceswap images from models trained using MSE as reconstruction loss and different α values (0.002, 0.005, and 0.007). The first 3 columns are from models trained without compression, followed by 3 columns from models trained with compression. Each row correspond different encoding qualities during testing, ranging from 0 to 100, with a last row of uncompressed (RAW) images.