Face Image Synthesis conditioned on Target Landmark Appearance

Group 17: Elias Kassapis (12409782), Stijn Verdenius (10470654), Klaus Ondrag (12265306)

UNIVERSITEIT VAN AMSTERDAM

1. Problem

• **Face manipulation** is the task of generating photorealistic, personalized head models capturing target expressions. Conditional **Generative Adversarial Networks** (cGANs) are widely used for this.

• cGANs yield high quality faces, however, they have **difficulties in preserving the identity**.

• The generator **loss function** of cGANs is designed to optimize features pertaining to the task (eg. identity, quality). Typically it contains **pixel distance minimizing** terms and **feature distance minimizing** terms.

2. Aim and Hypothesis

• Our aim in this study was to dissect the loss function terms with respect to their contribution to identity preservation

• Our **hypothesis** was that **feature space terms are more important for identity preservation than pixel space terms**.

3. Approach

• We use cGANs to generate a face with the desired expression by conditioning on an **input image** (source), and **target landmarks**.

• We analyse the contribution of terms in the loss function to identity preservation using an **ablation approach**.

4. Training

We trained our Generator using Adam's optimizer, and our Discriminator using SGD on the 300VW dataset. Our Discriminator loss was the binary cross-entropy, whereas the **Total Generator Loss** is given by:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{pix}\mathcal{L}_{pix} + \lambda_{cons}\mathcal{L}_{self} + \lambda_{cons}\mathcal{L}_{triple}$$

$$+\lambda_{pp}\mathcal{L}_{pp} + \lambda_{id}\mathcal{L}_{id}$$

Pixel Space Terms	
Cons. Losses	$\lambda_{cons} = 100$
Pixel Loss	$\lambda_{pix} = 10$

Feat. Space Terms	
Perceptual Loss	$\lambda_{pp} = 10$
Id Loss	$\lambda_{id} = 1$

5. Model Architecture

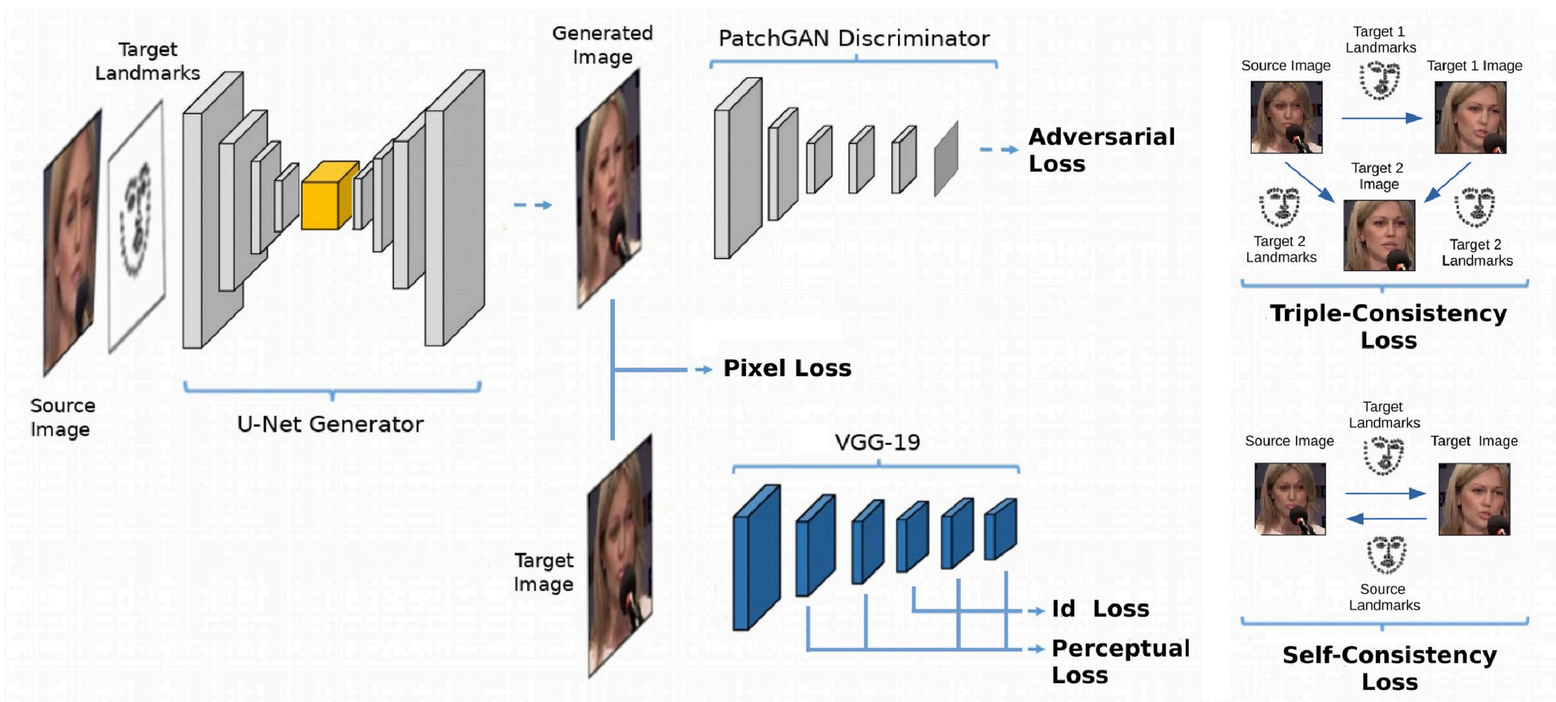
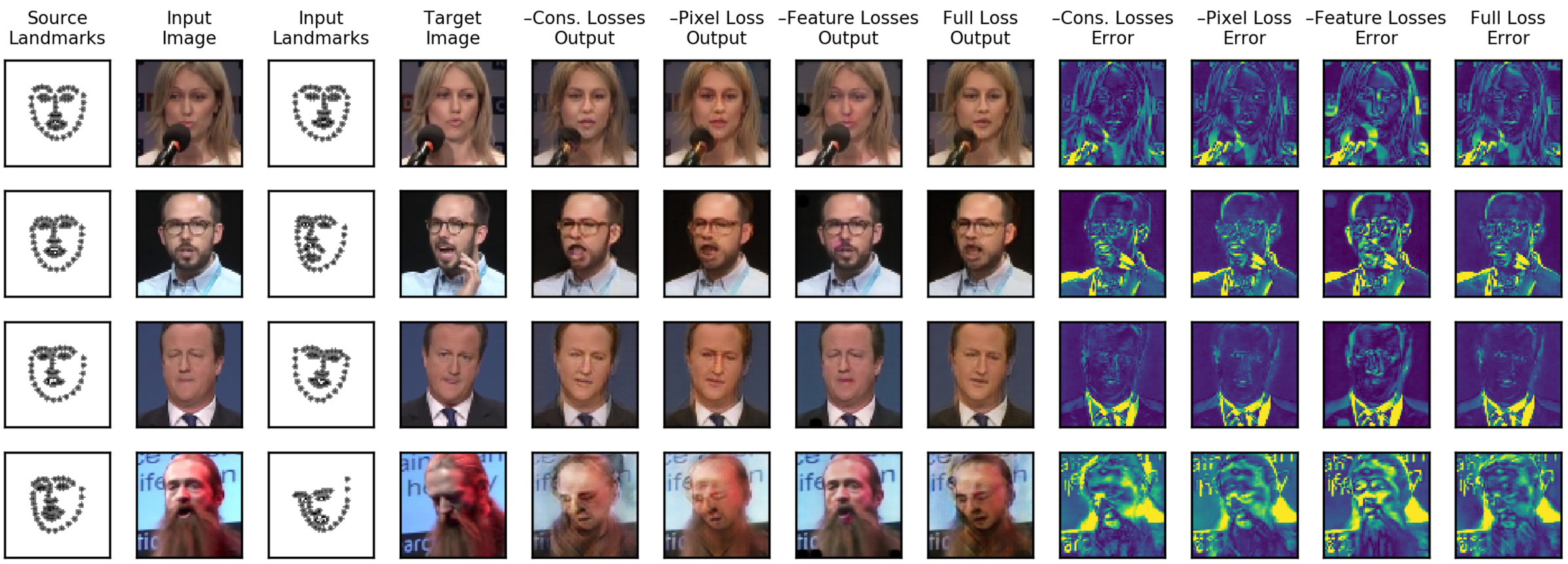
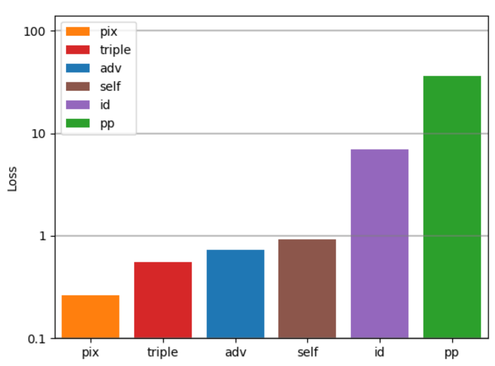


Figure 1: The panel on the **left** shows our general architecture and illustrates how the Pixel, Adversarial, Perceptual and Id losses are obtained. Pixel loss is computed by taking the L2 norm of the Generated Image and Target Image. For the Id loss and Perceptual loss, the L1 norm of the activations at selected layers of a VGG-19 is computed. The panel on the **right** shows the procedure used for computation of the Triple-Consistency loss (above), and Self-Consistency Loss (below).

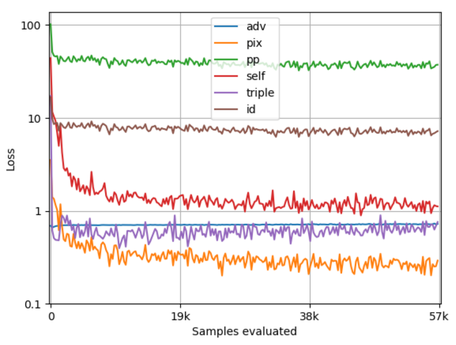
6. Results

Given our results we conclude that the **synergistic interaction between the feature space terms and pixel space terms promotes identity preservation**, as these appear to have complementary merits. Feature space terms appear to facilitate target conditioning, and pixel space terms appear to encourage the generator to regress to identity mapping.





Full Loss Function



Loss curves vs. Training time elapsed

Figure 2: The panel **above** shows generated images from different models. The first column displays the landmarks of the input image. This is not passed into the network. The second and third column are the image to condition on and the target landmarks. Both are the input to the network. In the fourth column, "Target Image", the ground truth image is shown, which displays the same person as in the input image and has the landmarks of the input landmarks. The next four columns show the generated images of the model trained with different loss functions. The last four columns show the absolute pixel-wise error averaged over the color channel, on a color scale where 0 is represented in blue and values above 96 in yellow.
The panel on the **left** shows the log of the magnitudes of the different loss terms at the last epoch of training, and the loss curves for each term during training.