Supplementary Material for
Domain Intersection and Domain Difference

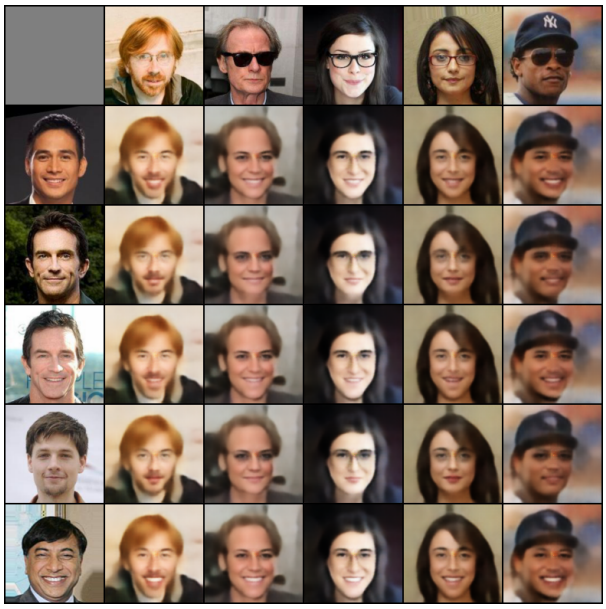

Figure 1. Translating from the domain of persons with glasses to the domain of smiling persons (reverse translation to Fig. 2 in main report)

Figure 2. Translating from the domain of persons with facial hair to the domain of smiling persons.

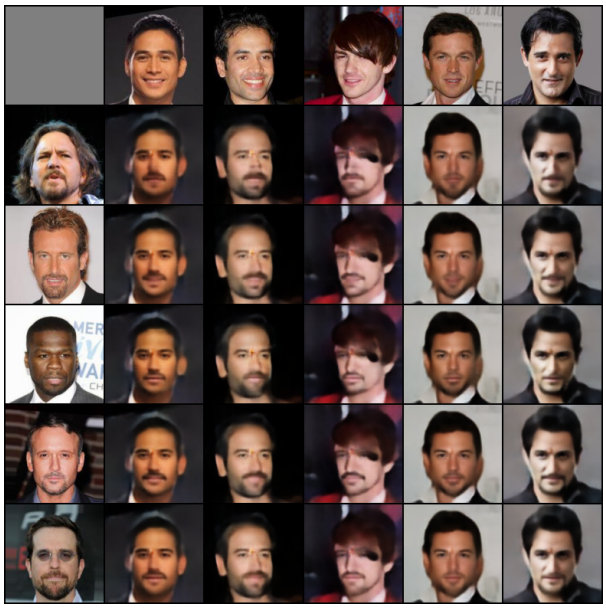

Figure 3. Reverse translation from the domain of smiling persons to the domain of persons with facial hair.

A. Additional Guided Translation Results

We provide the reverse translation to that given in Fig. 2 of the main report as well as additional cross domain translations in Fig. 1, 2, 3, 4, 5, 6, 7, 8 and 9.

Both forward and reverse directions are trained simultaneously using the same model as our model is symmetric. In the reverse direction, Given a sample $b \in B$ (top row) and a sample $a \in A$ (left column), each image constructed is of the form $G(E^c(b), E^s_A(a), 0)$

B. Architecture and Hyperparameters

We consider samples in A and B to be images in $\mathbb{R}^{3\times128\times128}$. The encoders E_c, E^A_s and E^B_s each consist of 6 convolutional blocks. Similarly, G consists of 6 deconvolutional blocks.

A convolutional block d_k consisting of: (a) 4×4 convolutional layer with stride 2, pad 1 and k filters (b) a spectral normalization layer (c) an instance normalization

Figure 4. Translating from the domain of persons with glasses to the domain of persons with facial hair.

Figure 5. Reverse translation from the domain of persons with facial hair to the domain of persons with glasses.

layer (d) a Leaky ReLU activation with slope 0.2. Similarly a de-convolutional block u_k consists of: (a) 4×4 de-convolutional layer with stride 2, pad 1 and k filters (b) a spectral normalization layer (c) an instance normalization layer (d) a ReLU activation.

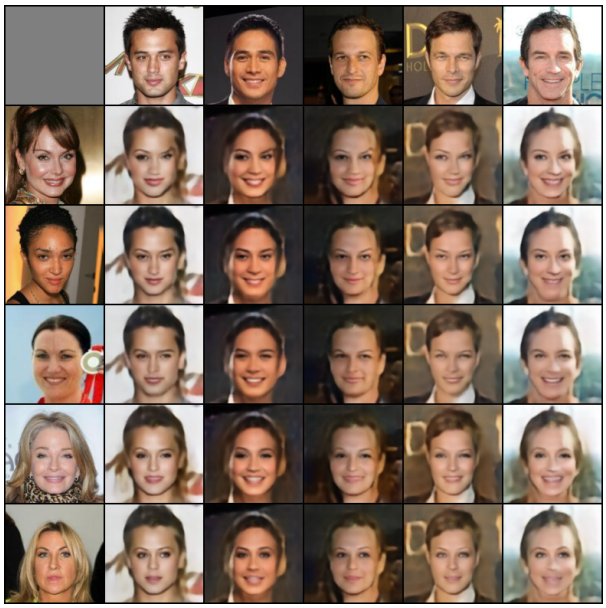

Figure 6. Translating from the domain of males to the females.

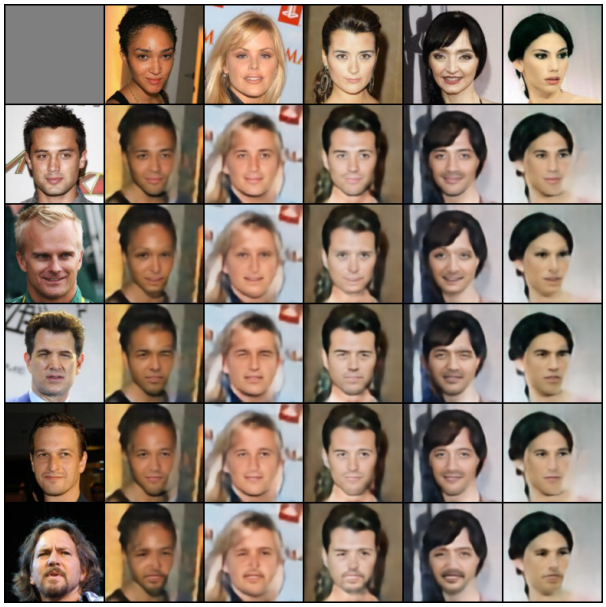

Figure 7. Reverse translation from the domain of females to the domain of females.

The structure of the encoders and generators is then:

E_c: $d_{32}, d_{64}, d_{128}, d_{256}, d_{512-sep}, d_{512-2 \cdot sep}$
E^s_A, E^s_B: $d_{32}, d_{64}, d_{128}, d_{128}, d_{128}, d_{sep}$
G: $u_{512}, u_{256}, u_{128}, u_{64}, u_{32}, u^*_3$

The last layer of G (u^*_3) differs in that it doesn't contain a spectral or instance normalization and that Tanh activation is applied instead of ReLU. sep is the dimension of the separate encoders, set to be 25 for all datasets.

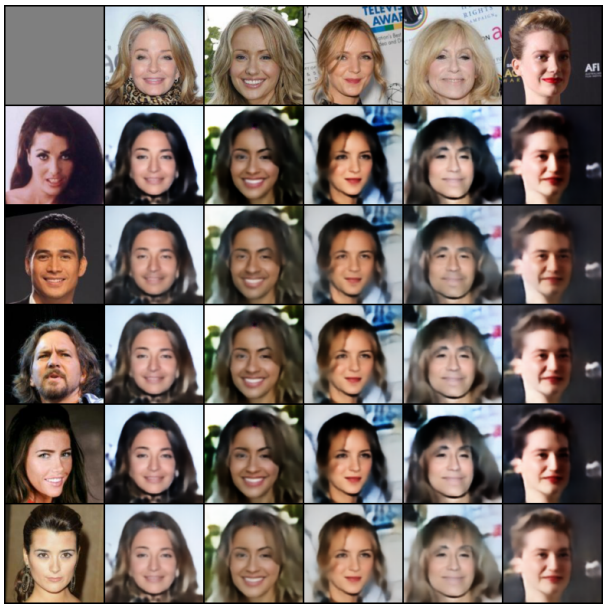

Figure 8. Translation from the domain of blond hair to the domain of black hair.

Figure 9. Reverse translation from the domain of black hair to the domain of blond hair.

The latent discriminator d consists of a fully connected layer of 512 filters, a Leaky ReLU activation with slope 0.2, second fully connected layer of 1 filters and a final sigmoid activation.

For the loss parameters specified in the equation 11 of the main report, λ_1 is set to 0.001 and λ_2 to 1. We use the Adam optimizer with $\beta_1 = 0.5, \beta_2 = 0.999$, and learning rate of 0.0002. We use a batch size of size 32 in training.

C. Theoretical Analysis

In this section we provide a formal version of Thm. 1 from the main text. For this purpose, we recall a few technical notations from [1]: the Shannon entropy (discrete or continuous) $H(X) := -\mathbb{E}_X[\log_2 \mathbb{P}[X]]$, the conditional entropy $H(X|Y) := H(X,Y) - H(Y)$, the (conditional) mutual information (discrete or continuous) $I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$. For clarity, we list a few important identities that are being used throughout the proofs in this section. For any two random variables X and Y, we have: $I(X;Y) = H(X) + H(Y) - H(X,Y)$. The data processing inequality, for any random variable X and two functions f and g, we have: $I(X;g(f(X))) \leq I(X;f(X))$.

In Sec. 2 in the main text, we represented our random variable $a \sim \mathbb{P}_A$ and $b \sim \mathbb{P}_B$ in the following forms $a = g(e^c(a), e^s_A(a), 0)$ and $b = g(e^c(b), 0, e^s_B(b))$, where $e^c(a) \perp\!\!\!\perp e^s_A(a)$, $e^c(b) \perp\!\!\!\perp e^s_B(b)$ and g is some invertible function. Our method learns three encoders $E(x) := (E^c(x), E^s_A(x), E^s_B(x))$ and a decoder G.

The following theorem is a formal version of Thm. 1 from the main text.

Theorem 1. *In the setting of Sec. 2 in the main text. Let $a \sim \mathbb{P}_A$ and $b \sim \mathbb{P}_B$ be two random variables distributed by discrete distributions $\mathbb{P}_A$ and $\mathbb{P}_B$. Assume that the representations $g(e^c(a), e^s_A(a), 0)$ and $g(e^c(b), 0, e^s_B(b))$ form an intersection between a and b, such that,*

$$H(E^s_A(a)) \leq H(e^s_A(a)) + \epsilon \tag{1}$$

In addition, assume that: $\mathbb{E}_a \|G(E^c(a), E^s_A(a), 0) - a\|_1 = 0$, $\mathbb{E}_b \|G(E^c(b), 0, E^s_B(b)) - a\|_1 = 0$ and $\mathbb{P}_{E^c(A)} = \mathbb{P}_{E^c(B)}$, i.e., the distribution of $E^c(A)$ is equal to the distribution of $E^c(B)$. Then, we have the following:

- $I(E^c(a); E^s_A(a)) \leq \epsilon$.
- *$E^c(a)$ is a function of $e^c(a)$.*
- $H(E^c(a)) \geq H(e^c(a)) - \epsilon$.

In this theorem, we make a few assumptions. The first assumption concerns the modeling of the data, the second is regarding the separate encoder E^s_A and the last one concerns the losses.

Our first assumption asserts that the ground truth representation (see Sec. 2) of the random variables $a = g(e^c(a), e^s_A(a), 0)$ and $b = g(e^c(b), 0, e^s_B(b))$ forms an intersection between them. Put differently, we can partition the information of a and b into independent features $e^c(a)$, $e^s_A(a)$ for a and $e^c(b)$, $e^s_B(b)$ for b, such that, the information of $e^c(a) \sim e^c(b)$ is maximal. Informally, any other partition into common and separate parts is unable to put more content information in the common part than the amount the ground truth representations do. For example, in the case

where A consists of images of persons with facial hair and B consists of images of persons with glasses, the assumption is verified, since, we cannot transfer information from the separate part (facial hair or glasses) into the common part (identity, pose, etc').

The second assumption asserts that the amount of information encoded in $E^s_A(a)$ is bounded by the amount of information encoded in $e^s_A(a)$. Differently viewed, since the function E^s_A is deterministic, we also have $I(E^s_A(a); a) = H(E^s_A(a))$, and therefore, the amount of mutual information between $E^s_A(a)$ and a is bounded as well. This implies that we cannot recover a given $E^s_A(a)$, since we cannot recover a from $e^s_A(a)$.

The third assumption is that several losses are minimized. In Sec. 3, we introduced reconstruction losses: $\mathcal{L}^A_{recon}$ and $\mathcal{L}^B_{recon}$ and an adversarial loss: $\mathcal{L}_{adv}$. These losses were measured on average with respect to the training set. In Thm. 1, the reconstruction losses $\mathcal{L}^A_{recon}$ and $\mathcal{L}^B_{recon}$ are replaced with their expected versions (we take expectations $\mathbb{E}_a$ and $\mathbb{E}_b$ instead of averages over the training sets $\mathcal{S}_A$ and $\mathcal{S}_B$), $\mathbb{E}_a\|G(E^c(a), E^s_A(a), 0) - a\|_1$ and $\mathbb{E}_b\|G(E^c(b), 0, E^s_B(b)) - b\|_1$. In the theorem, we assume that these losses are being minimized by E^c, E^s_A, E^s_B and G. In addition, the expected version of $\mathcal{L}_{adv}$ is $\sup_d \{\mathbb{E}_a l(d(E^c(a)), 1) + \mathbb{E}_b l(d(E^c(b)), 1)\}$ which is minimized by any encoder E^c that provides $\mathbb{P}_{E^c(A)} = \mathbb{P}_{E^c(B)}$ (see Prop. 2 in [2]), i.e., the distribution of $E^c(a)$ is equal to the distribution of $E^c(b)$. In Thm. 1, we assume that $\mathbb{P}_{E^c(A)} = \mathbb{P}_{E^c(B)}$ which implies that the adversarial loss is minimized as well. We note that in this analysis the zero-losses are not a requirement. It is also depicted in our ablation study that the zero-losses are not a requirement but slightly improve the results.

The consequences of the theorem are: (i) the encodings $E^c(a)$ and $E^s_A(a)$ are (almost) independent, (ii) $E^c(a)$ is a function of $e^c(a)$ and (iii) $E^c(a)$ holds most of information in $e^c(a)$. The second and third consequences provide that $E^c(a)$ and $e^c(a)$ encode the same information. We note that, given these consequences, we could also claim that $E^s_A(a)$ and $e^s_A(a)$ hold the same information. Therefore, we conclude that under the proposed assumptions, the learned encodings $E^c(a)$ and $E^s_A(a)$ capture the same information as $e^c(a)$ and $e^s_A(a)$ (resp.).

Finally, for clarity, we note that by symmetric arguments, we could arrive at the same conclusions for $E^c(b)$ and $E^s_B(b)$.

D. Proof of Thm. 1

Proof of Thm. 1. First, we consider that by $I(X;Y) = H(X) + H(Y) - H(X,Y)$, we have:

$$\begin{aligned} I(E^c(a); E^s_A(a)) =& H(E^c(a)) + H(E^s_A(a)) \\ & - H(E^c(a), E^s_A(a)) \end{aligned} \tag{2}$$

Since $\mathbb{E}_a\|G(E^c(a), E^s_A(a), 0) - a\|_1 = 0$, we have:

$$I(G(E^c(a), E^s_A(a), 0); a) = I(a; a) = H(a) \tag{3}$$

Next, by the data processing inequality, we have: $I(X; g(f(X))) \leq I(X; f(X))$. Therefore, by selecting $g(\cdot) := G(\cdot, 0)$ and $g(\cdot) := (E^c(\cdot), E^s_A(\cdot))$ and $X := a$, we have:

$$\begin{aligned} H(a) &= I(G(E^c(a), E^s_A(a), 0); a) \\ &\leq I(E^c(a), E^s_A(a); a) \end{aligned} \tag{4}$$

Since $a = g(e^c(a), e^s_A(a), 0)$, where $e^c(a)$ and $e^s_A(a)$ are assumed to be independent (see Sec. 2) and g to be is invertible, we have:

$$\begin{aligned} H(a) &= H(g(e^c(a), e^s_A(a), 0)) \\ &= H(e^c(a), e^s_A(a)) \\ &= H(e^c(a)) + H(e^s_A(a)) \end{aligned} \tag{5}$$

We assumed that the representations $g(e^c(a), e^s_A(a), 0)$ and $g(e^c(b), 0, e^s_B(b))$ form an intersection between a and b. In addition, $G(E^c(a), E^s_A(a), 0) \sim \mathbb{P}_A$, $G(E^c(b), 0, E^s_B(b)) \sim \mathbb{P}_B$ and $E^c(a) \sim E^c(b)$ (since we assumed that $\mathbb{P}_{E^c(A)} = \mathbb{P}_{E^c(B)}$). Therefore, for $G := \hat{g}$, $\hat{e}^c := E^c$, $\hat{e}^s_A := E^s_A$ and $\hat{e}^s_B := E^s_B$, by Def. 1:

$$H(E^c(a)) \leq H(e^c(a)) \tag{6}$$

By Eq. 1, we have:

$$H(E^s_A(a)) - \epsilon \leq H(e^s_A(a)) \tag{7}$$

By combining Eqs. 4, 5, 6 and 7, we have:

$$\begin{aligned} H(E^c(a), E^s_A(a)) &\geq H(a) \\ &= H(e^c(a)) + H(e^s_A(a)) \\ &\geq H(E^c(a)) + H(E^s_A(a)) - \epsilon \end{aligned} \tag{8}$$

By combining the last inequality with Eq. 2, we have:

$$I(E^c(a); E^s_A(a)) \leq \epsilon \tag{9}$$

Next, we define $\hat{e}^c(a) := (e^c(a), E^c(a))$, $\hat{e}^s_A(a) := (e^s_A(a), E^s_A(a))$, $\hat{e}^s_B(b) := (e^s_B(b), E^s_B(b))$ and g', such that, $g'(\hat{e}^c(a), \hat{e}^s_A(a), 0) = g(e^c(a), e^s_A(a))$ and $g'(\hat{e}^c(b), 0, \hat{e}^s_B(b)) = g(e^c(b), e^s_B(b))$. Since g is invertible for both domains, we conclude that g' is invertible as well. Therefore, by Def. 1, we conclude that $H(\hat{e}^c(a)) \leq H(e^c(a))$. But, $\hat{e}^c(a) = (e^c(a), E^c(a))$ and, therefore, we also have: $H(\hat{e}^c(a)) \geq H(e^c(a))$. In particular, $H(\hat{e}^c(a)) = H(e^c(a))$. We conclude that:

$$\begin{aligned} I(e^c(a); E^c(a)) =& H(e^c(a)) + H(E^c(a)) \\ & - H(e^c(a), E^c(a)) \\ =& H(E^c(a)) \end{aligned} \tag{10}$$

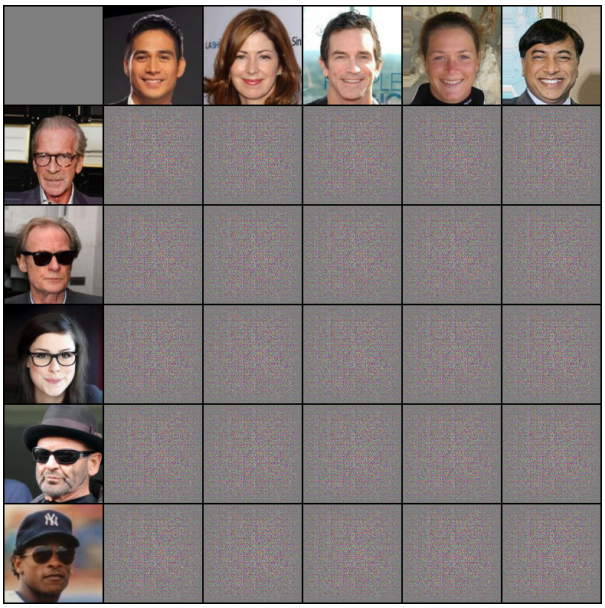

Figure 10. Translation from the domain of smiling persons to the domain of persons with glasses, when the reconstruction loss is removed

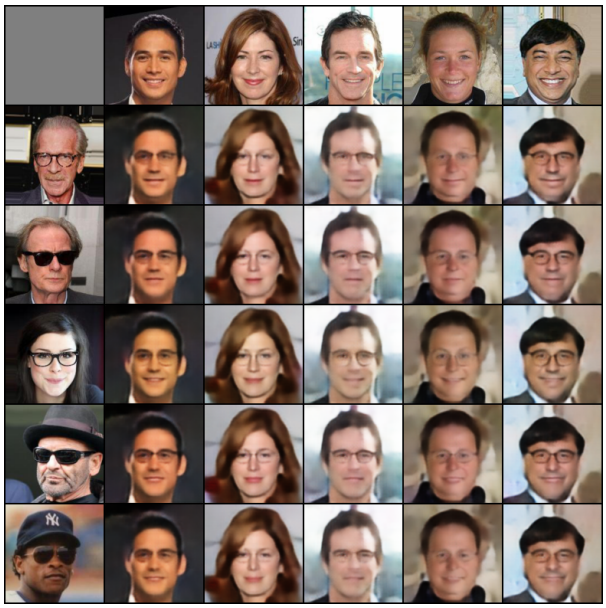

Figure 11. Translation from the domain of smiling persons to the domain of persons with glasses, when the adversarial loss is removed

Therefore, $E^c(a)$ is a function of $e^c(a)$. Finally, we consider that:

$$\begin{aligned} & H(E^c(a)) + H(e^s_A(a)) + \epsilon \\ \geq & H(E^c(a)) + H(E^s_A(a)) \\ \geq & H(a) \\ = & H(e^c(a)) + H(e^s_A(a)) \end{aligned} \tag{11}$$

In particular, $H(E^c(a)) \geq H(e^c(a)) - \epsilon$. □

E. Ablation Study Visual Results

In order to compare the effect of the different loss visually, we provide in Fig. 10, 11 and 12 the translation from smiling persons to persons with glasses, when each of the losses is removed. With no reconstruction loss the method is unable to create realistic face images, as the G is not affected by any of the losses remaining. With no adversarial loss the method is unable to add the glasses (separate part of domain B) to the given image. Without the zero-loss, results are only slightly worse numerically, and this is not observed visually.

F. Visual Comparison to Baseline Methods

In additional to the numerical comparison in tables 1 and 2 of the main report, we provide a visual comparison in Fig. 13, 14 and 15. For MUNIT and DRIT, the method is unable to change content in the source image, and so the smile (separate part of domain A) remains, and no glasses (separate part of domain B) are added. For Fader Networks, a generic glasses are added, and not the one specific to the image in domain B.

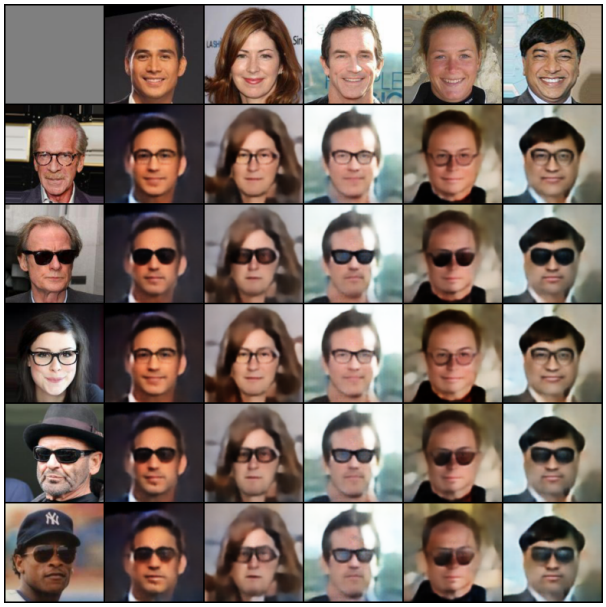

Figure 12. Translation from the domain of smiling persons to the domain of persons with glasses, when the zero loss is removed

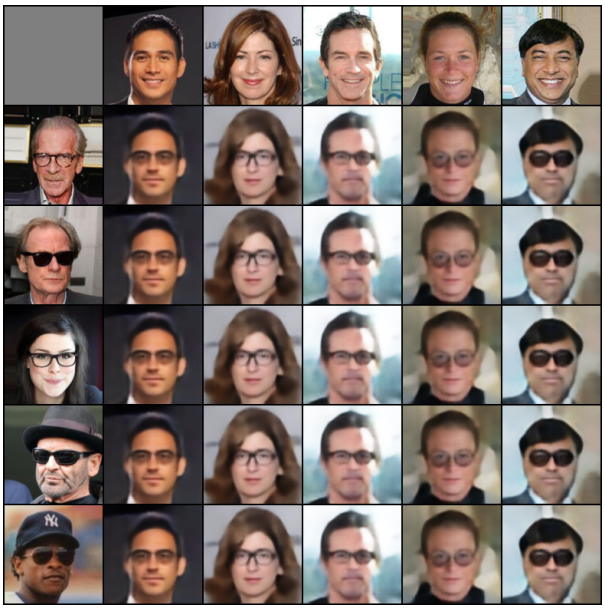

Figure 13. Translation from the domain of smiling persons to the domain of persons with glasses, using the Fader Networks method.

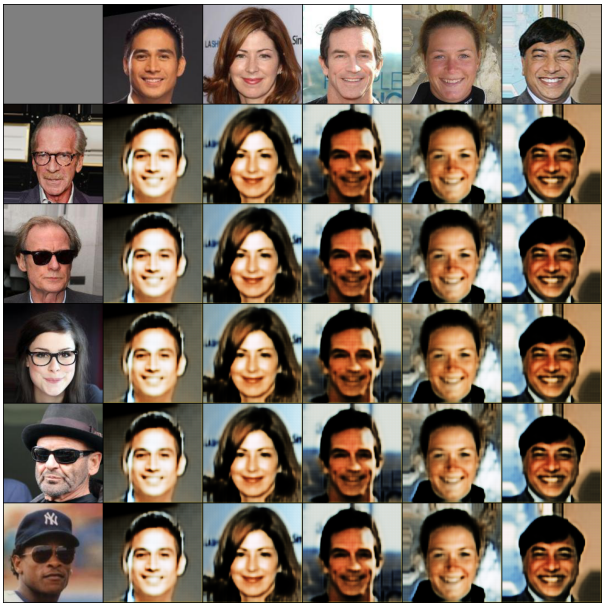

Figure 14. Translation from the domain of smiling persons to the domain of persons with glasses, using the DRIT method.

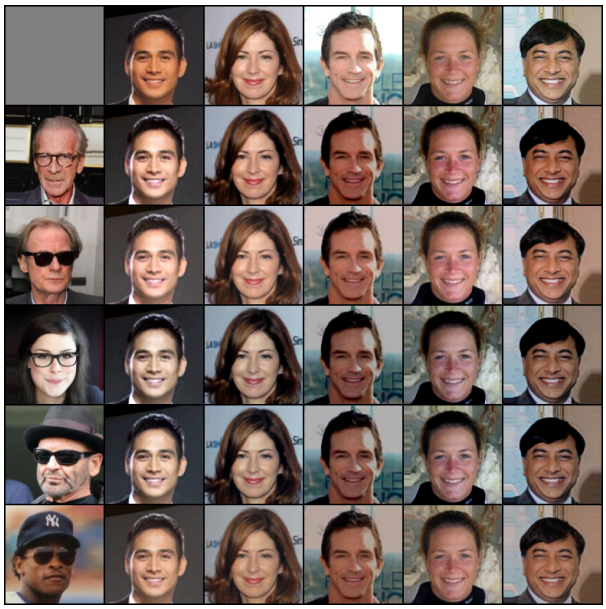

Figure 15. Translation from the domain of smiling persons to the domain of persons with glasses, using the MUNIT method.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. 3

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014. 4