Appendix

In addition to the appendix sections below, we provide our code publicly with the relevant instructions [5].

A. Impact of pre-trained networks on perceptual loss

We evaluate the impact of using the activations from different pre-trained convolutional backbones on our model. We assess four backbone models namely VGG [29], AlexNet [21], SqueezeNet [16], ResNet [13], all pretrained on ImageNet. The shape of the layers to calculate the losses in each backbone are included in Tables 4 - 7 respectively.

Fig. 13 shows the qualitative results of the human face and caricature domains. Table 2 shows the numerical results of this experiment on the hair segmentation model; all the metrics of this table are computed over the target domain. These experiments demonstrate that the pre-trained VGG [29] model offers the best performance for the perceptual loss in faces. Previous work [6, 7] only applied pre-trained VGG [29] model for perceptual loss, and in [38] the assumption for the perceptual loss was that the results from VGG [29] and AlexNet [21] are similar. We show qualitatively and quantitatively that some network backbones perform better than others.

We also demonstrate that this is domain specific in Fig. 14 and Fig. 15. For both the horse and the car domains, AelxNet offers significantly better results than VGG. This is an interesting observation given that the backbones are trained on the same data. We stipulate that the architecture design has been motivated by different tasks. Further exploration of the role of backbone in perceptual loss calculation is needed.

Table 2. Comparison of the performance of four backbones for perceptual loss. The result are on the natural faces and caricature domains, where the segmentation model is trained to segment hair.

		$DDS_{car.\to nat.}$			$DDS_{nat.\to car.}$		
		FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑
ResNet	$\mathcal{D}_t$	95.66	0.55	28.89	98.71	0.47	27.93
SqueezNet	$\mathcal{D}_t$	89.48	0.62	29.15	94.29	0.52	28.05
AlexNet	$\mathcal{D}_t$	88.63	0.66	29.36	93.27	0.71	28.11
VGG	$\mathcal{D}_t$	**87.28**	**0.71**	**29.58**	**76.96**	**0.72**	**28.39**

B. Dual-Domain Synthesis vs. Image Blending

To demonstrate the difference of our DDS with image blending approaches, we provide a comparison in Fig. 16 between Dual-Domain and recent Image Blending works (using the public codes of [30, 37]). Blending does not achieve better or even comparable results.

[5] https://github.com/denabazazian/Dual-Domain-Synthesis

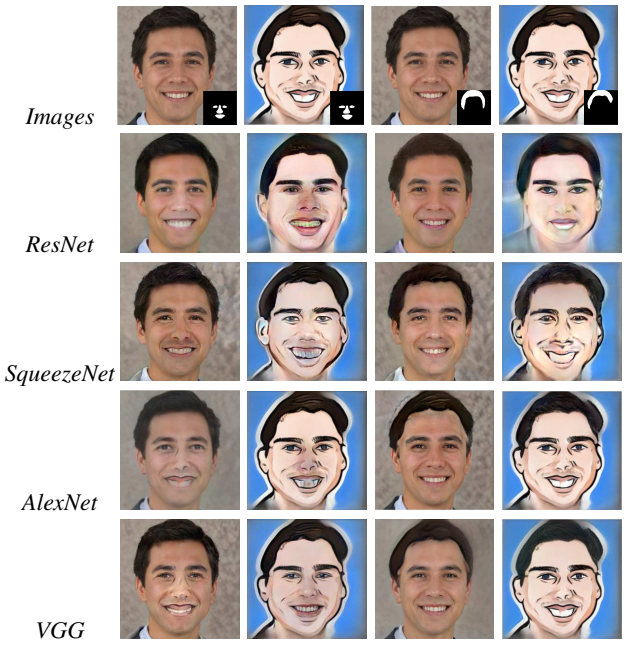

Figure 13. The impact of four pre-trained backbones for computing perceptual loss. The first two columns uses the eye/nose/mouth segmentation model, while the last two columns uses the hair segmentation model.

Figure 14. DDS results on horse domains. Odd columns are from natural horse domain, and even columns are from sketch horse domain. Segmentation masks are Incorporated. Second and third rows show the corresponding dual-domain images using the perceptual loss from the *conv* layers of VGG [29] and AlexNet [21] respectively.

C. DDS from few-shotGAN domains

We performed further experiments on synthesising images which contain features from both the caricature and sketch domains. Note that both GANs are trained us-

Figure 15. DDS results on car domains. Odd columns are standard car images, and even columns are abandoned cars. Segmentation masks are incorporated. Second and third row show the DDS results using the perceptual loss from the *conv* layers of VGG [29] and AlexNet [21] respectively.

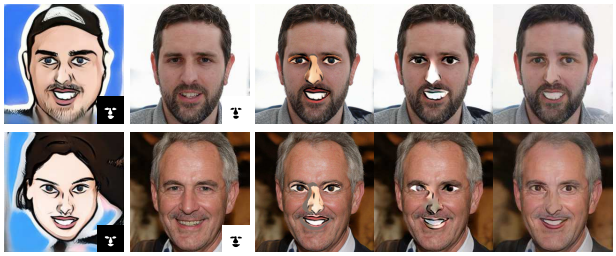

Figure 16. Comparing [30] (col3) and [37] (col4) to Dual-Domain (col5), for paired (top) and unpaired (bottom) examples.

ing few-shots, adapted independently from the natural face StyleGAN. Fig. 17 shows dual-domain images based on the integrating caricature and sketch features.

D. Final vs. intermediate latent space editing

In our results, we used the synthesised image from the last layer of StyleGAN. Instead, intermediate representations can be used to accommodate unpaired images of varying poses. We experiment with using the concatenation of hidden intermediate layer activations. In this experiment for the images with resolution 256×256, we get the 13 layers of features as described in Table 3. Hence, for computing the perceptual loss instead of passing the images to *conv* layers of *VGG*, we directly use the activations of intermediate latent space of styleGAN from both the target and source domains. Fig. 18 demonstrates a comparison of the results when we apply DDS framework and when we use the intermediate latent activations.

While intermediate activations can accommodate variations in pose, they cannot maintain the features from two domains.

Table 3. Shape of the intermediate StyleGAN features.

$\mathbf{w} \in \mathcal{W}$	*shape*
$\mathbf{w}_0$	$[512 \times 4 \times 4]$
$\mathbf{w}_1$	$[512 \times 8 \times 8]$
$\mathbf{w}_2$	$[512 \times 8 \times 8]$
$\mathbf{w}_3$	$[512 \times 16 \times 16]$
$\mathbf{w}_4$	$[512 \times 16 \times 16]$
$\mathbf{w}_5$	$[512 \times 32 \times 32]$
$\mathbf{w}_6$	$[512 \times 32 \times 32]$
$\mathbf{w}_7$	$[512 \times 64 \times 64]$
$\mathbf{w}_8$	$[512 \times 64 \times 64]$
$\mathbf{w}_9$	$[256 \times 128 \times 128]$
$\mathbf{w}_{10}$	$[256 \times 128 \times 128]$
$\mathbf{w}_{11}$	$[128 \times 256 \times 256]$
$\mathbf{w}_{12}$	$[128 \times 256 \times 256]$

Table 4. Shape of the *VGG-16* layers that used for perceptual loss.

Layer	*shape*
$conv_{1\_1}$	$[64 \times 256 \times 256]$
$conv_{1\_2}$	$[64 \times 256 \times 256]$
$conv_{2\_2}$	$[256 \times 64 \times 64]$
$conv_{3\_3}$	$[512 \times 32 \times 32]$

Table 5. Shape of the *AlexNet* layers that used for perceptual loss.

Layer	*shape*
$relu_1$	$[64 \times 55 \times 55]$
$relu_2$	$[192 \times 27 \times 27]$
$relu_3$	$[384 \times 13 \times 13]$
$relu_4$	$[256 \times 13 \times 13]$
$relu_5$	$[256 \times 13 \times 13]$

Table 6. Shape of the *SqueezeNet* layers that used for perceptual loss.

Layer	*shape*
$relu_1$	$[64 \times 127 \times 127]$
$relu_2$	$[128 \times 63 \times 63]$
$relu_3$	$[256 \times 31 \times 31]$
$relu_4$	$[384 \times 15 \times 15]$
$relu_5$	$[384 \times 15 \times 15]$
$relu_6$	$[512 \times 15 \times 15]$
$relu_7$	$[512 \times 15 \times 15]$

Table 7. Shape of the *ResNet-18* layers that used for perceptual loss.

Layer	*shape*
$relu_1$	$[64 \times 128 \times 128]$
$conv_2$	$[64 \times 64 \times 64]$
$conv_3$	$[128 \times 32 \times 32]$
$conv_4$	$[256 \times 16 \times 16]$
$conv_5$	$[512 \times 8 \times 8]$

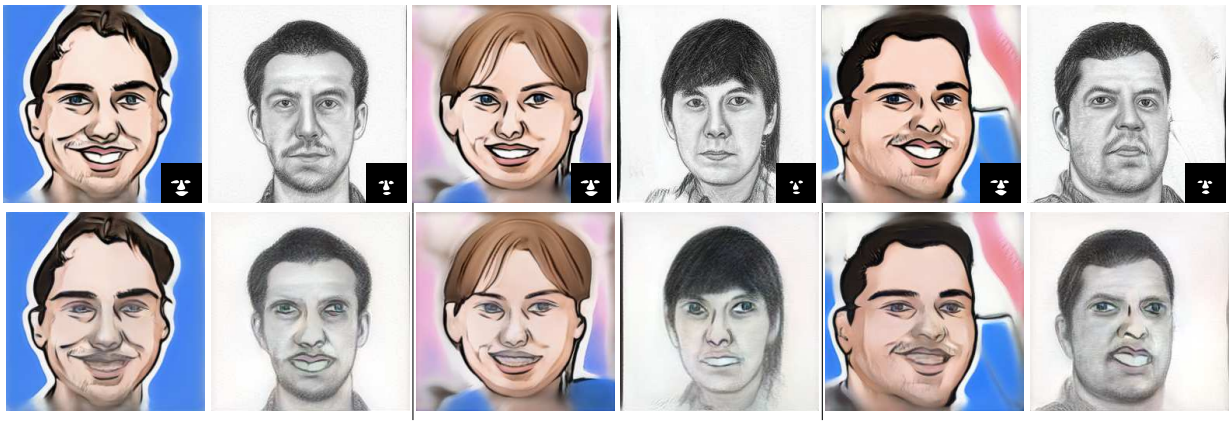

Figure 17. DDS results on caricature and sketch domains. The first row show a pair of images from the domains of caricature and sketch. The eyes/nose/mouth segmentation model used in these experiments, and all the correspondences masks are shown in the first row. The second row shows the dual-domain images containing the features of caricature and sketch based on the segmentation model. The changes in the shape of the nose and mouth and colour of the eyes can be observed from the dual-domain images.

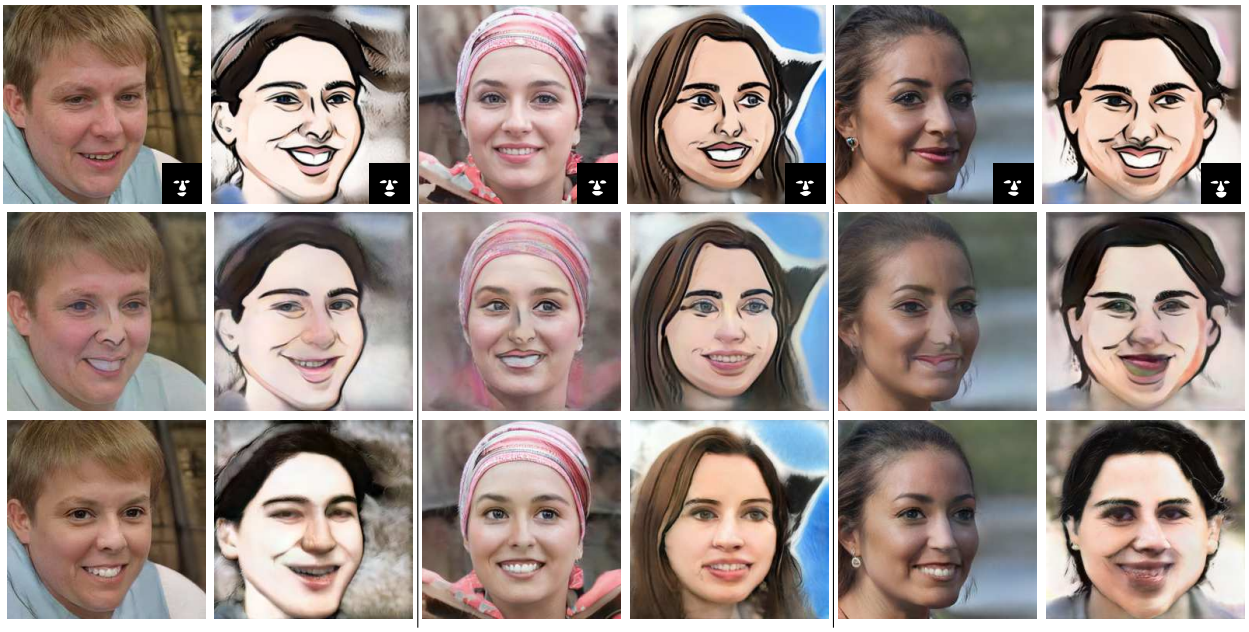

Figure 18. Comparison of the results of using the intermediate latent activations. First row shows unpaired images of two domains, masks are incorporated. Second row shows the DDS results, where each image maintain distinct features from the two domains. Third row shows the results when we use the intermediate latent space features for computing the perceptual loss. While image are integrated, they do not maintain the distinctness of the two domains. For example, in the first image (row 3, col 1) the output is purely a natural face.

E. Sketch horse data

Fig. 19 shows the images of sketch horses that used for training few-shotGAN.

Figure 19. Training samples for generating sketch horse images.