# InsetGAN for Full-Body Image Generation: Supplementary Materials

Anna Frühstück[1, 2] Krishna Kumar Singh[2] Eli Shechtman[2]
Niloy J. Mitra[2, 3] Peter Wonka[1] Jingwan Lu[2]
[1] KAUST [2] Adobe Research [3] University College London
anna.fruehstueck@kaust.edu.sa, {krishsin, elishe}@adobe.com, {niloym, pwonka}@gmail.com, jlu@adobe.com


Figure 1. **Optimizing two Insets.** InsetGAN successfully merges a canvas with two distinct insets (shoes and face).

## 1. Additional Results and Discussion

We use a video to illustrate our entire InsetGAN pipeline and to show additional results including a joint latent space walk (see the description below). The video is available at the project webpage `afruehstueck.github.io/insetgan`.

**Two Insets.** We demonstrate in Fig. 1 that our technique is able to generate good results for insets of another domain beyond faces. To that end, we trained a shoe generator at 256×256px resolution on shoes cropped from the same dataset and use the generated shoe outputs to improve problematic areas in the canvas domain with higher-quality shoe insets exhibiting more detailed and natural features than the full-body GAN. These results also demonstrate that our technique can jointly optimize more than one inset: in this example, we select a target face, and target shoes, and find an appropriate body, optimizing all three generators to create a seamless output.

**Face Orientations.** While our results exhibit somewhat limited body poses due to the entanglement of body pose variability and deterioration in image quality, we show that we are able to still capture a variety of face orientations and our technique can match the face orientation with a plausible looking body oriented correctly based on the face target, as shown in the results in Fig. 2.

**Face-Body Montage.** We show some additional results using a human generator trained on our custom dataset as well as another generator trained on the DeepFashion dataset as seen in Figures 4 and 5. The generator trained on Deep Fashion is able to synthesize bodies and garment details with good quality, but the generator is overfitted to the limited quantity and variety of the input data and is thus not very flexible in harmonizing skin tone differences (columns 2 and 3 in Fig. 5). Note how the results in Fig. 4 adapt the hair and body composition, even generating reasonable results for the short-haired woman in the rightmost column.

**Latent Space Walk.** We are able to create a joint latent space walk (see supplementary video) by linearly interpolating both the face and the body latents and use them as initialization in our joint optimization framework to combine the face and body seamlessly in a temporally coherent way. We explain the process to obtain a seamless interpolation in $n$ steps based on two key frames $K_{start}$ and $K_{end}$ and their corresponding optimized face and body

Figure 2. **Orientations.** We demonstrate that our technique can capture a wide range of face orientations and generate natural-looking face-body compositions that are oriented appropriately for each respective input face.

latent pairs ($\mathbf{w}_{A start}$, $\mathbf{w}_{B start}$) and ($\mathbf{w}_{A end}$, $\mathbf{w}_{B end}$). The naive solution is to simply linearly interpolate $n$ times between these latent pairs to obtain the inbetweens. However, interpolating in each of the two latent spaces independently does not yield a seamlessly merged boundary region between the canvas and the inset. In order to improve this boundary, we consider the latent space walk of the canvas as fixed and define the optimization of the inset as an interpolation problem where we optimize frame by frame and do the following at frame $i$:

(1) Consider the previous frame (initially, this is $K_{start}$) and obtain the next frame as the linear interpolation given by $f = \frac{1}{n-i}$ and $\mathbf{w}_{B next} = (1-f) \times \mathbf{w}_{B previous} + f \times \mathbf{w}_{B end}$

(2) To avoid unwanted jittering, we no longer reevaluate the face bounding box per frame but linearly interpolate from $\mathcal{B}_{start}$ to $\mathcal{B}_{end}$ to obtain a smooth transition from one inset position to the next.

(3) Optimize $\mathbf{w}_{B next}$ for a small number of iterations (e.g. 100 optimization steps) with a set of losses optimizing for (a) the edge coherence with the canvas, (b) the identity preservation with the starting point of $\mathbf{w}_{B next}$ and (c) the minimization of the edge region changes with respect to the last frame $K_{previous}$.

We use this method to insert about 20 to 40 interpolated frames between two given keyframes and render the resulting animation to a video at 16-20fps. By replicating the first keyframe as the last, the latent space walk can loop infinitely.

**Custom Face Generator.** Most results in the paper and supplementary use a face generator trained on the same data used to train our human GAN. We crop the faces and resample them to 256×256px resolution and train a face generator using the StyleGAN2 architecture. We show some generated face samples in Fig. 3. The visual quality is much higher than that of the faces generated by our full-body generator. Compared to the FFHQ face model, our custom face generator can be used to obtain nicer joint optimization results that better preserve the input face characteristics (ethnicity, skin tone, etc.) without distribution shift.

Figure 3. **Face Synthesis.** We show unconditionally generated results at 256×256px resolution from a face generator trained on the same data as our full-body human generator.

## 2. Evaluation

**Precision and Recall Scores.** In addition to calculating the FID scores to quantitatively evaluate our InsetGAN improved results, we also followed Kynkanniemi et al. [3] and evaluated the precision and recall score. Precision and Recall provide a more disentangled way of mapping quality and variability of samples. These metrics are a very intuitive quantitative evaluation tool for GANs. Of the two calculated values, precision describes a measurement of image quality (higher=better quality), and recall quantifies the variability of the generated images (high=more variability). Both metrics are scaled between 0 and 1.

We evaluate these scores on more ($t$ = 0.4) and less ($t$ = 0.7) truncated results to observe the impact of improving overall full-body generation quality at the cost of lowering the variability. All evaluations are performed both on the full-body image as well as on a crop area around the face region that includes a border around the pasted region to evaluate the image coherence. These generated images are compared to the dataset to evaluate a precision and recall score. We calculate the baseline as the precision & recall of unconditional generation of our model (1), and then evaluate the scores for two different datasets used for improving the face region: (2) the pretrained FFHQ face generator and (3) our custom face dataset trained on the same data as the full-body generator as shown in Fig. 3.

| $t$ = 0.7 | **Full-body Image** | | **Face Crop Area** | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| (1) unconditional | 0.6958 | **0.3280** | 0.7980 | **0.2570** |
| (2) FFHQ | 0.8293 | 0.3126 | 0.8522 | 0.1624 |
| (3) our dataset | **0.8364** | 0.3076 | **0.8891** | 0.1576 |
| $t$ = 0.4 | **Full-body Image** | | **Face Crop Area** | |
| | Precision | Recall | Precision | Recall |
| (1) unconditional | 0.9247 | 0.0334 | 0.9206 | **0.0552** |
| (2) FFHQ | **0.9333** | 0.0336 | 0.9386 | 0.0180 |
| (3) our dataset | 0.9298 | **0.0362** | **0.9541** | 0.0182 |

We can see that our method is able to achieve a significant improvement in precision throughout all experiments. The increase in precision is particularly large for the less-truncated ($t$ = 0.7) experiments exhibiting more artifacts in the unconditional generator, where we are able to improve the precision by a large margin using our own model. We can also achieve a comparable improvement using the FFHQ model, which is somewhat surprising since the training data is not based on the same input distribution as the full-body generator. This shows (a) that the generative capabilities of well-trained GANs are providing powerful generalizable models of their domain and (b) that our method is able to encourage good results even for specialized part generators that are trained on a completely different distribution. Note that our improvements come at a cost of a

small drop in recall, which denotes that the variability of the samples goes down a little in most instances. The drop is insignificant when measured on the full body, yet noticeable when calculating the metrics on the cropped face area. We attribute this drop in recall to the fact that we reduce artifacts caused by outliers and unusual generated samples in the images, which decreases the variability in the samples. Our results on larger truncation ($t = 0.4$) paint a similar picture, albeit with smaller margins in the improvement, as the generator is significantly more restricted at this setting. We can also see that the Recall values at this truncation level are already extremely low.

**Issues of FID and Training Data.** As shown in the main paper, the FID score [1] of unconditionally generated samples is similar to that of InsetGAN improved samples. We think there are two main reasons: (1) Even though FID is the most commonly used metric for evaluating image generation quality, it does not correlate well with human perceptual quality and cannot effectively capture subtle visual differences as discussed in [5]. Our user study results also contradict results based on FID because in the user study our InsetGAN results are clearly preferred by the users but FID cannot properly reflect the quality improvements; (2) Our dataset contains photographs of varying quality and resolution, ranging from high-resolution studio quality photographs to low-lighting cellphone snapshots. Additionally, human subjects might only occupy small regions of the original photographs. After cropping and resampling, artifacts can be quite visible and sometimes magnified. For instance, as shown in Fig. 6 left, we observe JPG artifacts, motion blur and noisiness caused by low-lighting condition. Our face GAN trained on cropped face regions from this dataset can alleviate some of these artifacts when used to improve the generated faces from the trained human GAN as shown in Fig. 6 right. We notice a good number of our randomly-sampled 4K training images used for FID evaluation contain artifacts. The human GAN generated faces that contain more artifacts might accidentally have more similar distribution to the training set than the nice clean faces generated by the face GAN used in our joint optimization.

**Additional Comparison with CoModGAN.** In Fig. 7, we show an additional comparison of our method to CoModGAN where we evaluate the quality of our "inpainting" capabilities after removing the constraint that the output face needs to be similar to the underlying input face. We only optimize the face latent code based on the edge coherence term and keep the body latent code fixed. This makes the comparison fairer, since CoModGAN invents completely new faces based only on the context pixels outside the input bounding box and does not alter the pixels of the body. We show that we are able to generate plausible and coherent results without using the input face as guidance and without joint optimization.

Figure 4. **Face Body Montage.** Given faces *(top row)* generated by a pretrained FFHQ model and bodies *(left column)* synthesized by our full-body human generator, we apply *joint* latent optimization to find compatible face and human latent codes that are combined to produce coherent full-body humans.

**User Study Details.** We provide additional details about the user studies we conducted on Amazon Mechanical Turk. We adopt a forced choice paired comparison procedure where the participant is shown a pair of images at a time and is asked to "select in which of the two images the person looks more plausible and real" as seen in Fig. 8. For each HIT (Human Intelligence Task), we randomize the pair order and whether our result is on the left or right.

Figure 5. **Face Body Montage.** Given faces *(top row)* generated by a pretrained FFHQ model and bodies *(left column)* synthesized by full-body DeepFashion [4] generator, we apply *joint* latent optimization to find compatible face and human latent codes that are combined to produce coherent full-body humans.

We performed four different independent studies:

(1) Compare unconditionally generated samples (truncated with $t = 0.4$) with images in our training set.

(2) Compare unconditionally generated samples ($t = 0.4$) with the results of joint InsetGAN optimization for face refinement.


Figure 6. **Dataset Quality.** We show a comparison of faces cropped from our dataset *(left)* with faces sampled from unconditionally generated and InsetGAN-improved humans *(right)*. Zoom in to observe the variable quality of the input data.

Figure 7. **Improved Comparison with CoModGAN.** We remove the conditional constraint on the face and allow for unconditional (only edge-conditional) face insertion. In contrast to Fig. 9 of the main paper, we see that the face is allowed to diverge from the face input. We also keep the body latent fixed so that the body pixels in both our results and CoModGAN results remain unchanged.

(3) Compare unconditionally generated samples ($t = 0.4$) with the results of using CoModGAN for face regeneration.

(4) Compare our InsetGAN results with CoModGAN results directly.

In study (1) the images are unpaired since there is no correspondence between any generated image and any training image. Given two images, we show the first one for a second and then the other one for another second. The images are rendered at 384×768px resolution so that they fit into the browser without the need of scrolling.
For studies (2), (3) and (4), we show 512×1024px center-cropped images side-by-side to the participants so that they can focus on the differences of the image details. The selection buttons above the image pairs are faded in after a 6-second delay, so that users are encouraged to carefully study the image differences before making their selections. We collect 5 votes per image pair and choose the winner image that receives 3 or more votes. We summarize the results into the following table:

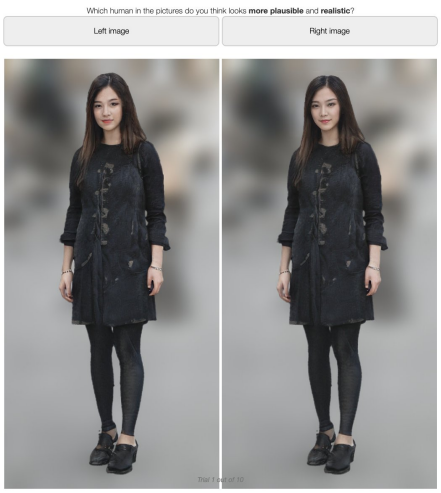


Figure 8. **User Study Interface.** We show the web interface presented to participants of our user study for task (2).

| Study | A | | B | |
|---|---|---|---|---|
| Real (A) / Generated (B) | 438 | 87.6% | 62 | 12.4% |
| Generated (A) / InsetGAN (B) | 10 | 2.0% | 490 | 98.0% |
| Generated (A) / CoModGAN (B) | 465 | 93.0% | 35 | 7.0% |
| InsetGAN (A) / CoModGAN (B) | 495 | 99.0% | 5 | 1.0% |

## 3. Implementation Details

**Unconditional Generation and Adaptive Truncation.** Since our generator is trained on very diverse data, we can observe a wide range of image quality when generating untruncated output. When truncating the generated results as described in the original StyleGAN2 paper [2] by linearly interpolating from the sample position in $\mathbf{w}$ space to the average latent $\mathbf{w}_{avg}$ we can drastically reduce artifacts in pose and details. However, this trick also reduces the diversity in the sample output, and notably reduces the color vibrancy of the output images, as outfit colors are interpolated towards an averaged greyish hue. In our approach, whenever possible (i.e. whenever we are not constrained to operate in the $\mathbf{w}$ space), we use a layer-adaptive truncation scheme to generate visually pleasing result of improved perceptual quality while preserving as many diverse features as possible from the untruncated samples, as shown in Fig. 9.

To achieve this, when generating unconditional samples, we use the $\mathbf{w}^+$ space and define a separate truncation value for each layer. In our generator, we have 18 layers, and we define the layer-wise truncation values as

$$t = [0.35, 0.25, 0.25, 0.70, 0.75, 0.65, 0.65, 0.40, 0.40, \\ 0.35, 0.25, 0.15, 0.15, 0.05, 0.05, 0.05, 0.05, 0.05]$$

The values were chosen through experimentation where we truncate individual layers separately to identify the ones that cause the most artifacts. Note that we apply almost no truncation on later layers, as they can be used to generate desirable clothing details, vibrant colors and accessories and do not cause significant artifacts. We observe in our experiments that latent codes for the middle layers (4-7) of the network are most responsible for artifacts, so we truncate them the most.

Figure 9. **Adaptive Truncation.** We show a set of untruncated samples from our human generator exhibiting unrealistic poses and unwanted artifacts. Standard truncation ($t$=0.6, *bottom row*) reduces artifacts, but also removes desirable clothing details and reduces the color vibrancy. Our adaptive truncation *(center row)* better preserves colors, texture details and accessories.

We also measure the Fréchet Inception Distance (FID) of 4K random results generated using our adaptive truncation scheme and observe a significantly lower FID (53.26) as compared to using regular truncation at $t$=0.6 (71.89). We would like to point out that we did not use the adaptive truncation trick when we performed the quantitative evaluations in the main paper, both for clarity and simplicity and because we were optimizing in $\mathbf{w}+\delta_i$ space, which restricts the effect of adaptive truncation.

**Optimization details.** All our results were optimized using ADAM. We usually stop the optimization when the edge loss falls below a certain threshold (usually defined as $L_1(border)(\mathbf{w}_B) < 0.09$) or after the number of iterations exceed a threshold (typically 1000 optimization steps). When performing joint optimization, we define two distinct optimizers for $\mathbf{w}_A$ and $\mathbf{w}_B$ and switch the optimization target every 50 iterations. Depending on the application, we can start with the canvas optimizer or the inset optimizer. We choose different learning rates for the canvas optimizer: $lr = 0.05$ and for the inset optimizer: $lr = 0.002$. We reevaluate the bounding box every 25 iterations during optimization for a certain number of iterations (typically 150 iterations during body generation, 75 iterations during face refinement.) before keeping the bounding box fixed. We observe that reevaluating the bounding box too often or too long makes the optimization unstable.

**Lambda weights for Losses.** We report the $\lambda$ weight combination we use for the face body montage application. In this use case, we have losses for improving the coherence of $\mathcal{G}_A$ and $\mathcal{G}_B$ from the perspective of each GAN, as well as losses for controlling the appearance of each output, either by constraining closeness of the center image region (face) to a target or some outer image region (body) to adhere to a specific body.

| $\lambda$ | Loss Description | $L_{body}$ | $L_{face}$ |
|---|---|---|---|
| $\lambda_1$ | $L_1$ | 500 | 500 |
| $\lambda_2$ | $L_{lpips}$ | 0.05 | 0.05 |
| $\lambda_3$ | $L_{lpips}(border)$ | - | 0.1 |
| $\lambda_4$ | $L_1(border)$ | 2500 | 10000 |
| $\lambda_{r1}$ | $L_{reg}$ | 25000 | - |
| $\lambda_5$ | $L_1(target\_body)$ | 9000 | - |
| $\lambda_6$ | $L_{lpips}(target\_body)$ | 0.1 | - |
| $\lambda_7$ | $L_1(target\_face)$ | - | 5000 |
| $\lambda_8$ | $L_{lpips}(target\_face)$ | - | 1.75 |

We define several different optimization targets that require custom parameters and setup:

1. **Improving the face area of a given human image.** We either run one-way optimization or take a small learning rate for the human optimizer to allow for a small wiggling of the canvas area around the inset, which generally improves the coherence of inset and canvas. We start with optimizing the inset from a random starting point for a certain number of iterations (e.g. $n$=100) and then fix the inner face area by adding an additional loss constraint keeping the face close to the remembered state. This allows more iterations for the boundary area to improve but prevents overfitting to unwanted artifacts in the face region of the input canvas or deterioration of the facial quality due to over-optimization.

2. **Finding suitable bodies for a given input face.** We start with optimizing the canvas from a random starting point for a certain amount of iterations (e.g. $n$=150), allowing the body generator to roughly hallucinate a person with similar facial structure as the target. Then, we switch to an alternating optimization schedule, allowing the appearance of body and face to gradually resemble each other in the boundary regions. We can regularize the appearance of the body using a similar strategy as described above to avoid over-optimization.

3. **Seamlessly combine a given face and body** In order to maintain the appearance of both the input face and the body, we constrain the joint optimization so that the face GAN result stays close to the input face and the body GAN result outside of the face stays close to the input body.

## References


[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. 5

[3] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[5] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via Co-Modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021. 3