SmileGAN-PTI: Counterfactual Image Generation for Visual Attribute Editing

Author: Mohammad Mosaffa (mm3322)

Introduction:

This report presents the counterfactual image generation framework that underpins our broader research agenda on measuring political polarization in visual media (Mosaffa et al., 2025). While our overarching goal is to estimate how specific visual attributes influence the polarized portrayal of political figures, this methodological component focuses on generating controlled counterfactuals—images in which a single attribute (such as a smile) is systematically modified while all other aspects remain fixed.

Conventional approaches often rely on the two-step approach: first, extracting features from images using pre-trained computer vision models, and then applying reduced-form regressions to estimate the parameter of interest. However, this approach risks substantial information loss, bias, and confounding, particularly when working with high-dimensional, unstructured image data. To address these limitations, our proposed algorithm, *Polarization Measurement Using Counterfactual Image Generation (PMCIG)*, leverages Generative Adversarial Networks (GANs) to directly create counterfactual image pairs.

Identification Strategy:

In this part, we briefly outline the identification strategy that underlies our measurement framework, focusing on how we isolate the effect of specific visual attributes on media portrayal.

To quantify visual polarization, we formalize the utility difference that news outlets derive from the visual presentation of a politician. Let Z denote the original image, and let T be a binary indicator representing the presence ($T = 1$) or absence ($T = 0$) of a specific visual attribute, such as a smile. For a given politician p and news outlet y, we define the utility difference associated with turning the attribute "on" versus "off" as:

$$\Delta_T u(p,y) = \mathbb{E}_X \left[u\big(Z(T=1), X, p, y\big) - u\big(Z(T=0), X, p, y\big) \,\big|\, p, y \right],$$

where X captures additional contextual variables (e.g., title, date of publication). We then express the visual polarization between two outlets y_1 and y_2 as:

$$\rho_T(p, y_1, y_2) = \Delta_T u(p, y_1) - \Delta_T u(p, y_2),$$

which quantifies how differently the two outlets value the presence of the target attribute in their portrayal of the politician.

A critical element of our framework is the ability to generate realistic counterfactual images. For an input image Z and binary attribute T, we define two operators:

- π_1 **Operator:**

  $$\pi_1(Z) = \big(T = 1, Z(-T)\big) \equiv Z_1,$$

  which produces a counterfactual image where the attribute is activated (e.g., the politician smiles), while holding all other features $Z(-T)$ fixed.

- π_0 **Operator:**

$$\pi_0(Z) = \big(T = 0, Z(-T)\big) \equiv Z_0,$$

which returns the baseline image without the attribute (i.e., the neutral expression).

By comparing the change in the log-odds of outlet choice probabilities between Z_1 and Z_0, we obtain an empirical estimate of the utility shift induced by the attribute:

$$\rho_T(p, y_1, y_2) \approx \log\left(\frac{g(y_1 \mid Z_1, X, p)}{g(y_2 \mid Z_1, X, p)}\right) - \log\left(\frac{g(y_1 \mid Z_0, X, p)}{g(y_2 \mid Z_0, X, p)}\right).$$

This setup ensures that the only systematic difference between the two images is the activation of T, providing a robust basis for isolating and measuring visual polarization effects. A detailed explanation of the operators π_1, along with their technical implementation, follows in the next section.

1 Operator π_1:

The purpose of the π_1 operator is to transform a neutral facial image into a smiling counterpart while preserving all other visual attributes unchanged. In the following, we first outline the key technical challenges associated with implementing this operator in Section 1.1, and then introduce our proposed solution in Section 1.2.

1.1 Challenges:

Adding a smile to real-world images poses significant technical challenges. While GANs have become a foundational tool for realistic image editing, their direct application to attribute-specific manipulations—such as adding a smile to a face—introduces two central challenges that must be addressed.

- **Challenge 1:** First, while conditional GANs are capable of manipulating specific visual attributes such as facial expressions (Mirza and Osindero, 2014; Choi et al., 2018), they regenerate the entire image in the process—including both the target feature and the surrounding context. As a result, they are prone to introducing unintended changes to contextual elements such as background, lighting, or hairstyle alongside the facial edit. This poses a fundamental problem for our identification strategy, which relies on the assumption that only the treatment variable T (e.g., the presence of a smile) is altered, while all other image attributes $Z(-T)$ remain fixed. When contextual or non-target features are unintentionally modified, it violates this assumption and undermines the validity of the counterfactual comparison.

- **Challenge 2:** Second, the computational demands of well-trained cGANs pose a key bottleneck. High-resolution editing typically requires multi-stage procedures, including attribute manipulation and generator fine-tuning, all of which are computationally expensive and memory-intensive. For example, state-of-the-art models, such as StyleGAN2 (Karras et al., 2020), require over 300 million parameters and consume approximately 1–2 GB of GPU memory per image generation at 1024×1024 resolution, with latent inversion procedures often taking several minutes per image on a modern GPU (Abdal et al., 2019; Roich et al., 2022). These costs multiply when fine-tuning is incorporated, as optimization typically spans hundreds to thousands of gradient descent steps. Importantly, standard pipelines do not inherently guarantee that identity is preserved after editing, which is a critical requirement in our setting: to compare the treatment and control images of the same subject, we must ensure that only the target attribute changes while all other personal and contextual features remain fixed.

1.2 Our Approach, SmileGAN-PTI:

To overcome challenges in Subsection 1.1, we propose SmileGAN-PTI, which is a two-phase framework designed to overcome the challenges of semantic control and computational efficiency in generating counterfactual smile edits. An overview of the algorithm is shown in Figure 1.

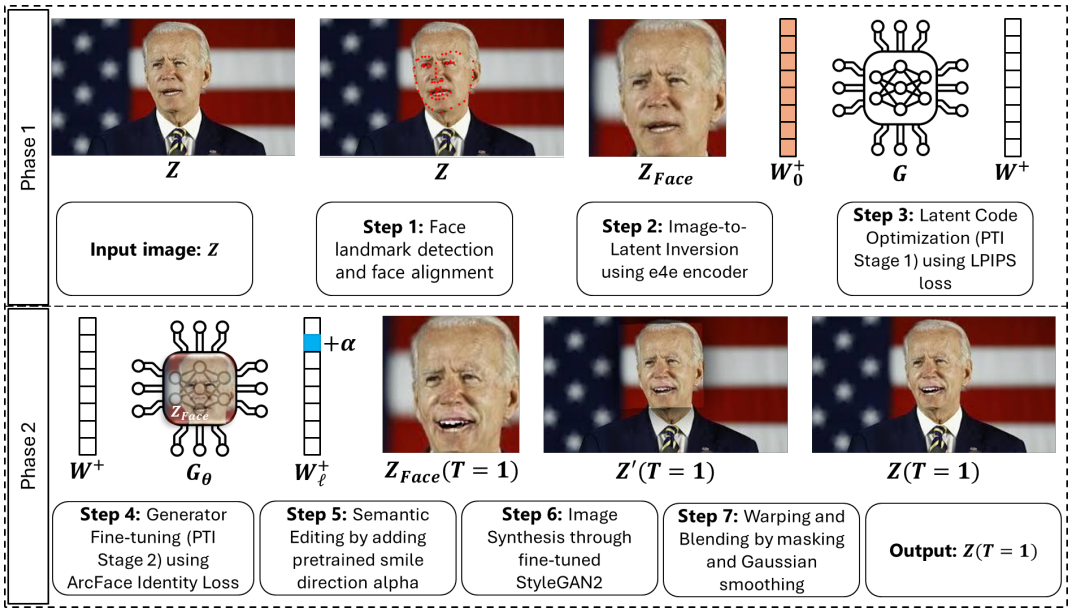

Figure 1: SmileGAN-PTI Model

In the first phase, we apply a robust face alignment and cropping procedure using a landmark-based alignment model (Bulat and Tzimiropoulos, 2017; Zhu et al., 2016) to normalize the input images. This alignment ensures that the facial region is consistently centered and scaled, effectively eliminating background and contextual distractions. By tightly cropping around the face, we address the first challenge of preserving contextual information: we explicitly minimize the risk that downstream edits will inadvertently modify non-target regions such as the background or scene layout, which is critical for maintaining the validity of counterfactual comparisons.

In the second phase, we use a pre-trained StyleGAN2 generator (Karras et al., 2020), a leading architecture known for producing high-resolution, photorealistic faces with disentangled latent spaces. We first perform latent inversion with the e4e encoder (Tov et al., 2021), which maps real images into the StyleGAN2 latent space. Next, we edit the latent codes along a learned smile direction to modify facial expression. To preserve identity, we apply Pivotal Tuning Inversion (PTI) (Roich et al., 2022), which fine-tunes the generator on each input image. This two-phase framework offers high semantic precision and avoids training models from scratch. It also keeps computational costs low by limiting fine-tuning to small parts of the network. As a result, our method generates high-quality smiles, preserves non-target features, and scales well to large datasets—ideal for counterfactual image generation in our setting. Next, we discuss the algorithm in detail.

Algorithm 1 SmileGAN-PTI (Operator π_1)

Input: Input image Z

Output: Final counterfactual image $Z(T=1)$ with edited smiling face

Step 1: Input and Alignment

Detect facial landmarks $L = \{l_i\}_{i=1}^{68}$ using a pretrained detector.

Compute similarity transform $\mathscr{T}$ to align L to canonical landmarks L^*.

Warp cropped face region:

$$Z_{\text{Face}} = \text{warp}(Z, \mathscr{T}^{-1}).$$

Step 2: Image-to-Latent Inversion

Encode aligned face region into latent space:

$$w_0^+ = E_{\text{e4e}}(Z_{\text{Face}}).$$

Step 3: Latent Code Optimization

Refine latent code to match the aligned face:

$$\min_{w^+} \|G(w^+) - Z_{\text{Face}}\|_2^2 + \lambda_{\mathscr{L}} \cdot \mathscr{L}(2G(w^+) - 1, 2Z_{\text{Face}} - 1).$$

Step 4: Generator Fine-tuning

Optimize generator weights for accurate reconstruction:

$$\begin{aligned}\min_{\theta_G} &\|G_{\theta_G}(w^+) - Z_{\text{Face}}\|_1 + \lambda_{\mathscr{L}} \cdot \mathscr{L}(2G_{\theta_G}(w^+) - 1, 2Z_{\text{Face}} - 1) \\ &+ \lambda_{\text{mouth}} \|M \odot (G_{\theta_G}(w^+) - Z_{\text{Face}})\|_1 + \lambda_{\mathscr{I}} \cdot \mathscr{I}\big(F(G_{\theta_G}(w^+)), F(Z_{\text{Face}})\big),\end{aligned}$$

Step 5: Semantic Editing

Introduce smile edit by modifying latent code:

$$\forall \ell \in [4,6], \quad w_\ell^+ \leftarrow w_\ell^+ + \alpha \cdot v_{\text{smile}}.$$

Step 6: Image Synthesis

Generate edited smiling face crop:

$$Z_{\text{Face}}(T=1) = G_{\theta_G}(w^+).$$

Step 7: Warping and Blending

Compute inverse warp $\mathscr{T}^{-1}$ to map $Z_{\text{Face}}(T=1)$ back to the original image space.

Generate Gaussian-smoothed spatial mask m.

Blend into original image:

$$Z(T=1) = (1-m) \odot Z + m \odot \text{warp}(Z_{\text{Face}}(T=1), \mathscr{T}).$$

return $Z(T=1)$

1.3 Algorithm:

Algorithm 1 provides an overview of the proposed SmileGAN-PTI. In the following, we describe each step of the algorithm in detail, outlining the techniques employed, their purpose, and their contribution

to the overall framework.

- *Step 1: Input and Alignment (Face Alignment):*

  Given an input image Z, we first detect 2D facial landmarks $L = \{l_i\}_{i=1}^{68}$ using a pretrained facial landmark detector. To ensure that downstream processes, particularly the StyleGAN2 generator, receive inputs with consistent pose, scale, and orientation, we compute a similarity transform $\mathscr{T} : \mathbb{R}^2 \to \mathbb{R}^2$ that aligns the detected landmarks L to a canonical landmark configuration L^*. We then extract and warp the face region into this canonical space:

  $$Z_{\text{Face}} = \text{warp}(Z, \mathscr{T}^{-1}),$$

  where $\text{warp}(\cdot)$ applies the inverse transform. The purpose of this alignment step is twofold: (i) to standardize face geometry across inputs, reducing variability in pose and scale that could confound latent inversion and editing, and (ii) to minimize unintended changes to the non-target regions of the image, thereby satisfying the identification assumption that the contextual background $Z(-T)$ remains fixed when modifying only the target attribute T (smile).

- *Step 2: Image-to-Latent Inversion (Latent Inversion with e4e):*

  Given the aligned face region Z_{Face}, we project it into the latent space $\mathscr{W}^+$ of StyleGAN2 using the e4e encoder E_{e4e} such that:

  $$w_0^+ = E_{\text{e4e}}(Z_{\text{Face}}),$$

  where w_0^+ denotes the per-layer latent codes across the 18 layers of the generator. The purpose of this stage is to embed the real face image into a latent representation where precise and semantically meaningful edits can be applied. The e4e encoder is designed to balance reconstruction fidelity and editability, producing latent codes that preserve the identity and appearance of the input while enabling downstream manipulations. This initialization avoids optimization from random codes, improves computational efficiency, and provides a high-quality starting point for later optimization and editing stages.

- *Step 3: Latent Code Optimization (PTI Stage 0):*

  Starting from the initial latent code w_0^+, we refine the latent representation by minimizing a combined objective of pixel-wise and perceptual similarity losses:

  $$\min_{w^+} \|G(w^+) - Z_{\text{Face}}\|_2^2 + \lambda_{\mathscr{L}} \cdot \mathscr{L}(2G(w^+) - 1, 2Z_{\text{Face}} - 1),$$

  where G is the fixed StyleGAN2 generator, Z_{Face} is the aligned face region, and $\mathscr{L}(\cdot,\cdot)$ is the Learned Perceptual Image Patch Similarity (LPIPS) loss (Zhang et al., 2018) computed over $[-1,1]$-normalized images. The hyperparameter $\lambda_{\mathscr{L}}$ balances pixel-level fidelity and perceptual quality. Optimization is performed using AdamW.

  Pivotal Tuning Inversion (PTI) (Roich et al., 2022) is a two-stage optimization framework that first refines the latent code and then fine-tunes the generator to better match a target image, enabling precise editing while preserving identity. LPIPS is included because standard L2 losses alone fail to capture perceptual differences meaningful to human observers; LPIPS computes distances in deep feature space, providing a better measure of high-level visual similarity.

- *Step 4: Generator Fine-tuning (PTI Stage 2):*

  In this stage, we fine-tune the generator weights θ_G to improve the reconstruction fidelity of the aligned face image Z_{Face}, using the optimized latent code w^+. The optimization objective is formulated as:

$$\min_{\theta_G} \|G_{\theta_G}(w^+) - Z_{\text{Face}}\|_1 + \lambda_{\mathscr{L}} \cdot \mathscr{L}(2G_{\theta_G}(w^+) - 1, 2Z_{\text{Face}} - 1) + \lambda_{\text{mouth}} \|M \odot (G_{\theta_G}(w^+) - Z_{\text{Face}})\|_1 + \lambda_{\mathscr{I}} \cdot \mathscr{I}\big(F(G_{\theta_G}(w^+)), F(Z_{\text{Face}})\big),$$

  where $G_{\theta_G}(w^+)$ is the generator output, M is a binary mouth-region mask, $\mathscr{L}(\cdot,\cdot)$ is the perceptual LPIPS loss, and $\mathscr{I}(\cdot,\cdot)$ denotes the identity loss computed using ArcFace IR-SE50 embeddings $F(\cdot)$. The weights $\lambda_{\mathscr{L}}, \lambda_{\text{mouth}}, \lambda_{\mathscr{I}}$ control the balance between terms. The pixel-wise ℓ_1 loss encourages accurate low-level reconstruction, helping to preserve edge sharpness and fine details. The mouth-region loss, guided by the mask M, restricts optimization to the area of semantic change, preventing unwanted edits to background and identity. Finally, the identity loss $\mathscr{I}$, computed using ArcFace embeddings (Deng et al., 2019), enforces that the generator preserves the subject's identity—an essential requirement for valid counterfactual comparisons. This optimization is performed using AdamW, allowing the generator to adapt specifically to the input image while maintaining both local accuracy and global consistency.

- *Step 5: Semantic Editing (Smile Editing in Latent Space):*

  In this step, we modify the optimized latent code w^+ to introduce a smile by adding a pretrained smile direction vector v_{smile} to selected latent layers. Formally, for each layer ℓ in the target range,

$$w^+_\ell \leftarrow w^+_\ell + \alpha \cdot v_{\text{smile}},$$

  where α is a scalar controlling the intensity of the smile transformation. This targeted latent manipulation enables precise semantic editing while preserving non-target attributes such as identity, pose, and background. By restricting the edit to mid-level layers related to the smile manipulation (typically layers 4–6), the method ensures that global structure and fine details remain intact.

- *Step 6: Image Synthesis (High-resolution Image Generation):*

  In this stage, we generate the edited face image by performing a forward pass through the fine-tuned generator. Specifically, we compute:

$$Z_{\text{Face}}(T{=}1) = G_{\theta_G}(w^+),$$

  where G_{θ_G} is the generator with fine-tuned weights and w^+ is the latent code modified with the smile edit. The purpose of this stage is to translate the semantic edit in latent space into a high-resolution, photorealistic image of the face with the smile attribute $T = 1$. The forward pass ensures efficient image synthesis while preserving the fine details, facial identity, and background consistency achieved in earlier stages.

- *Step 7: Warping and Blending (Face-to-original Mapping and Blending):*

  Finally, we project the edited smiling face $Z_{\text{Face}}(T{=}1)$ back into the original image context Z. We compute the inverse similarity transform $\mathscr{T}^{-1}$ to map the edited face from the aligned space to the original image space. A spatial mask m is generated, typically centered on the face region, and

smoothed using a Gaussian filter to ensure soft transitions at the boundaries. The final composite image is computed via alpha blending:

$$Z(T{=}1) = (1-m) \odot Z + m \odot \text{warp}(Z_{\text{Face}}(T{=}1), \mathscr{T}),$$

where Z is the original input image, and $\text{warp}(Z_{\text{Face}}(T{=}1), \mathscr{T})$ maps the smiling face back into the original coordinate space. This stage ensures that the edited face is seamlessly integrated into the full-resolution photograph, preserving the original background and contextual cues. The combination of geometric warping and smooth alpha blending minimizes visual artifacts, producing a photorealistic final image that retains both local edits and global consistency.

1.4 Results:

We present two sets of empirical results to evaluate the effectiveness of the proposed *SmileGAN-PTI* framework. The first set consists of qualitative visual outputs, showcasing the algorithm's ability to add realistic smiles to the faces of prominent political figures. The second set provides a quantitative illustration of how the strength of the smile manipulation scales with the latent space editing parameter α.

1.4.1 Qualitative Evaluation of Smile Edits:

Figure 2 displays a set of 6 politicians, where the left image in each pair shows the original neutral face Z, and the right image shows the counterfactual smiling face $Z(T=1)$ generated by our algorithm. The dataset includes a diverse range of political figures, including Marco Rubio, Nancy Pelosi, Joe Biden, Ted Cruz, Kamala Harris, Donald Trump, and Barack Obama. Across all examples, *SmileGAN-PTI* successfully applies the smile transformation with $\alpha = 1$ while preserving key non-target attributes, such as identity, pose, hairstyle, clothing, and background.

The results demonstrate the robustness and generalizability of the framework across demographic and photographic variability. Notably, the generated smiles are photorealistic, with fine-grained detail around the mouth and eyes, avoiding common pitfalls such as artifacts, identity drift, or context distortion. This visual consistency is particularly important for downstream tasks that rely on counterfactual analysis, as it ensures that only the target attribute T (smile) is manipulated while $Z(-T)$ remains stable.

1.4.2 Quantitative Analysis of Smile Intensit:

To further assess the controllability of the smile transformation, we conducted an experiment varying the latent manipulation strength parameter α. Figure 3 visualizes how increasing α progressively amplifies the smile intensity, generating a smooth continuum from neutral to highly expressive smiles.

The α parameter controls the magnitude of the shift along the pretrained smile direction v_{smile} in latent space:

$$w_\ell^+ \leftarrow w_\ell^+ + \alpha \cdot v_{\text{smile}},$$

for selected layers ℓ. As shown in the figure, small values of α yield subtle expression changes, while larger values induce broader smiles with more pronounced mouth and eye movements. Importantly, the generator maintains photorealism and identity preservation across a wide range of α, highlighting the stability of the editing mechanism. However, we observe that very high α values push the edited images toward out-of-distribution regions, occasionally producing visually abnormal results such as exaggerated mouth deformations or unnatural facial asymmetry. This phenomenon reflects the limits of linear manipulation in latent space, where extreme edits can move samples beyond the data manifold learned during

Figure 2: Qualitative examples of smile edits using SmileGAN-PTI. Left: original neutral image Z; right: smiling version $Z(T = 1)$.

GAN training. These findings emphasize the importance of calibrating α carefully in applied settings to balance the strength of the semantic transformation against the risk of introducing unrealistic artifacts.

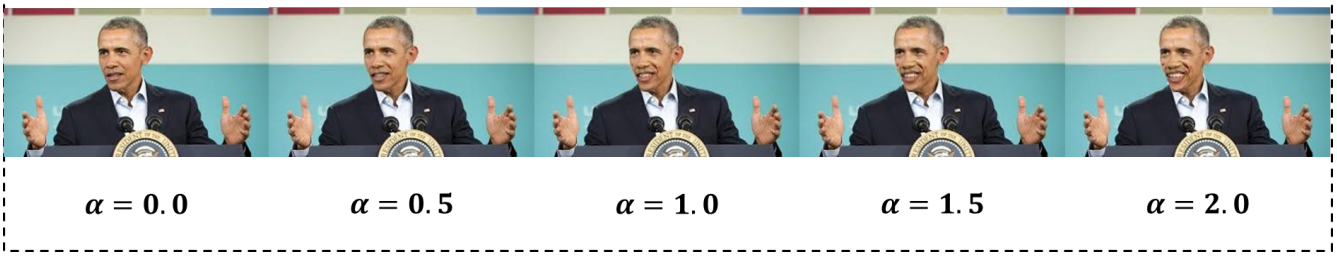

Figure 3: Effect of varying α on smile intensity. Leftmost: neutral face; rightward progression: increasing smile strength.

1.5 Discussion:

The proposed *SmileGAN-PTI* framework produces high-quality, photorealistic smile edits while preserving identity, pose, and background. This is essential for valid counterfactual analysis in image-based research. A key novelty is the integration of face alignment, e4e inversion, PTI fine-tuning, and latent space editing into a unified framework. While the current focus is on smiles, the framework is general and can handle other facial attributes T by learning suitable latent directions. The framework has some advantages. It balances semantic precision and computational efficiency without requiring full model retraining. It works well across diverse subjects and images. However, it has limitations. Linear latent edits can produce unrealistic artifacts at extreme levels. The PTI step, although efficient, still adds some computational cost, especially for large datasets. Future work will explore nonlinear or diffusion-based editing, automatic tuning of the α parameter, and multi-attribute counterfactual generation.

Codes can be found in: https://colab.research.google.com/drive/1nqji3ymQvORaAYd5AGduIubQ9Oc-cuV8?usp=sharing

References

Abdal, R., Qin, Y., and Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? In *International Conference on Computer Vision*.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699.

Karras, T., Laine, S., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. In *Advances in neural information processing systems*.

Mosaffa, M., Rafieian, O., and Yoganarasimhan, H. (2025). Visual polarization measurement using counterfactual image generation. *arXiv preprint arXiv:2503.10738*.

Roich, D., Mokady, R., Bermano, A. H., and Cohen-Or, D. (2022). Pivotal tuning for latent-based editing of real images. In *ACM Transactions on Graphics*.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., and Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. In *SIGGRAPH*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595.

Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition*.