Towards Quantitative Evaluation Metrics for Image Editing Approaches

Dana Cohen Hochberg Oron Anschel Alon Shoshan
Igor Kviatkovsky Manoj Aggarwal Gérard Medioni

Amazon

{danacoh, oronans, alonshos, kviat, manojagg, medioni}@amazon.com

Abstract

In the rapidly evolving field of Generative AI, this work takes initial steps towards establishing a systematic approach for comparing image editing methods. Currently, there is a lack of quantitative metrics for evaluating image editing tasks, with new methods being evaluated mostly qualitatively. Our methodology involves three key components: 1) The creation of a large synthetic dataset using GAN-Control, which enables the generation of ground-truth images for consistent edits across different facial identities; 2) A matching procedure that pairs the edited images with their corresponding ground-truth; and 3) Application of the Perceptual Distance metric to matched pairs. We assessed the effectiveness of our proposed framework through a user study and a set of simulation experiments. Our results indicate that our approach can rank image-editing methods in a way that aligns with human judgment. This research seeks to lay the foundation for a comprehensive evaluation framework for image editing techniques in subsequent studies, initiating a dialogue on this topic.

Evaluation Methods	Consistency	Editing	Scalability
LPIPS	✓	✗	✓
L2	✓	✗	✓
ID Similarity	✓	✗	✓
Attribute Classification	✗	✓	✓
User Study	✓	✓	✗
Ours	✓	✓	✓

Table 1. Comparison of various metrics against consistency, editing, and scalability criteria.

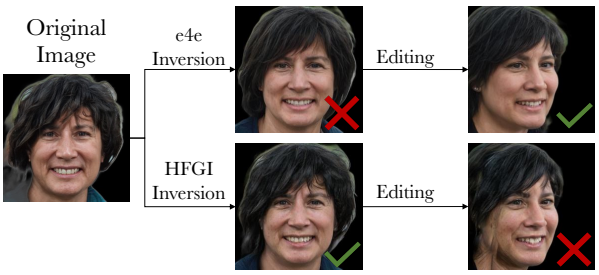

Figure 1. **Inversion vs. Editing Tradeoff.** e4e [14] inversion result is inferior to HFGI's [16]. However, e4e editing is of higher quality than HFGI's.

1. Introduction

In the realm of image synthesis and editing, the advances of generative models, particularly Generative Adversarial Networks (GANs) [5], have enabled a new era of capabilities for creating and altering realistic images. Notably, GAN Inversion techniques [3, 4, 11, 14, 16] have emerged as a pivotal area of research, offering the ability to edit real images by projecting them into a GAN's latent space. This enables adjustments to be made to specific attributes of an image, such as pose, smile, and age, while aiming to preserve the integrity of the original image.

Nowadays, GAN Inversion and editing research faces a significant challenge: the absence of a quantitative framework to objectively assess and compare the performances of these editing techniques, as discussed in [18]. In previous work, and even in recent ones, researchers often resort to selecting favorable examples, conducting user studies, or using proxy evaluation metrics such as identity preservation or measuring attribute-specific classification accuracy between original and edited images. Although user studies provide valuable insights, they lack scalability, are not feasible for every researcher to conduct and are hard to reproduce or extend. Moreover, proxy evaluation metrics may offer some level of evaluation, with some focusing on the consistency of the images and others on measuring the edit degree. However, they often fall short in facilitating direct comparisons or rankings among multiple editing methods, as they do not assess both aspects simultaneously.

Table 1 summarizes the pros and cons of different evaluation methods. In addition in Figure 1 one can observe the difficulty in evaluating different editing methods, and the inherit tradeoff between inversion and editing.

Recognizing the essential role of an evaluation metric in advancing a research field, this paper initiates a conversation on potential solutions and introduces an initial method to address this issue. We propose a framework designed to automatically and systematically compare and rank image-editing methods. To showcase our claims, this work specifically concentrates on facial images and several commonly used editing tasks to demonstrate and benchmark various methods against one another.

Our approach consists of three phases. First, we create a vast synthetic dataset by altering specific attributes (*e.g.*, pose, smile, age) using the GAN-Control's [13] capabilities, resulting in numerous identities exhibiting identical changes, including gradual transitions from unedited to fully edited states. This dataset serves as our reference ground truth. Next, we apply the evaluated image editing technique to generate a spectrum of edited images for a given attribute, adjusting the attribute scale to produce a set of images with smooth edits for each identity. The second phase involves aligning the two datasets (ground truth and edited) using attribute-specific classification networks, matching, for example, GAN-Control images with a 15-degree pose adjustment to the closest corresponding angle in the edited images. The final phase involves applying a metric to the aligned images and averaging the results to obtain a single score for each editing degree.

Our proposed evaluation framework is tested through a series of experiments that simulate editing scenarios with progressively degraded quality, encompassing both the fidelity of inversion and the precision of edits. This systematic degradation approach allows us to validate the sensitivity and reliability of our metric, showcasing its capacity to discern subtle differences in editing quality.

We demonstrate our proposed evaluation framework through experiments on three well-known image editing techniques: e4e [14], HFGI [16], and HyperStyle [4]. We then validate the latter comparison by a user study, which reveals a strong correlation between the rankings generated by our method and those derived from human judgments.

We would like to emphasis, that this work does not claim to fully resolve the challenge of comparing image-editing methods but aims to kick-start a conversation on this crucial topic. We believe that opening this dialogue is vital for further advancements in the field. Furthermore, we introduce a preliminary solution for evaluating and ranking methods, along with a thorough discussion of its strengths and limitations. Finally, we plan to release the dataset generated by GAN-Control to enable reproducibility in future works.

Our contributions are summarized as follows:

- We introduce a framework for assessing image-editing algorithms with certain limitations.
- We conduct a user study, demonstrating a strong correlation between our proposed metric and human preferences.
- We initiate a discussion on developing a structured methodology for evaluating image-editing methods.
- We intend to release a dataset for future study reproducibility.

2. Related Works

2.1. GAN Inversion Techniques

The advances in image synthesis and manipulation have been largely driven by the capabilities of GANs. GAN inversion [21], involves finding a latent representation of a real image within the pretrained GAN's domain. StyleGAN [9, 10], in particular, has achieved prominence for its semantically rich latent space, enabling detailed attribute-based image edits. There are generally three approaches to GAN inversion: direct optimization of the latent vector to minimize reconstruction error for a given image, training an encoder to map an image to its latent representation over a large dataset, and a hybrid approach [20] that combines elements of both direct optimization and encoding.

Direct optimization methods [1, 2] focus on iteratively refining the latent code for each specific image, often leading to highly accurate inversions at the cost of computational efficiency. On the other hand, encoder-based methods [3, 4, 11, 14, 16] aim to learn a more general mapping from image space to latent space, offering faster inversions at the risk of reduced fidelity for individual images.

Encoder-based methods in the context of GAN inversion, represent a strategy where the networks are trained to map an image directly to the latent space of a pretrained GAN. A landmark work within encoder-based methods is e4e [14], which paved the way by focusing on editability, facilitating subtle modifications while preserving the original image's integrity. However, these advancements also brought to light the inversion-editing tradeoff, a balancing act between the accuracy of the inversion and the extent of possible edits, which remains a pivotal consideration in the efficacy of image editing processes (refer to Figure 1 for an example).

Building on the foundation set by e4e, HFGI (High-Fidelity GAN Inversion) [16] was introduced to address the inversion-editing tradeoff. HFGI incorporates a feature distillation approach which involves extracting high-frequency features from the original image throughout the inversion process to preserve high-level details during the inversion process. Meanwhile, HyperStyle [4] introduced a novel concept by utilizing a hypernetwork to dynamically

adapt the weights of a StyleGAN generator, tailoring them to the inverted image. These approaches collectively push the boundaries of what is achievable in terms of both image quality and editability, yet they also highlight the field's need for a nuanced understanding of how to balance the competing objectives of preserving the original image's authenticity and achieving the desired modifications.

In this paper, we exclusively concentrate on encoder-based methods for two main reasons. Firstly, there appears to be a consensus within the research community favoring encoder-based methods, owing to their superior editing performance. Secondly, the inference speed for both inversion and editing processes is much higher. This notably enhanced speed makes encoder-based methods more practical for research and study purposes, offering a considerable advantage over the several minutes per image required by optimization-based methods.

2.2. Editing Evaluation Metrics

In the field of GAN inversion-based image editing, numerous studies have aimed to demonstrate the superiority of their editing methods through various evaluation methods, often accompanied by compelling visual comparisons. A common approach among many previous works [4, 11] involves the use of identity (ID) similarity metrics [7]. These methods involve assessing the cosine-similarity between the original and edited images to evaluate identity preservation. However, while these metrics account for identity loss, they do not address the quality of the edit, the maintenance of original image details, or the presence of image artifacts.

Another metric employed is classification consistency, as discussed in [15], which evaluates whether the intended attribute modification was successfully applied. This is done by using a classifier specifically trained for detecting the edit and its extend. This form of evaluation focuses solely on the accuracy of the specific editing action, without considering unintended changes to the image.

Moreover, metrics such as PSNR, SSIM [17], LPIPS [19], and FID [6], and Perceptual Distance [8], have been utilized in [4, 11, 14, 16] to assess the quality of the inversion. It is important to note that our work also incorporates the Perceptual Distance metric. However, as we will detail in subsequent sections, we apply it in a different manner, comparing edited images with corresponding ground-truth images to provide a more comprehensive evaluation of editing efficacy.

User studies are another common approach for evaluating GAN-based editing techniques, as seen in works such as [11, 16, 18]. In these studies, participants are presented with two or more edited images and are asked to select the one that represents the best editing quality. However, these studies typically cover a limited selection of images, attributes, and editing intensities. In contrast, in our work we chose to use synthetic dataset as a ground-truth benchmark, enabling the evaluation on potentially unlimited number of identities, attributes, and editing degrees.

Although user studies yield important feedback on the subjective quality of edits, their lack of uniform standards and the high costs involved restrict their practicality as a widespread benchmark for assessing editing approaches. Moreover, these studies merely rank the evaluated methods in a relative manner and fail to deliver comparative metrics usable for subsequent studies.

3. Method

Our project introduces a novel evaluation framework designed specifically to assess the effectiveness of GAN inversion methods in editing image attributes such as pose, smile, and age. The core of our method is in utilizing per-attribute synthetic datasets, generated using GAN-Control, as benchmarks for evaluating editing quality. These synthetic datasets, referred to as ground truth (GT) datasets, serve as a standard against which the edited images are compared to gauge the editing quality. An overall illustration of our method is presented in Figure 2.

3.1. Ground Truth Datasets Generation

In the absence of real datasets that isolate changes to a single attribute, we generate synthetic datasets using GAN-Control, which allows for explicit control of specific image attributes. For each individual identity within these datasets, we create multiple images that systematically vary a specific attribute such as pose, smile, or age while holding other attributes constant. Additionally, the image background is removed using face segmentation [12] in order to mitigate ambiguous background changes that are not controllable. This approach allows us to simulate a diverse set of controlled conditions for each attribute, providing a consistent baseline for evaluating the performance of various editing methods. These GT images are crucial as they furnish a standard against which the precision of attribute-specific edits can be measured. Figure 3 displays examples of synthetic images produced by GAN-Control for each attribute.

3.2. Evaluation Method

Our evaluation framework is a multi-step process designed for evaluating the fidelity of attribute-specific edits. It begins with the inversion of a base image representing the subject in a neutral state with respect to the targeted attribute (see Figure 3). This image is projected into the latent space using a selected GAN inversion method to establish a starting point for subsequent edits.

Next, the base image is edited to generate a series of images, each varying in the targeted attribute. This creates a spectrum of edited images from a single base identity, re-

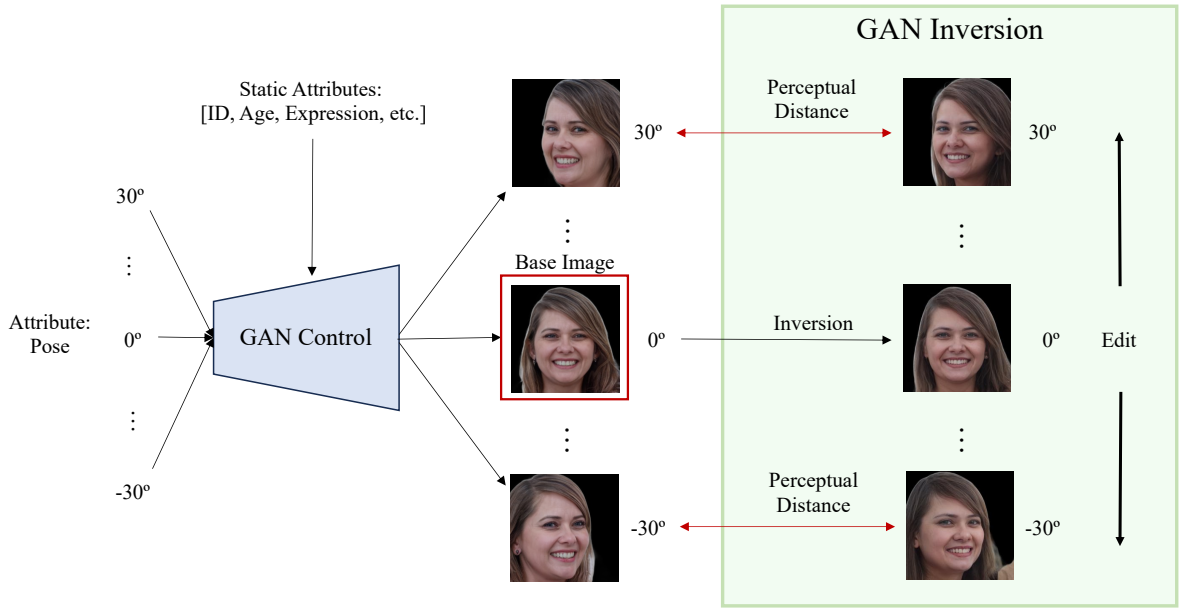

Figure 2. **Evaluation Framework.** GT images are generated using GAN-Control at varying poses (yaw degrees). The base image, specifically noted as yaw$= 0°$, is then inverted and edited using a known GAN inversion method. Finally, we compare the Perceptual Distance between each GT image and its corresponding edited image, *i.e.*, the image with the same attribute as the GT (calculated via an attribute classifier).

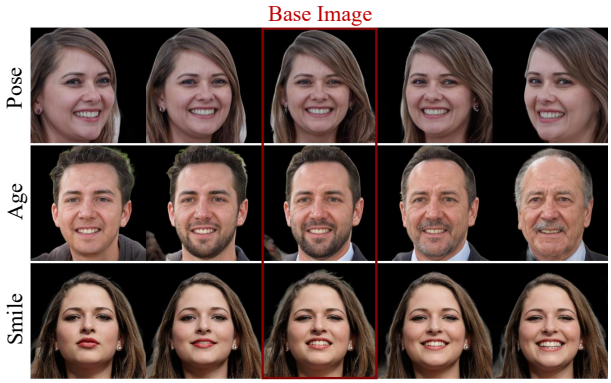

Figure 3. Examples of ground truth images synthesized using GAN-Control, depicting variations of each attribute, including different poses, ages, and expressions. The base image, marked by a red rectangle, serves as the reference for inversion and base for editing within each attribute.

sulting in a diverse set of potential outcomes for comparison.

The algorithm then utilizes an attribute classification network, which assesses the edited images to identify the one that closely resembles the attribute state in the GT image. This correspondence matching procedure is critical as it determines the most accurate representation of the intended attribute change.

The final step involves computing the Perceptual Distance between this selected edited image and the GT image. We note here that the Perceptual Distance is computed between corresponding images, which have similar edits and therefore it captures both the edit quality, and image consistency.

3.3. Method Drawbacks

We identify three major drawbacks in our method:

- GAN-Control inaccuracies.
- Attribute ambiguity.
- Image to image comparison using spatial based metrics.

First, our method relies on GAN-Control for generating multiple images of the same person while a single attribute is varied (*e.g.*, same person with yaw$=-45°, -40°, ..., 45°$) and all other attributes are static (*e.g.*, 45 years old, large smile, brown hair). While GAN-Control exhibits a high level of disentanglement, it is not perfect and editing one attribute might slightly change others, leading to imperfections in the dataset. These slight variations could potentially affect the purity of our ground truth images, introducing a degree of uncertainty in our evaluations. However, we recognize that across a large dataset, these minor inconsistencies tend to average out, thus preserving the overall integrity and reliability of our evaluations. The controlled environment provided by GAN-Control, despite these challenges, enables systematically studying the effects

of specific attribute manipulations.

Second, even if GAN-Control would have perfect control capabilities, there is an inherent ambiguity when editing images. For example given an image of a person at age 20, there could be many possibilities of how this person will look like at age 40. This means that even if both GAN-Control and the method under evaluation have perfect editing capabilities two images produced by both pipelines will probably not be exactly the same.

Finally, comparing images is known to be a complex task [19]. Coupled with the aforementioned issue of ambiguity in image editing, using a metric that fails to consider this ambiguity can be problematic. Even though Perceptual Distance offers a comparison beyond mere pixel-level analysis, it still involves a spatial assessment of images that may significantly differ.

Taking all these three drawbacks into account, there are many reasons for such a framework to fail at its desired purpose of ranking the quality of image editing algorithms. However, in the experiment section we show that despite the drawbacks, the proposed framework is able to correctly rank methods with known quality differences (Section 4.1) and is compatible with human preferences (Section 4.3).

4. Experiments and Results

In this section, we detail the results of two sets of experiments. Initially, to ensure our method offers a reasonable and justified ranking, we artificially conducted a simulated degradation process. This experiment assures that as the quality of the inversion and editing technique decreases, its ranking accurately reflects this decline. Subsequently, we apply our framework to three GAN inversion and editing algorithms, ranking them via the framework and verifying their alignment with human judgment through an extensive user study.

4.1. Ranking Objectively Simulated Methods

It is challenging to validate the effectiveness of the proposed method since there is no set of objectively ranked editing image algorithms. If such a set would exist, we could evaluate it using the new method and check if the results are consistent between the objective ranking and the ranking achieved using our approach, *i.e.*, validate if the results reflect the expected hierarchical order. In this experiment we simulate such a case by artificially degrading the inversion and editing capability of an established editing method. By systematically degrading the method, we effectively create a set of algorithms 'objectively' ranked by their degradation level; the higher the degradation, the lower the method's ranking. We introduce degradation of two distinct forms, which are designed to mimic the effects of varying qualities in inversion and editing techniques and their consequent impact on images. For our evaluation, we use the HyperStyle

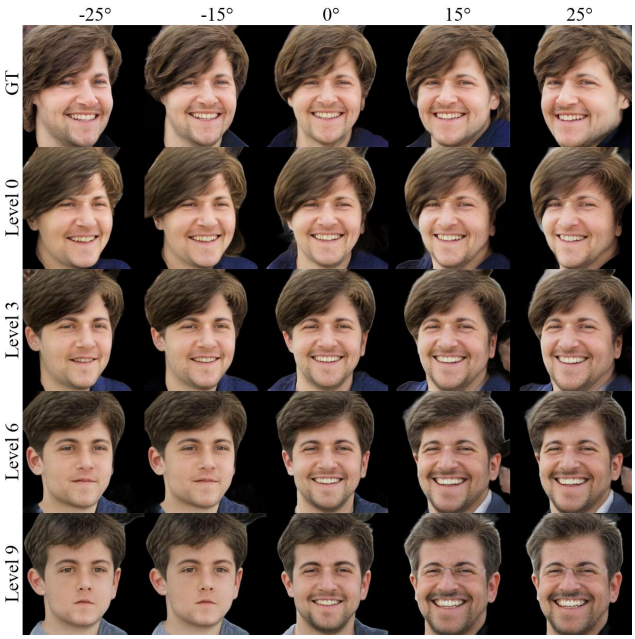

Figure 4. **Degradation in Pose Attribute Inversion and Editing.** The top row features Ground Truth (GT) images that exhibit pose variations, with the image labeled as 0°in the center serving as the inversion baseline. The images surrounding it illustrate varying degrees of editing. At Level 0, the outcomes of both inversion and editing are presented without degradation. With each subsequent level moving downwards, there is a noticeable increase in degradation affecting both inversion and editing images. Specifically, the middle column highlights the progression of inversion degradation, while the columns on either side demonstrate the combined impact of degradation on both inversion and editing.

method [4] and perform the degradation in the following manner:

Inversion Degradation: In this phase, we methodically merge two different ground truth (GT) images within the latent space of GAN-Control, and then perform the inversion of the blended image back into the latent space of the GAN inversion technique. One of the GT images is assigned as the 'base image,' and we progressively integrate it with the other within GAN-Control's latent space to achieve incremental alterations. Essentially, each deteriorated version of HyperStyle is inverting a composite image derived from the base image, rather than the original base image itself, to mimic a progressive decline in the quality of inversion. An illustrative example of this inversion degradation is depicted in Figure 4, within the middle column labeled as 0°. Here, the GT image positioned at 0°serves as the base image. The notation 'Level 0' corresponds to the initial inversion absent any degradation, whereas subsequent levels, arranged vertically from top to bottom, depict escalating degrees of degradation of the base image.

Editing Degradation: GT images underwent editing by

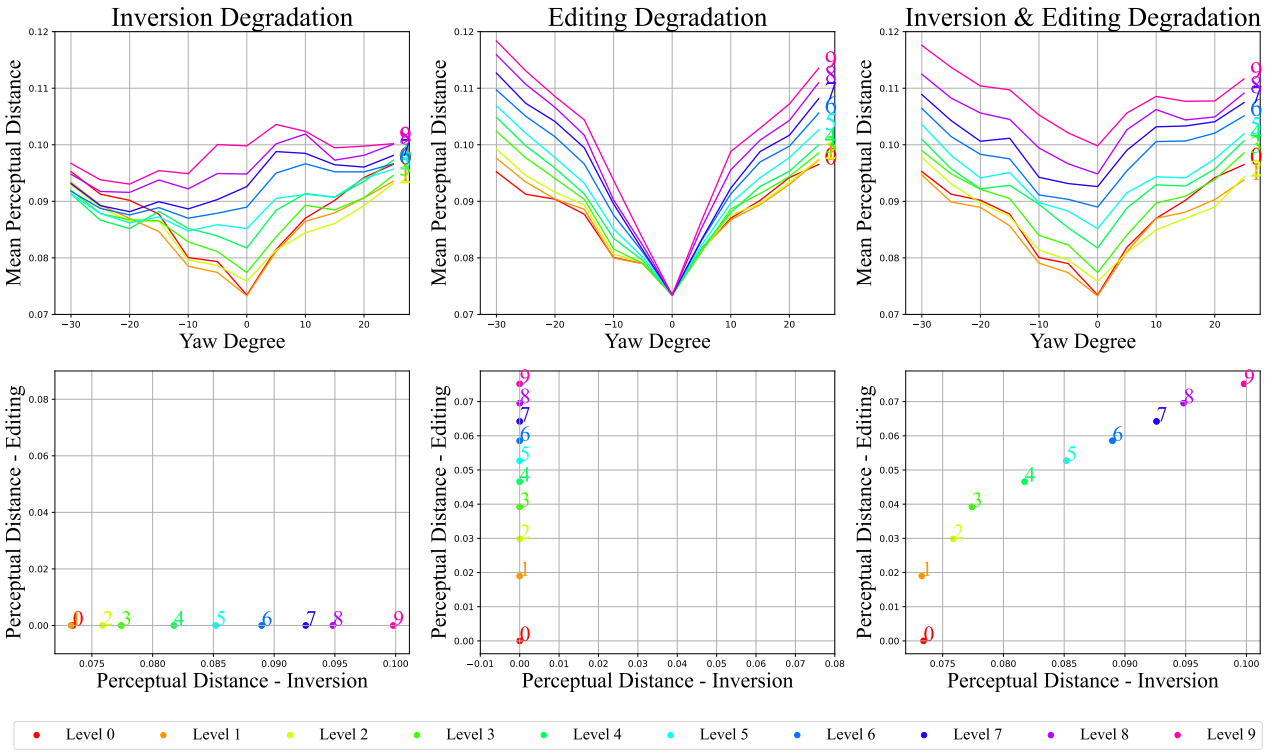

Figure 5. **Simulated Degradation Plots for the Pose Attribute.** Top row, from left to right: Inversion Degradation, Editing Degradation, and Combined Editing and Inversion Degradations, showcasing the Perceptual Distance variations. Bottom row: detailed visualizations of the degradation processes for inversion (left), editing (middle), and their combination (right), demonstrating the progressive intensity of degradations. Degradation levels are indicated on a scale from 0 to 9, with 0 indicating no degradation and 9 denoting the highest level of degradation.

traversing along specific attribute direction vectors within the latent space. Rather than rotating the direction vector, we incrementally increased the steps taken towards other well-defined latent directions, introducing varying levels of degradation. This gradual increase in step size allowed us to simulate a spectrum of editing intensities, from subtle to more pronounced changes. The extent of degradation was then quantified by measuring the Perceptual Distance between these progressively edited images and the baseline non-degraded edited image (noted as 'Level 0').

Figure 5 illustrates the results of our methodology in distinguishing between varying degradation intensities. In the plots across the top row, it is observable that with an increase in degradation level, there is a corresponding rise in Perceptual Distance, which is in line with expectations. Additionally, in accordance with predictions, the Perceptual Distance escalates as the editing intensity increases. These findings underscore the capability of our approach in identifying and quantifying the quality of inversion and editing through different degrees of degradation. Looking at the plot for editing degradation, it's evident that at $0°$, all methods exhibit the same Perceptual Distance, indicating no degradation applied to the inversion. Yet, as editing intensifies, the differential ranking among the methods becomes apparent.

4.2. Ranking GAN Inversion Methods

We apply our evaluation framework to three prominent GAN inversion methods: e4e [14], HFGI [16], and HyperStyle [4]. Figure 6 illustrates edits in pose, age, and smile from each method and Figure 7 presents the performance results of our approach on these attributes across 1000 distinct identities. In this figure, for each attribute, the point of minimum Perceptual Distance represents the baseline inversion, serving as a reference for the unedited state in our method. The varying degrees along the curves denote the results of subsequent edits, with the Perceptual Distance providing a quantitative measure of the editing impact as per our evaluation framework.

Aligned with our expectations based on the reported in published research, e4e demonstrated superior performance in image editing despite its comparatively lower inversion

Figure 6. Side-by-Side Comparison of Editing Techniques Across Various Attributes. Arranged from top to bottom: Ground Truth (GT), followed by e4e, HFGI, and HyperStyle methods, respectively. For each attribute, the central column features the inverted images, while the images to the sides demonstrate the edit results for each attribute.

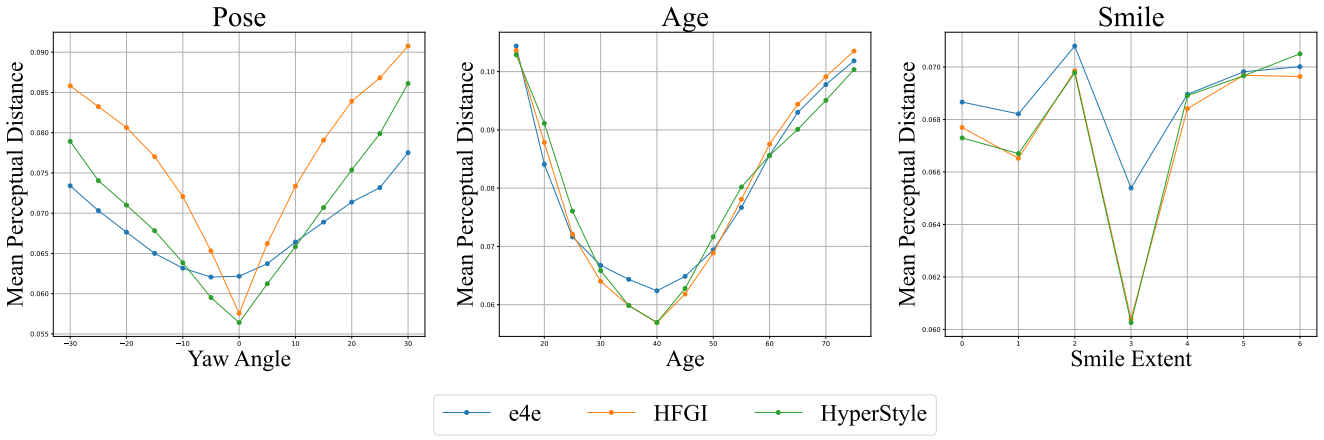

Figure 7. Comparative analysis using our framework across GAN Inversion methods. The plots, sequentially showcasing pose, age, and smile attributes from left to right, present the mean Perceptual Distance as a function of the edited attribute for each method. Each plot features performance curves for e4e (blue), HFGI (orange), and HyperStyle (green), displaying the trade-offs of each method in editing the specified attributes. The point of minimum Perceptual Distance in each plot marks the baseline inversion, which is the base point for the unedited state (with inversions occurring at yaw=0 for pose, age 40, and smile degree 3).

precision. This is particularly apparent in the pose editing scenario, as illustrated by the left plot in Figure 7. Here, e4e distinctly shows the greatest resilience in maintaining editing accuracy at higher editing degrees, evidenced by the lowest Perceptual Distance relative to its counterparts. Contrastingly, HFGI and HyperStyle, which offer superior inversion quality, display a decline in editing precision at higher degrees of attribute modification. This is showcased in all attribute plots, where the Perceptual Distance for these methods increases more rapidly with the degree of editing compared to e4e. These observations affirm the anticipated inverse relationship between inversion quality and editing intensities.

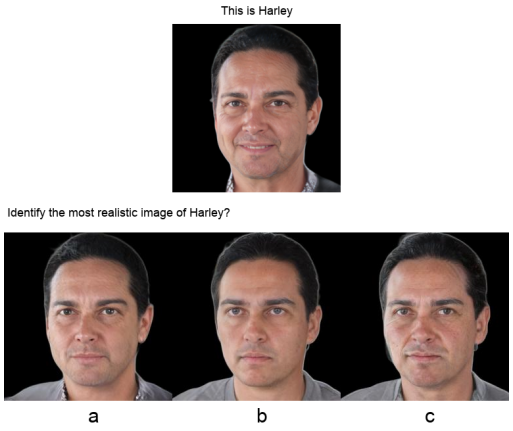

Figure 8. User Study Evaluation for Smiling Attribute: Participants are provided with a reference image with a neutral expression (top) and asked to choose the most realistic representation of the image from three edited images below (labeled a, b, c) without a smile, each produced by a different GAN Inversion method.

4.3. User Study

To empirically validate the efficacy of our proposed approach and its applicability in real-world image editing scenarios, we conducted a comprehensive user study. This study encompassed 19 distinct experiments, each involving 1000 unique identities modified to exhibit specific degrees of the target attributes: pose, age, and smile. Specifically, we explored 9 different yaw degrees for pose, 5 extents of smiling for the smile attribute, and 5 distinct age levels, ensuring a wide spectrum of attribute manipulations.

4.3.1 Study Design

Study participants were shown edited images produced by three different editing techniques: e4e, HFGI, and HyperStyle. For each image, feedback was gathered from five separate annotators. Their task involved choosing the edit that appeared most realistic compared to the original base image, described as "Select the most realistic image of [name]," with [name] indicating the face in the specific ground-truth base image. Figure 8 illustrates a sample question from the user study for the smile attribute. The objective was to evaluate the images based on both their overall quality and the fidelity of the edits. Overall, the study involved assessing 95,000 samples, calculated as 1000×19×5. We chose SageMaker Ground-Truth as the platform to conduct the study.

4.3.2 Study Results

The core of our analysis was to measure the agreement between the rankings derived from our evaluation algorithm and the choices made by participants in the user study. To guarantee the reliability of the study's results, we only considered outcomes with a confidence level greater than 0.7, as per the SageMaker Ground-Truth computed confidence. This step was taken to focus on samples with a high level of consensus and eliminate any ambiguous cases. After this filtering, the dataset included roughly 300-400 images per attribute experiment, providing a substantial sample size for a significant analysis.

The consensus results between our algorithm and user study, are presented in Table 2. These findings indicate a notable correspondence between our algorithmic assessments and human evaluations, suggesting the effectiveness of our proposed framework as a method for evaluating image editing quality. Where all experiments resulted in 80% agreement or higher.

Attribute	**Agreement (%)**
Pose	85.2
Age	80.0
Smile	86.7

Table 2. The percentage agreement between the user study rankings and the proposed method's rankings for each attribute.

5. Conclusions

In this work, we initiate a crucial dialogue, in our opinion, on the assessment of image editing methods, an underdiscussed aspect in the current discourse on image editing. Our approach introduces an initial step toward addressing this issue through a multi-step framework that employs a synthetic dataset created via GAN-Control, a matching process, and the use of Perceptual Distance as a metric for comparing pairs of corresponding ground-truth and edited images. The conducted experiments quantitatively demonstrated the inherent compromises in GAN-based image editing, particularly emphasizing the challenge of balancing high-quality inversion with achieving quality image edits. Furthermore, we presented outcomes from our proposed framework for three image-editing methods, with results aligning with the findings reported in the papers that introduced these methods. The validity of our approach received additional support from a user study. We highlight the constraints of our proposed framework, considering it a preliminary solution to the issue at hand. A thorough discussion of these limitations can be found in the "Method Drawbacks" section. Ultimately, we aspire for this work to encourage further research in this area and lay the groundwork for a systematic evaluation framework for image-editing techniques.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 2

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 1, 2

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022. 1, 2, 3, 5, 6

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[6] 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

[7] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 3

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, 2016. 3

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[10] 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 conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[11] Hongyu Liu, Yibing Song, and Qifeng Chen. Delving stylegan inversion for image editing: A foundation latent space viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10082, 2023. 1, 2, 3

[12] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 3

[13] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14083–14093, 2021. 2

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 1, 2, 3, 6

[15] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020. 3

[16] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 1, 2, 3, 6

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[18] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3121–3138, 2022. 1, 3

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3, 5

[20] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 2

[21] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016. 2