Towards Fair and Robust Face Parsing for Generative AI: A Multi-Objective Approach

Sophia J. Abraham[1], Jonathan D. Hauenstein[2], Walter J. Scheirer[1]
[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556
[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556

***Abstract*— Face parsing is a fundamental task in computer vision, enabling applications such as identity verification, facial editing, and controllable image synthesis. However, existing face parsing models often lack fairness and robustness, leading to biased segmentation across demographic groups and errors under occlusions, noise, and domain shifts. These limitations affect downstream face synthesis, where segmentation biases can degrade generative model outputs. We propose a multi-objective learning framework that optimizes accuracy, fairness, and robustness in face parsing. Our approach introduces a homotopy-based loss function that dynamically adjusts the importance of these objectives during training. To evaluate its impact, we compare multi-objective and single-objective U-Net models in a GAN-based face synthesis pipeline (Pix2PixHD). Our results show that fairness-aware and robust segmentation improves photorealism and consistency in face generation. Additionally, we conduct preliminary experiments using ControlNet, a structured conditioning model for diffusion-based synthesis, to explore how segmentation quality influences guided image generation. Our findings demonstrate that multi-objective face parsing improves demographic consistency and robustness, leading to higher-quality GAN-based synthesis.[1]**

I. Introduction

Facial parsing—the segmentation of fine-grained facial components such as *eyes, nose, mouth, and hair*—is a fundamental task in computer vision, supporting applications in *face recognition* [34], *augmented reality* [17], and *facial expression analysis* [5]. Recent advances in deep learning have significantly improved segmentation accuracy [3], [18], yet existing models primarily optimize for benchmark performance while often neglecting key concerns such as: (1) *fairness* across demographic groups, (2) *robustness* to noise, occlusions, and domain shifts, and (3) the *impact* of segmentation on downstream generative models. While face parsers may perform well on clean, well-represented data, they often degrade sharply for underrepresented demographics [2], [11], [25] or in challenging real-world conditions [9], [8], [12]. Such biases and fragility not only reduce trust and usability in applications like *identity verification* and *facial editing* but also propagate into *generative synthesis*, amplifying disparities in downstream tasks.

Recent efforts have explored *multi-objective optimization* for general segmentation [15], [30], [28] and fairness-aware approaches in facial analysis [21], [24]. However, a unified strategy that jointly optimizes *accuracy, fairness, and robustness* in face parsing remains underexplored. Furthermore, integrating fair and robust segmentation with *generative models* introduces additional complexities: state-of-the-art **GAN-based** [10] and **diffusion-based** models [39] rely on semantically structured segmentation maps [26], [35] to generate realistic and controllable faces. If the segmentation model introduces bias or lacks robustness, these deficiencies are propagated—and often amplified—by generative models, leading to unnatural or demographically skewed outputs [20], [32], [7]. This issue is particularly pronounced in **GAN-based synthesis**, where segmentation errors cause unnatural facial reconstructions, and in **diffusion-based models** such as ControlNet, where inaccurate parsing reduces *semantic alignment* and *editability*.

To address these challenges, we propose a **homotopy-based multi-objective learning framework** for face parsing that explicitly balances *accuracy, fairness, and robustness*. Our method dynamically adjusts training objectives over time, shifting from an accuracy-first paradigm in early training to a balanced trade-off incorporating fairness and robustness. This approach enables stronger segmentation performance across diverse demographic groups while improving resilience to *occlusions, noise, and domain shifts*. Unlike prior works that optimize for fairness or robustness in isolation, our framework *unifies these perspectives* within a single pipeline and systematically evaluates their impact on generative face synthesis.

To validate our approach, we integrate **multi-objective and single-objective U-Net models** into a **GAN-based face synthesis pipeline (Pix2PixHD)** and assess their impact on generative quality. We further conduct **preliminary experiments** with **ControlNet**, a structured diffusion model, to examine how segmentation quality affects *guided image generation*. Our evaluations span real-world perturbations—including Gaussian noise, occlusions, blur, and lighting shifts—as well as multiple demographic groups, measuring both segmentation performance (*mIoU, fairness variance*) and generative quality (*Fréchet Inception Distance (FID), LPIPS similarity* [38], [40]). Our key contributions are as follows:

(1) Fairness-Aware Face Parsing: We introduce a *multi-objective learning framework* that explicitly optimizes *accuracy, fairness, and robustness*.

(2) Systematic Fairness & Robustness Evaluation: We

[1]For anonymity during the review process, source codes and trained model weights are omitted but will be made publicly available upon publication.

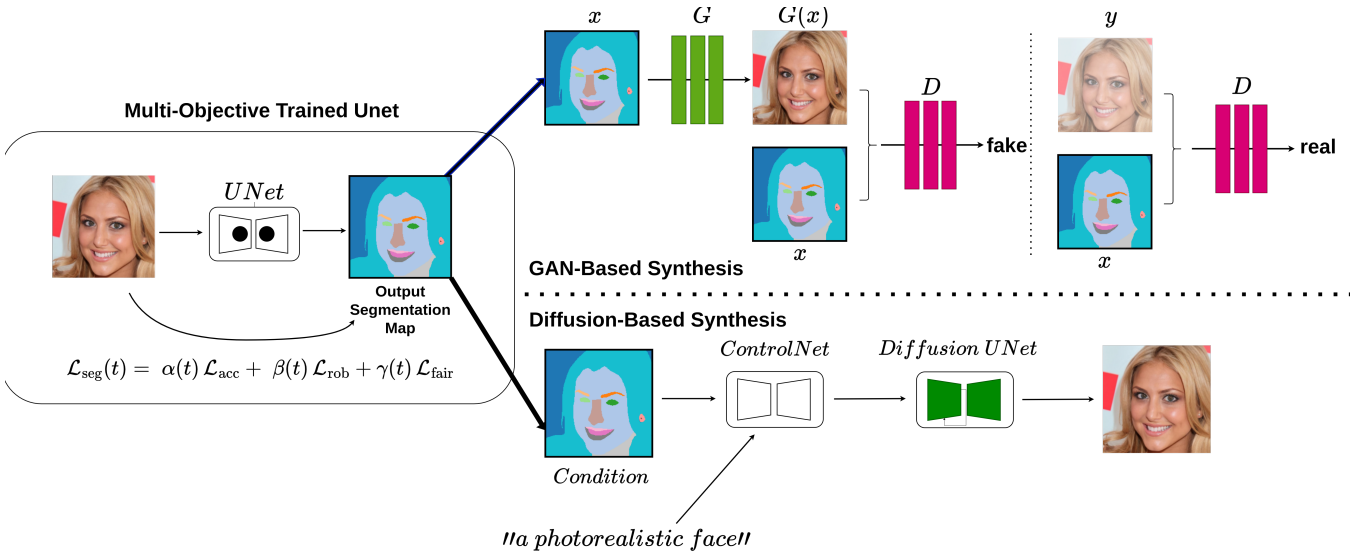

Fig. 1. **Overview of Our Multi-Objective Face Parsing and Synthesis Framework.** Our proposed *homotopy-based multi-objective learning framework* optimizes **accuracy** (L_{acc}), **robustness** (L_{rob}), and **fairness** (L_{fair}). This framework produces **fairness-aware and robust segmentation maps**, which are used to train two generative pipelines: (1) a **GAN-based synthesis model (Pix2PixHD)**, where improved segmentation enhances *photorealism and demographic consistency*, and (2) a **diffusion-based synthesis model (ControlNet)**, where structured parsing maps guide *semantic alignment and editability*. The improved segmentation quality enhances photorealism, fairness, and robustness in generative models. Key improvements include reduced bias in GAN-generated faces and more stable semantic conditioning in diffusion synthesis.

quantify *segmentation fairness* via *mIoU variance* and assess robustness under *occlusions, noise, and domain shifts*.

(3) Impact on GAN-Based Face Synthesis: We show that fairness-aware segmentation improves GAN-generated face quality, reducing *FID scores* and enhancing perceptual realism (*lower LPIPS scores*).

(4) Preliminary Diffusion-Based Analysis: We explore how segmentation quality influences *diffusion-based synthesis (ControlNet)*.

Our findings highlight the importance of **fair and robust face parsing** in developing *bias-aware generative models* with applications in *face editing, identity verification, and ethical AI deployment*. The remainder of this paper is structured as follows: we discuss *related work* in Section II, present our *proposed method* and *experiments* in Section III, describe *results* in Section IV, and conclude with *implications and future research directions* in Section V.

II. Related Work

A. Multi-Objective Optimization in Computer Vision

Multi-objective optimization is widely used in computer vision to balance competing objectives such as accuracy, efficiency, and robustness [29]. Traditional methods rely on fixed weighting schemes for loss functions, limiting adaptability across different tasks. More recent techniques, such as *homotopy-based optimization*, introduce dynamic weighting mechanisms that shift priorities during training [4]. These methods have shown promise in solving complex optimization problems, particularly in high-dimensional polynomial systems [23]. However, their application to specialized areas such as *face parsing and generative modeling* remains largely unexplored. Our work extends homotopy-based optimization to structured face parsing, explicitly integrating *accuracy, fairness, and robustness* into a single multi-objective framework.

B. Generative Adversarial Networks and Multi-Objective Training

Generative Adversarial Networks (GANs) are widely used for tasks such as *face generation, editing, and domain adaptation* [10]. Unlike diffusion models, which rely on iterative denoising, GANs synthesize high-quality images in a single forward pass, making them efficient for applications such as *interactive facial editing* [14]. GAN-based architectures provide structured control over facial attributes through techniques such as *semantic segmentation-guided generation* [26] and *latent space manipulation* [36].

Multi-objective training of GANs has been explored through *multi-discriminator architectures* to improve stability and diversity [1] and *evolutionary optimization approaches* for adversarial training [33]. However, GANs remain susceptible to *mode collapse*, demographic biases, and robustness issues, particularly when trained on imbalanced datasets [32]. Our work introduces a homotopy-based optimization framework that explicitly balances *perceptual realism, semantic alignment, and demographic fairness* by leveraging segmentation maps as conditioning inputs. This allows us to systematically evaluate how fairness-aware parsing influences structured image synthesis. Additionally, we extend our analysis to *diffusion-based synthesis (ControlNet)*, enabling a direct comparison between GANs and diffusion models in terms of *controllability, fairness, and robustness*.

C. Diffusion Models and Structured Conditioning

Diffusion models have emerged as powerful alternatives to GANs for high-resolution image generation, achieving state-of-the-art performance in photorealistic synthesis [13]. Structured conditioning mechanisms, such as *ControlNet* [39], improve controllability by integrating external control signals, including segmentation or edge maps. While prior work has demonstrated the effectiveness of diffusion models for face synthesis [27], their dependence on structured inputs has not been systematically examined in the context of fairness-aware segmentation pipelines. Our study investigates the role of multi-objective face parsing in guiding diffusion-based synthesis and compares its impact against traditional GAN-based conditioning.

D. Fairness in Face Parsing and Generative Models

Fairness has been extensively studied in face recognition and classification, where demographic biases in deep learning models have been well documented [2]. However, fairness-aware segmentation remains underexplored [11], despite evidence that segmentation models exhibit higher error rates for underrepresented demographic groups [6]. These disparities can propagate into downstream applications, such as attribute editing and face synthesis, amplifying biases in generative outputs.

While existing work has introduced fairness-aware regularization for generative models [32], few studies explicitly examine how segmentation biases affect generative synthesis pipelines. Our approach addresses this gap by incorporating fairness as an explicit training objective in face parsing and evaluating its effect on both GAN- and diffusion-based synthesis. By demonstrating how fairness-aware segmentation improves photorealism and demographic consistency, we establish a framework for more equitable face generation.

E. Face Parsing for Generative Synthesis

Face parsing, which involves segmenting facial components such as eyes, lips, and hair, plays a critical role in tasks like face editing, synthesis, and attribute manipulation [19]. Previous work has explored using segmentation maps to enhance GAN-based face editing [26]. For instance, regional GAN inversion techniques leverage parsing maps to enable fine-grained control over facial feature editing [37]. However, existing methods prioritize accuracy without explicitly addressing fairness or robustness.

Our framework extends face parsing for generative synthesis by integrating segmentation and GAN training into a unified multi-objective pipeline. Using homotopy-based optimization, we balance *realism, semantic alignment, and fairness*, ensuring that segmentation maps remain robust to variations in demographic attributes and imaging conditions.

F. Robustness and Cross-Domain Generalization

Robustness to noise, occlusion, and domain shifts remains a key challenge in vision models [8], [12]. While segmentation models are often evaluated under controlled conditions, their deployment in real-world applications requires generalization across diverse datasets and imaging environments [22]. Domain adaptation methods, such as *CycleGAN*, have been explored for cross-domain transfer [41], but they do not explicitly enforce robustness constraints in segmentation.

Our work incorporates robustness as an explicit optimization objective in face parsing, ensuring consistent performance under occlusions, noise, and cross-domain shifts. We validate our models across multiple datasets, including CelebAMask-HQ [16], demonstrating that multi-objective optimization enhances generalization and resilience.

G. Contributions of Our Work

While prior work has explored elements of *multi-objective optimization, fairness-aware segmentation, and GAN-conditioned synthesis*, our approach is the first to integrate these into a *unified homotopy-based framework*. By dynamically balancing *accuracy, fairness, robustness, and semantic fidelity*, we improve face parsing performance and demonstrate its downstream impact on generative tasks.

Unlike prior methods that address fairness at the synthesis stage, our approach ensures *fairness and robustness at the segmentation level*, reducing biases before they propagate into generative models. We systematically evaluate how segmentation quality affects both **GAN-based synthesis (Pix2PixHD)** and **diffusion-based synthesis (ControlNet)**, showing improvements in photorealism, demographic consistency, and structured conditioning. Our findings highlight the importance of fairness-aware segmentation for bias-aware generative modeling in face editing and synthesis.

III. Proposed Method

In this section, we introduce our homotopy-based multi-objective framework for face parsing and its integration with both **GAN-based** and **diffusion-based** face editing models. We outline the problem formulation, dataset preparation, model architecture, training strategy, and evaluation pipeline, emphasizing **fairness**, **robustness**, and **semantic alignment**.

A. Problem Formulation

We define the dataset $\mathbf{X} = \{x_i\}$, where each face image is paired with a segmentation mask $y_i \in \mathbf{Y}$, mapping to 19 facial components (e.g., hair, eyes, mouth). Demographic attributes are denoted as $\mathbf{a}$ (e.g., `Male`, `Young`, `Wearing Hat`). Our objective is to train a segmentation function $f_\theta(\cdot)$ that predicts $\hat{y}_i$ while optimizing for accuracy, fairness, and robustness. Accuracy is maximized by aligning $\hat{y}_i$ with y_i using Dice loss [31]. Fairness is enforced by minimizing variance $\mathrm{Var}(\mathrm{mIoU}_g)$ across demographic groups, ensuring equitable segmentation quality. Robustness is maintained by penalizing performance degradation (mIoU drop) under input perturbations such as noise and occlusion.

B. Dataset Preparation

We employ the CelebAMask-HQ dataset [16], divided into training, validation, and test sets. Each image and mask

Algorithm 1 Multi-Objective Face Parsing (Pseudo-code)

[t]

Require: Homotopy function $h(t)$ providing (α, β, γ) for epoch t

for epoch $t = 1 \dots T$ **do**
  $(\alpha, \beta, \gamma) = h(t)$
  for each batch in DataLoader **do**
    Load images $\{x\}$, masks $\{m\}$, attributes $\{a\}$
    outputs $= f_\theta(x)$ ▷ U-Net forward pass
    $\mathcal{L}_{\text{acc}} = \text{DiceLoss}(outputs, m)$
    $\text{outputs}_{\text{noisy}} = outputs + \text{random\_noise}()$
    $\mathcal{L}_{\text{rob}} = -\text{mIoU}(\text{softmax}(outputs_{\text{noisy}}), m)$
    $\mathcal{L}_{\text{fair}} = \text{Var}\left[\text{mIoU}_g\right]$
    $\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{acc}} + \beta\,\mathcal{L}_{\text{rob}} + \gamma\,\mathcal{L}_{\text{fair}}$
    Backward and **update** θ
  end for
end for

are resized to 256×256 for compatibility with our U-Net architecture. Demographic attributes are extracted from annotations to compute fairness metrics.

C. Model Architecture

Our segmentation model utilizes a U-Net architecture with a ResNet-34 encoder pre-trained on ImageNet. It outputs 19 channels corresponding to distinct facial regions, balancing computational efficiency with high segmentation accuracy.

D. Multi-Objective Training

We train the U-Net segmentation models by optimizing a weighted sum of accuracy, fairness, and robustness losses, dynamically adjusted using homotopy-based scheduling. The training process is outlined in Algorithm 1.

a) Loss Components:

- **Accuracy Loss ($\mathcal{L}_{\text{acc}}$):** Dice loss measures the overlap between predicted and ground truth masks.
- **Robustness Loss ($\mathcal{L}_{\text{rob}}$):** Negative mIoU under perturbed predictions to ensure stability.
- **Fairness Loss ($\mathcal{L}_{\text{fair}}$):** Variance of mIoU across demographic groups to promote equitable performance.

Alternative Fairness Computation: We also compute per-group mIoU for each demographic attribute, enabling detailed analysis of performance disparities (see Section IV-C.1).

E. Homotopy-Based Loss Scheduling

We dynamically balance the three loss components using epoch-dependent weights $\alpha(t)$, $\beta(t)$, and $\gamma(t)$, ensuring $\alpha(t) + \beta(t) + \gamma(t) = 1$. Initially, accuracy is prioritized, with weights shifting towards robustness and fairness over time. We explore three scheduling strategies:

- **Linear:** $\alpha(t)$ decreases linearly, while $\beta(t)$ and $\gamma(t)$ increase proportionally.
- **Sigmoid:** Smooth logistic transitions for gradual emphasis shifts.

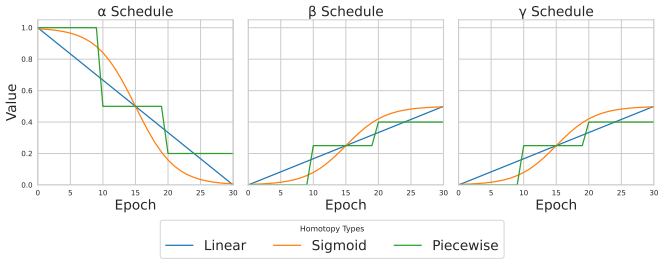

Fig. 2. Comparison of α, β, and γ schedules across three homotopy methods (Linear, Sigmoid, and Piecewise) over 30 epochs. Each subplot illustrates the evolution of a parameter (α, β, or γ) as it adapts during training, highlighting the differences in transition dynamics across homotopy strategies. The legend below the figure identifies the homotopy method for each curve.

- **Piecewise:** Abrupt changes in weight distribution at predefined training stages.

Figure 2 illustrates the evolution of these weights across training epochs for each homotopy method.

F. Integration with Generative Models

1) GAN-Based Face Editing: We utilize the trained U-Nets to generate segmentation maps for the training and validation sets, which are then used to train a Pix2PixHD GAN. The GAN architecture comprises:

- **Generator** G: Transforms segmentation maps into RGB images.
- **Discriminator** D: Distinguishes real images from generated ones.

The GAN training involves a combination of adversarial loss and pixel-level L_1 reconstruction loss:

$$\mathcal{L}_{\text{GAN}} = \mathcal{L}_{\text{adv}}(G, D) + \lambda \left\|\hat{x} - x\right\|_1,$$

where $\hat{x} = G(\text{segmentation\_map})$ and x is the real image.

During testing, the GAN generates images using segmentation maps from the test set produced by both single-objective and multi-objective U-Nets, enabling evaluation of how segmentation quality impacts generative performance.

2) ControlNet-Based Face Editing: In addition to GANs, we integrate **ControlNet** [39] for diffusion-based face editing. ControlNet leverages segmentation maps to guide the diffusion process, enhancing image fidelity and semantic alignment. Our setup includes:

- **ControlNet Model:** Pre-trained on Stable Diffusion, fine-tuned on our segmentation maps.
- **Diffusion Pipeline:** Combines ControlNet with a text encoder and U-Net backbone to generate photorealistic faces conditioned on segmentation maps.

Training Procedure: ControlNet is fine-tuned for a single epoch using segmentation maps from the training set. In diffusion-based experiments, we compare only the single-objective model with the multi-objective linear homotopy model to manage computational resources effectively. The training minimizes the standard denoising loss:

$$\mathcal{L}_{\text{ControlNet}} = \mathcal{L}_{\text{denoise}},$$

where $\mathcal{L}_{\text{denoise}}$ is the Mean Squared Error between predicted and actual noise. During testing, ControlNet generates images using test set segmentation maps from both U-Net models, allowing assessment of segmentation quality's effect on diffusion-based generation.

G. Evaluation Metrics and Setup

a) Segmentation Metrics: We evaluate segmentation performance using the mean Intersection-over-Union (mIoU) across 19 facial classes. Fairness is quantified by the variance $\text{Var}(\text{mIoU}_g)$ across demographic groups, and robustness is assessed through performance under Gaussian noise, occlusions, and blur.

b) Generative Metrics: For GAN outputs, we evaluate image quality using **Fréchet Inception Distance (FID)**, which quantifies realism by comparing feature distributions between generated and real images. Additionally, **Learned Perceptual Image Patch Similarity (LPIPS)** measures perceptual similarity, where lower scores indicate greater visual resemblance to real images.

c) Implementation Details: All experiments are implemented in PyTorch and trained on four NVIDIA A10 GPUs using the Adam optimizer with a learning rate of 10^{-4}. For ControlNet, we fine-tune the pre-trained `control_v11p_sd15_seg` model based on Stable Diffusion v1.5. Our pipeline supports gradient accumulation and mixed precision (FP16) for computational efficiency. Homotopy-based loss scheduling is configurable (`linear`, `sigmoid`, `piecewise`). Detailed training configurations will be released alongside our code and models to ensure reproducibility.

d) Workflow Summary:

1) **Train U-Nets:** Train single-objective and multi-objective U-Nets on the training set, validate on the validation set.
2) **Generate Segmentation Maps:** Use trained U-Nets to produce segmentation maps for training, validation, and test sets.
3) **Train GAN:** Train the Pix2PixHD GAN using segmentation maps from the training and validation sets.
4) **Fine-Tune ControlNet:** Fine-tune ControlNet on training set segmentation maps for one epoch.
5) **Generate and Evaluate Images:** Generate images using GAN and ControlNet with test set segmentation maps from both U-Net models; evaluate using FID and LPIPS.

In the following section, we present quantitative and qualitative results demonstrating the effectiveness of our approach across various conditions and demographic groups.

IV. Results & Discussion

This section presents a comprehensive evaluation of our segmentation models and their impact on both **face parsing** and **generative synthesis** (GAN and diffusion-based). We compare single-objective and multi-objective training strategies across robustness, fairness, and perceptual quality.

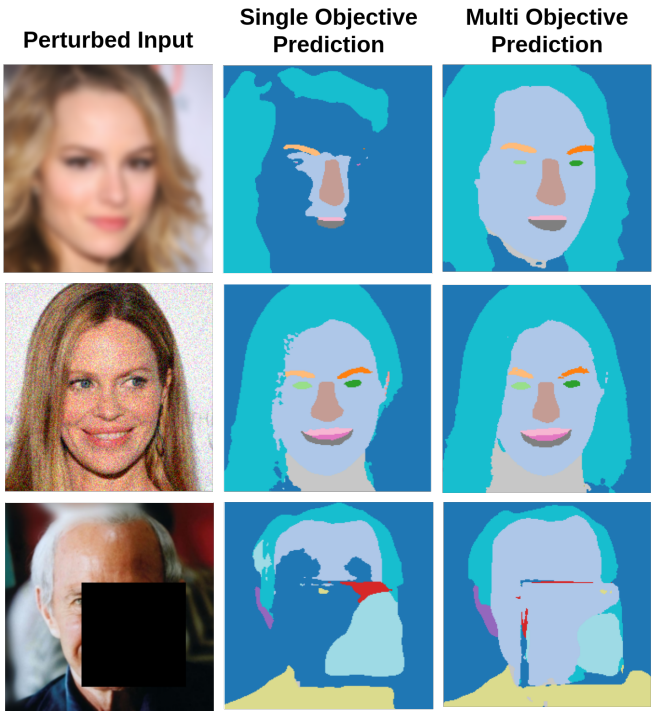

Fig. 3. **Qualitative comparison of Single-Objective and Multi-Objective models under perturbations.** Blur (severity = 0.3), Gaussian Noise (severity = 0.1), and Occlusion (severity = 0.5) are applied to input images (first column). The Single-Objective model produces fragmented and inaccurate segmentations, especially in occluded and blurred regions. In contrast, the Multi-Objective model exhibits greater robustness, preserving facial structure despite degradations, with improved stability under occlusion.

TABLE I

COMPARISON OF SEGMENTATION OBJECTIVES ON U-NET. QUANTITATIVE RESULTS COMPARING SINGLE-OBJECTIVE AND MULTI-OBJECTIVE TRAINING STRATEGIES (LINEAR, SIGMOID, AND PIECEWISE HOMOTOPY) BASED ON MEAN INTERSECTION OVER UNION (MIOU) AND DICE COEFFICIENT.

Objective	mIoU (%)	Dice (%)
Single Objective	73.87	94.46
Multi-Objective (Linear)	**74.21**	94.28
Multi-Objective (Sigmoid)	73.50	94.35
Multi-Objective (Piecewise)	73.80	94.47
Multi-Objective (Alt. Fairness)	73.81	**94.49**

A. Segmentation Performance

Despite dedicating training capacity to multiple competing objectives (fairness and robustness) rather than solely optimizing for accuracy, the multi-objective models achieve segmentation performance that remains on par with or even surpasses the single-objective baseline (Table I). This suggests that our homotopy-based optimization effectively balances competing goals without significantly compromising segmentation accuracy, demonstrating the feasibility of integrating fairness and robustness without sacrificing core performance.

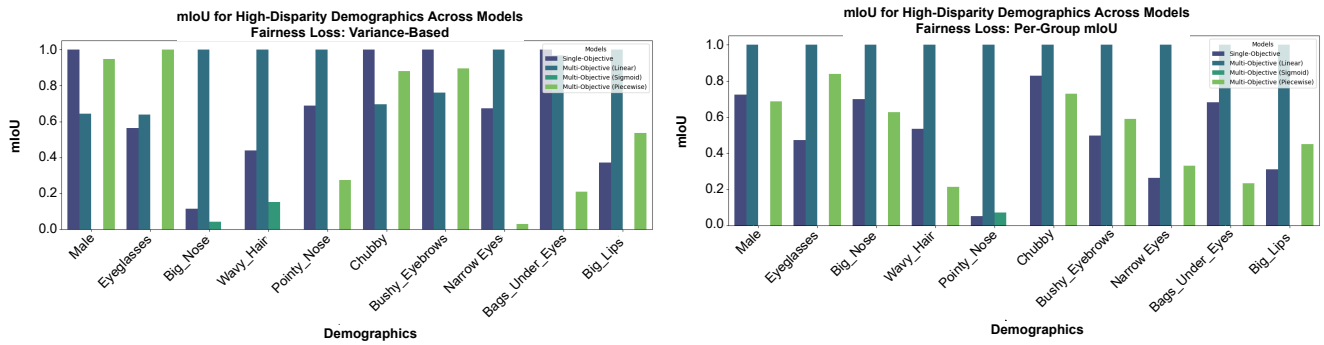

Fig. 4. **Comparison of Fairness Loss Strategies on High-Disparity Demographics.** The left plot represents the fairness variance-based approach, which minimizes the variance of per-group mIoU scores, indirectly reducing fairness gaps across demographic attributes. The right plot represents the per-group mIoU fairness loss, which explicitly tracks and optimizes fairness at a finer granularity. While the variance-based approach smooths out overall disparities, the per-group fairness loss provides better control over specific demographic attributes, ensuring higher consistency across subpopulations. Multi-objective models (Linear, Sigmoid, Piecewise) tend to provide more equitable segmentation across demographics compared to the Single-Objective baseline, though certain attributes still show variability in performance.

B. Robustness Analysis: Performance Under Perturbations

To evaluate the robustness of our segmentation models, we introduce perturbations including **Gaussian noise, blur, occlusion, and salt-and-pepper noise**. We then measure mIoU degradation under increasing severity.

Table II reports the robustness of each U-Net variant to common perturbations (Gaussian noise, blur, brightness increase, darkness) within the semantic-to-image GAN framework. While the single-objective baseline exhibits high FID scores across most perturbations, it maintains moderate LPIPS values, indicating that at least part of its generated diversity may stem from artifacts or less coherent facial/gestural details. In contrast, the multi-objective piecewise and linear homotopy models generally achieve lower FID scores—particularly under noise and brightness shifts—suggesting more robust and realistic outputs. Their LPIPS values remain in a comparable range, indicating that these methods retain perceptual diversity without succumbing to as many mode distortions. Interestingly, the sigmoid variant shows strong performance against blur in terms of FID (208.30) but the highest LPIPS under darkness (0.456), possibly reflecting that it generates more varied (yet not always consistently realistic) outputs under certain perturbations.

Figure 3 illustrates segmentation performance under perturbations. The first column presents perturbed inputs, followed by predictions from the Single-Objective U-Net and the Multi-Objective U-Net (Linear). Under blur (top row), the Single-Objective model loses fine facial details, especially around the eyes and nose, leading to misalignment, whereas the Multi-Objective model maintains more cohesive structures. Gaussian noise (middle row) introduces artifacts and noisy edges in the Single-Objective model, while the Multi-Objective approach yields smoother and more stable segmentations. Occlusion (bottom row) severely disrupts Single-Objective predictions, often causing key facial regions to disappear, whereas the Multi-Objective model preserves identifiable structures, mitigating segmentation failures. These

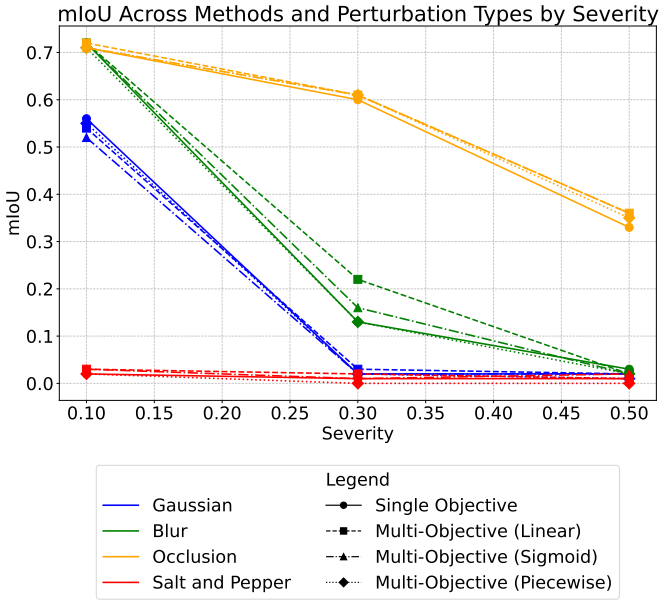

Fig. 5. **Performance comparison of mIoU across methods and perturbation types under varying severities**. The plot illustrates the sensitivity of Single Objective and Multi-Objective methods (Linear, Sigmoid, and Piecewise) to perturbations, categorized by Gaussian noise, blur, occlusion, and salt-and-pepper noise. Each method is distinguished using different line styles and markers, while colors indicate the perturbation types.

results confirm that Multi-Objective training improves segmentation resilience against real-world distortions.

In Figure 5 the Multi-Objective (Linear) method marginally outperforms Single Objective at mild and moderate severities. Under severe occlusion (0.5), both methods experience significant performance drops, but Multi-Objective retains a slight advantage. Salt and pepper noise sees marginal gains for Multi-Objective (Linear) at 0.1 severity (mIoU 0.03 vs. 0.02 for Single Objective), with all methods converging to very low mIoU ($\sim$0.00) at higher severities (0.3, 0.5). While Single Objective excels in specific mild perturbations, Multi-Objective methods, particularly the Linear variant, exhibit better robustness under moderate

TABLE II

ROBUSTNESS OF U-NET VARIANTS UNDER DIFFERENT PERTURBATIONS (GAN RESULTS). EACH CELL SHOWS FID ↓ / LPIPS ↓. LOWER FID INDICATES MORE REALISTIC OUTPUTS, WHEREAS LPIPS REFLECTS PERCEPTUAL DISTANCE (WHICH CAN IMPLY DIVERSITY OR ARTIFACTS).

Model	Gaussian Noise		Blur		Brightness		Darkness		Notes
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	
U-Net (Single)	363.06	0.435	259.12	0.403	319.57	0.407	367.75	0.431	Baseline segmentation
U-Net (Linear + Alt. Fairness)	373.83	0.419	211.52	0.390	298.46	0.384	293.71	0.417	Linear homotopy w/ alternate fairness
U-Net (Linear)	322.23	0.434	236.44	0.386	313.02	0.433	285.24	0.425	Linear homotopy approach
U-Net (Sigmoid)	349.30	0.437	208.30	0.412	286.38	0.444	331.36	0.456	Sigmoid multi-objective
U-Net (Piecewise)	307.16	0.435	216.98	0.384	330.10	0.439	326.82	0.430	Piecewise multi-objective

conditions, showcasing their adaptability to more challenging scenarios.

C. Fairness Evaluation

We measure fairness by computing performance disparities across demographic attributes. The class-wise mIoU results presented in Table III highlight the consistent advantages of incorporating fairness-based multi-objective training into U-Net models for face parsing. Across all 19 facial components, the multi-objective model outperforms the single-objective counterpart. Slight improvements are observed for all regions. Even for less distinctive or ambiguous classes like Class 7 (Eyebrows) and Class 15 (Accessories/Background), the multi-objective model achieves modest higher scores, with Class 15 improving from 73.21% to 73.49%.

1) Comparison of Fairness Approaches: In addition to our original fairness variance objective, we tested a refined approach that calculates per-group mIoU more explicitly (see Section III-D). Figure 4 show side-by-side comparisons of Single-Objective vs. Multi-Objective models across high-disparity demographic attributes. In both figures, our Multi-Objective (Linear) method generally achieves higher performance on underrepresented attributes (e.g., `Wearing_Lipstick`, `Chubby`, `Big_Lips`) compared to the Single-Objective baseline, though the exact mIoU values vary slightly under the new per-group logging scheme. Under this updated fairness measurement, we observe clearer demographic separations, particularly for attributes like `Eyeglasses` and `Smiling`, which highlights the subpopulations that benefit most from the Linear Homotopy weighting. Interestingly, while some categories (`Big_Nose`, `Bags_Under_Eyes`) display marginally lower absolute mIoU scores, the gap between Single-Objective and Multi-Objective models becomes smaller. Finally, despite refining group-wise performance tracking, the Multi-Objective (Linear) model maintains comparable overall accuracy, indicating that this focus on fairness variance does not unduly compromise mean IoU on standard benchmarks.

D. Impact on Generative Face Synthesis: GAN-Based Evaluation

To analyze the downstream impact of segmentation quality, we use the generated segmentation maps as inputs to a Pix2PixHD GAN. Although we employed the same Pix2Pix-like GAN architecture for all experiments, the segmentation maps fed into the GAN were derived from U-Nets trained

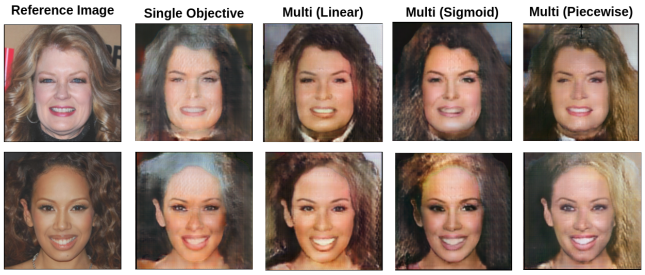

Fig. 6. **Impact of Segmentation Maps on GAN-Based Face Synthesis.** Segmentation maps from a **Single-Objective U-Net** and **Multi-Objective U-Nets** (Linear, Sigmoid, Piecewise) serve as inputs to a **Pix2Pix GAN**. Single-objective segmentation introduces inconsistencies, distorting facial details. In contrast, multi-objective segmentation improves structural coherence, yielding more natural and perceptually accurate face synthesis.

under distinct objectives. In the *Single-Objective* case, which optimizes for raw segmentation accuracy alone, the parser occasionally misaligned facial contours—particularly around the eyes and mouth—producing vague or blurred regions in the final GAN-synthesized faces (Figure 3). By contrast, our *Multi-Objective* U-Nets (Linear, Sigmoid, Piecewise) integrated fairness and robustness considerations, leading to segmentation maps with sharper boundaries and more consistent labeling of challenging facial attributes (e.g., hairlines or eyeglasses).

When these cleaner, more robust segmentation maps were passed to the same GAN, the generator more reliably reconstructed key features, yielding fewer artifacts and better overall realism. For instance, faces derived from the *Multi-Objective (Sigmoid)* model exhibited reduced color bleed around the hair boundary, while the *Piecewise* schedule often improved the jawline and cheek areas. Despite occasional local artifacts (e.g., minor tonal shifts), the multi-objective parses generally offered stronger geometric and semantic cues, enabling the GAN to produce final images that more faithfully mirrored the real reference. This underscores how upstream segmentation quality can impact downstream generative performance.

Table IV reports FID and LPIPS values for our U-Net-based segmentation models applied to face and gesture synthesis. While the baseline model exhibits the highest LPIPS (0.4419), a closer inspection reveals that much of this increased perceptual distance stems from undesirable artifacts and inconsistencies in facial structure, rather than meaningful diversity. This aligns with its lower overall

TABLE III

CLASS-WISE MEAN MIOU COMPARISON FOR U-NET MODELS. THE TABLE COMPARES THE SEGMENTATION PERFORMANCE OF THE SINGLE-OBJECTIVE AND MULTI-OBJECTIVE TRAINED U-NET MODELS FOR EACH CLASS IN THE CELEBAMASK-HQ DATASET. METRICS ARE REPORTED AS MEAN INTERSECTION-OVER-UNION (MIOU) FOR 19 FACIAL COMPONENTS. HIGHER VALUES FOR EACH CLASS ARE **BOLDED**.

Model	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Single Objective	73.87	73.87	73.87	73.87	73.87	73.87	75.02	73.89	73.88	73.71	73.87	73.87	73.87	73.87	73.87	73.21	73.89	73.87	75.17
Multi-Objective (Alt. Fairness)	**74.21**	**74.21**	**74.21**	**74.21**	**74.21**	**74.21**	**75.06**	**74.23**	**74.22**	**74.09**	**74.21**	**74.18**	**74.21**	**74.21**	**74.21**	**73.49**	**74.23**	**74.21**	**75.24**

TABLE IV

COMPARISON OF SEGMENTATION MODELS ON GAN-BASED AND DIFFUSION-BASED FACE SYNTHESIS.

GAN-Based Face Synthesis		
Segmentation Source	**FID ↓**	**LPIPS ↓**
Single-Objective U-Net	117.93	0.4419
Multi-Objective (Linear)	99.93	0.4269
Multi-Objective (Linear + Alt. Fairness)	107.92	0.4198
Multi-Objective (Sigmoid)	117.57	0.4378
Multi-Objective (Piecewise)	**98.87**	**0.4222**
Diffusion-Based Face Synthesis		
Segmentation Source	**FID ↓**	**LPIPS ↓**
Single-Objective U-Net	261.01	0.7867
Multi-Objective U-Net (Linear)	**257.18**	**0.7848**

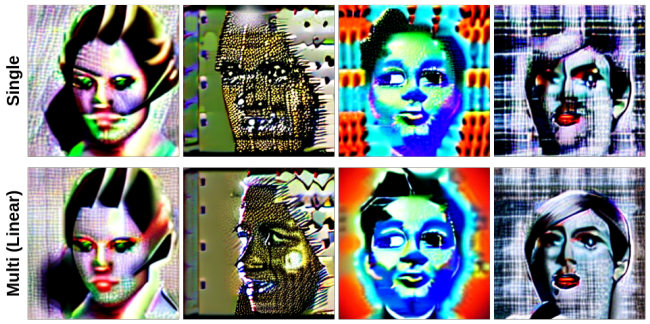

Fig. 7. **Effect of Segmentation Quality on Diffusion-Based Synthesis After One Epoch of Fine-Tuning.** The top row shows images generated using segmentation maps from the Single-Objective U-Net, while the bottom row corresponds to the Multi-Objective (Linear) U-Net. Despite being fine-tuned for just one epoch, the Multi-Objective model produces structurally coherent and visually consistent images, reducing artifacts and distortions in facial features. In contrast, the Single-Objective model exhibits irregular textures and geometric inconsistencies.

fidelity (FID = 117.93), suggesting that higher LPIPS in this case correlates with noisy, less coherent generations rather than enhanced variation.

In contrast, the Piecewise (FID = 98.87; LPIPS = 0.4222) and Linear Homotopy (FID = 99.93; LPIPS = 0.4269) models achieve lower LPIPS while maintaining strong realism, producing more structurally accurate and perceptually consistent outputs. These models strike a crucial balance—maintaining sufficient diversity in synthesis while avoiding excessive distortions.

E. Impact on Diffusion-Based Face Synthesis

To assess the role of segmentation quality in image synthesis, we integrate our segmentation maps into **ControlNet** and evaluate their impact on diffusion-based face generation. Unlike GAN-based synthesis, diffusion models rely on iterative denoising, making them particularly sensitive to structured conditioning inputs such as segmentation maps.

We conducted an initial experiment using Stable Diffusion with ControlNet, conditioned on segmentation maps from both **Single-Objective** and **Multi-Objective** U-Nets. Due to computational constraints, training was limited to **one epoch**, making this an exploratory assessment. Results in Table IV suggest that segmentation maps from Multi-Objective models yield slight but consistent improvements in **FID** and **LPIPS**, indicating higher perceptual realism and stability. While the performance gains are modest, they demonstrate that fairness- and robustness-aware segmentation models provide more reliable conditioning for diffusion-based face generation. As this study was limited to a single training epoch, further research is needed to refine these insights.

V. LIMITATIONS AND FUTURE DIRECTIONS

Despite notable improvements in fairness, robustness, and segmentation quality, several challenges remain, presenting opportunities for further research. First, the CelebAMask-HQ dataset, while diverse, remains imbalanced across demographic groups, which may limit generalization. Addressing this requires more strategic data augmentation, active reweighting, or leveraging larger, demographically-balanced datasets to further mitigate bias and enhance equitable performance. Second, our current framework treats GANs as passive consumers of segmentation maps. Incorporating *bi-directional optimization*, where segmentation feedback influences GAN training, could improve both parsing fidelity and generative realism. Such an approach could be extended to diffusion models, where structured conditioning remains underexplored in fairness-aware synthesis.

Additionally, while our method is broadly applicable beyond facial segmentation, extending it to domains such as medical imaging, autonomous perception, or video-based synthesis may require task-specific adaptations. Future research should explore domain-aware multi-objective formulations that account for context-specific biases and robustness challenges. Finally, while homotopy scheduling improves optimization efficiency, fairness-aware training introduces additional computational overhead due to subgroup evaluations. Exploring adaptive sampling strategies or efficient approximations could make large-scale deployments more

feasible, especially for real-time applications.

Our findings underscore that multi-objective training does not impose rigid trade-offs—adaptive optimization can integrate fairness and robustness without sacrificing accuracy. By extending these ideas to broader datasets, generative frameworks, and real-world applications, future research can drive the development of more equitable and resilient vision models for AI-driven image synthesis and recognition.

ETHICAL IMPACT STATEMENT

Our research focuses on fairness-aware and robust face parsing for generative AI, addressing biases in segmentation models and their downstream impact on generative synthesis. While our work aims to mitigate demographic disparities and improve model resilience, we acknowledge potential ethical concerns related to dataset biases, misuse, and unintended societal impact.

Potential Risks and Negative Impacts: Face parsing and generative models can be misused for unethical applications, such as surveillance, deepfake generation, or reinforcing demographic stereotypes. Despite our efforts to improve fairness, residual biases in datasets (e.g., CelebAMask-HQ) may persist, potentially leading to unequal model performance across demographic groups. Additionally, robustness improvements could inadvertently be leveraged to enhance adversarial facial synthesis, raising concerns about identity fraud.

Risk-Mitigation Strategies: To mitigate these risks, we employ fairness-aware multi-objective training to reduce demographic disparities and systematically evaluate robustness against real-world perturbations. Our methodology prioritizes transparency and reproducibility—our dataset choices, fairness metrics, and evaluation protocols will be made publicly available to facilitate scrutiny and improvement. Furthermore, we emphasize ethical use cases, discouraging applications in deceptive or harmful generative AI practices.

Human Subject and Data Ethics: Our study does not involve human subjects or personally identifiable information (PII). The datasets used (CelebAMask-HQ) are publicly available, and we adhere to all ethical guidelines concerning their use. While we acknowledge that publicly available datasets can contain biases, our methodology explicitly addresses this issue through fairness-aware training and demographic evaluation.

Future Ethical Considerations: Future research should extend fairness-aware segmentation to more diverse and representative datasets, ensuring broader applicability and minimizing demographic bias. Additionally, interdisciplinary collaborations with ethicists, policymakers, and domain experts will be crucial to guiding responsible deployment and regulation of AI-generated content.

By integrating fairness and robustness into face parsing, we aim to contribute to the development of ethical, bias-aware AI models that enhance inclusivity and reliability in computer vision applications.

REFERENCES

[1] I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas. Multi-objective training of generative adversarial networks with multiple discriminators. In *International Conference on Machine Learning*, pages 202–211. PMLR, 2019.

[2] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[4] C.-H. Chien, H. Fan, A. Abdelfattah, E. Tsigaridas, S. Tomov, and B. Kimia. Gpu-based homotopy continuation for minimal problems in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15765–15776, 2022.

[5] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[6] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa. Pass: protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15087–15096, 2021.

[7] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[8] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[9] G. Ghiasi, Y. Yang, D. Ramanan, and C. C. Fowlkes. Parsing occluded people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2401–2408, 2014.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] P. Grother, M. Ngan, and K. Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[12] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. 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.

[15] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[16] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] H.-M. Lee. Implementing augmented reality by using face detection, recognition and motion tracking. *Journal of the Korea society of computer and information*, 17(1):97–104, 2012.

[18] J. Lin, Y. Yuan, T. Shao, and K. Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5891–5900, 2020.

[19] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2480–2487. IEEE, 2012.

[20] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

[21] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey.

IEEE transactions on pattern analysis and machine intelligence, 44(7):3523–3542, 2021.

[23] A. Morgan and A. Sommese. Computing all solutions to polynomial systems using homotopy continuation. *Applied Mathematics and Computation*, 24(2):115–138, 1987.

[24] A. Navon, I. Achituve, H. Maron, G. Chechik, and E. Fetaya. Auxiliary learning by implicit differentiation. *arXiv preprint arXiv:2007.02693*, 2020.

[25] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022.

[26] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[29] S. Sharma and V. Kumar. A comprehensive review on multi-objective optimization techniques: Past, present and future. *Archives of Computational Methods in Engineering*, 29(7):5605–5633, 2022.

[30] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

[31] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[32] S. Tan, Y. Shen, and B. Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

[33] C. Wang, C. Xu, X. Yao, and D. Tao. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, 23(6):921–934, 2019.

[34] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[35] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[36] Z. Wu, D. Lischinski, and E. Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12863–12872, 2021.

[37] Z. Xu, X. Yu, Z. Hong, Z. Zhu, J. Han, J. Liu, E. Ding, and X. Bai. Facecontroller: Controllable attribute editing for face in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3083–3091, 2021.

[38] Y. Yu, W. Zhang, and Y. Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3, 2021.

[39] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.