Study on Face Hallucinating Features

Group: 36 Course Project

Om Kumar (B22CS081) Vinit Thakur (B22ES026) Shyam Sathvik (B22EE036)

Neermita Bhattacharya (B22CS092)

Abstract

Facial Super-Resolution (FSR) techniques aim to reconstruct high-resolution (HR) face images from low-resolution (LR) inputs. While state-of-the-art deep learning models generate visually plausible results, they often introduce ”hallucinations” – features not present in the original ground truth. This study focuses specifically on identifying *which facial features* are most commonly hallucinated by FSR models. To investigate this, we subjected HR images from the FFHQ dataset to various degradations (downsampling, blur, noise, compression) and then applied three prominent FSR models (GFPGAN, GPEN, CodeFormer) to generate super-resolved outputs. We employed a multi-modal evaluation strategy combining SIFT keypoint/landmark analysis, LPIPS perceptual scores, and a detailed Visual Question Answering (VQA) approach with semantic comparison to pinpoint discrepancies between original and generated faces. Our primary aim is to characterize the nature and location of common facial hallucinations, providing insights for the research community into the specific challenges and artifacts inherent in current FSR methods. A secondary outcome is a comparison of how these hallucination patterns manifest across the different models studied. To the best of our knowledge, this work represents a focused effort to catalog specific hallucinated facial features in FSR.

In a complementary investigation, we explored hallucination within the context of prompt-guided facial expression generation, using samples from the DeeperForensics dataset. Starting from neutral face images, we tasked three diffusion-based models (InstructPix2Pix, Stable Diffusion XL Refiner 1.0 and Dreamshaper Inpainting) with generating target emotions (happy, sad, surprise, angry). Evaluating these outputs against ground-truth expressions using SIFT, Dlib landmarks, color histograms, and CLIP semantic similarity revealed distinct failure modes. Key challenges included accurately translating semantic prompts into correct **facial geometry** (generated faces often remained too neutral), achieving reliable **semantic representation** of the target emotion (poor CLIP alignment), and avoiding the **hallucination of missing features** (e.g., absent teeth) or **introduction of artifacts** (e.g., unnatural textures, color shifts, blending errors). This secondary analysis characterizes the specific difficulties and hallucination types encountered when modifying facial appearance based on textual prompts, contrasting them with FSR challenges.

Contents

I. Introduction

The goal of Facial Super-Resolution (FSR) is the reconstruction of high-resolution (HR) facial images from low-resolution (LR) counterparts. Recent advances, particularly using deep generative models, have shown remarkable progress in producing high-quality outputs [5, 6, 7]. However, a significant challenge remains: these models often ”hallucinate” details. Facial hallucination refers to the generation of facial features that are visually plausible but are not accurate representations of the original HR ground truth, potentially inventing or distorting details like skin texture, wrinkles, eye characteristics, or hair patterns.

While existing research often evaluates FSR models based on overall image quality metrics (e.g., PSNR, SSIM, LPIPS) or perceptual realism, there has been less focus on systematically identifying *which specific facial features* are most prone to hallucination. Understanding the nature and common locations of these artifacts is crucial for improving model robustness, assessing reliability in sensitive applications (e.g., forensics, biometrics), and guiding future FSR research.

This report details an investigation aimed squarely at identifying and characterizing common facial hallucination artifacts. We utilize three state-of-the-art FSR models – GFPGAN [5], GPEN [6], and CodeFormer [7] – not primarily to rank them, but as tools to generate super-resolved faces under various controlled degradation scenarios (downsampling, Gaussian blur, Gaussian noise, JPEG compression) applied to images from the Flickr-Faces-HQ (FFHQ) dataset [2]. By systematically comparing the generated outputs against the original HR images using a multi-faceted evaluation approach, we aim to pinpoint recurring discrepancies and identify which anatomical facial features are most frequently distorted or invented. This study seeks to provide the research community with a clearer understanding of *what* gets hallucinated in FSR, contributing knowledge potentially valuable for developing more faithful reconstruction techniques.

This report also describes an investigation using samples from the DeeperForensics dataset, focusing on a controlled setting involving four individuals expressing five emotions: angry, happy, sad, neutral, and surprise. For the first task, we introduced noise specifically to the neutral facial images and employed two diffusion models—InstructPix2Pix and Stable Diffusion XL Refiner 1.0—to generate the remaining four emotional expressions from three variants of the neutral face: noisy, denoised, and original. To assess the fidelity of the generated images, we used Scale-Invariant Feature Transform (SIFT) to compare them against the ground-truth emotional expressions. Furthermore, we extracted facial keypoints using python's face_alignment, dlib and SIFT, and analyzed which landmarks appeared consistently in both generated and original emotional expressions. This approach allowed us to explore how different types of input degradation affect emotion transformation quality, and which facial regions or features are most prone to hallucination during expression synthesis from noisy and denoised inputs to diffusion models. For the second task, we focused on generating the same target expressions (happy, sad, surprise, angry) but starting from the original, clean neutral face image of selected subjects (specifically Persons 3 and 4 from the dataset). This allowed us to isolate the generation quality without the confounder of input noise. Two distinct generative approaches were evaluated: direct image-to-image generation using Stable Diffusion XL Refiner 1.0 and masked inpainting using the `Lykon/dreamshaper-8-inpainting` model, where masks highlighting expressive regions were derived from the difference between neutral and ground-truth expressions. The quality of these generated expressions was assessed through a comprehensive multi-modal evaluation against the ground-truth expression images. This included detailed geometric analysis via Dlib facial landmark comparisons, textural and local feature fidelity using SIFT keypoint matching, assessment of color distribution preservation through color histogram similarity, and semantic evaluation using CLIP to measure the alignment between the generated image and the intended emotion's textual description. This detailed analysis aimed to identify specific generation failures, artifacts, and hallucinated or altered features introduced during the expression generation process from a clean starting point.

The code and detailed results of this study are available on Github.

II. Objective

The primary objective of this study is to **identify and characterize specific facial features that are commonly hallucinated** by modern FSR and Diffusion models when reconstructing faces from degraded inputs.

Secondary objectives include:

- To assess how different types and levels of input degradation influence the occurrence and nature of facial hallucinations.
- To comparatively analyze the hallucination patterns exhibited by GFPGAN, GPEN, and CodeFormer, understanding if different architectures lead to different types of feature invention or distortion.
- To evaluate the effectiveness of a combined evaluation methodology (SIFT+landmarks, LPIPS, VQA-based semantic analysis) for the specific task of detecting and describing hallucinated facial features.
- To investigate how facial expression generation from noisy inputs affects the preservation of identity and feature realism by generating emotional variants (angry, happy, sad, surprise) from neutral faces using Diffusion models, and comparing their keypoint and region-level fidelity using SIFT and DLIB-based landmark analysis.

III. Dataset and Degradation Methodology

A Source Dataset

Two datasets were used in this study to support complementary tasks involving face restoration and emotional expression synthesis:

- **FFHQ Dataset**: High-resolution (HR) ground truth images (1024×1024 pixels) were selected from the diverse Flickr-Faces-HQ (FFHQ) dataset [2]. To manage computational demands while allowing for in-depth analysis, this study utilized a focused subset of **3 original face images**.
- **DeeperForensics Dataset** [1]: A separate set of samples was curated from this dataset for a controlled emotion transformation experiment. Specifically, **4 individuals** - P1 to P4, were selected, each exhibiting **5 distinct facial expressions**: *angry*, *happy*, *sad*, *neutral*, and *surprised*.

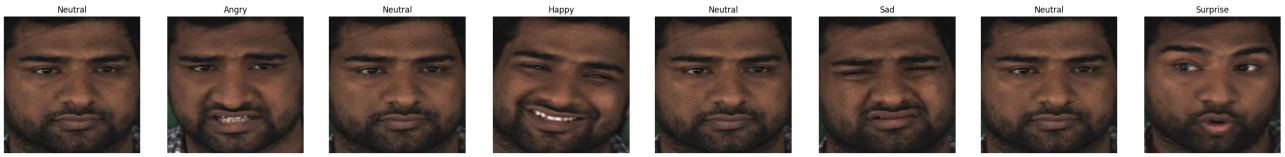


Figure 1: P1 - Angry **Figure 2:** P1 - Happy **Figure 3:** P1 - Sad **Figure 4:** P1 - Surprise

Figure 5: Facial expressions of participant P1 showing different emotions

Figure 6: P2 - Angry **Figure 7:** P2 - Happy **Figure 8:** P2 - Sad **Figure 9:** P2 - Surprise

Figure 10: Facial expressions of participant P2 showing different emotions

(a) 32×32 **(b)** 64×64 **(c)** 128×128 **(d)** 256×256 **(e)** Original

Figure 11: Example image degradation at various resolutions compared to the original high-resolution image.

B Input Degradation Process

To simulate realistic low-quality image capture scenarios, input images were degraded. The following process, derived from the provided script, details how a 256×256 LR image was generated from a 1024×1024 HR image:

1. **Gaussian Blurring:** The original HR image (1024×1024) is first smoothed using a Gaussian filter with a kernel size of 5×5 (`BLUR_KERNEL = 5`).
2. **Downsampling:** The blurred HR image is then downsampled by a factor of 4 (`SCALE_FACTOR = 4`) using area interpolation (`cv2.INTER_AREA`) to 256×256 pixels.
3. **Gaussian Noise Addition:** Gaussian noise (mean 0, std dev 5, `NOISE_SIGMA = 5`) is added to the 256×256 image, clipping values to [0, 255].

4. **JPEG Compression:** Finally, the noisy 256×256 image is compressed using JPEG with quality factor 70 (`JPEG_QUALITY = 70`).

This specific script-based process was used to generate the 256×256 degraded inputs. *Generation of other LR resolutions (32×32, 64×64, 128×128) involved applying different scaling factors or potentially different degradation parameters not detailed in the example script.* Figure B shows examples of the 1024×1024 outputs generated by the FSR models when applied to these different LR inputs derived from a single source image.

Figure 12: Visual comparison of FSR model outputs (1024×1024) generated from differently degraded inputs derived from a single source image. Rows correspond to models: GPEN, CodeFormer (CF), GFPGAN (top to bottom). Columns correspond to input resolutions: 32×32, 64×64, 128×128, 256×256 (left to right).

For the **DeeperForensics** subset, degradation focused specifically on the **neutral facial expression** for each individual. Three variants of the neutral face were prepared:

- **Original**: The clean, unaltered image.
- **Noisy**: Corrupted by adding Gaussian noise (mean 0, std 1) to simulate visual degradation.
- **Denoised**: The noisy version was processed and denoised through OpenCV's `fastNlMeansDenoisingColored` function.

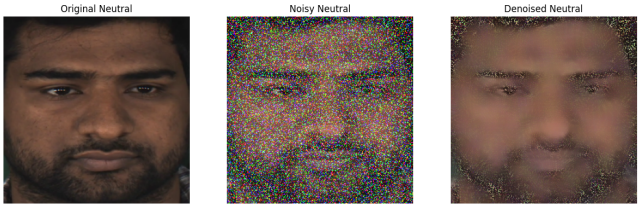


Figure 13: Addition of gaussian noise to neutral P1 image.

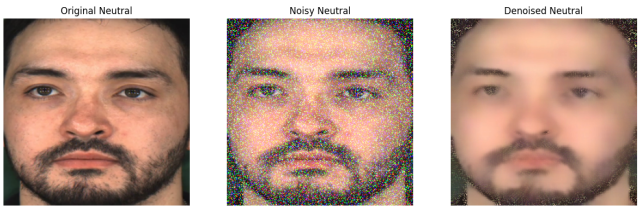


Figure 14: Addition of gaussian noise to neutral P2 image.

IV. Super-Resolution Models as Analysis Tools

FSR Models For Feature Hallucination Analysis: To generate hallucinated features for analysis, we employed three well-regarded blind face restoration models. These models serve as the generators of the data (super-resolved images) that we analyze for hallucinations:

- **GFPGAN (Generative Facial Prior GAN):** Known for leveraging strong facial priors from pretrained GANs for restoration [5].
- **GPEN (GAN Prior Embedded Network):** Integrates GAN priors directly into the network architecture [6].
- **CodeFormer:** Employs a transformer architecture with a codebook lookup mechanism [7].

Each model processed the degraded LR inputs to produce SR outputs at a target resolution of 1024×1024. Input scaling was performed as required by each model's architecture. The variation in their approaches allows us to observe if different mechanisms lead to different hallucination tendencies.

V. Diffusion Models as Analysis Tools

To investigate how diffusion models handle expression synthesis from degraded inputs, we utilized two recent state-of-the-art conditional image generation models. These models served as synthesis engines for transforming degraded neutral faces into expressive facial images:

- **InstructPix2Pix**: A diffusion-based model guided by text and image instructions. It enables expression transformations conditioned on prompts (e.g., "make the face look angry"), making it ideal for testing prompt-based editing robustness under degradation.
- **Stable Diffusion XL Refiner 1.0**: A high-fidelity image generation model that improves fine-grained visual details, particularly suited for tasks requiring photorealistic editing at higher resolutions.
- **DreamShaper 8 Inpainting**: An inpainting-capable diffusion model known for its stylistic versatility and fidelity in localized edits. It was employed to reconstruct and modify facial regions while preserving contextual coherence, making it especially useful for evaluating how well localized facial expressions can be synthesized from degraded or partially missing input data.

Each model was used to generate four emotional expressions (*angry*, *happy*, *sad*, *surprise*) from three variants of the neutral expression: **original**, **noisy**, and **denoised**. All the images were of size 256×256. This setup enabled analysis of how different input conditions affect the accuracy and realism of the synthesized emotional content. Prompts were standardized across models to isolate the effect of input degradation rather than prompt variability.

VI. Evaluation Methodology for Hallucination Detection

To systematically identify and analyze hallucinated features, we compared the original HR images with the SR outputs, and compared the original expression images with the generated expression images using the following methods:

A Geometric and Local Feature Discrepancy Analysis (SIFT + Dlib)

This method targets inconsistencies in local image structure and geometry, often indicative of hallucination or distortion.

- **Feature Extraction:** SIFT keypoints [4] and Dlib facial landmarks (68 points) [3] were detected on both original HR and generated SR images.

- **Matching & Visualization:** SIFT keypoints were matched. Visualizations were created comparing the images side-by-side, highlighting:
  - Good SIFT matches (connecting lines).
  - Unmatched SIFT keypoints on the generated image (e.g., red dots) – potentially indicating hallucinated or significantly altered texture/details.
  - Dlib landmark positions on both images – assessing geometric consistency of major features.
- **Hallucination Indication:** Dense clusters of unmatched keypoints (red dots) in specific regions (e.g., skin, hair, beard, around eyes) suggest areas where the SR model generated novel or inconsistent details not supported by corresponding features in the original. Misaligned landmarks would indicate larger geometric distortions.

B Perceptual Similarity Assessment (LPIPS)

To quantify the overall perceptual difference between original high-resolution (HR) images and the generated super-resolved (SR) outputs, we use the Learned Perceptual Image Patch Similarity (LPIPS) metric [**?**]. LPIPS aims to better reflect human perception of image similarity compared to traditional metrics like PSNR or SSIM. It achieves this by comparing deep features extracted from multiple layers of a pre-trained convolutional neural network (e.g., AlexNet, VGG).

The distance $d(x, x_0)$ between a generated image x and the original reference x_0 is calculated by computing the weighted L2 distance between their channel-normalized feature activations $(\hat{y}^l, \hat{y}^l_0)$ across several layers (l), averaged spatially:

$$d(x, x_0) = \sum_l \frac{w_l}{H_l W_l} \sum_{h,w} \|\hat{y}^l_{hw} - \hat{y}^l_{0,hw}\|_2^2 \tag{1}$$

Where w_l are learned weights per layer, H_l, W_l are feature map dimensions, and $\hat{y}^l_{hw}$ are the normalized feature activations at spatial location (h, w) in layer l. Lower LPIPS scores indicate greater perceptual similarity.

In the context of this study:

- **Calculation:** LPIPS scores were computed between each original HR image and its corresponding generated SR pairs.
- **Purpose:** LPIPS serves as a quantitative gauge of overall visual deviation. While not pinpointing specific hallucinations, higher scores suggest larger perceptual differences, potentially correlating with more significant alterations, thus complementing the local SIFT/landmark and semantic VQA analyses.

C Semantic Feature Description and Comparison (VQA)

This novel approach directly targets the identification of hallucinated features through descriptive analysis.

- **Structured Description Generation:** Google's Gemini 2.5 Pro model prompted with a detailed, structured template (see Section C.1) to describe specific anatomical features of both the original HR and generated SR images.
- **Semantic Comparison:**
  - **Similarity Scoring:** 'SentenceTransformer' ('all-MiniLM-L6-v2') embeddings and cosine similarity calculated for corresponding description sections (e.g., comparing Section 7 'Skin & Wrinkles' from original vs. generated). Low similarity suggests semantic divergence, possibly due to hallucination.
  - **Feature Term Extraction:** 'spaCy' ('en_core_web_sm') used to extract key descriptive terms (nouns, adjectives related to anatomy) from each section. Comparing these terms directly reveals differences in described features (e.g., 'smooth skin' vs 'visible pores', 'sharp jawline' vs 'rounded jawline').
- **Hallucination Identification:** Discrepancies identified through low similarity scores or differing feature terms directly point to potential hallucinated or altered anatomical details.

C.1 VQA Prompt Template

The following template was used for the VQA task with Gemini 2.5 Pro:

VQA Prompt Template

Carefully analyze the face and describe its features in detail. Focus only on facial anatomy—ignore clothing and background.
1. Face Shape: Shape (oval, round, etc.), symmetry, balance. 2. Jaw & Cheekbones: Jawline contour (sharp, rounded), cheekbone prominence and placement. 3. Forehead: Size, shape, any lines, spots, or skin texture (e.g., pimples, blemishes). 4. Eyes & Brows: Eye size, shape, position; eyebrow thickness, arch; note fine eyelid structure. 5. Nose: Bridge (high, flat, etc.), tip (rounded, pointed), and nostril shape. 6. Lips & Mouth: Lip shape, fullness, position, and even faintly visible teeth. 7. Skin & Wrinkles: Texture (smooth, rough), scratches, pores, wrinkles, or acne. 8. Tone & Lighting: Skin tone consistency and how lighting affects shadows/highlights. (Describe perceived tone/lighting) 9. Hairline & Hair on Face: Hairline shape, and how hair (if any) falls on the forehead or face (e.g., beard texture). 10. Ears & Expression: Shape and visibility of ears. 11. Expression: Emotional expression and facial tension.
Respond using numbered sections as above.

D Evaluation of Emotion Generation Fidelity (DeeperForensics)

We conducted a complementary evaluation on expression generation fidelity using the DeeperForensics dataset. This analysis focused on four identities, each displaying five distinct expressions: angry, happy, sad, neutral, and surprise.

Input Conditions: For each identity, the neutral facial image was used as a base and modified in three ways: original, noisy (via Gaussian noise), and denoised (via fastNlMeansDenoising). Each variant was then used to generate the remaining emotional expressions using two diffusion-based generative models: InstructPix2Pix and Stable Diffusion XL Refiner 1.0.

Keypoint and Landmark-Based Hallucination Detection:

- SIFT keypoints [4] and face_alignment feature landmarks were extracted from both the generated emotion images and the corresponding real expression images.
- Visual comparison included:
  - Good matches based on SIFT keypoints — shown with lines between the real and generated images.
  - Keypoint size and orientation distribution.
  - UMAP plot of SIFT descriptors of the generated image.
  - Keypoint density heatmap.
  - Keypoints per facial feature for both original and generated expression (from noisy, denoised, and clean neutral image).
- This setup allowed us to analyze how noise and denoising affect the quality of expression synthesis, and to identify regions more prone to hallucination (e.g., around the mouth or eyes or jaw during expressive change).

E Expression Generation via Masked Inpainting

To explore an alternative generation strategy focused on localized changes, we employed an inpainting strategy to incorporate expressions into the original face image via the model (`dreamshaper-8-inpainting`).

- For each target expression (e.g., ”happy”), a mask was created by computing the absolute pixel difference (`cv2.absdiff`) between the ground-truth expression image and the corresponding neutral image from the dataset. This difference map was thresholded, dilated, and blurred to create a mask highlighting regions that change significantly with the expression (e.g., mouth corners, cheeks). Images were resized to ensure consistent dimensions prior to difference calculation.
- The generated mask (inverted, as required by the model) was applied to the neutral base image. The inpainting pipeline was then prompted with text like ”Turn the expression on this person’s face into a [target expression] expression” to fill in the masked regions.
- The generated inpainted images were compared against the ground-truth expression images using SIFT feature matching and color histogram similarity, mirroring the evaluation methods used for the direct image-to-image generation approach discussed in the earlier sections. This allowed comparison of hallucination patterns and fidelity between the two generation methods.

F Semantic Fidelity Assessment via CLIP

To quantitatively assess whether the generated expressions semantically match the intended emotion, we utilized the Contrastive Language-Image Pre-Training (CLIP) model.

- **Models:** Expression generation was performed using Stable Diffusion XL Refiner 1.0, while semantic analysis employed the `openai/clip-vit-base-patch32` model and processor.
- **Embedding Extraction:** CLIP embeddings were extracted for the neutral base image, the ground-truth expression image, and the generated expression image. Additionally, embeddings were generated for a set of descriptive text prompts corresponding to the target emotion (e.g., "a person with a happy expression", "a person smiling").
- **Similarity Calculation:** Cosine similarity was computed between each image embedding (neutral, real, generated) and the average embedding of the corresponding text prompts.
- **Metrics & Visualization:**
  - Average similarity scores were compared to gauge how well each image aligned semantically with the target expression.
  - Semantic improvement scores were calculated (real vs. neutral, generated vs. neutral) to measure the change in semantic alignment relative to the base neutral face.
  - The percentage of the real image's semantic improvement achieved by the generated image was calculated.
  - t-SNE was used to visualize the relative positions of the neutral, real, and generated image embeddings in a lower dimension CLIP feature space.

G Color Distribution Analysis (Histograms)

To assess potential color shifts or unrealistic rendering introduced during generation, we analyze the color distributions of the images using histograms.

- **Histogram Calculation:** For each pair of real and generated images, 3D color histograms were computed using OpenCV's `cv2.calcHist` function in the BGR color space, typically with 8 bins per channel.
- **Normalization:** The resulting histograms were normalized using `cv2.normalize` to ensure fair comparison irrespective of minor brightness variations.
- **Similarity Metric:** The correlation between the normalized histograms of the real and generated images was calculated using `cv2.compareHist` with the `cv2.HISTCMP_CORREL` method. This yields a score between -1 and 1, where 1 indicates perfect correlation (identical color distributions) and values closer to 0 or -1 indicate significant differences.
- **Purpose:** While not directly identifying anatomical hallucinations, a low histogram correlation score signals substantial deviations in the overall color profile of the generated image compared to the original. This can indicate issues like unnatural skin tones, color casts, or other artifacts related to color rendering, providing another dimension for evaluating generation fidelity.

VII. Results: Identifying Hallucinated Features

This section details the specific facial features identified as prone to hallucination, based on the evaluation methods.

A Geometric and Textural Discrepancies (SIFT + Dlib Analysis)

The SIFT/Dlib visualizations were key in pinpointing regions with generated details inconsistent with the original.

- **Consistent Landmarks:** Major facial landmarks (eyes, nose, mouth) generally showed good alignment across models, indicating successful gross structural restoration.
- **Regions Prone to Unmatched Keypoints:** The highest density of unmatched SIFT keypoints (red dots on generated images), indicating hallucinated or altered local texture/details, consistently appeared in:
  - **Skin Regions:** Forehead, cheeks, chin – especially areas with subtle texture like pores, fine lines, or blemishes in the original. Models often generated overly smooth or uniformly textured skin.
  - **Hair and Eyebrows:** Individual strands, hair texture, and eyebrow density were often misrepresented, leading to numerous unmatched keypoints. Models might generate generic hair patterns.
  - **Periocular Area:** Fine details around the eyes, such as small wrinkles (crow's feet), eyelid structure, and reflections in the iris, were frequently altered or invented.
  - **Facial Hair (if present):** Beard or mustache texture was difficult to reproduce faithfully, resulting in significant local differences.

  - **Teeths:** Very little visible teeth are completely removed in the generated images, a major change.
  - **Ears and Nose:** Shape of ears and nose are altered, but not much detected by sift.

- **Model Variations in Texture Generation:**
  - *GFPGAN:* Often produced sharp but potentially artificial-looking textures, leading to many unmatched points in skin/hair compared to the original's natural variations.
  - *GPEN:* Showed variable results; sometimes preserved texture better, other times introduced blocky or patch-like artifacts causing clusters of mismatches.
  - *CodeFormer:* Tended towards smoother textures, reducing the number of keypoints overall but leading to mismatches where original fine texture existed.

- **Example Visualization:** Figure 27 illustrates typical areas with unmatched keypoints.

(a) codeformer **(b)** gpen

Figure 15: Comparison for SIFT/Dlib Analysis. (a) Visualization showing unmatched SIFT keypoints (red dots) overlaid on the generated image, indicating potential texture hallucinations. (b) The corresponding context image (e.g., the generated SR image without overlays).

The semantic landmark comparison for the image shown in Figure 27 revealed significant discrepancies. Widespread inconsistencies in landmark positions (with varying distance errors) were found between the original and generated images across the jawline, eyebrows, nose structure, eye positions, and mouth shape. Notably, some landmarks were entirely missing in the generated image (e.g., parts of the right eyebrow and outer lip), suggesting potential hallucination or failure to reconstruct these features accurately. Several landmarks were also undetected in the original low-resolution input, further complicating the comparison for those specific points.

B Overall Perceptual Fidelity (LPIPS)

To quantitatively assess the perceptual similarity between the restored images and the original high-resolution ground truth, we employed the Learned Perceptual Image Patch Similarity (LPIPS) metric [**?**]. Lower LPIPS scores indicate a better perceptual match. The average LPIPS scores for each model across different input resolutions are presented in Table 1.

Key observations from the LPIPS analysis include:

- **Correlation with Degradation:** As expected, higher LPIPS scores (indicating lower perceptual similarity) were observed for lower input resolutions (e.g., 32×32) compared to higher ones (e.g., 256×256) for both CodeFormer and GPEN. This confirms that perceptual fidelity generally improves as more information is available in the input image, strongly affirming the dependence of restoration quality on input resolution. Both CodeFormer and GPEN demonstrated marked improvements in perceptual similarity as the input image resolution increased from 32×32 up to 256×256 pixels, highlighting the benefit of richer input information.

- **Model Ranking (Perceptual):** Based on the average LPIPS scores obtained in this specific experimental setup (Table 1), GPEN consistently achieved lower (better) scores than CodeFormer across all tested input resolutions (32×32, 64×64, 128×128, 256×256), achieving the lowest LPIPS scores, particularly when restoring from 128×128 and 256×256 inputs. The GFPGAN model also demonstrated strong performance with an average LPIPS score of 0.0554, which is slightly higher than GPEN's best score (at 256×256 input) but significantly better than CodeFormer's score under similar high-resolution input conditions.

Table 1: Average LPIPS Scores (Lower is Better) for Different Upsampling Models vs. Input Resolution. All outputs are 1024×1024, compared against the original 1024×1024 images.

Model	Input 32×32	Input 64×64	Input 128×128	Input 256×256
CodeFormer	0.2599	0.1422	0.0960	0.0853
GPEN	0.2217	0.1362	0.0755	0.0489
GFPGAN	–	–	–	0.0554

Values show mean LPIPS scores computed across the test set for each condition. The '–' indicates data was not generated or provided for GFPGAN under the same staged downsampling conditions used for CodeFormer and GPEN in this table.

In summary, the quantitative evaluation using the LPIPS metric revealed significant insights into the performance of the models under varying input degradation levels. Based purely on this metric, GPEN exhibited the highest overall perceptual fidelity in restoring images from the downsampled inputs, followed closely by GFPGAN, with CodeFormer ranking third among the evaluated models in this test. While LPIPS provides a valuable measure of perceptual similarity, it is important to note that it may not fully capture specific artifacts or hallucination characteristics. Therefore, these findings should ideally be corroborated with qualitative visual analysis to fully understand the nature of the restorations, including potential artifacts or undesirable hallucinations, especially when dealing with severely degraded inputs.

C Semantic Feature Discrepancies (VQA Analysis)

Visual Question Answering (VQA) analysis was employed to directly reveal hallucinated or altered features by comparing textual descriptions generated for the original images versus their restored variants. This analysis was performed across 3 original source images (IDs: 22610, 22679, 22614) and their corresponding restored versions (4 input resolutions for GPEN, 4 for CodeFormer, 1 for GFPGAN per source image), totaling 27 comparisons.

- **Quantifying Semantic Divergence (Average Scores):** Cosine similarity scores were calculated between the VQA description of each original image and its corresponding generated variants using SentenceTransformer embeddings. Table 2 presents the average similarity scores calculated across the 3 source images for each model and input resolution condition.
  - *Trend with Resolution:* Generally, average similarity scores tended to increase (improve) as the input resolution increased from 64x64 to 256x256 for both GPEN and CodeFormer. However, results for the 32x32 and 64x64 inputs showed significant variability and sometimes lower scores than expected, particularly for the 64x64 input which yielded the lowest average scores for both models.
  - *Model Comparison (Average Scores):* On average, CodeFormer achieved slightly higher cosine similarity scores than GPEN across most input resolutions in this analysis. GFPGAN, evaluated only on the 256x256 input, had a lower average similarity score compared to GPEN and CodeFormer at the same input resolution based on these three images.

Table 2: Average Cosine Similarity Scores (Higher is Better) between VQA Descriptions (Original vs. Generated), Averaged Across 3 Source Images.

Model	Input 32×32	Input 64×64	Input 128×128	Input 256×256
GPEN	0.886	0.802	0.887	0.886
CodeFormer	0.897	0.823	0.895	0.888
GFPGAN	-	-	-	0.836

Scores represent the mean cosine similarity computed across 3 original images for each condition. Corresponding standard deviations are omitted for brevity but showed considerable variance, especially at lower resolutions.

- **Qualitative Patterns (Adjective Shifts):** Examining the specific adjectives added and removed across all 27 comparisons revealed consistent patterns indicative of common alterations:
  - *Frequently Added Adjectives:* Terms emphasizing smoothness, clarity, sharpness, realism, similarity, or perceived accuracy were commonly added (e.g., 'smooth', 'clear', 'sharper', 'visible', 'similar', 'accurate', 'consistent', 'original', 'present', 'defined', 'realistic', 'natural', 'symmetrical'). In cases with lower similarity scores, terms indicating artifacts or unnaturalness often appeared (e.g., 'blurry', 'simplified', 'artificial', 'unnatural', 'distorted', 'softened', 'uniform', 'less defined').

  - *Frequently Removed Adjectives:* A wide range of adjectives describing specific nuances present in the original descriptions were frequently lost in the generated ones. These often related to subtle textures ('slight', 'minor', 'faint'), specific shapes/proportions ('broad', 'narrow', 'lower', 'rectangular', 'proportional'), lighting/tone ('warm', 'bright', 'Dark'), specific details ('small', 'individual', 'flyaway', 'nasolabial', 'pigment'), or expressions ('relaxed', 'playful', 'neutral', 'happy', 'engaging').

This analysis consistently highlights a tendency for FSR models to simplify descriptions by removing nuanced vocabulary and adding terms related to generic quality (smoothness, clarity) or, occasionally, artifacts. This loss of descriptive detail often correlates with the hallucination of texture or the normalization of unique features.

- **Model Tendencies in VQA Descriptions:**
  - *GPEN:* Showed variable similarity scores, sometimes high, sometimes low (especially at 64x64 input). Adjective lists reflected this, sometimes matching well, other times adding terms like 'simplified' or removing many specific details.
  - *CodeFormer:* Generally yielded the highest average similarity scores, suggesting better preservation of the overall described semantics. However, the adjective analysis still revealed significant removal of nuanced terms and addition of 'smooth'/'clear' type adjectives, indicating potential texture smoothing despite high overall similarity.
  - *GFPGAN:* Had the lowest average similarity score in this limited comparison (at 256x256 input). Adjective lists showed considerable removal of specific descriptors and addition of terms like 'softer' or 'smoother'.

Conclusion on VQA Analysis: The VQA analysis across multiple images confirms its utility in characterizing semantic discrepancies introduced by FSR models. The average cosine similarity scores provide a quantitative overview, showing general improvement with input resolution (except for inconsistencies at lower resolutions) and indicating CodeFormer slightly outperformed GPEN and GFPGAN in preserving overall descriptive semantics in this test set. More importantly, the consistent patterns observed in the added/removed adjectives across all images and models strongly suggest common hallucination modes: loss of specific textural and shape nuances ('removed' adjectives) often replaced by generic descriptors of smoothness or clarity ('added' adjectives), or sometimes by terms indicating artifacts. This combined quantitative and qualitative VQA approach offers valuable insights into *how* generated images differ semantically from originals, complementing perceptual and local feature metrics.

D Evaluation of Emotion Generation Fidelity (DeeperForensics)

Humans evaluation can provide meaningful scores and insights too. We use 4 such metrics (on a scale of 1-5) along with SIFT keypoints matched to certain facial features using dlib to quantify the outputs generated.

Evaluation Metrics Used: Each output is compared based on:

- Expression Accuracy (EA) – How well the expression matches the prompt
- Identity Retention (IR) – How well the original face is preserved
- Image Quality (IQ) – Realism and visual clarity
- Prompt Dependency (PD) – How much the output relies on the text prompt alone (lower is better)

D.1 InstructPix2Pix

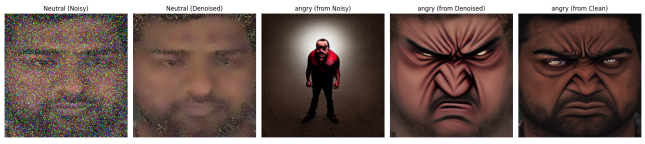


Figure 16: Outputs of InstructPix2Pix for "make him angry" prompt-P1

Figure 17: Outputs of InstructPix2Pix for "make him happy" prompt-P1

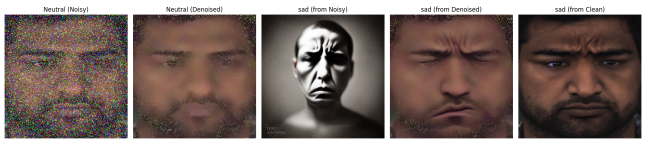


Figure 18: Outputs of InstructPix2Pix for ”make him sad” prompt-P1

Figure 19: Outputs of InstructPix2Pix for ”make him surprised” prompt-P1

- It is clear that the diffusion model fails to understand the given conditional text input if the image itself is noisy (even with a mean of 0 and std deviation of 1). It produces a completely random image with that expression. InstructPix2Pix relies heavily on the prompt to guide expression generation. Hence, outputs are often exaggerated or distorted, because the model is “guessing” what to do based only on noise + prompt.
- The ”anger” emotion usually produces cartoonish images. During denoising, the prompt guides the model to accentuate aggressive traits: Furrowed brows, narrowed eyes, frown lines. The neutral image serves as anchor, but denoising amplifies tension and contrast in features. The severity of anger correlates with how much the noise got shaped during guided denoising.
- The denoised version produces images that are a little better than the one’s produced by the noisy images. Similarly, the clean neutral images produce the best expression images. Model just applies small, targeted changes based on the prompt on the clean input.
- For ”make him angry” - P1; From Noisy: EA–5, IR–1, IQ–1, PD–5; Denoised Generation: EA–5, IR–3, IQ–3, PD–3; Clean Input: EA–5, IR–4, IQ–4, PD–2.
- For ”make him happy” - P1; From Noisy: EA–2, IR–2, IQ–1, PD–4; Denoised Generation: EA–4, IR–4, IQ–4, PD–2; Clean Input: EA–5, IR–5, IQ–5, PD–1.
- For ”make him sad” - P1; From Noisy: EA–5, IR–1, IQ–1, PD–5, Denoised Generation: EA–4, IR–3, IQ–4, PD–2, Clean Input: EA–4, IR–5, IQ–5, PD–1
- For ”make him surprised” - P1; From Noisy: EA–5, IR–1, IQ–2, PD–5, Denoised Generation: EA–4, IR–3, IQ–3, PD–2, Clean Input: EA–3, IR–5, IQ–5, PD–1

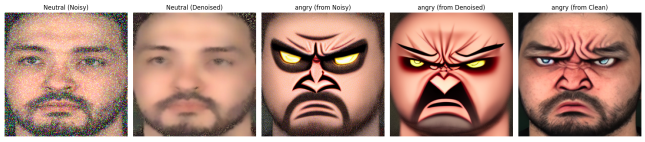


Figure 20: Outputs of InstructPix2Pix for ”make him angry” prompt-P2

Figure 21: Outputs of InstructPix2Pix for ”make him happy” prompt-P2

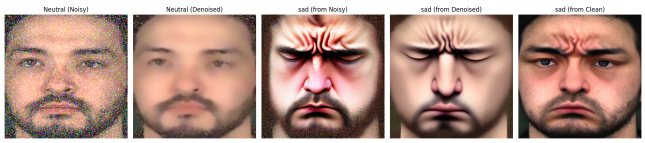


Figure 22: Outputs of InstructPix2Pix for ”make him sad” prompt-P2

Figure 23: Outputs of InstructPix2Pix for ”make him surprised” prompt-P2

- For ”make him angry” - P2; From Noisy: EA–5, IR–1, IQ–1, PD–5; Denoised Generation: EA–5, IR–3, IQ–1, PD–4; Clean Input: EA–5, IR–4, IQ–4, PD–2.
- For ”make him happy” - P2; From Noisy: EA–5, IR–3, IQ–3, PD–3; Denoised Generation: EA–5, IR–2, IQ–4, PD–2; Clean Input: EA–4, IR–5, IQ–4, PD–1.
- For ”make him sad” - P2; From Noisy: EA–5, IR–3, IQ–2, PD–3, Denoised Generation: EA–4, IR–3, IQ–4, PD–2, Clean Input: EA–4, IR–5, IQ–5, PD–1
- For ”make him surprised” - P2; From Noisy: EA–3, IR–2, IQ–1, PD–5, Denoised Generation: EA–4, IR–2, IQ–1, PD–3, Clean Input: EA–3, IR–5, IQ–4, PD–1

D.2 StableDiffusionXLRefiner

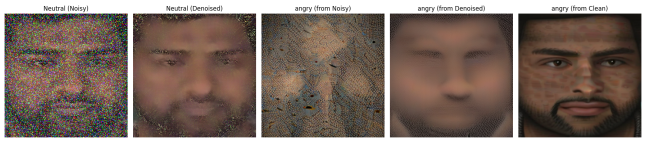


Figure 24: Outputs of StableDiffusionXLRefiner for ”make him angry” prompt-P1

Figure 25: Outputs of StableDiffusionXLRefiner for ”make him happy” prompt-P1

Figure 26: Outputs of StableDiffusionXLRefiner for ”make him sad” prompt-P1

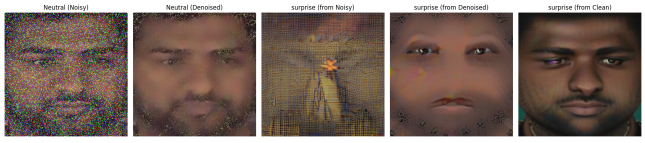


Figure 27: Outputs of StableDiffusionXLRefiner for ”make him surprised” prompt-P1

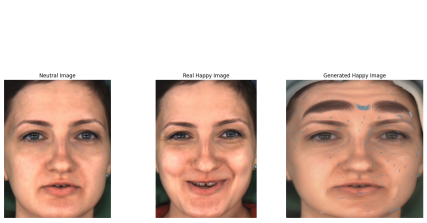
- Noticeably, this model performs worse than InstructPix2Pix.
- It produces completely random noise from the noisy input, with only some features magnified (In Fig 26, the eyebrows are significant).
- Denoised inputs enable the model to slightly learn the facial structure, so the IR increases.
- Clean neutral inputs performs the best with high IQ, IR, and low PD. The ERs are not as high as InstructPix2Pix's. This suggest that InstructPix2Pix depends on the textual prompts more and is good at generating expressions.

D.3 Expression Inpainting (Dreamshaper-8-inpainting)

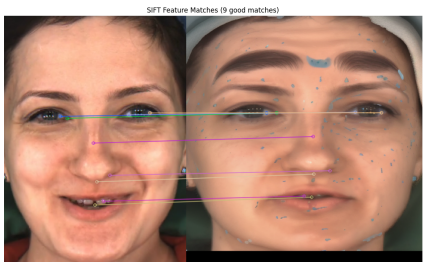
We qualitatively evaluated the expression generation results from the masked inpainting approach (Section E) using the `Lykon/dreamshaper-8-inpainting` model, focusing on textural fidelity and hallucinated features.

Analysis of Person 3, Happy expression (representative example):

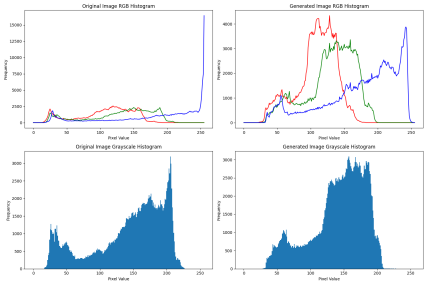
- The generated image (Figure 28b) exhibits noticeable artifacts, particularly bluish patches around the hairline, eyebrows, and chin. These seem directly related to the boundaries of the inpainted regions derived from the mask, indicating imperfect blending or color matching during the inpainting process.
- **Textural Alterations:**
  - *Skin Texture:* While potentially less smoothed than the SDXL Refiner output, the inpainted skin still appears somewhat artificial or 'airbrushed' compared to the natural texture variations in the real image (Figure 28b). The generated keypoint count (127) is closer to the original (143) than SDXL Refiner's output (89), suggesting some texture preservation, but the very low SIFT match count (9, see Figure 28b) confirms poor local feature correspondence.
  - *Eyebrows:* The eyebrows appear significantly thickened and darkened, with an unnaturally uniform texture, a clear textural hallucination within the inpainted region.
- The color histogram similarity score (0.894) is notably lower than that achieved by the SDXL Refiner (0.982), indicating more substantial shifts in the overall color distribution. This is visually confirmed by comparing the RGB and grayscale histograms (Figure 28c), where the generated image histograms differ significantly in shape and peak positions from the original. The generated image appears to have a slightly different, perhaps cooler, skin tone.
- **Expression Geometry Failure:** Similar to the SDXL Refiner, the inpainting model failed to reproduce the geometric characteristics of the happy expression accurately. While the mouth shape is altered, it lacks the pulled-back corners and upward curve of a genuine smile. The teeth visible in the real image are again missing in the generated version. The SIFT matches (Figure 28b) primarily connect regions around the eyes and nose tip, failing to capture correspondences in the expressive mouth region.
- **Interpretation:** The inpainting approach, while intended to target specific expressive regions identified by the mask, struggled with both seamless blending (leading to artifacts) and accurate feature generation within the masked areas. It hallucinated textures (eyebrows, skin smoothness) and failed to replicate key geometric elements (smile shape, teeth), resulting in an unnatural appearance. The low SIFT match count and reduced color similarity further highlight the significant local deviations from the ground truth, despite the guidance provided by the mask.



(a) Image Comparison



(b) SIFT Feature Matches



(c) Color and Grayscale Histograms

Figure 28: Inpainting analysis for Person 3 (Happy). (a) Comparison between Neutral Input, Real Happy Target, and Generated Happy via Inpainting, (b) SIFT Feature Matches (9 good matches) between Real and Generated expressions, (c) Color and Grayscale Histograms showing distribution differences (color similarity score: 0.894).

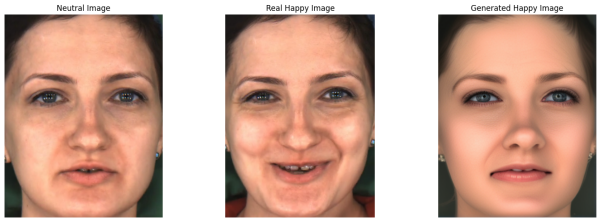
E Semantic Expression Fidelity (CLIP Analysis)

To evaluate whether the generated facial expressions successfully conveyed the intended emotion at a semantic level, we employed CLIP analysis (Section F). This involved comparing CLIP embeddings of the neutral, real (ground truth), and generated expression images against embeddings of descriptive text prompts for the target emotion.

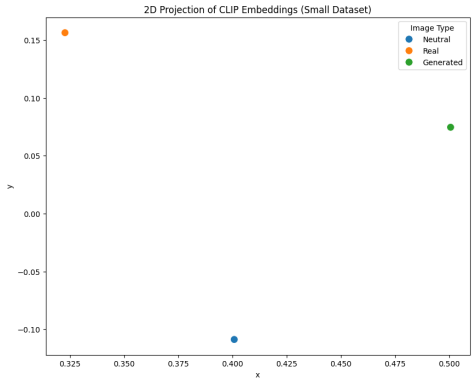
Key findings from the analysis of Person 3 (summarizing results across happy, sad, surprise, angry expressions):

- **Quantifying Semantic Shift:** Average cosine similarity scores between image embeddings and corresponding text prompts were calculated. For example, for the 'happy' expression, the scores were: Neutral (0.235), Real (0.270), Generated (0.242). Similar calculations were performed for sad, surprise, and angry expressions. (See example Figure 29).

- **Generated vs. Real Semantic Improvement:** We calculated the ”semantic improvement” achieved by both the real and generated images relative to the neutral base image. The percentage of the real image’s semantic improvement captured by the generated image varied significantly across expressions:
  - *Happy:* Generated image achieved approx. 20% of the semantic improvement seen in the real happy image.
  - *Sad:* Generated image achieved approx. 71% of the semantic improvement, indicating relatively good semantic alignment for this expression.
  - *Surprise:* Generated image showed a *negative* improvement (-35%), meaning its embedding moved further away from the ’surprise’ text prompts compared to the neutral image, failing to capture the intended semantics.
  - *Angry:* Generated image achieved approx. 25% of the semantic improvement.
- **Embedding Space Visualization:** Due to comparing only three embeddings (Neutral, Real, Generated) per expression, a direct 2D projection was used instead of t-SNE. Figure 29 (b) illustrates this, showing the relative positions. For ’happy’, the ’Generated’ point lies closer to ’Neutral’ than to ’Real’, visually reflecting the modest semantic improvement score.
- **Interpretation:** The CLIP results suggest that while the model (SDXL Refiner in this case) attempts to modify the expression based on the prompt, its success in capturing the target *semantics* is inconsistent. It performed reasonably well for ’sad’ but poorly for ’surprise’, with moderate results for ’happy’ and ’angry’ for this specific subject. This quantitative semantic assessment complements visual inspection and low-level feature analysis, providing insight into whether the *meaning* of the expression is successfully generated.



(a) Image Comparison (Neutral, Real Happy, Generated Happy) for Person 3.



(b) 2D Projection of CLIP Embeddings for the Happy Expression.

Figure 29: Example CLIP Analysis Outputs for the ’Happy’ expression (Person 3). (a) Shows the input neutral, real target, and generated images. (b) Visualizes the CLIP embeddings in 2D space, indicating the semantic proximity between the images.

F Geometric Expression Fidelity (SIFT + Dlib Analysis)

We evaluated the geometric accuracy of the generated expressions using SIFT keypoint matching and Dlib facial landmark comparisons between the neutral, real, and generated images (using SDXL Refiner) and between the original expression images and the generated images (using InstructPix2Pix).

F.1 InstructPix2Pix

Key findings for Person 2, Angry expression (representative example):

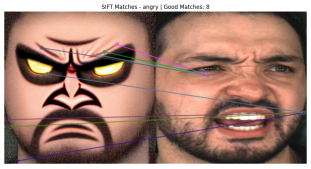


Figure 30: Good matches between generated image from noisy input for angry expression and actual angry expression - P2

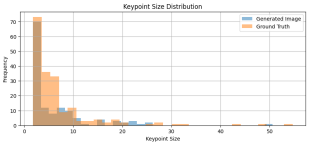


Figure 31: Keypoints sizes of generated image from noisy input for angry expression and actual angry expression - P2

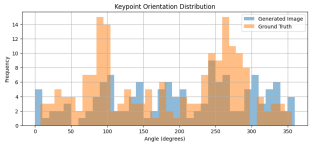


Figure 32: Keypoints orientations of generated image from noisy input for angry expression and actual angry expression - P2

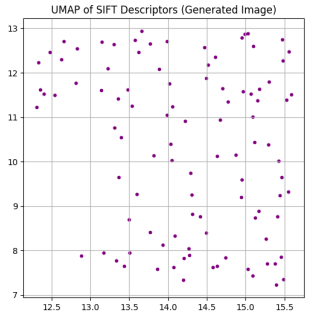


Figure 33: UMAP of SIFT descriptors of generated image from noisy input for angry expression - P2

Figure 34: Outputs of SIFT on noisy input to InstructPix2Pix - P2

Figure 35: Good matches between generated image from denoised input for angry expression and actual angry expression - P2

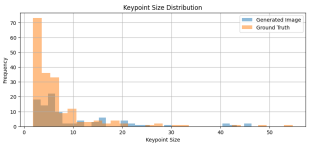


Figure 36: Keypoints sizes of generated image from denoised input for angry expression and actual angry expression - P2

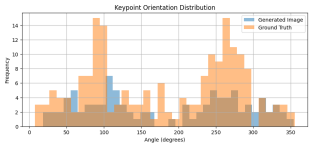


Figure 37: Keypoints orientations of generated image from denoised input for angry expression and actual angry expression - P2

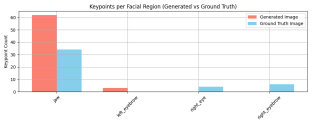


Figure 38: Keypoints per facial region of generated image from denoised input for angry expression - P2

Figure 39: Outputs of SIFT on denoised input to InstructPix2Pix - P2

Figure 40: Good matches between generated image from clean neutral input for angry expression and actual angry expression - P2

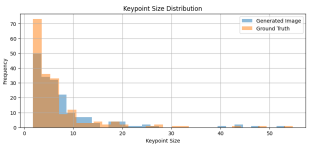


Figure 41: Keypoints sizes of generated image from clean neutral input for angry expression and actual angry expression - P2

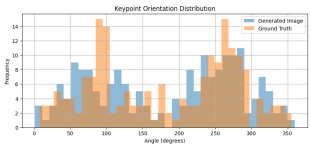


Figure 42: Keypoints orientations of generated image clean neutral input for angry expression and actual angry expression - P2

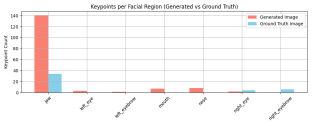


Figure 43: Keypoints per facial region of generated image from clean neutral input for angry expression - P2

Figure 44: Outputs of SIFT on clean neutral input to InstructPix2Pix - P2

- It is interesting to note that in Fig 34, no facial landmarks were detected in the generated image.
- There are many keypoints in the **jaw** region mostly for generated images. This suggests that the jaw region undergoes the most transformation or distortion during generation, possibly due to its role in conveying emotional cues like smiles, frowns, or expressions of anger and surprise.
- Overgeneration of Jaw Keypoints: The landmark analysis reveals a disproportionate number of detected keypoints in the jaw region for generated images, as compared to ground truth. This indicates that the model disproportionately modifies the lower facial region when attempting to encode expressions, regardless of emotion type.
- Disrupted Expression Geometry: The jaw-heavy modifications suggest the model may be relying on exaggerated jaw contouring to depict expressions, instead of accurately manipulating holistic expression geometry (e.g., coordinated changes in eyes, eyebrows, and mouth curvature). This could result in expressions that feel exaggerated or artificial.

- Underrepresentation in Upper Facial Regions: Regions such as the left/right eyes and eyebrows show significantly fewer keypoints in generated images than in real ones. This indicates a lack of expressive detail where it matters—especially for emotions like sadness or surprise that require eyebrow motion or eyelid widening. The resulting expressions may thus lack emotional nuance or appear "blank."
- Overreliance on Lower Face: The model appears to overfit expression synthesis to jawline reshaping, likely due to its easier detectability in loss gradients or dataset bias. This overreliance may cause it to neglect subtler but more semantically rich features in the upper face, leading to less believable outputs.

Interpretation: The landmark distribution analysis, particularly via Dlib, reveals a clear overconcentration of keypoints in the jaw region for generated images compared to ground-truth expressions. This suggests that the model attempts to express emotion primarily through jawline reshaping, rather than making the distributed geometric adjustments necessary for authentic expressions. While this may maintain a visually smooth and consistent face, it fails to evoke the multi-region facial dynamics essential for emotional realism—such as eyebrow movement, eye crinkling, or lip curvature. The lack of keypoints in upper facial areas (eyebrows and eyes) underscores the model's neglect of fine-grained expression cues, potentially due to training bias or an overemphasis on preserving facial identity over expression authenticity. This pattern indicates a systemic issue: the model exhibits reluctance to deviate from neutral geometry, and instead overcompensates with jaw distortions that do not generalize well across emotions.

F.2 SDXL Refiner

Key findings for Person 3, Happy expression (representative example):

- **Low-Level Feature Mismatch (SIFT):** A very low number of good SIFT matches (7) was found between the real happy image (143 keypoints) and the generated happy image (89 keypoints). This indicates significant differences in local textures, edges, and fine geometric details between the ground truth expression and the generated output. The high color histogram similarity (0.982) suggests the overall color palette was preserved, but local features diverged substantially.
- **Landmark Deviation Analysis:**
  - *Neutral vs. Real:* Showed large mean landmark differences (16.33 pixels), confirming that the real happy expression involves substantial geometric shifts from the neutral pose, particularly in the nose, eyebrows, and lip regions (see Figure 45 for regional differences).
  - *Real vs. Generated:* Also exhibited large mean differences (15.40 pixels), nearly identical to the Neutral vs. Real difference. This signifies that the generated image is geometrically very dissimilar to the actual happy expression it was intended to replicate.
  - *Neutral vs. Generated:* Crucially, this comparison revealed very *small* mean differences (3.88 pixels). The largest regional difference was only 6 pixels (jaw), with most regions showing less than 5 pixels deviation.
- **Identified Hallucinations and Generation Failures:**
  - **Failure to Generate Expression Geometry:** The landmark analysis strongly indicates the primary failure: the generated "happy" image remains geometrically very close to the neutral input image, despite the prompt. The model failed to induce the necessary landmark shifts (e.g., raising cheekbones, pulling lip corners up and back, crinkling eyes) characteristic of a genuine smile.
  - **Altered/Missing Features:** Visual inspection (Figure 29a) confirms the landmark findings. The generated face lacks a convincing smile; the mouth shape is altered slightly but doesn't convey happiness. Notably, the teeth visible in the real happy image are entirely absent (hallucinated away or smoothed over) in the generated image.
  - **Texture Smoothing:** Consistent with SIFT mismatches, the skin texture in the generated image appears significantly smoother and less detailed than in both the neutral and real images, suggesting hallucination of an idealized texture. Eyebrow shape also appears slightly altered.
- **Interpretation:** The combination of SIFT and Dlib analysis reveals that for this case, the SDXL Refiner model failed to translate the "happy" text prompt into corresponding geometric changes on the face. While maintaining high color similarity and staying geometrically close to the input neutral face, it produced significant local feature discrepancies compared to the ground-truth happy expression and hallucinated away details like teeth and natural skin texture. This highlights a potential limitation where the model prioritizes input image consistency over drastic, prompt-guided geometric transformation for expressions.

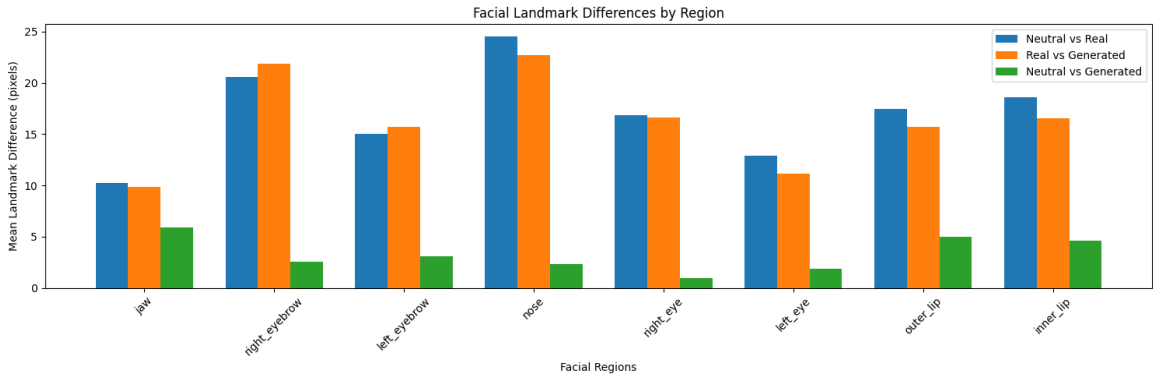


Figure 45: Mean Facial Landmark Differences by Region for Person 3, Happy Expression. Note the large 'Neutral vs Real' and 'Real vs Generated' differences, contrasted with the small 'Neutral vs Generated' differences, indicating the generated image stayed close to the neutral input.

VIII. Discussion: Characterizing Facial Hallucinations

A Hallucinations in Face Super-Resolution (FSR)

By integrating the findings from SIFT/landmark analysis (Section **??**), LPIPS perceptual metrics (Section B), and the VQA-based semantic comparisons (Section C), this discussion synthesizes the evidence to characterize the specific facial features most commonly hallucinated by the GFPGAN, GPEN, and CodeFormer FSR models under the tested conditions.

Key Hallucinated Features Identified: The combined results consistently reveal that certain facial characteristics are frequently altered or invented during super-resolution:

1. **Skin Texture:** This is a primary area of hallucination. Natural variations like pores, fine lines, and minor blemishes are often lost, replaced by textures described via VQA as 'smooth', 'clear', or 'uniform'. This is corroborated visually by dense unmatched SIFT keypoints in skin regions across models.

2. **Hair Details (Scalp, Eyebrows, Facial Hair):** Reproducing hair complexity remains challenging. Models often simplify individual strand structure and texture, rendering generic patterns instead. This is evident from SIFT mismatches within hair masses and VQA analysis showing loss of terms like 'individual strands' or 'flyaways'.

3. **Periocular Details:** The complex region around the eyes frequently shows inconsistencies beyond major landmarks. VQA analysis effectively identified alterations in fine wrinkles (crow's feet), subtle eyelid structures, and the generation of non-authentic iris reflections or 'sparkle'. SIFT analysis confirmed textural invention here.

4. **Micro-expressions and Subtle Lines:** Faint lines (e.g., nasolabial folds, forehead lines) and subtle expression cues are often diminished or entirely removed. This was particularly clear from the VQA results, showing the systematic loss of descriptive adjectives like 'faint', 'slight', 'subtle', and specific expression terms.

5. **Minor Asymmetries and Unique Imperfections:** Features like small moles, faint scars, pigment variations, or slight asymmetries tend to be normalized or removed, likely due to strong model priors favouring regularity. This aligns with the removal of unique VQA descriptors ('pigment', 'small') and potential SIFT keypoint loss. The complete removal of faintly visible teeth is a stark example.

6. **Structural Elements (Subtler Changes):** While gross structure held, SIFT analysis suggested alterations in ear and nose shapes. Furthermore, detailed landmark comparisons revealed significant geometric inconsistencies in some cases (jawline, eyebrows, etc.), indicating hallucination can extend beyond texture.

Why These Features? The susceptibility of these features likely stems from multiple factors:

- **High Frequency Information Loss:** Details like texture and fine lines contain high-frequency information fundamentally lost during degradation (Section B). Reconstruction thus heavily relies on model priors, increasing the chance of deviation from ground truth.

- **High Natural Variability:** Skin, hair, and minor imperfections vary greatly. Faced with ambiguous LR inputs, models may generate 'average' or 'idealized' features based on statistical priors, rather than reconstructing unique (but lost) details.

- **Influence of Generative Priors:** Models like GFPGAN and GPEN leverage strong GAN priors (Section IV), which impose learned styles and can override subtle details. CodeFormer's mechanism also generalizes in ways that can smooth or replace unique textures with learned representations.

Model Behavior Comparison (in context of Hallucination): Comparing model behaviors revealed potential trade-offs in how hallucinations manifest:

- **GPEN:** Achieved the best average LPIPS scores (Section B, Table 1), indicating strong perceptual fidelity overall. However, its VQA cosine similarity was more variable, sometimes lower than CodeFormer's, with analysis showing instances of simplification or artifacts (Section C, Table 2).
- **CodeFormer:** Showed the highest average semantic similarity via VQA scores (Table 2), suggesting better preservation of described structure. This often coincided with texture smoothing (evident in VQA adjectives and SIFT, Section **??**) and less competitive LPIPS scores (Table 1).
- **GFPGAN (256x256 input):** Produced sharp outputs via strong priors, but textures were sometimes described as artificial or overly smooth (VQA/SIFT analysis), contributing to a lower VQA similarity score in this test (Table 2). Its LPIPS score remained competitive (Table 1).

This suggests a trade-off: optimizing for perceptual similarity (low LPIPS) may not guarantee semantic/textural fidelity, while preserving semantic structure might involve texture simplification.

Contribution to Research Community: This study provides a focused analysis identifying *specific features* commonly hallucinated in FSR, moving beyond global metrics. Cataloging vulnerable characteristics (skin texture, hair, periocular details, etc.) offers concrete targets for improving model faithfulness. The VQA methodology proved valuable for achieving interpretable descriptions of these feature-level discrepancies.

B Hallucinations and Failures in Expression Generation

Analysis of expression generation models (SDXL Refiner, Dreamshaper Inpainting, InstructPix2Pix, StableDiffusionXL-Refiner) revealed distinct challenges and failure modes compared to FSR, primarily centred around accurately modifying facial geometry and semantics based on text prompts.

Key Failures and Hallucinated Features in Expression Generation:

1. **Failure of Geometric Transformation:** This was a critical failure mode. Landmark analysis (Section F) for SDXL Refiner showed minimal geometric change between the neutral input and the generated "happy" output (mean difference: 3.88 pixels), whereas the real happy expression exhibited substantial shifts (mean difference from neutral: 16.33 pixels). The generated face failed to adopt the characteristic shapes of the target expression (e.g., lacking smile curvature, raised cheeks, or eye crinkling), effectively remaining a slightly modified neutral pose.
2. **Inaccurate Semantic Representation:** Generated expressions often failed to convey the intended emotion semantically. CLIP analysis (Section E) quantified this, showing the SDXL Refiner's generated "happy" image achieved only 20% of the semantic improvement of the real image, while the "surprise" generation resulted in a negative semantic score (-35%), indicating it was less representative of surprise than the neutral input. This demonstrates a failure to translate the text prompt into features CLIP associates with that emotion.
3. **Hallucination of Missing/Altered Core Features:** Key anatomical features expected in certain expressions were frequently missing or altered. Most consistently, **visible teeth**, present in the ground-truth happy expression, were entirely absent in outputs from both SDXL Refiner and Dreamshaper (Figure 29a, 28). This represents a significant hallucination where an expected feature is deleted or occluded.
4. **Unnatural Textural Modifications:** Both models introduced textural inconsistencies. Dreamshaper, within its inpainted regions, generated **unnaturally thick, dark, and uniform eyebrow textures** (Figure 28b), a distinct textural hallucination. SDXL Refiner tended towards generalized skin smoothing, similar to FSR models but inappropriate for maintaining realism during expression changes. Low SIFT match counts (7 for SDXL, 9 for Dreamshaper vs. 143 keypoints in the original happy image) confirmed widespread local texture divergence from the ground truth for both models (Sections F, E).
5. **Color Inconsistency and Artifacts:** While SDXL Refiner maintained high color histogram similarity (0.982), the Dreamshaper inpainting approach resulted in lower similarity (0.894) and noticeable color shifts (Figure 28c). Furthermore, Dreamshaper introduced visually distracting **bluish patch artifacts** at the boundaries of the inpainted regions (e.g., hairline, chin), indicating imperfect blending (Figure 28b).

InstructPix2Pix Behavior

1. **Overemphasis on Jaw Region:** Keypoint analysis revealed that InstructPix2Pix disproportionately modifies the jaw region across all expressions. As seen in Figure 44, the number of keypoints on the jawline in generated images far exceeds those in ground-truth images. This indicates an over-reliance on lower facial geometry to encode expression, while upper-face regions (e.g., eyebrows, eyes) remain underutilized. However, they were still generated good enough to recognise the emotion.

2. **Success and Failure in Conveying Expression through Texture:** Despite structural changes, InstructPix2Pix frequently managed to generate necessary local texture cues. Smiles had teeth and cheek creases, and anger expressions showed overly smoothed or exaggerated eyebrows with fine-grained skin wrinkling. Whereas for StableDiffusionXL-Refiner it wasn't able to understand the prompt very well and relied heavily on the identity image.

3. **Possible Identity Overpreservation:** Unlike SDXL Refiner, InstructPix2Pix doesn't prioritize identity preservation over prompt-driven change. But, for StableDiffusionXLRefiner the facial expression transformations remain shallow, often preserving the input face structure too strictly at the cost of emotional authenticity.

Reasons for Expression Generation Failures:

- **Challenge of Semantic Control:** Translating abstract emotional concepts ("happy", "sad") from text prompts into precise, coordinated geometric and textural changes across multiple facial features remains a significant hurdle for current models.

- **Input vs. Prompt Conflict:** The observed tendency for generated expressions to remain geometrically close to the neutral input (especially for SDXL Refiner) suggests models may over-prioritize fidelity to the source image structure, resisting the substantial deformations required by the prompt.

- **Complexity of Expressive Deformation:** Accurately modeling the complex, non-linear muscle movements, skin stretching/wrinkling, and feature interactions (like lip corner retraction revealing teeth) that constitute realistic facial expressions is difficult, particularly when guided only by a high-level text prompt.

- **Limitations of Inpainting for Expression:** While masking targets specific areas, the inpainting model (Dreamshaper) struggled to generate contextually appropriate features (like teeth or natural eyebrows) within the mask and failed to blend the inpainted regions seamlessly, introducing color and texture artifacts.

Expression Model Behavior Comparison:

- **SDXL Refiner:** Demonstrated strong global consistency (color, overall structure) relative to the input but failed significantly in applying the requested geometric and semantic changes for expression, resulting in subtly modified neutral faces lacking emotional content and missing key features like teeth.

- **Dreamshaper Inpainting:** Attempted more localized changes guided by a mask but produced noticeable artifacts (color patches, blending issues), introduced specific textural hallucinations (eyebrows), exhibited color shifts, and similarly failed to reproduce accurate expression geometry or features (teeth).

- **StableDiffusionXL Refiner:** Despite being trained for high-fidelity refinement, the model prioritized identity preservation over prompt-driven transformation. It consistently failed to implement the geometric changes necessary to convey expressions like happiness or surprise, producing outputs that were visually close to the neutral input. While color and texture realism were preserved, the expression-specific modifications were largely absent or minimal, highlighting its limitation in prompt-following for emotional content.

- **InstructPix2Pix:** Exhibited the strongest alignment with the intended expression prompt among the tested models. It produced visibly different and semantically correct emotional expressions (e.g., smiling with raised cheeks or widened eyes in surprise). However, its heavy reliance on the prompt made it sensitive to prompt phrasing and introduced variability in output quality. Noisy or ambiguous prompts led to distorted or exaggerated features, occasionally resulting in unrealistic faces.

The core challenges identified in this expression generation task relate less to the fine-grained texture invention seen in FSR, and more to the fundamental difficulties in achieving accurate semantic control and complex, prompt-guided geometric manipulation of facial features without generating inconsistencies or artifacts.

IX. Conclusion

This investigation focused on identifying and characterizing specific facial features commonly hallucinated by state-of-the-art FSR models (GFPGAN, GPEN, CodeFormer) when restoring faces from degraded inputs (Section **??**). Our primary objective was to pinpoint *which* anatomical features are most prone to distortion. We utilized a multi-modal evaluation strategy, integrating SIFT/landmark analysis for local inconsistencies (Section **??**), LPIPS for perceptual difference (Section B), and a targeted VQA approach for semantic comparison (Section C).

Our main finding confirms a consistent pattern: certain features are highly susceptible to hallucination across models and conditions. These notably include **fine skin texture** (often replaced by unrealistic smoothness), **hair and eyebrow details** (simplified structure and texture), **fine periocular features** (altered wrinkles, eyelids, reflections), **subtle lines/expressions**, and **minor unique imperfections** (normalized or removed). Faintly visible **teeth** were also sometimes

eliminated. These alterations arise from information loss in LR inputs and the strong influence of model priors substituting plausible but inaccurate details.

Model comparisons highlighted differing tendencies and potential trade-offs. GPEN generally excelled in perceptual similarity (LPIPS), while CodeFormer showed better average semantic consistency (VQA similarity) in our tests, often accompanied by texture smoothing. GFPGAN produced sharp results but sometimes with artificial textures. The multi-modal evaluation proved effective, with VQA offering valuable, interpretable insights into specific anatomical changes.

Parallel analysis of the **expression generation** task revealed a different set of primary challenges and failure modes. While FSR primarily deals with reconstructing lost detail, expression generation involves controlled modification based on semantic prompts. Our key finding here is the significant difficulty models (SDXL Refiner, Dreamshaper Inpainting) face in accurately translating high-level emotional concepts into correct facial geometry and appearance. The most prominent failures included a lack of substantial **geometric transformation** (generated faces remaining too similar to the neutral input, confirmed by landmark analysis), poor **semantic fidelity** where the generated expression did not align with the intended emotion (quantified by inconsistent and sometimes negative CLIP scores), and the consistent **hallucination or removal of key expressive features**, such as visible teeth in smiles.

Furthermore, expression generation models introduced their own types of artifacts and textural inconsistencies, distinct from typical FSR smoothing. These included unnatural **textural modifications** like artificially uniform eyebrows (Dreamshaper), significant **color shifts** and **blending artifacts** (especially with inpainting), and poor local feature correspondence despite prompts (low SIFT matches). These failures likely stem from the inherent difficulty in semantic control, potential conflicts between maintaining input fidelity and applying prompt-based deformations, and the complexity of modeling realistic facial muscle movements. SDXL Refiner tended to under-deliver on transformation, while Dreamshaper introduced more visual inconsistencies.

Stable Diffusion-based models such as StableDiffusionXLRefiner also prioritized preserving global image structure and identity over making expressive, prompt-driven changes. In doing so, they frequently ignored critical facial deformations required by the prompt—failing to produce meaningful differences between ”neutral” and ”happy” or ”surprised” faces. This reflects a broader challenge within the Stable Diffusion architecture: balancing fidelity with creative transformation, where the model defaults toward photorealism and coherence, even at the cost of semantic alignment. By contrast, **InstructPix2Pix**, which relies heavily on the textual prompt and fine-tunes generation with strong semantic conditioning, showed better alignment with the intended expressions. Despite producing noisier and sometimes exaggerated outputs, it consistently delivered faces with correct emotional direction—such as smiles with raised cheeks or widened eyes in surprise—even when realism or identity fidelity was compromised. This demonstrates the effectiveness of prompt-sensitive editing in guiding expressive transformations, albeit with a trade-off in visual consistency.

Ultimately, this work contributes a differentiated catalog of commonly hallucinated facial features and failure modes for both FSR and expression generation. This provides tangible targets for future research aimed at enhancing the fidelity and trustworthiness specific to each task. Understanding precisely *what* gets hallucinated, and *why* it differs between reconstruction and editing paradigms, is essential for developing next-generation models that better preserve identity and truthful details, crucial for reliable applications in diverse domains.

X. Future Work

Building on this focused study of hallucinated features, future work could involve:

- Expanding the analysis to a larger, more diverse dataset to confirm the universality of the identified hallucinated features.
- Quantifying the *degree* of hallucination for each feature type more rigorously, perhaps developing specific metrics.
- Investigating the root causes within model architectures that lead to specific feature hallucinations.
- Directly linking hallucination patterns to the models' training data and priors.
- Assessing the impact of these specific feature hallucinations on downstream tasks like face recognition or emotion analysis.

XI. Contributions

- Om Kumar - FSR models (inferencing) and VQA based comparison approach
- Vinit Thakur - SIFT with dlib, and lpips comparison

- Shyam - Applied SIFT based analysis, Clip based semantic analysis, facial landmarks based analysis and color histogram based analysis on two different Diffusion based models (sdxl refiner and dream inpainting) for expression generation and inpainting.
- Neermita - Assessing how noise affects Diffusion models (InstructPix2Pix and StableDiffusionXLRefiner) for P1 and P2 from DeeperForensics Dataset. Quantifying SIFT keypoints in facial landmarks. Streamlit Application.

References

[1] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, 2020.

[2] 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.

[3] Davis King. Dlib-ml: A machine learning toolkit, 2009.

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Gfp-gan: Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9283–9292, 2021.

[6] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021.

[7] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, 2022.