PATMAT: Person Aware Tuning of Mask-Aware Transformer for Face inpainting

Saman Motamed[1,2], Jianjin Xu[1], Chen Henry Wu[1], Christian Häne[3], Jean-Charles Bazin[3], Fernando De la Torre[1]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
[2] INSAIT, Sofia University, Bulgaria
[3] Independent Researcher

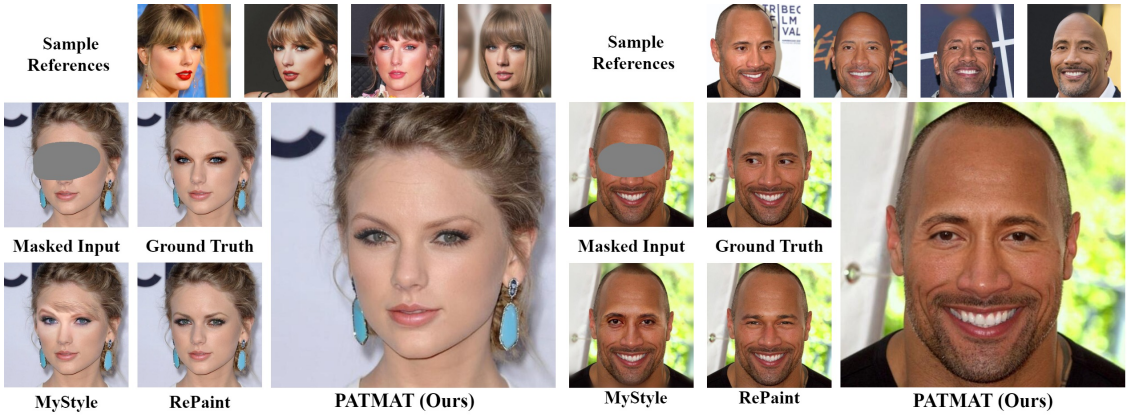

Figure 1: An illustration of PATMAT's identity preserving inpainting. By using reference images of Taylor Swift and Dwayne Johnson, PATMAT enables Mask-Aware Transformer to preserve the identity of the person. We also show results of recent methods such as MyStyle [25], trained with the same number of reference images as PATMAT and RePaint[22] that is not fine-tuned with any reference images.

Abstract

Generative models such as StyleGAN2 and Stable Diffusion have achieved state-of-the-art performance in computer vision tasks such as image synthesis, inpainting, and de-noising. However, current generative models for face inpainting often fail to preserve fine facial details and the identity of the person, despite creating aesthetically convincing image structures and textures. In this work, we propose Person Aware Tuning (PAT) of Mask-Aware Transformer (MAT) for face inpainting, which addresses this issue. Our proposed method, PATMAT [1], effectively preserves identity by incorporating reference images of a subject and fine-tuning a MAT architecture trained on faces. By using ~ 40 reference images, PATMAT creates anchor points in MAT's style module, and tunes the model using the fixed anchors to adapt the model to a new face identity. Moreover, PATMAT's use of multiple images per anchor during training allows the model to use fewer reference images than competing methods. We demonstrate that PATMAT outperforms state-of-the-art models in terms of image quality, the preservation of person-specific details, and the identity of the subject. Our results suggest that PATMAT can be a promising approach for improving the quality of personalized face inpainting [2].

1. Introduction

The objective of image inpainting is to generate plausible content to complete missing regions within an image. Preserving contextual integrity of the inpainted image is a crucial factor, where the reconstructed regions must conform

[1] All experiments and data processing activities were conducted at Carnegie Mellon University.

[2] The code will be available at https://github.com/humansensinglab/PATMAT. For any questions, please reach out to sam(dot)motamed(at)insait(dot)ai.

to reasonable structure and texture based on local and non-local priors in the image. This task becomes increasingly challenging when addressing larger and irregular missing regions in the image. In particular, facial inpainting poses a significant challenge as it requires maintaining fine facial details and the subject's identity, which are essential for various applications including security (e.g., inpainting a face behind a mask or sunglasses), entertainment (e.g., seeing a person while they are wearing a virtual reality headset), or photo restoration. In each of these tasks, it is essential to keep the subject's identity and fine facial details intact. However, recent inpainting methods have largely focused on generating high-quality images with little attention given to preserving the identity of the subject. Therefore, we explore the feasibility of inpainting techniques that aim to maintain the subject's identity and facial characteristics while generating high-quality, photo-realistic images.

Recent works have pushed the boundaries of large-hole image inpainting [17, 36, 35, 23]. Mask-Aware Transformer (MAT) [17] used vision transformers, with a multi-head contextual attention mechanism with shifting windows [20] to build long-range dependency priors and achieved great inpainting results compared to competing methods such as CoModGAN [43], ICT [36] and LaMa [35]. While these models are able to synthesize high quality inpainted images, the reconstruction result sometimes fails to recover the details of the ground truth. Particularly, in the case of face inpainting, existing models (e.g., MAT, CoModGAN, RePaint) cannot recover the same subtle facial details as the ground truth (e.g., brows shape, eye look, beard shape,...). As a result, they do not preserve the subject's identity after inpainting (see Fig. 1). For tasks such as image editing and restoration, especially in the domain of faces, identity preservation is a requirement, yet the problem of identity preserving image inpainting is relatively unexplored.

This work proposes PATMAT for personalized face inpainting. Given a few reference images of a person, PATMAT integrates this information into a pre-trained MAT model by tuning the network parameters, conditioned on the style vectors within MAT. We compare two style conditioning methods and propose a regularization loss to prevent over-fitting to the reference images during the tuning process. To evaluate our method, we curated images of seven public figures due to the lack of diverse and sufficiently large datasets (no public datasets have sufficient number of images per identity such as CelebA-HQ [12] and VFHQ [39]). Our qualitative and quantitative experiments demonstrate that PATMAT outperforms several contemporary state-of-the-art inpainting models in terms of quality and identity-preservation.

Our contributions are:

1. We propose PATMAT, a tuning method based on creating anchors in MAT's style space that allows for high quality personalized inpainting of the face.

2. We propose a regularization method that controls overfitting during tuning and further improves the quality of personalized inpainting.

3. In order to make our personalized method more practical, our effort reduces the number of required reference images from $\sim 100 - 200$ in previous works [25] to ~ 40 images. algorithm

2. Related Work

Image inpainting has been a long standing task in computer vision. Early pixel-matching and diffusion-based inpainting methods [45, 2, 3], which propagated neighbouring pixel information to the masked region lacked a high level understanding of the image that hindered them from generating semantically reasonable images and have since been replaced by more sophisticated deep learning approaches.

Pathak *et al.*'s [27] use of an encode-decoder architecture, equipped with adversarial training and a pixel-wise reconstruction loss objective achieved photo-realistic inpainting results and has since been the basis of more follow-up works [11, 18, 26, 41, 29].

With recent advances of diffusion models in image synthesis [15, 7, 31, 24, 32], RePaint [22] leveraged the expressiveness of a pre-trained Denoising Diffusion Probabilistic model [10] and used it as a prior in their inpainting model.

Success of GANs in synthesizing high quality images [13, 14] brought forward different methods and variations [19, 5, 4], adapting them for image inpainting. Most GAN-based inpainting methods are prone to deterministic transformations due to limited control during synthesis [38]. UTCGAN [42] and Zhao *et al*. [44] proposed VAE-based networks to combat this issue. By projecting images and masked counterparts onto a low-dimensional manifold space and optimizing the KL-divergence, UTCGAN [42] achieved pluralistic generation capabilities for mask filling. CoModGAN [43] used a co-modulation layer in order to improve reconstruction and diversity of their inpainting.

To the best of out knowledge, PTI [30] and MyStyle[25] are the only works that focused on personalizing models for editing out-of-domain (OOD) images. MyStyle built on the work presented by pivot tuning (PTI) of StyleGAN's [13] latent space. PTI was proposed as a method for fine-tuning a pre-trained Generative Adversarial Network that enabled the model to generate OOD images. Given an OOD image of interest (x_i), PTI tunes the Generator to generate x_i. First, by projecting x_i onto the pre-trained GAN's latent space, a pivot latent code w_i is acquired. While $G(w_i)$ has some characteristics of x_i, due to the image being OOD, it cannot fully recover the same identity of the person. PTI, while keeping w_i unchanged, trains G to minimize $\mathcal{L}(G(w_i), x_i)$

where $\mathcal{L}$ is the reconstruction objective. MyStyle used pivot tuning and $\sim 100-200$ reference images in order to create a personalized manifold in the latent space of a StyleGAN and showed identity preserving image editing capabilities such as inpainting and super resolution in this manifold.

PTI and MyStyle work with the latent space of GANs and require a large number of reference images to learn a personalized manifold. MAT however, does not operate with a GAN-like latent space. Instead, it uses a noise-style space and a style manipulation module that enables pluralistic generation. We draw inspiration from PTI's use of anchors and condition the style manipulation module by defining anchors in the noise-style space. We show that our method is able to preserve the identity for inpainting and achieves high quality image synthesis, competing with recent SOTA inpainting models. PATMAT's use of multiple images per anchor during training allows the personalization of the model by using as few as ~ 40 reference images (see PATMAT-C's improvement in identity preserving, that uses multiple images per anchor over PATMAT-S that uses one image per anchor in table 2).

3. Method

Given an image x of person with identity p, a binary mask image b, and a set of reference images $\mathcal{X} = \{x_i\}_{i=1}^{N}$ of person p, PATMAT aims to personalize face inpainting by fine-tuning a pre-trained MAT, such that $x \odot b$ can be filled with plausible content that preserves the identity of p.

3.1. MAT Architecture

The recently proposed MAT [17] achieved great quality inpaintings of images with large holes, compared to an ensemble of recent models [43, 36, 40, 42]. MAT used vision transformer blocks [8] to model long range dependencies in the image. In addition, inspired by [14], MAT introduced a style manipulation module that enabled pluralistic mask filling for a single masked image.

Given an input image x, a binary mask image b, and an unconditional noise-style code $\mathbf{s}_u \in \mathbb{R}^{512}$, MAT generates an inpainted image $\hat{x} = \text{MAT}(x \odot b, \mathbf{s} = \mathcal{A}(\mathbf{s}_u, \mathbf{s}_c))$. $\mathbf{s}_u$ is used in formulating $\mathbf{s}$ (3), that is tasked with manipulating MAT's weight normalization of convolution layers in the image reconstruction layers, such that using different $\mathbf{s}_u$ inputs results in different inpaintings of the masked image input. To enhance the representation ability of $\mathbf{s}_u$, MAT fuses ($\mathcal{A}$) the noise-style code $\mathbf{s}_u$ with an image-conditional style code $\mathbf{s}_c \in \mathbb{R}^{1024}$. $\mathbf{s}$ and $\mathbf{s}_c$ are formulated as:

$$\mathbf{X}^{'} = \mathbf{B} \odot \mathbf{X} + (1 - \mathbf{B}) \odot \text{Resize}(\mathbf{s}_u) \quad (1)$$

$$\mathbf{s}_c = \mathcal{F}(\mathbf{X}^{'}) \quad (2)$$

$$\mathbf{s} = \mathcal{A}(\mathbf{s}_u, \mathbf{s}_c) \quad (3)$$

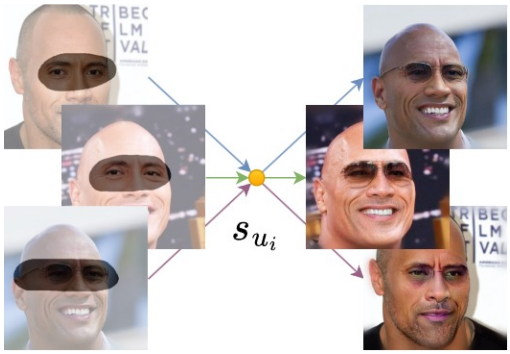

Figure 2: An illustration of how using the same pivot to inpaint images with and without sunglasses leads to the features of sunglasses leaking to other images.

Where $\mathcal{F}$ and $\mathcal{A}$ are mapping functions, $\mathbf{B}$ is a random binary mask and $\mathbf{s}_c$ is conditioned on both the image features $\mathbf{X}$ (output of transformer blocks given image x) and the unconditional noise-style $\mathbf{s}_u$. The resulting style code $\mathbf{s}$ then changes the inpainted image by manipulating the weights $\mathbf{W}$ of convolution layers in the reconstruction network:

$$\mathbf{W}'_{ijk} = \mathbf{W}_{ijk} \cdot \mathbf{s}_i \quad (4)$$

where i, j, k denote the input channels, output channels and spatial footprint of the convolution. PATMAT uses the unconditional noise-styles $\mathbf{s}_u$ in order to personalize MAT.

3.2. Person Aware Tuning

PATMAT allows tuning of MAT through its style manipulation framework using a few reference images, enabling the inpainting of faces while preserving the identity. We draw inspiration from PTI's [30] idea of fixing a pivot in a GAN's latent space and tuning the generator to generate an out-of-distribution (OOD) image using the fixed pivot.

PATMAT conditions the $\mathbf{s}_u$ noise-style codes by treating them as anchors and fixing them during training, such that the inpainted image $\hat{x}_i = MAT_\theta(x_i \odot b, \mathcal{A}(\mathbf{s}_{c_i}, \mathbf{s}_{u_i}))$ minimizes the reconstruction objective formulated as:

$$\mathcal{L}_\theta = \mathcal{L}_P + \mathcal{L}_2(\hat{x}_i, x_i) \quad (5)$$

where θ represents the parameters of MAT, $\mathcal{L}_P = ||\phi_i(\hat{x_i}) - \phi_i(x_i)||_1$ is the perceptual loss [9], $\phi_i(.)$ denotes the layer activation of a pre-trained VGG-19 [34] network and $\mathcal{L}_2$ is the Mean Squared Error.

3.2.1 Style Conditioning

We propose two different methods for tuning MAT, using reference images $x_i \in \mathcal{X}$ and corresponding noise-style anchors $\mathbf{s}_{u_i}$. PATMAT-S uses a single image per anchor in its tuning mechanism (similar to how PTI uses one latent

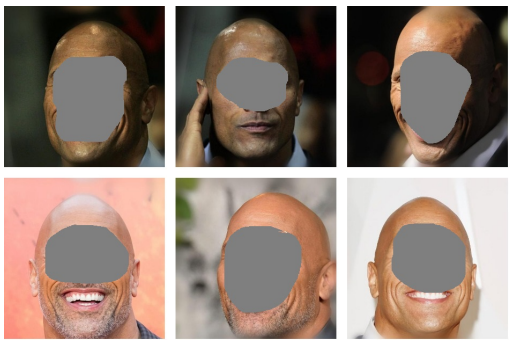

Figure 3: An illustration of how different lighting conditions in the training images can drastically affect the skin-tone of the person.

code per image) while PATMAT-C uses multiple images per anchor. We explored two methods (random vs optimized) for calculating initial anchors in section 4.2 and found both to be equally effective.

PATMAT-S. For each reference image x_i and its corresponding anchor $\mathbf{s}_{u_i} \in \mathcal{S}_u$, while keeping the anchor fixed, we tune MAT such that inpainted image $\hat{x}_i$ minimizes $\mathcal{L}(x_i, \hat{x}_i(\theta))$ (5),

$$\theta = \text{argmin}_\theta \sum_{i=1}^{N} \mathcal{L}(x_i, \text{MAT}_\theta(x_i \odot b, \mathcal{A}(\mathbf{s}_{u_i}, \mathbf{s}_{c_i}))), \quad (6)$$

where θ is the parameters of MAT and N is the number of reference images. While this improved identity preserving ability of MAT for inpainting, we found that tuning with multiple images per anchor better utilizes the limited number of reference images and further improved the inpainting by taking advantage of the ability that MAT's style manipulation module has in using a single $\mathbf{s}_u$ to inpaint multiple images.

PATMAT-C. Observe that MAT can use multiple reference images x_j per style anchor $\mathbf{s}_{u_i}$. In other words, we can tune the network using $\mathcal{L}(x_i, MAT_\theta(x_i \odot b, \mathcal{A}(\mathbf{s}_u, \mathbf{s}_{c_i})))$. The main difference with PATMAT-S is that in PATMAT-C, the $\mathbf{s}_u$ anchors stay the same for all the reference images. Unlike latent vectors in a GAN's latent space that are unique per each image, MAT's style code $\mathbf{s}$ (3) is comprised of both the noise-style anchor $\mathbf{s}_{u_i}$ and an image-conditioned style $\mathbf{s}_{c_i}$. Given two images $x_i, x_j \in \mathcal{X}$, $\mathbf{s}_{u_i}$ can be used to inpaint both images since $\mathbf{s}_{c_i}$ will be different from $\mathbf{s}_{c_j}$. Instead of tuning MAT to inpaint each image using its unique style anchor, we can use each style anchor $\mathbf{s}_{u_i} \in \mathcal{S}_u$ and enforce the network to generate all reference images $x_j \in \mathcal{X}$ using that single style anchor.

However, enforcing MAT to reconstruct every image $x_j \in \mathcal{X}$ using a single anchor s_{u_i} is too restrictive. We observed that during tuning, using the same anchor to inpaint images with distinct accessories will lead to features of those accessories propagating to other images. Figure 2 shows an example of this phenomena where an anchor $\mathbf{s}_{u_i}$ is used to inpaint both images with and without sunglasses, leaving the inpainted results with glasses-like shadow over the eyes where there are no sunglasses in the ground truth. This kind of effect can also happen to images with drastic difference in skin tone, as shown in figure 3. To mitigate this type of effect, we perform a manual grouping of the reference images. Most importantly, we separate images with glasses and sunglasses, each in their own group. Furthermore, in the case of having images with noticeably different skin tone as shown in figure 3, we group such images as well. This step not only improves the inpainting results, but also speeds up PATMAT by reducing the number of image - anchor pairs used in training. This results in M groups of reference images, $\mathcal{X}_m$, and the anchor for each group, $\mathcal{S}_{u_m} = \{\mathbf{s}_{u_m}\}$. As a whole, we have $\mathcal{X} = \cup_{m=1}^{M} \mathcal{X}_m$, and $\mathcal{S}_u = \cup_{m=1}^{M} \mathcal{S}_{u_m}$. In our experiments, $M \sim 3$. Now, for all groups of images $\mathcal{X}_m$ and their corresponding anchors $\mathbf{s}_{u_m}$, we tune MAT by inpainting each $x_i \in \mathcal{X}_m$ using all $s_{u_i} \in \mathcal{S}_{u_m}$. Our results in table 2 showed that PATMAT-C performed better in preserving the identity compared to PATMAT-S and our qualitative study (4.5.2) further confirmed this finding.

Furthermore, by adding random noise to anchors and creating anchor clusters, we observed more stable inpainting results over different runs of the algorithm. In this setting, for each group of images $\mathcal{X}_m$ and their corresponding style anchors $\mathcal{S}_m$, we fixed $|\mathcal{X}_m| - 1$ new anchors around each $s_{u_i} \in \mathcal{S}_m$ where each new anchor $s_{u_{i,m}}$ is calculated as $s_{u_i} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. PATMAT-C then uses unique anchors for each reference image, such that x_i is inpainted using s_{u_i} and $s_{u_{j,m=i}}$ for $j \in \{1, 2, ..., |\mathcal{X}_m|\}$ where $j \neq i$. PATMAT-C can be described as

$$\text{argmin}_\theta \sum_{m=1}^{M} \sum_{x_i \in \mathcal{X}_m} \mathcal{L}(x_i, \hat{x}_i(\theta)), \quad (7)$$

$$\hat{x}_i(\theta) = \text{MAT}_\theta(x_i \odot b, \mathcal{A}(\mathbf{s}_{u_{j,m}} + \epsilon_i, \mathbf{s}_{c_i})), \quad (8)$$

where $\hat{x}_i(\theta)$ is the inpainted image, $\mathbf{s}_{u_{j,m}}$ stands for the j^{th} shared cluster anchor for group m, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise added for each image, $\mathbf{s}_{c_i}$ is the image conditional style vector for image x_i.

Note that training without fixed anchors faces the same issues as we showed in figure 2. See Appendix D for more details.

3.2.2 Regularization via Anchors

Tuning MAT using fixed anchors exhibits over-fitting similar to those observed in generative models[25, 30]. This occurs in the form of cascading the features of person p not only local to anchor style codes, but throughout the network.

Figure 4 - c shows the inpainting result of a masked image after PATMAT tuning the network with reference images of Dwayne Johnson. While we want to use PATMAT for personalized inpainting and do not worry about images of other identities being misrepresented, figure 5 - top row shows examples of this over-fitting interfering with inpainting images of person p where there is misalignment in the inpainted teeth and eyebrows. To mitigate this, we use a small number (~ 30) of randomly sampled images from MAT's original training data $\mathcal{X}_T$ (or any other set of random images that are not of person p). We regularize PATMAT using $\mathcal{X}_T$. We fix a unique style anchor $\mathbf{s}_{u_t}$ for each $x_t \in \mathcal{X}_T$ and at each iteration, minimize $\mathcal{L}_{reg} = \mathcal{L}(\hat{x}_r, \hat{x}_r^{'})$ (5) where:

$$\hat{x}_r = \text{PATMAT}_{\theta'}(x_t \odot b,\ \mathbf{s}_{u_t})$$
$$\hat{x}_r^{'} = \text{MAT}_{\theta}(x_t \odot b,\ \mathbf{s}_{u_t})$$

In other words, we regularize the tuning by enforcing the new PATMAT network to generate the same inpainting on a random set of images $\mathcal{X}_T$ that the pre-trained MAT would generate, using the same anchor point. Figure 4-d shows the regularization effect where PATMAT tuning of Dwayne Johnson does not interfere with somewhat reasonable inpainting of another face. Regularization improves the inpainting results for person p. Figure 5 shows the results of two networks, trained for personalizing Barack Obama and Dwayne Johnson. The regularized networks (bottom row) better synthesises color consistency and facial feature alignment compared to the unregularized networks (top row) where there is misalignment in the teeth and eye-brows.

Algorithm 1 gives an overview of PATMAT-C. For PATMAT-S, Cartesian products "$\times$" are replaced by dot products "$*$".

Algorithm 1 PATMAT

initialize $\text{PATMAT}_{\theta'} = \text{MAT}_{\theta}$
for $(\mathcal{X}_m, \mathcal{S}_{u_m})$ in $\mathcal{X} \times \mathcal{S}_u$ **do**
  for (x_i, s_{u_i}) in $\mathcal{X}_m \times \mathcal{S}_{u_m}$ **do**
    $\theta^{'} \leftarrow \text{argmin}_{\theta'}\ \mathcal{L}(\text{PATMAT}_{\theta'}(x_i \odot b, s_{u_i}),\ x_i)$
    for $(x_j^{'}, s_{u_j}^{'})$ in $\mathcal{X}_T * \mathcal{S}_T$ **do**
      $\theta^{'} \leftarrow \text{argmin}_{\theta'}\ \mathcal{L}(\text{PATMAT}_{\theta'}(x_j^{'} \odot b^{'}, s_{u_j}^{'}),$
      $\text{MAT}_{\theta}(x_j^{'} \odot b^{'}, s_{u_j}^{'}))$
    end for
  end for
end for
return $\text{PATMAT}_{\theta'}$

3.3. Inpainting

After tuning the network with PATMAT, given a masked image $x_{ib} = x_i \odot b$ to be inpainted, we optimize over noise-style codes in order to refine the inpainting results. While we used a combination of perceptual and $\mathcal{L}_2$ loss to fine-tune the model, pixel-wise metrics are not appropriate for optimizing an identity preserving operations since the same identity can have different facial expressions. We opt to use the ArcFace [6] face identity detection model ($\phi(.)$) instead. We optimize s_u with the objective to minimize the following loss during inpainting:

$$s_u = \text{argmin}_{s_u} \mathcal{L}_{\text{ArcFace}}, \quad (9)$$
$$\mathcal{L}_{\text{ArcFace}} = 1 - \cos(\phi(\text{PATMAT}_{\theta'}(x_{ib}, s_u)), \phi(x_p)), \quad (10)$$

where x_p is a random image from reference images $\mathcal{X}$.

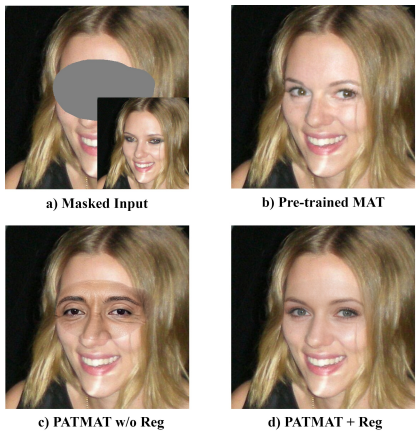

Figure 4: An illustration of PATMAT's over-fitting. **b** shows a pre-trained MAT's inpainting results for a given masked image (**a**). **c** and **d** show the result using the same input **a**, with PATMAT tuned for personalizing images of Dwayne Johnson, without and with regularization, respectively. Note that these images are not a scenario where we would use PATMAT since the image to be inpainted is not of Dwayne Johnson and this figure is only used to illustrate the over-fitting and regularization effects.

Figure 5: An illustration of the effects of Style Regularization on PATMAT's performance. Regularization (bottom row) improves structural integrity of the inpainting compared to no regularization (top row).

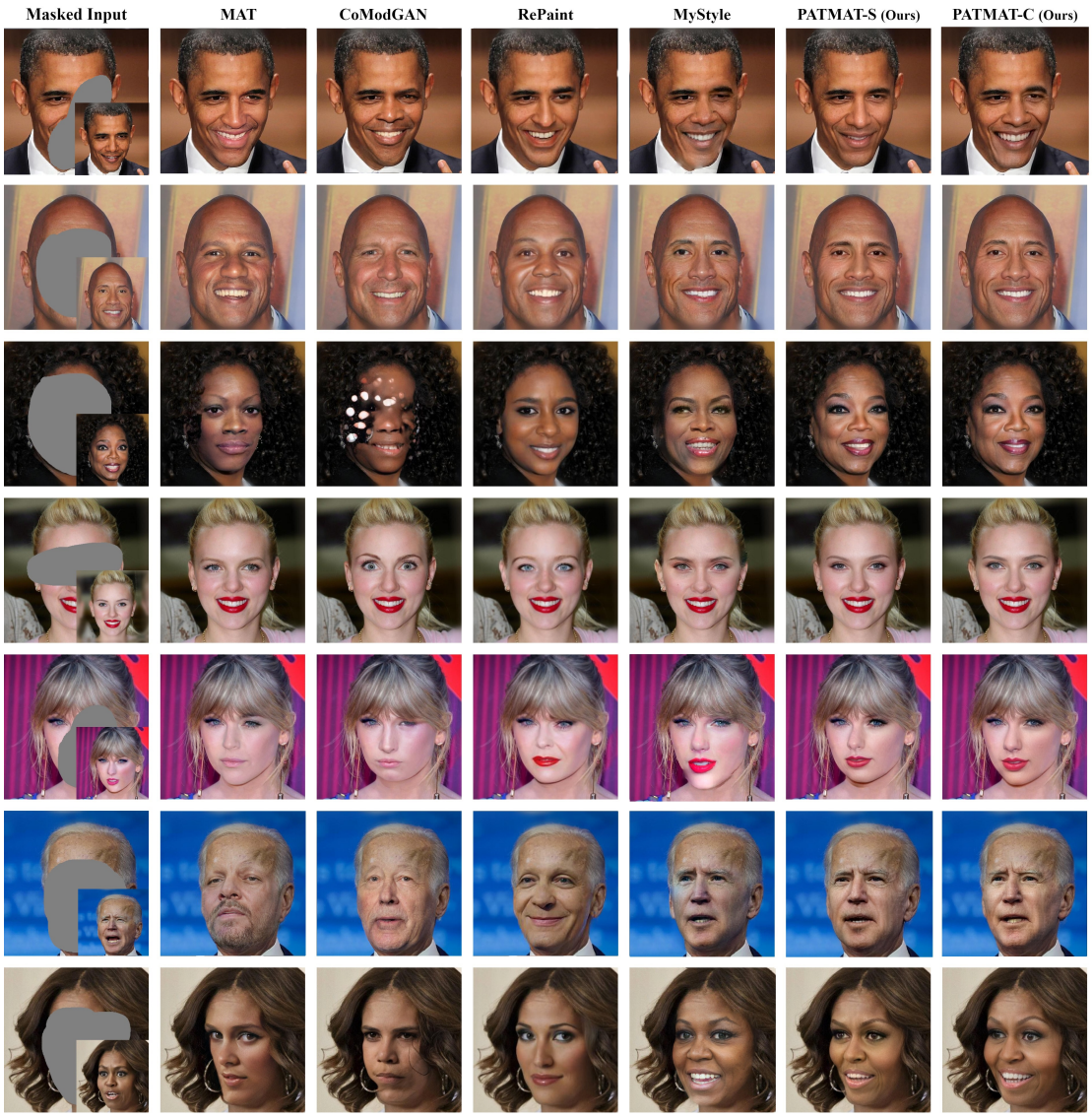

Figure 6: A qualitative comparison of different inpainting models, with images of seven identities used in this study (from top to bottom: Barack Obama, Dwayne Johnson, Oprah, Scarlett Johansson, Taylor Swift, Joe Biden and Michelle Obama). Only PATMAT and MyStyle used reference images for tuning. MAT and RePaint were pre-trained on CelebA-HQ while RePaint and MyStyle was pre-trained on FFHQ. Please zoom in to see details.

4. Experiments

4.1. Datasets and Metrics

There is a lack of high-quality public datasets that could cater to the problem of identity preserving image editing. For this reason, studies resort to using images of celebrities and public figures [25]. However, such images are not representative of day to day photos, since celebrities are often posing and smiling for photos, often in bright light. Due to this limitation, in this paper we show our method's identity preserving inpainting performance using images of seven known figures (Michelle and Barack Obama, Joe Biden, Taylor Swift, Oprah, Scarlett Johansson and Dwayne Johnson) (Appendix A). We selected these well known figures to facilitate our perceptual study, since most people know the looks of these celebrities. We evaluated the inpainting results in terms of quality and identity preservation. For quality, we opt to use the **FID** [9] score as it has been widely adopted for reporting the quality of generative model samples [14, 17, 35, 43] and has shown to correlate well with human perception. Pixel-wise **L1** distance, **SSIM**[37] and **PSNR** have a weak correlation with human perception re-

garding image quality [33, 16]. For identity preservation, we rely on a pre-trained face recognition network, ArcFace [6]. We calculated the cosine similarity of features between the inpainted image and the ground truth. We used different models in inpainting and evaluation of our methods. For evaluation, we used the R50 model trained on the MS1MV3 dataset; for inpainting, we used the R100 model trained on the Glint360K dataset [1].

Our experiments showed that ArcFace disregards some degree of misalignment and color discrepancy between inpainted areas of the image and the outer mask. For instance, ArcFace would give a high similarity score to PATMAT's inpainting of Barack Obama in figure 8, where a human judge would immediately see the image as abnormal. We further conducted a user study to measure the performance of PATMAT with human judges.

4.2. Finding Style Anchors

MAT uses the same mapping network as StyleGAN [14], mapping Gaussian noise $\mathbf{z}_i$ to noise-style code $\mathbf{s}_{u_i}$. PATMAT uses an initial style anchor $\mathbf{s}_{u_i}$ per reference image x_i. We experimented with two methods for obtaining an initial style anchor for each image.

4.2.1 Random Style Anchors

In one set of experiments, for each reference image x_i, we picked a random noise z_i, mapped to s_{u_i}.

4.2.2 Optimized Style Anchors

In the second set of experiments, we optimized $\mathcal{L}_{\mathbf{z}}(\hat{x}_i, x_i)$ such that the noise vector $\mathbf{z}$ minimizes the similarity between a reference image and its inpainting results, using the pre-trained MAT network and a random mask (similar to projecting images onto a GAN's latent space).

We did not observe a significant difference between the inpainting performance, using the two different methods for assigning anchors.

4.3. Implementation Details

We followed the same architecture design proposed in MAT. Images were resized to 512×512 resolution and were aligned using a face key-point detection model, consistent with those used in FFHQ [13] and CelebA [21] datasets. The initial calculation of style codes were performed by optimizing $\mathbf{z}$ for 200 steps (4.2.2), using the reconstruction loss described in equation 5. For inpainting, we optimized for $\mathbf{s}_u$ via ArcFace loss (9) for 50 steps. During inpainting and quantitative analysis of the inpainting results, we used different pre-trained weights for ArcFace.

Figure 7: An illustration of Poisson blending where PATMAT's raw output (left) shows RGB inconsistencies in the inpainted image, and blending (right) can mitigate this undesired effect.

4.3.1 Poisson Blending

A caveat of using limited reference images is that the images do not cover the many lighting conditions that affect the skin-tone. For that reason, there could be color inconsistencies between the masked region and the prior. To mitigate this undesired effect, we performed a simple yet effective Poisson blending [28] of the output as shown in figure 7.

4.4. Comparison with State of the Art

We compare the proposed PATMAT-C and PATMAT-S with a number of contemporary approaches. For a fair comparison, we use publicly available models to test on the same masks. MyStyle [25], pre-trained on FFHQ [13] dataset, is the only model that is tuned besides PATMAT, using the same reference images. RePaint [22], MAT [17] and PATMAT are trained on CelebA-HQ [21] and CoModGAN is trained on FFHQ. Table 1 shows the tuning (where applicable) and inference cost of different models in seconds/minutes (averages). Experiments were carried out on a single NVIDIA RTX A4000 GPU.

Method	**Tuning**	**Test** (per image)
PATMAT-C	$65'$	$0.04''$
PATMAT-S	$11'$	$0.04''$
MyStyle	$53'$	$42''$
RePaint	✗	$397''$
CoModGAN	✗	$0.08\ ''$

Table 1: ($'$) and ($''$) denote minute and second respectively.

4.5. Results

4.5.1 Quantitative Analysis

Table 2 shows the FID [9] measurement of Ground Truth (train + test) images against different inpainting results on the test images. MAT achieved the best FID, with PATMAT-S and PATMAT-C achieving the second and third best FID scores. To measure identity preservation in the inpainted images, we calculated the average cosine similarity of each inpainted image, against all ground truth (train + test) images.

PATMAT-C achieved the best identity preservation of **0.67**, with ground truth images achieving the threshold of 0.70. MAT, CoModGAN and RePaint achieved low ID scores since they are not tuned to preserve identity. While MyStyle does better than these models, with as few as ~ 40 reference images as PATMAT uses, it often achieves lower quality inpainting compared to PATMAT (see figure 6 and table 2).

4.5.2 Qualitative Analysis

We conducted a perceptual user study to answer the following two questions:

- Which PATMAT model (**S** vs **C**) better preserves the identity?
- Are users able to detect images generated by PATMAT from real images?

To answer the first question, we randomly selected 5 images per identity (35 total) and used the same mask to inpaint the images with both PATMAT-S and PATMAT-C. We asked 19 participants to choose the image that most closely resembled the celebrity without providing them with a reference image. The results of the study showed that 62% of the participants preferred PATMAT-C as a better representation of the person compared to PATMAT-S.

In the second survey, we evaluated the perceptual quality of PATMAT-C's inpainted images by asking human judges to distinguish between real images and inpainted images. 28 participants were presented with sets of images consisting of three real images and one inpainted image or one real image and one inpainted image per question. The judges were then asked to pick the image they believed was inpainted. 10 questions used three real images and one inpainted image and 10 other questions used one real and one inpainted image per question. Participants were able to pick PATMAT's output from three real images with an average 44% accuracy. Given two choices, one real and one inpainted image, the average accuracy increased to 56% (close to random chance).
We excluded obvious failure cases as shown in figure 8 and randomly picked images from the remaining results. The surveys did not use any repeating images (see Appendix B for more details).

4.6. Error Analysis

When the collective of reference images used to train PATMAT fails to cover specific poses, accessories, lighting, etc. that appear in an image to be inpainted, PATMAT can fail to properly inpaint the image. We showed an example of how lighting affects the results and proposed Poisson blending which works well in mitigating this undesired effect. Other failures however, are not easily fixed without retraining the models with more reference images. Figure 8 shows a few examples of such failures, with CoModGAN and MAT's outputs for comparison. Most of PATMAT failed inpainting attempts are on images with poses that are close to a side-view of a face (figure 8 - row 1). The second row of figure 8 shows how PATMAT inpaints part of Taylor Swift's face with makeup on, while the unmasked part does not have makeup. This is due to almost all the reference images having makeup during tuning. Last row shows how PATMAT cannot replicate the shape of the missing sunglasses properly, due to these specific sunglasses not represented in the training data.

	$FID \downarrow$	$ArcFace \uparrow$
Ground Truth	–	0.70
MAT [17]	**22.9***	0.35
MyStyle [25]	24.3	0.55
RePaint [22]	31.0	0.32
CoModGAN [43]	26.0	c 0.34
PATMAT-S (ours)	**23.5**†	**0.63**†
PATMAT-C (ours)	23.7	**0.66***

Table 2: Comparison of FID and face identity similarity between the generated images and the ground truth images using PATMAT and competing methods.
$*$ and † denote the best and second best result respectively

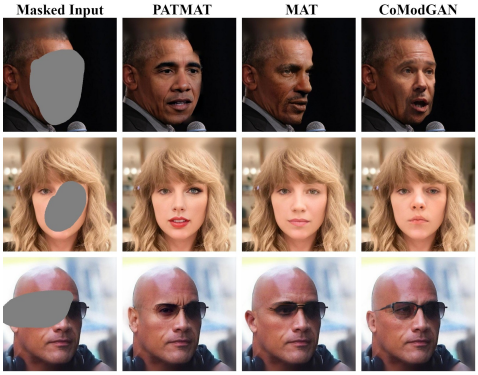

Figure 8: Examples of PATMAT failing to properly inpaint an image. These cases are representative of general types of failures that occur.

5. Conclusion, Limitation and Future Work

This work proposed PATMAT, a tuning method that takes advantage of Mask-Aware Transformer's multi image per style-code inpainting ability. We showed that using as few as ~ 40 reference images, we were able to create a personalized MAT that competes with SOTA inpainting methods in terms of image quality, while preserving the identity of a person of interest. Our study showed that while PATMAT-S has a slight advantage over PATMAT-C in terms of FID, PATMAT-C outperformed PATMAT-S in qualitative inpainting and

identity preservation, as evaluated by the ArcFace model.

Limitation and future work: We provided an insight into PATMAT's limitations in figure 8. PATMAT reflects the limitations in the reference images it uses to create a personalized inpainting space. For instance, we showed that poses where the reference images do not cover, lead to poor inpainting results. We are interested in further exploring the role that style codes play in MAT's image synthesis properties. While PATMAT showed both random (4.2.1) and structured (4.2.2) anchors perform well preserving the identity, it would be interesting to treat the style-space as a GAN-like latent space and test for different properties such as image editing capabilities. Our method currently relies on a manual data separation step where we separate images with glasses and sunglasses and different lighting conditions. We would like to explore end-to-end methods to automate this step and finally, we would like to generalize PATMAT to non-face images (buildings, pets, etc).

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4051, June 2022.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[3] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

[5] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[11] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*, 2020.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[13] 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.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[17] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022.

[18] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, pages 725–741. Springer, 2020.

[19] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9371–9381, 2021.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[22] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[23] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Aishan Liu, Dacheng Tao, and Edwin R Hancock. Regionwise generative adversarial image inpainting for large missing areas. *IEEE Transactions on Cybernetics*, 2022.

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[25] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-or. Mystyle: A personalized generative prior, 2022.

[26] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Chajin Shin, Taeoh Kim, Hanbin Son, et al. Aim 2020 challenge on image extreme inpainting. In *European Conference on Computer Vision*, pages 716–741. Springer, 2020.

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[28] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[29] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019.

[30] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[32] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[33] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500, 2017.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.

[36] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021.

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

[38] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando De la Torre. Generative visual prompt: Unifying distributional control of pre-trained generative models. *arXiv preprint arXiv:2209.06970*, 2022.

[39] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022.

[40] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020.

[41] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[42] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5741–5750, 2020.

[43] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

[44] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[45] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

Supplemental Materials: PATMAT

A. Dataset

We used images of seven public figures and celebrities to train our models. For evaluation, we used 35 images per person, with the number of training images shown in table 3. Dwayne Johnson and Oprah Winfrey were the only people which had images with glasses and sunglasses. For that reason, we manually split them into train and test sets to have images with glasses in both sets. For the rest of the celebrities, we randomly split them into train and test sets.

Celebrity	Training size	Test size
Barack Obama	47	35
Dwayne Johnson	46	35
Oprah Winfrey	43	35
Scarlett Johansson	45	35
Taylor Swift	41	35
Joe Biden	42	35
Michelle Obama	40	35

Table 3: Training and test sizes used in our experiments.

B. Qualitative Study

We created three surveys as follows. survey **A** had 19 participants and 35 questions. Each question gave participants the choice to pick an image that best represented one of the seven people we used in our study (table 3). Both given options were inpaintings of the same image, with the same mask, one using PATMAT-S and one using PATMAT-C. We asked 5 questions per identity, totalling 35 questions.

These questions relied on the participants knowing and remembering what each figure looked like. We provided 3 real images of each person, not appearing anywhere else in our study and survey questions, as a reference. 62% of the answers picked PATMAT-C as a better representation of the celebrities.

Survey **B** had 28 participants and 10 questions. Each question had 4 different images of a person. Each participant had to pick what they thought was an inpainted image from three real images and one inpainted image using PATMAT-C. Survey **C** followed the same structure as survey **B**, with 21 participants, except in each question, the participants were given two images of a person; one real and one inpainted. In Survey **B**, 43% of the answers were able to tell an image was inpainted. In survey **C**, 56% were able to correctly pick the inpainted image.

A **B**

Figure 9: Additional error analysis examples of PATMAT.

Figure 10: Tuning MAT without anchors for Oprah, inpaints images with glasses like features.

C. Additional Error Analysis

In figure 9, we show additional scenarios where PATMAT fails to properly inpaint an image. Figure 9-**A** shows examples where PATMAT attempts to generate / reconstruct sunglasses, yet fails to do so in a convincing fashion. Our experiments showed that this happens when the reference images do not contain the same type of glasses, hence PATMAT inpaints the missing part of the glasses with what is has seen in the references. Figure 9-**B** shows examples of PATMAT failing to properly align the inpainted area with the local and non-local priors.

D. Tuning Without Anchors

Anchors give structure to tuning MAT. We showed how using one anchor point to inpaint images with and without glasses will result in glasses-like features spreading to other images. Training with no anchors will have a similar effect, where in each iteration, random s_u noise-style codes are picked for tuning with reference image x_i. Without any structure to s_u codes, separating the features, figure 10 shows similar behaviour to what we saw in figure 2.

E. Additional Qualitative Results

Figure 11 shows additional inpainting examples using PATMAT-S, PATMAT-C and other competing methods.

Figure 11: Additional qualitative results. Please zoom in for details.

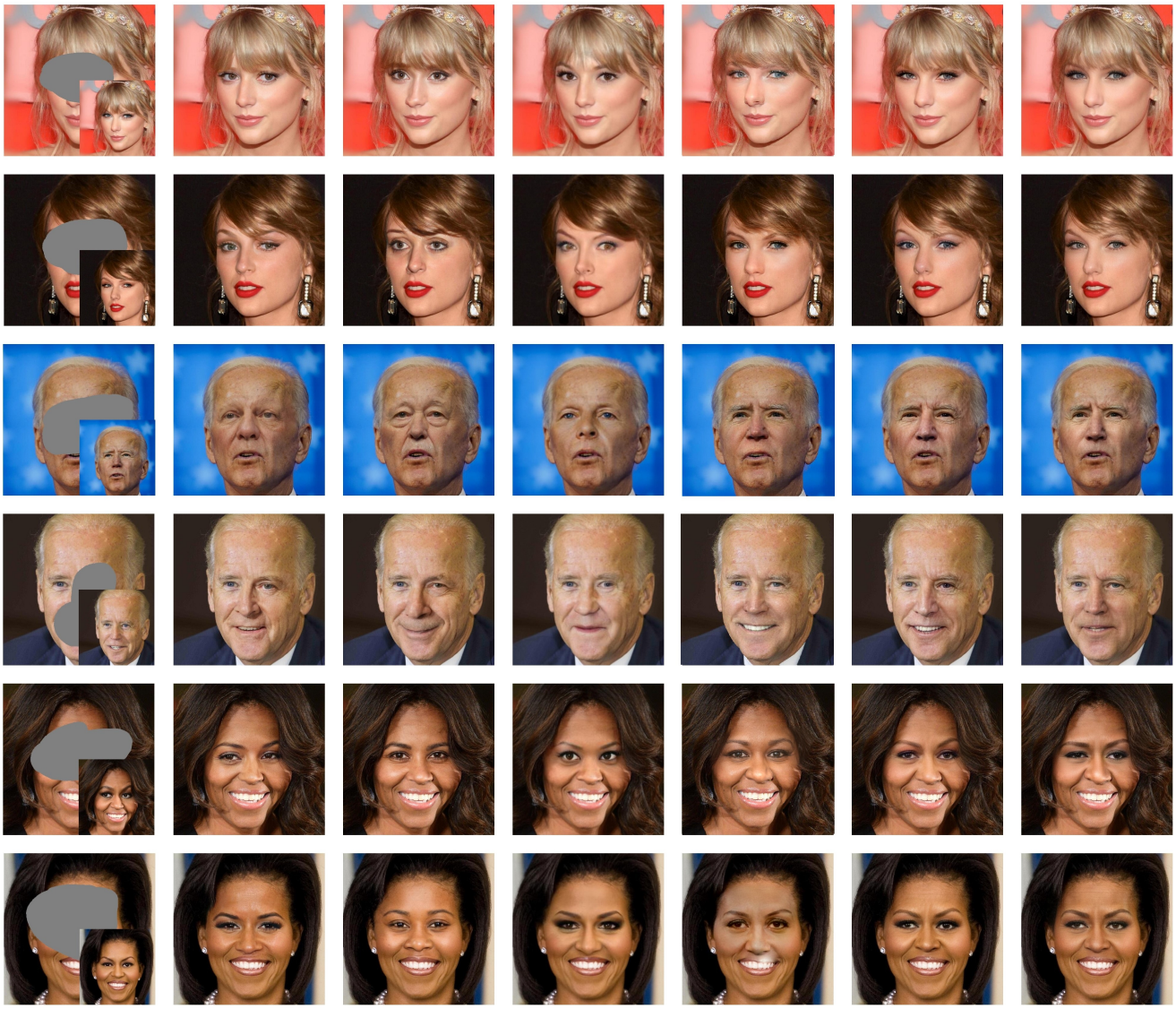

Figure 11: Additional qualitative results. Please zoom in for details.