UNIVERSITY OF CALIFORNIA SAN DIEGO

Controllable and Efficient Visual Generation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Zheng Ding

Committee in charge:

Professor Zhuowen Tu, Chair
Professor Ravi Ramamoorthi
Professor Hao Su
Professor Xiaolong Wang

2025

The Dissertation of Zheng Ding is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

DEDICATION

To my family.

EPIGRAPH

The only true wisdom is in knowing you know nothing.

Socrates

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My PhD journey starts in 2020 with lots of uncertainties. At that time, I was very concerned about how my PhD would go. It feels very different now when I take a look back. And I would like to express my deepest gratitude to all those who have made this happen.

First and foremost, I would like to express my gratitude for the guidance and support provided by my advisor, Professor Zhuowen Tu. I felt very lucky to have the supervision from him during the summer of 2019 when I was still an undergraduate and he introduced me to the research field of computer vision and deep learning. Later he became my PhD advisor and continued supporting me in my research and career. The insights and advice he provided on research and other fields are invaluable.

I am also indebted to the members of my committee including Ravi Ramamoorthi, Hao Su, and Xiaolong Wang for their expertise and thoughtful suggestions that have enhanced the quality of this thesis. Their research on different fields also inspired me a lot.

I would also like to extend my gratitude to the industry teams that provided internship opportunities during my PhD years. My time at Adobe, Google, and Nvidia broadened my horizons and enriched my practical understanding of industry applications. I would first like to thank Zhihao Xia and Cecilia Zhang who have been my mentors for over two years and not only provided me advice and suggestions on my research but also on lots of other fields. I would also like to thank Xiuming Zhang, Lars Jebe, Marc Levoy, Zhichao Yin, Rui Huang, Manika Puri, Shichen Liu, Brendan Shillingford for their support and helpful guidance. These experiences have been invaluable in connecting theoretical insights with real-world impact.

I would also like to thank my lab mates for their support during my PhD career, Weijian Xu, Kwonjoon Lee, Yifan Xu, Justin Lazarrow, Xiang Zhang, Zeyuan Chen. I would also like to thank some of friends who we collaborated or discuss a lot on research: Weirui Ye, Zhanghao Sun, Jackie Wang, Huaijin Wang, Haiwen Feng, Tianyi Xiong, Xin Xu.

Finally, on a personal note, I owe endless thanks to my family, friends and loved ones for their steadfast support and understanding through the long hours and inevitable setbacks

that come with pursuing a PhD. Your encouragement and belief in me have been my anchor, providing the strength and resilience to keep moving forward during challenging times.

Thank you all for being part of this extraordinary journey.

This dissertation is supported by IIS-2127544.

Chapter 2, in full, is a reprint of the material as it appears in "Open-Vocabulary Universal Image Segmentation with MaskCLIP". Ding, Zheng; Wang, Jieke; Tu, Zhuowen, International Conference on Machine Learning (ICML), 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "DiffusionRig: Learning Personalized Priors for Facial Appearance Editing". Ding, Zheng; Zhang, Cecilia; Xia, Zhihao; Jebe, Lars; Tu, Zhuowen; Zhang, Xiuming, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in "Patched Denoising Diffusion Models For High-Resolution Image Synthesis". Ding, Zheng; Zhang, Mengqi; Wu, Jiajun; Tu, Zhuowen, International Conference on Learning Representations (ICLR), 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in "Restoration by Generation with Constrained Priors". Ding, Zheng; Zhang, Cecilia; Tu, Zhuowen; Xia, Zhihao, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. The dissertation author was the primary investigator and author of this paper.

VITA

2020	Bachelor of Engineering, Tsinghua University
2024	Master of Science, University of California San Diego
2025	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Zheng Ding, Cecilia Zhang, Zhuowen Tu and Zhihao Xia. "Restoration by Generation with Constrained Priors", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zheng Ding*, Mengqi Zhang*, Jiajun Wu and Zhuowen Tu. "Patched Denoising Diffusion Models For High-Resolution Image Synthesis", *International Conference on Learning Representations (ICLR)*, 2024.

Zheng Ding, Cecilia Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu and Xiuming Zhang. "DiffusionRig: Learning Personalized Priors for Facial Appearance Editing", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zheng Ding, Jieke Wang and Zhuowen Tu. "Open-Vocabulary Universal Segmentation with MaskCLIP", *International Conference on Machine Learning (ICML)*, 2023.

ABSTRACT OF THE DISSERTATION

Controllable and Efficient Visual Generation

by

Zheng Ding

Doctor of Philosophy in Computer Science

University of California San Diego, 2025

Professor Zhuowen Tu, Chair

This dissertation presents several contributions aimed at enhancing visual generation by focusing on controllability and efficiency within computer vision systems. Before delving into the field of visual generation, we will first introduce MaskCLIP, which efficiently leverages pretrained vision-language models for open-vocabulary image segmentation tasks. Following that, we will discuss DiffusionRig, PatchDM, and Gen2Res to showcase the advancements we have made in controllable and efficient image generation. Collectively, the works presented in this dissertation strive to establish vision systems that are both controllable and efficient, facilitating improvements in visual generation as well as understanding.

Chapter 2 introduces a novel task, open-vocabulary universal image segmentation, which

aims for semantic, instance, and panoptic segmentation on arbitrarily described categories at inference. We first build a baseline using pre-trained CLIP models and then propose MaskCLIP—a Transformer-based approach featuring a MaskCLIP Visual Encoder that integrates mask tokens with a pre-trained ViT CLIP model. This design enables segmentation and class prediction while efficiently leveraging CLIP's dense features without resource-intensive student-teacher training.

Chapter 3 introduces DiffusionRig for personalize facial appearance editing. By employing a diffusion model conditioned on rough 3D face models derived from in-the-wild images, it maps simple CGI renderings to realistic images of an individual. DiffusionRig is trained in two stages: first learning generic facial priors from a large-scale dataset, then fine-tuning on limited person-specific photos. This strategy robustly edits facial features while preserving identity and high-frequency details.

Chapter 4 describes Patch-DM, a denoising diffusion model that generates high-resolution images (e.g., 1024×512) using small image patches (e.g., 64×64) during training. It alleviates boundary artifacts in patch-based synthesis by using a novel feature collage strategy that crops and combines overlapping features from neighboring patches to seamlessly predict shifted patches.

Chapter 5 proposes a method for adapting pretrained denoising diffusion models to image restoration tasks. The approach restores images by adding noise to degraded inputs and then denoising them using the pretrained model. By fine-tuning the model on selected anchor images that preserve the input's characteristics, the constrained generative space ensures high-quality restoration that maintains the original identity and overall quality.

Chapter 1

Introduction

In recent years, the fields of computer vision and generative modeling have witnessed transformative advancements, driven largely by the emergence of deep learning methodologies. From generating realistic human faces to creating intricate artworks, the capability of machines to produce high-quality visuals has expanded the horizon of both academic research and practical applications. However, as the complexity and scale of visual data continue to grow, so do the challenges associated with generating images that are not only aesthetically appealing but also controllable and computationally efficient.

The need for controllability in visual generation stems from the desire to steer the output of generative models in accordance with specific, user-defined attributes or semantic conditions. Whether it is adjusting the style of an artwork, tailoring the features of a synthetic face, or ensuring coherent adaptation of scenes in virtual environments, the ability to exert precise control over generated imagery is critical for bridging the gap between algorithmic creativity and human intent.

Equally important is the aspect of efficiency. In practical applications, resource constraints and the need for real-time performance require models that not only produce quality outputs but also maintain low computational costs. Efficient visual generation encompasses both the architectural design and the learning paradigms that minimize the burden on hardware while ensuring that models scale gracefully to larger, more diverse datasets. This balance between

efficiency and quality is paramount in contexts such as mobile applications, interactive media, and large-scale simulation environments, where processing power and latency are often limited.

Although the central focus of this thesis is on developing methods for controllable and efficient visual generation, a supporting theme of visual understanding is interwoven throughout the work. Visual understanding plays an important role in influencing how generated content can be more effectively aligned with semantic cues and context, thereby enhancing both the interpretability and practical utility of vision systems.

First, we addressed open-vocabulary universal segmentation, which unifies semantic segmentation for "stuff" (background regions) and instance segmentation for "things" (foreground objects). Traditional methods rely on a fixed set of categories with discrete labels, but recent developments in computer vision are pushing toward open-world and zero-shot settings where models recognize categories beyond what they were trained on. Taking advantage of pre-trained CLIP models that embed both images and text in the same feature space, we first establish an open-vocabulary panoptic segmentation baseline without additional training. We then propose MaskCLIP—a Transformer-based algorithm that efficiently leverages partially dense CLIP features through minimal re-training.

Second, we tackled the problem of the controllability of the facial image generation. More specifically, we studied the challenging problem of photorealistically editing portrait photos—adjusting lighting, expression, head pose, etc.—while preserving a person's identity and fine facial details. Traditional approaches use zero-shot learning on large datasets to generalize across identities, but this often loses the high-frequency facial nuances specific to an individual. To overcome this, we propose a two-stage method: we first learn generic facial priors from a large-scale face dataset. Then, using about 20 images of the target person (e.g., from personal photo albums), these generic priors are fine-tuned to capture the individual's unique high-frequency details. By first utilizing general facial priors and then fine-tuning with a small set of personalized images, we demonstrate that the model can achieve impressive photorealistic face editing with 3D understanding, effectively preserving identity and detailed facial characteristics.

Next, we studied the problem of efficiency in diffusion models. While diffusion models have recently garnered attention for their high-quality outputs despite being computationally intensive due to pixel-space optimization and multi-timestep training. To overcome these limitations in high-resolution image generation, current methods either depend on super-resolution techniques or latent space optimization, both of which require large models and substantial memory. In contrast, we propose Patch-DM to introduce a patch-based denoising diffusion model that generates full-size, high-resolution images directly without boundary artifacts by employing a novel feature collage strategy. This strategy uses a sliding-window, shifted patch generation process to ensure seamless feature sharing and consistency across neighboring patches without adding extra model complexity, thereby offering an efficient and promising direction for high-resolution generative diffusion modeling with lightweight architectures.

Last, we studied a problem on personalized facial image restoration where we leveraged powerful facial generation models to restore high-quality images from degraded ones. Standard discriminative methods learn an inverse mapping from paired data but are confined to the degradations seen during training. In contrast, model-based approaches learn image priors assuming a known degradation process at inference, often hindering real-world applicability. Our method relies solely on a trained denoising diffusion model, bypassing any assumption on the degradation process. We first project a low-quality image into the diffusion process by adding Gaussian noise to mimic clean-image distributions, then constrain generation to preserve key features (e.g., identity) via fine-tuning with anchor images. When explicit anchors are absent, we generate a "generative album" from soft-guided diffusion results that resemble the input, ensuring the restored image retains its essential characteristics.

Chapter 2

Open-Vocabulary Universal Image Segmentation with MaskCLIP

2.1 Introduction

Panoptic segmentation [KHG+19] or image parsing [TCYZ05] integrates the task of semantic segmentation [Tu08] for background regions (e.g. "stuff" like "road", "sky") and instance segmentation [HGDG17] for foreground objects (e.g. "things" such as "person", "table"). Existing panoptic segmentation methods [KHG+19, KGHD19, LCZ+19, XLZ+19, LLST20] and instance segmentation approach [HGDG17] deal with a fixed set of category definitions, which are essentially represented by categorical labels without semantic relations. DETR [CMS+20] is a pioneering work that builds a Transformer-based architecture for both object detection and panoptic segmentation. Under a more general setting, the tasks of semantic [Tu08], instance [HGDG17], and panoptic [KHG+19] can be unified under a universal image segmentation paradigm [CMS+22].

The deep learning field is moving rapidly towards the open-world/zero-shot settings [BB15] where computer vision tasks such as classification [RKH+21a], object detection [LZZ+22, ZRHC21, ZLZ+22, GLKC22, CKR+22], semantic labeling [LWB+22, GGCL22], and image retrieval [BB15, HS18, ZRHC21, HS18, KSL+21] perform recognition and detection for categories beyond those in the training set.

In this paper, we take advantage of the existence of pre-trained CLIP image and text

embedding models [RKH+21a], that are mapped to the same space. We first build a baseline method for open-vocabulary panoptic segmentation using CLIP models without training. We then develop a new algorithm, MaskCLIP, that is a Transformer-based approach efficiently and effectively utilizing pre-trained partial/dense CLIP features without heavy re-training. The key component of MaskCLIP is a Relative Mask Attention (RMA) module that seamlessly integrates the mask tokens with a pre-trained ViT-based CLIP backbone. MaskCLIP is distinct and advantageous compared with previous approaches in three aspects: 1) A canonical background and instance segmentation representation by the mask token representation with a unique encoder-only strategy that tightly couples a pre-trained CLIP image feature encoder with the mask token encoder. 2) MaskCLIP avoids the challenging student-teacher distillation processes such as OVR-CNN [ZRHC21] and ViLD [GLKC22] that face a limited number of teacher objects to train; 3) MaskCLIP also learns to refine masks beyond simple pooling in e.g. OpenSeg [GGCL22].

The contributions of our work are listed as follows.

- We develop a new algorithm, MaskCLIP, to perform open-vocabulary universal image segmentation building on top of canonical background and instance mask representation with a cascade mask proposal and refinement process.
- We device the MaskCLIP Visual Encoder under an encoder-only strategy by tightly coupling a pre-trained CLIP image feature encoder with the mask token encoder, to allow for the direct formulation of the mask feature representation for semantic/instance segmentation+refinement, and class prediction. Within the MaskCLIP Visual Encoder, there is a new module called Relative Mask Attention (RMA) that performs mask refinement.
- MaskCLIP expands the scope of the CLIP models to open-vocabulary universal image segmentation by demonstrating encouraging and competitive results for open-vocabulary semantic, instance, and panoptic segmentation.

Table 2.1. Comparison for recent open-vocabulary approaches for object detection, semantic segmentation, instance segmentation, and panoptic segmentation. GLIP [LZZ+22]; OVR-CNN [ZRHC21]; ViLD [GLKC22]; RegionCLIP [ZYZ+22]; OV-DETR [ZLZ+22]; LSeg [LWB+22]; OPenSeg [GGCL22]; DenseCLIP [RZC+22]; XPM [HKL+22]. ✗ indicates that the corresponding method is loosely following the definition. Dense Clip features refer to the use of pixel-wise/local features. Note that OpenSeg uses its ALIGN [JYX+21], which is an alternative to CLIP.

Task	Method	Arbitrary Online	Segmentation		Dense CLIP	Training	Annotation
		Inference	semantic	instance	features	data	type
Object Det.	GLIP	✔				FourODs, GoldG, Cap24M	labels + bbox + captions
	OVR-CNN	✔				COCO base, CC3M	bbox + captions
	ViLD	✔				COCO	labels + bbox
	RegionCLIP	✔				CC3M, COCO	captions
Semantic Seg.	LSeg	✗	✔			COCO + others	labels + segmentations
	OpenSeg	✔	✔	✗	✔	COCO, LocalizedNarratives	masks + captions
	DenseCLIP		✔		✔	COCO	labels + segmentations
Instance Seg.	XPM	✗		✔		COCO, CC3M	labels + masks + captions
Panoptic Seg.	MaskCLIP (ours)	✔	✔	✔	✔	COCO	labels + masks

2.2 Related Work

Open vocabulary. The open vocabulary setting is gaining increasing popularity lately as traditional fully supervised settings cannot handle unseen classes during testing, while real-world vision applications like scene understanding, self-driving and robotics are commonly required to predict unseen classes. Previous open-vocabulary attempts have been primarily made for object detection. ViLD [GLKC22] trains a student model to distill the knowledge of CLIP. RegionCLIP [ZYZ+22] finetunes the pretrained CLIP model to match the image areas with corresponding texts. OV-DETR [ZLZ+22] uses CLIP as an external model to obtain the query embedding from CLIP model. Recently there is also work made for open-vocabulary semantic segmentation [GGCL22].

Universal segmentation. Previously semantic/instance/panoptic segmentation tasks have been treated as different tasks using different methods. With the recent trends in computer vision, the formulation and methods of the three segmentation tasks have gradually been uniformed [CSK21, CMS+22]. Instead of separately dealing with the stuff/instance, those methods treat them as the same one and output masks for each stuff/instance and do a post-process on the

output masks for different segmentation tasks.

Open-vocabulary universal segmentation: an emerging task. As open-set, open-world, zero-shot, and open-vocabulary are relatively new concepts that have no commonly accepted definitions, thus, different algorithms are often not directly comparable with differences in problem definition/setting, training data, and testing scope. Table 2.1 gives a summary for the recent open-vocabulary applications. XPM [HKL+22] utilizes vision-language cross-modal data to generate pseudo-mask supervision to train a student model for instance segmentation, and thus, it may not be fully open-vocabulary to allow for arbitrary object specifications in the inference time. LSeg [LWB+22] also has a limited open-vocabulary aspect as the learned CNN image features in LSeg are not exposed to representations beyond the training labeling categories. OpenSeg [GGCL22] is potentially applicable for instance/panoptic segmentation, but OpenSeg is formulated to be trained on captions that lack instance-level information that is fundamental for panoptic segmentation. The direct image feature pooling strategy in OpenSeg is potentially another limiting factor towards the open-vocabulary universal segmentation. Nevertheless, no results for open-vocabulary panoptic/instance segmentation are reported in [GGCL22].

Class-agnostic segmentation. Most closed-vocabulary segmentation methods are class-ware i.e. predicting the classes along with the corresponding labels [HGDG17, CSK21, CMS+22]. However, in tasks involving open-vocabulary or open-world scenarios where novel classes may appear during testing, it is common to use class-agnostic segmentation methods for generating masks[JYX+21, QKW+22, XZW+22]. The difference in methodology between class-aware and class-agnostic segmentation methods is typically not substantial. Class-aware methods often incorporate a class-prediction head, whereas class-agnostic methods do not. In our method, we adopt a class-agnostic segmentation model by removing the class-prediction head from previous class-aware class-aware segmentation methods.

CLIP model distillation/reuse. After its initial release, the CLIP model [RKH+21a] that is learned from large-scale image-text paired captioning datasets has received a tremendous amount of attention. Some other similar vision-language models have also been proposed

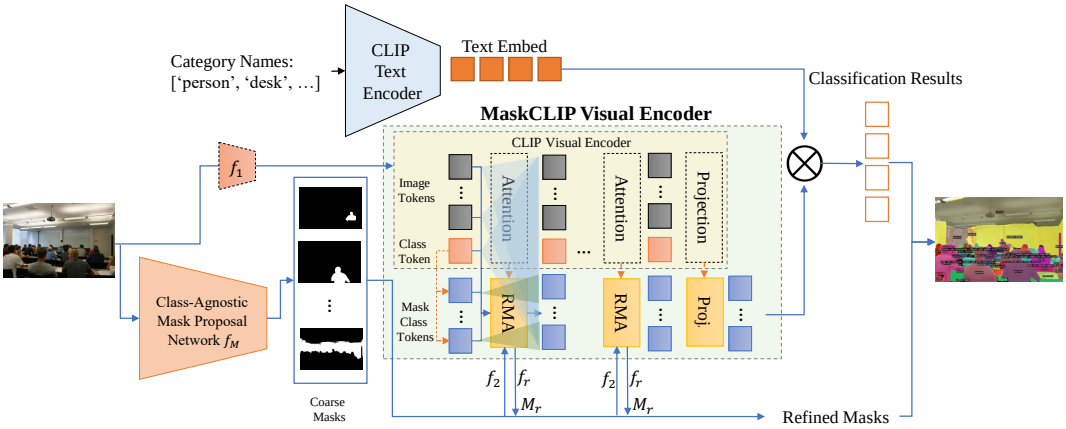


Figure 2.1. Illustration of the pipeline. Our pipeline contains two stages. The first stage is a class-agnostic mask proposal network and the second stage is built on the pretrained CLIP ViT model. All the weights of the CLIP ViT model during training are fixed. Arrows in orange denote weight sharing. The embeddings' weights of Mask Class Tokens are shared by Class Tokens in the CLIP ViT model and are fixed. RMA represents Relative Mask Attention which is built based on the CLIP ViT attention layer. RMA contains all the weights from CLIP ViT attention layer which are all fixed during training. Additional weights are added in RMA for further mask information utilization and mask refinement. The demo image we use here is from ADE20K [ZZP+19].

later e.g. ALIGN [JYX+21], GLIP [LZZ+22]. Many algorithms have been developed lately [ZLZ+22, WCC+22, ZYZ+22, LJZ+21, PWS+21, SLT+22] trying knowledge distillation from the CLIP model to benefit the down-stream tasks one way or the other by leveraging the rich semantic language information paired in the images. Here, we directly adopt the backbone of CLIP image model to train for open-vocabulary panoptic segmentation. There have been attempts [RZC+22, ZLD22] that use the partial/dense CLIP features to represent pixel-wise features as teacher model to train student model for semantic segmentation.

2.3 Method

Our pipeline, shown in Figure 2.1, contains two stages. The first stage is a class-agnostic mask proposal network. The second stage is MaskCLIP Visual Encoder which is built on the CLIP [RKH+21a] ViT architecture. It takes the images and the coarse masks from the first stage as the input and outputs refined masks along with the corresponding partial/dense image features

for further classification.

2.3.1 Class-Agnostic Mask Proposal Network

Our Class-Agnostic Mask Proposal Network is built on instance/segmentation models such as MaskRCNN[HGDG17] and Mask2Former[CMS+22]. To make the model class-agnostic, we remove the class supervision during training. The classification head thus becomes a binary classification for either positive or negative in these models.

2.3.2 MaskCLIP Visual Encoder

Similar to CLIP, our MaskCLIP Visual Encoder also predicts the image features. Unlike the CLIP Visual Encoder which only uses one class token to output the feature of the whole image. Our MaskCLIP Visual Encoder uses another M Mask Class Tokens to output the partial/dense features for each corresponding area of the image given the masks. The Mask Class Tokens use attention masks and Relative Mask Attention to obtain the partial/dense features which we discuss in the following two parts.

Mask Class Tokens.

In order to obtain partial/dense image features for the corresponding masks or bounding boxes for further recognition or distillation, an easy way to do this is simply masking or cropping the image and then sending the obtained image to the pretrained image encoder. This method has been widely used in several open vocabulary detection/segmentation methods [ZYZ+22, GLKC22, XZW+22]. The problem is that it's not computation efficient (N masks/boxes will lead to N images and they will be computed through the image encoder independently) and also loses the ability to see the global image context information which is very important for recognizing some objects and stuff. For masking, another problem is that masks are in different shapes and simply masking the image will cause the resulting image to have a transparent background which usually doesn't exist in real images that are used for training in large language-vision models e.g., CLIP.

To solve this, we propose Mask Class Tokens for efficient feature extraction from images without losing the global image context information. In the original CLIP ViT-based visual encoder framework, the input of the network is N image tokens and 1 class token. The final output of the class token will be used for the relation computation with the text embeddings. Our newly introduced M Mask Class Tokens will be alongside the image tokens and the class token. The embeddings' weights of the Mask Class Token are provided by the class token in the pretrained CLIP ViT model and are fixed. Each Mask Class Token will output a corresponding partial/dense image feature similar to the class token which outputs the feature of the whole image. To achieve this, we design an attention mask as follows

$$\mathscr{M} = \begin{bmatrix} \mathscr{F}_{(N+1)\times(N+1)} & \mathscr{T}_{(N+1)\times M} \\ \mathscr{M}'_{M\times N} \quad \mathscr{F}_{M\times 1} & \mathscr{T}_{M\times M} \end{bmatrix} \tag{2.1}$$

in which M is the number of Mask Class Tokens, N is the number of image tokens, $\mathscr{T}_{m\times n}$ is an $m \times n$ True matrix, $\mathscr{F}_{m\times n}$ is an $m \times n$ False matrix and $\mathscr{M}'$ is defined as following:

$$\mathscr{M}'_{i,j} = \begin{cases} \text{False} & \text{if mask}_i \text{ contains at least one pixel in patch}_j \\ \text{True} & \text{otherwise.} \end{cases} \tag{2.2}$$

where True means that this position is masked out i.e. not allowed to attend and False otherwise.

In our mask attention matrix $\mathscr{M}$, $\mathscr{F}_{(N+1)\times(N+1)}$ shows the N Image Tokens and one Class Token are attending each other as in the original CLIP. $\mathscr{T}_{(N+1)\times M}$ shows that the N Image Tokens and one Class Token are not attending the M Mask Class Tokens. $\mathscr{M}'_{M\times N}$ shows that the Mask Class Tokens are attending the Image Tokens given the corresponding masks. $\mathscr{F}_{M\times 1}$ shows that the M Mask Class Tokens are attending the Class Token. $\mathscr{T}_{M\times M}$ shows that the M Mask Class Tokens are not interacting with each other.

In this way, each Mask Class Token will learn from the corresponding mask area of the images. The image tokens are also interacting with each other which means the global

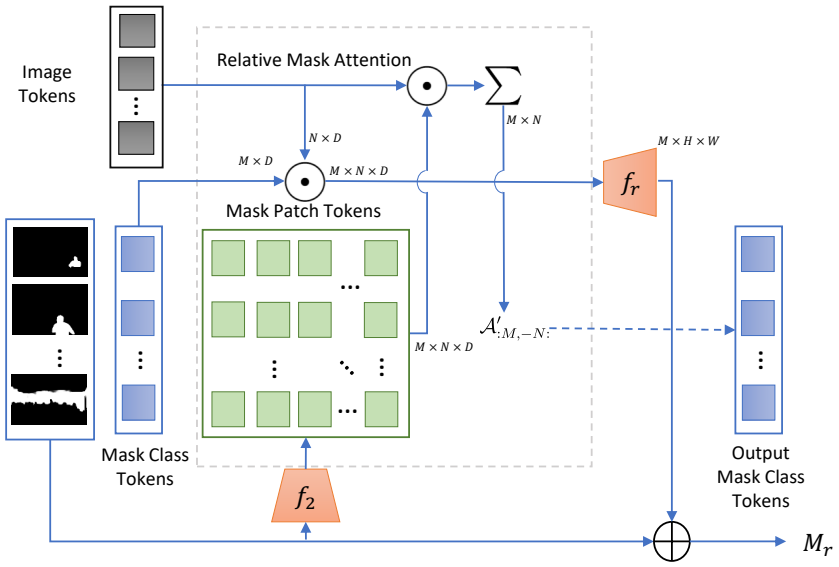


Figure 2.2. Relative Mask Attention. Our Relative Mask Attention mechanism adds another attention matrix $A'_{:M,-N:}$ to the original attention matrix. The newly added attention matrix is computed using the Image Tokens and the Mask Patch Tokens. The mask patch tokens are acquired by patchifying the masks using a similar way for the images as shown here. Moreover, the masks are refined by using M_r in Eq. 2.5 which is computed by Image Tokens and Mask Class Tokens.

information won't lose. And it's also very efficient since we don't need to do redundant computing for each mask or finetune the pretrained model. However, the mask information is not fully utilized and it cannot be refined either. But we will see in the experiments later that simply adopting Mask Class Tokens to the pretrained CLIP model without any finetuning will already serve as a competitive baseline.

Relative Mask Attention.

To further utilize the mask information and refine the coarse masks, we propose Relative Mask Attention mechanism in our transformer. Our key design principle is to try not to change the CLIP features directly as this would destroy the learned relationship between the image features and text features in the CLIP model. Therefore, we adopt a way to only change the attention matrix in the transformer to learn a better linear combination of the values in the attention layers according to the mask information. As in Figure 2.2, our proposed Relative Mask

Attention Mechanism only changes the attention matrix and refines the masks. M_r is defined in Eq. 2.5. $\mathscr{A}'_{:M,-N:}$ is defined in Eq. 2.3. f_M is the class-agnostic mask proposal network. f_1 and f_2 are two downsampling networks that encode the images/masks to image tokens/mask patch tokens sharing the same architecture. f_r is a two-layer convolutional network that maps the attention matrix to a mask residual.

Similar to relative positional encoding, we use a relative attention mechanism here. Let D be the dimension of the token embedding, for each Mask Class Token $T_i^{\mathrm{MC}} \in \mathbb{R}^D$ with a corresponding mask $K_i \in \mathbb{R}^{H\times W}$ whose shape is the same as the image, we use a similar way as for the images to get mask patch tokens $T^{\mathrm{MP}} \in \mathbb{R}^{M\times N\times D}$ in the computation of the attention. In our attention matrix, the Mask Class Tokens attending the image tokens part will then be as follows:

$$\mathscr{A}'_{:M,-N:} = \sum_c^D (\phi_{Q_m}(T^{\mathrm{MP}}) \odot \phi_{K_m}(T^{\mathrm{IM}}))_c \tag{2.3}$$

$$\mathscr{A}_{:M,-N:} = \frac{\phi_Q(T^{\mathrm{MC}}) \cdot \phi_K(T^{\mathrm{IM}}) + \mathscr{A}'_{:M,-N:}}{2\sqrt{D}} \tag{2.4}$$

where $T^{\mathrm{IM}} \in \mathbb{R}^{N\times D}$ is image tokens, $T^{\mathrm{MC}} \in \mathbb{R}^{M\times D}$ is Mask Class Tokens, $T^{\mathrm{MP}} \in \mathbb{R}^{M\times N\times D}$ is Mask Patch Tokens $\phi_Q, \phi_K, \phi_{Q_m}, \phi_{K_m}$ are linear transformations, $\odot$ is element-wise product and $\sum_c^D(\cdot)_c$ is the sum of the embedding dimension. $\phi_{K_m}(T^{\mathrm{IM}}) \in \mathbb{R}^{N\times D}$ will first be broadcast to $\mathbb{R}^{M\times N\times D}$ before doing element-wise production.

The attention will also in turn be used for the refinement of the masks. The vanilla attention can be seen as a relationship between each mask area and all the image patches. Thus we utilize this to help our coarse masks be more accurate. The updating process of the masks is as following:

$$M_r = \sigma(\sigma^{-1}(M_c) + f_r(\phi_Q(T^{\mathrm{MC}}) \odot \phi_K(T^{\mathrm{IM}}))) \tag{2.5}$$

where $M_c, M_r \in \mathbb{R}^{N\times H\times W}$ denotes the coarse mask and refined mask respectively, f_r is a learnable

non-linear function that maps the attention matrix to a mask residual, σ and σ^{-1} are sigmoid and inverse sigmoid functions respectively.

The RMA method aims to leverage detailed mask information and refine masks by utilizing CLIP's features. Without RMA, the method would only utilize the mask information in the attention mask (which is just a low-resolution mask) and cannot refine the mask using CLIP's features. In order to utilize the detailed information of masks, we add another attention matrix, which is obtained from the Mask Patch Tokens and the Image Tokens, to the original attention matrix in the CLIP ViT model so that the new attention matrix could be aware of the detailed mask information and thus the Mask Class Tokens could attend the information more accurately. Furthermore, we use the information from the original attention matrix, which is obtained from the Mask Class Tokens and the Image Tokens, to refine the mask.

2.4 Experiments

In this part, we train our proposed MaskCLIP method using COCO [LMB+14] training data and test on other datasets (ADE20K [ZZP+19, ZZP+17], PASCAL Context [MCL+14], LVIS) under the open vocabulary setting. We report our results on semantic/instance/panoptic segmentation tasks to evaluate the performance of out model's universal segmentation.

2.4.1 Datasets

COCO: COCO [LMB+14] includes 133 classes where 80 classes are things and 53 classes are stuff or background. There are 118k training images and 5k validation images. In our experiments, we first train the class-agnostic mask proposal network on COCO training dataset using the annotations of panoptic masks. Then we train our models on COCO training images in a supervised manner.

ADE20K: ADE20K [ZZP+19, ZZP+17] contains 20,210 images and annotations for training and 2000 images and annotations for validation. It serves both panoptic segmentation and semantic segmentation. The full version (A-847) [ZZP+19] includes 847 classes and the

Table 2.2. Results on open-vocabulary semantic segmentation. A-150 and A-847 represent the ADE20K dataset with 150 classes and 847 classes respectively. P-459 and P-59 represent PASCAL Context dataset with 459 classes and 59 classes respectively. All results use the mIoU metric. All methods presented here don't use extra data other than COCO for training.

Method	COCO Training Data	A-150 ↑	A-847 ↑	P-459 ↑	P-59 ↑
ALIGN [JYX+21]	None	10.7	4.1	3.7	15.7
ALIGN w/ proposals [JYX+21]	Masks	12.9	5.8	4.8	22.4
LSeg+ [LWB+22]	Masks + Labels	18.0	3.8	7.8	46.5
OpenSeg [GGCL22]	Masks + Captions	21.1	6.3	9.0	42.1
SimSeg [XZW+22]	Masks + Labels	20.5	7.0	-	**47.7**
CLIP Baseline	Masks	13.8	5.2	5.2	25.3
MaskCLIP w/o RMA	Masks	14.9	5.6	5.3	26.1
MaskCLIP (MaskRCNN)	Masks + Labels	22.4	6.8	9.1	41.3
MaskCLIP	Masks + Labels	**23.7**	**8.2**	**10.0**	45.9

short version (A-150) [ZZP+17] includes 150 classes. We use the validation set in ADE20K for testing without any training on this dataset in which case we can test our model's capability of open vocabulary segmentation.

PASCAL Context: PASCAL Context [MCL+14] contains 10,103 per-pixel annotations for images of PASCAL VOC 2010 [EVGW+], where 4998 for training and 5105 for validation. The full version (P-459) includes 459 classes and the short version includes 59 classes. This dataset serves as another benchmark testing our model's open vocabulary segmentation ability.

LVIS: LVIS [GDG19] contains 100,170 images for training and 19,809 images for validation. It extends COCO [LMB+14] but contains 1,203 categories. It is considered one of the most challenging benchmark for instance segmentation because of its large vocabulary, long-tailed distribution, and fine-grained classification. We report our model's performance of open vocabulary instance segmentation on the validation dataset.

2.4.2 Implementation Details

Class-Agnostic Mask Proposal Network. In our first stage, we train a class-agnostic mask proposal network using MaskRCNN [HGDG17] and Mask2Former [CMS+22] on COCO training data. The experiment setting we use for MaskRCNN is R50-FPN-1x. The backbone we use in Mask2Former is ResNet-50. All the training setting follows the default in their models.

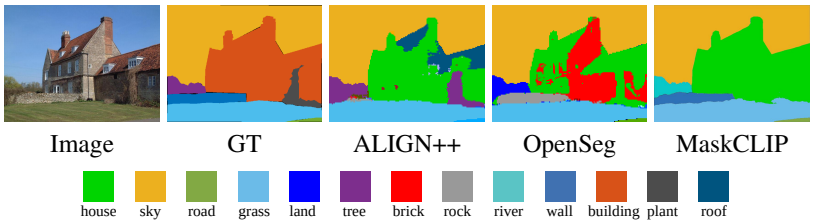


Figure 2.3. Comparison on open-vocabulary semantic segmentation. The input image and the results for GT, ALIGN++, OpenSeg are from [GGCL22].

CLIP Baseline. We design our first baseline by directly using the class-agnostic mask proposal network from the first stage and the pretrained CLIP model. We mask the images according to the masks from the class-agnostic mask proposal network and send the masked images to the CLIP model to get classification results. The pretrained CLIP model we use is ViT-L/14@336px and the text inputs we use are simply the category names defined by each dataset. Those two settings keep the same with the following two methods for a fair comparison.

MaskCLIP w/o RMA Baseline. Our second baseline is based on the Mask Class Tokens which doesn't use the Relative Mask Attention mechanism. Instead of masking the images and sending the resulting images directly to the CLIP model for feature extraction, we use Mask Class Tokens to acquire the corresponding partial/dense image features. The obtained image features will then be used for further open vocabulary classification.

The two baselines above don't need any training in the second stage and can be used to directly perform the open vocabulary tasks. We will demonstrate that the second baseline is better at feature extraction in both quantitative results and qualitative results under the open vocabulary setting and show the effectiveness and efficiency of the proposed Mask Class Tokens.

MaskCLIP. In our MaskCLIP method, we still use the CLIP ViT-L/14@336px pretrained model as with the previous two. This model has 24 attention layers and we add Relative Mask Attention in four of them which is 6, 12, 18, 24. We use AdamW [LH19] as our optimizer and the learning rate is set to 0.0001. We train our model on COCO training data for 10k iterations with a batch size of 8. The training takes around 3h on 8 Nvidia A5000 GPUs.

Loss Function. The loss function is $\mathscr{L} = \lambda_{\text{ce}}\mathscr{L}_{\text{ce}} + \lambda_{\text{dice}}\mathscr{L}_{\text{dice}} + \lambda_{\text{bce}}\mathscr{L}_{\text{bce}}$, where $\mathscr{L}_{\text{ce}}$ is the loss for classification, $\mathscr{L}_{\text{dice}}$ and $\mathscr{L}_{\text{bce}}$ are the losses for mask localization. In our experiments, We set $\lambda_{\text{ce}} = 2, \lambda_{\text{dice}} = 5, \lambda_{\text{bce}} = 5$.

In the next three parts, we evaluate our methods on open vocabulary semantic, panoptic segmentation, and instance segmentation tasks. The class-agnostic mask proposal networks we use in those methods are trained using Mask2Former other than noted.

2.4.3 Open-Vocabulary Semantic Segmentation

First, we use our method to compare with open-vocabulary semantic segmentation as in Table 2.2. We train our method on the COCO dataset and evaluate on another four different datasets. On the four datasets we test, MaskCLIP outperforms the two baselines we described in the implementation details which demonstrates that our feature extraction method is better than the vanilla way in this setting. It extracts the features without the need to change the input and can simultaneously extract multiple mask area features easily. For 100 masks' feature extraction in a single image, the CLIP baseline takes about 3s on a single 3090 GPU while the MaskCLIP w/o RMA baseline only takes ˜0.6s which is about 4x faster. Our MaskCLIP beats both baselines significantly as it utilizes accurate mask information and refines the masks during the feature extraction process. Furthermore, our proposed method also reaches state-of-the-art results on three of the benchmarks with only P-59 slightly lower than LSeg+[LWB+22].

To compare with previous methods, we also provide a semantic segmentation comparison in Figure 2.3. Results on ALIGN++ and OpenSeg are directly from [GGCL22] and we run the same image using our MaskCLIP model. It can be seen that due to the open vocabulary setting, some similar classes may be mistakenly classified e.g. all three methods predict the house in this image while the ground truth is building.

Table 2.3. Results on open-vocabulary panoptic segmentation using the ADE20k validation dataset. th and st represent thing and stuff classes respectively.

Method	PQ ↑	PQ^{th} ↑	PQ^{st} ↑	SQ ↑	SQ^{th} ↑	SQ^{st} ↑	RQ ↑	RQ^{th} ↑	RQ^{st} ↑
CLIP Baseline	8.207	8.473	7.675	53.124	52.661	54.048	10.534	10.883	9.835
MaskCLIP w/o RMA	9.565	8.922	10.852	62.507	62.268	62.985	12.645	11.758	14.418
MaskCLIP (MaskRCNN)	12.860	11.242	16.095	64.008	64.183	63.658	16.803	14.968	20.473
MaskCLIP	**15.121**	**13.536**	**18.290**	**70.479**	**70.021**	**71.396**	**19.211**	**17.448**	**22.737**

2.4.4 Open-Vocabulary Panoptic Segmentation

Next, we compare our MaskCLIP with the two baselines on ADE20K validation set under the open vocabulary panoptic segmentation setting. The results are presented in Table 2.3. As can be seen from the table, the MaskCLIP w/o RMA baseline performs better on all the metrics in panoptic segmentation setting which further demonstrates our method's effectiveness.

We also show two sets of images to demonstrate our model capability. The first is the qualitative results on ADE20K. We compare our method with the two baselines in Figure 2.4. It can be seen that our method performs much better than the two baselines. The results from the first column show that due to the lack of global information, CLIP baseline fails to predict "floor". Instead, it predicts "skyscraper". While the MaskCLIP w/o RMA baseline and MaskCLIP model can predict the floor correctly as it does not lose the global context information.

The second set of images we're presenting is in Figure 2.5. These figures show our capability of specifying any arbitrary classes in performing panoptic segmentation task. The results show that though we train a new model based on the CLIP model without any distillation methods, we can still preserve the CLIP image features very well. Our model doesn't have a clear bias towards the base classes in the training set and could tell the difference very well that have no chance to learn in the COCO training: e.g toy vs real and filled vs empty.

2.4.5 Open-Vocabulary Instance Segmentation

Cross-Dataset Setting. We present the results on open vocabulary instance segmentation in Table 2.4 under the cross-dataset setting. Since instance segmentation can be regarded as

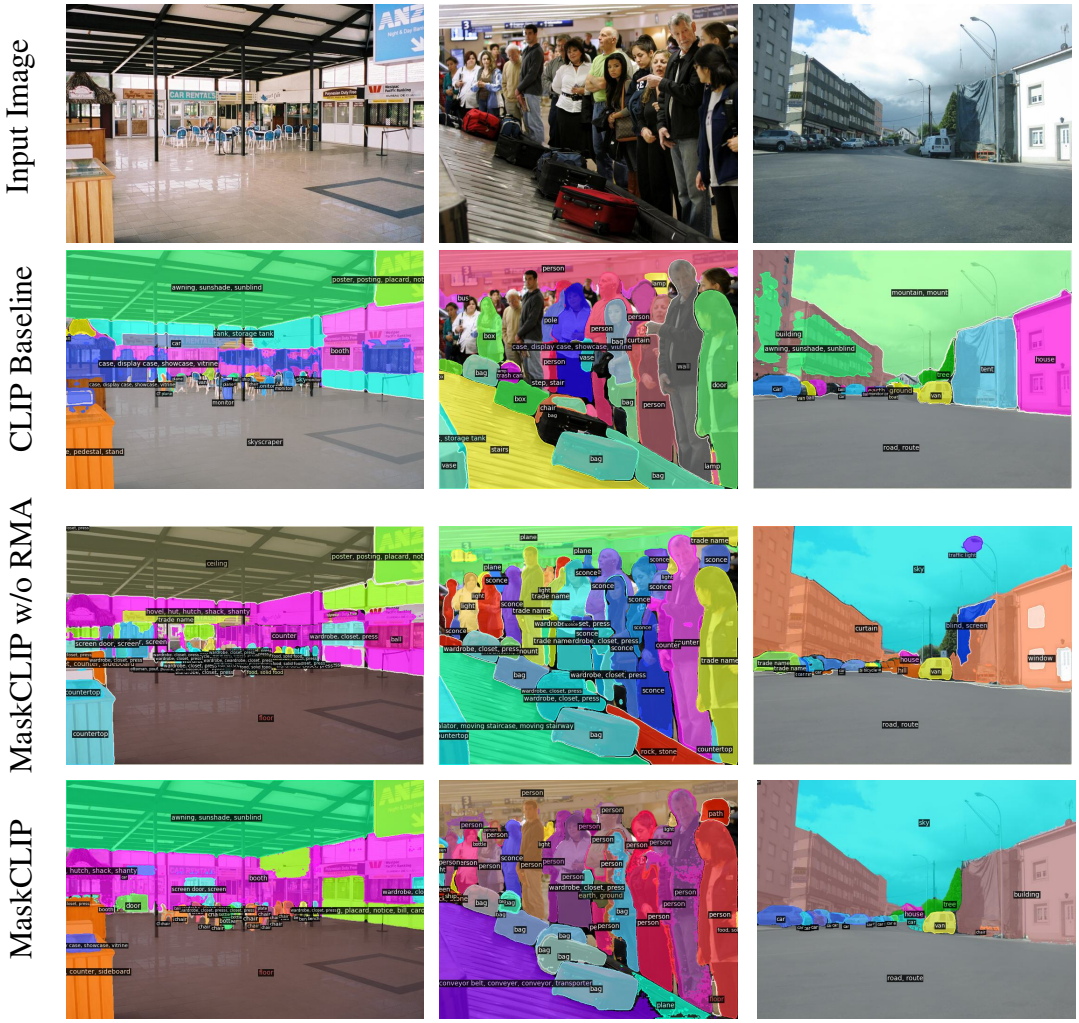


Figure 2.4. Qualitative results on ADE20K panoptic segmentation. The images are taken from the ADE20K validation set. We use the class names directly from the ADE20K 150 classes as the text inpputs. Three images are presented here using our MaskCLIP model along with the two baselines.

(a) "**toy rabbit**", "**real rabbit**", "**background**"

(b) "horse", "**donkey**", "sky", "grass"

(c) "**empty bottle**", "**filled bottle**", "door", "wall", "ground"

Figure 2.5. User-specified class panoptic segmentation. The labels above are the text inputs we used for testing the images. Texts in bold are novel classes i.e. don't exist in the labels of COCO training data. (a) Our model is able to distinguish object properties of real rabbit and toy rabbit. (b) This example shows that our model is potential for fine-grained classifications and does not have bias toward the base classes. (c) Our results show that it can tell the difference between the filled status and empty status of bottles.

Table 2.4. Results on open-vocabulary instance segmentation under the cross-dataset setting.

Method	ADE20K			LVIS		
	AP ↑	AP^{50} ↑	AP^{75} ↑	AP ↑	AP^{50} ↑	AP^{75} ↑
CLIP Baseline	3.974	6.090	4.288	4.989	7.244	5.227
MaskCLIP w/o RMA	4.263	6.696	4.402	5.762	8.202	6.169
MaskCLIP (MaskRCNN)	**6.164**	**12.072**	5.775	6.431	**12.753**	5.777
MaskCLIP	5.989	9.739	**6.209**	**8.404**	12.190	**8.810**

Table 2.5. Results on open-vocabulary instance segmentation under the COCO split setting.

Method	Constrained		Generalized		
	Base	Target	Base	Target	All
Soft-Teacher[XZH+21]	41.8	14.8	41.5	9.6	33.2
Unbiased-Teacher[LMH+21]	41.8	15.1	41.4	9.8	33.1
XPM[HKL+22]	42.4	**24.0**	41.5	21.6	36.3
MaskCLIP	**42.8**	23.2	**42.6**	**21.7**	**37.2**

"thing-only" panoptic segmentation, we directly apply our model trained on COCO panoptic dataset to the instance segmentation task. MaskCLIP with different class-agnostic mask proposal networks performs better than CLIP Baseline and MaskCLIP w/o RMA in general.

COCO Split Setting. Besides the cross-dataset setting, we also follow the COCO Split Setting in XPM[HKL+22] to perform the instance segmentation in Table 2.5. On the generalized setting which is a more challenging setting, we outperform previous results in base, target, and all categories. In the constrained setting, we also show competitive results in both base and target categories.

2.4.6 Efficiency Analysis

We further provide efficiency analysis in Table 2.6 to demonstrate the efficiency of our feature extraction method. Previous methods usually perform a crop/mask operation on the input images and send the resulted images to CLIP to obtain the partial/dense features for classification which is rather slow. In contrast, our proposed method employs Mask Class

Table 2.6. FLOPs Comparison. We use the CLIP ViT-L/14 model in all methods for fair comparison and 640x640 as the input resolution.

Method	TFLOPs
RegionCLIP[ZYZ+22]	9.5
ZegFormer[DXXD22]	10.3
SimSeg[XZW+22]	9.6
CLIP Baseline	10.5
MaskCLIP(Ours)	**0.3**

Table 2.7. Incorporating GT Masks. Results on using GT masks as mask proposals for open-vocabulary panoptic segmentation and semantic segmentation.

	PQ ↑	mIoU ↑
OpenSeg [GGCL22]	-	21.1
MaskCLIP	15.1	23.7
OpenSeg + GT masks [GGCL22]	-	27.5
MaskCLIP + GT masks	35.8	31.7

Tokens to obtain the partial/dense features for classification. By doing so, our method can extract partial/dense features more efficiently (instead of running CLIP N times, our method only requires running CLIP one time with N more Mask Class Tokens) and is also aware of the global context information.

2.5 Ablation Study

2.5.1 Incorporating GT Masks.

Since our model can decouple the mask proposal process and the classification process, we could also use the ground truth mask proposals which can be regarded as a "perfect" mask proposal network in our method. In this way, we can eliminate the effects of the quality of the mask proposals and inspect the method's classification capabilities. In Table 2.7. We can see that the performance could gain a lot from the "perfect" mask proposals. And our MaskCLIP method also outperforms OpenSeg in this setting.

Table 2.8. Ablation Study on Mask Refinement. Results on ADE20K validation set are reported here. Both methods are trained on COCO and tested on ADE20K validation dataset. MR refers to mask refinement.

	PQ ↑	PQ^{Th} ↑	PQ^{St} ↑	**SQ↑**	$\mathbf{SQ^{Th}}$ ↑	$\mathbf{SQ^{St}}$ ↑
MaskCLIP w/o MR	13.624	13.253	14.368	66.361	67.715	63.653
MaskCLIP	15.121	13.536	18.290	70.479	70.021	71.396

2.5.2 Mask Refinement.

In our Relative Mask Attention part, the attention layer will use the accurate mask information to learn a better attention matrix and the mask will also use the attention information to gradually refine itself. In this ablation study, we only let the attention matrix learn from the mask without any mask refinement. And we get the results in Table 2.8. Since the SQ reflects the segmentation quality, we care more about SQ here. It can be seen that MaskCLIP performs slightly better than that without the mask refinement which demonstrates the effectiveness of the mask refinement.

2.6 Conclusion

In this paper, we have presented a new algorithm, MaskCLIP, to tackle an emerging computer vision task, open-vocabulary universal image segmentation. MaskCLIP is a Transformer-based approach using mask queries with the ViT-based CLIP backbone to efficiently and effectively utilize pre-trained partial/dense CLIP features. MaskCLIP consists of a Relative Mask Attention (RMA) module that is seamlessly integrated with a pre-trained CLIP. MaskCLIP is distinct compared with prior approaches in open-vocabulary semantic segmentation/object detection by building an integrated encoder module for segmentation mask refinement and image feature extraction with a pre-trained CLIP image model. Encouraging experimental results on open-vocabulary semantic/instance/panoptic segmentation have been obtained.

2.7 Acknowledgments

This work is supported by NSF Award IIS-2127544. We thank Xiang Zhang and Boyi Li for helpful discussions.

This chapter, in full, is a reprint of the material as it appears in "Open-Vocabulary Universal Image Segmentation with MaskCLIP". Ding, Zheng; Wang, Jieke; Tu, Zhuowen, International Conference on Machine Learning (ICML), 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Learning Personalized Priors for Facial Appearance Editing with DiffusionRig

3.1 Introduction

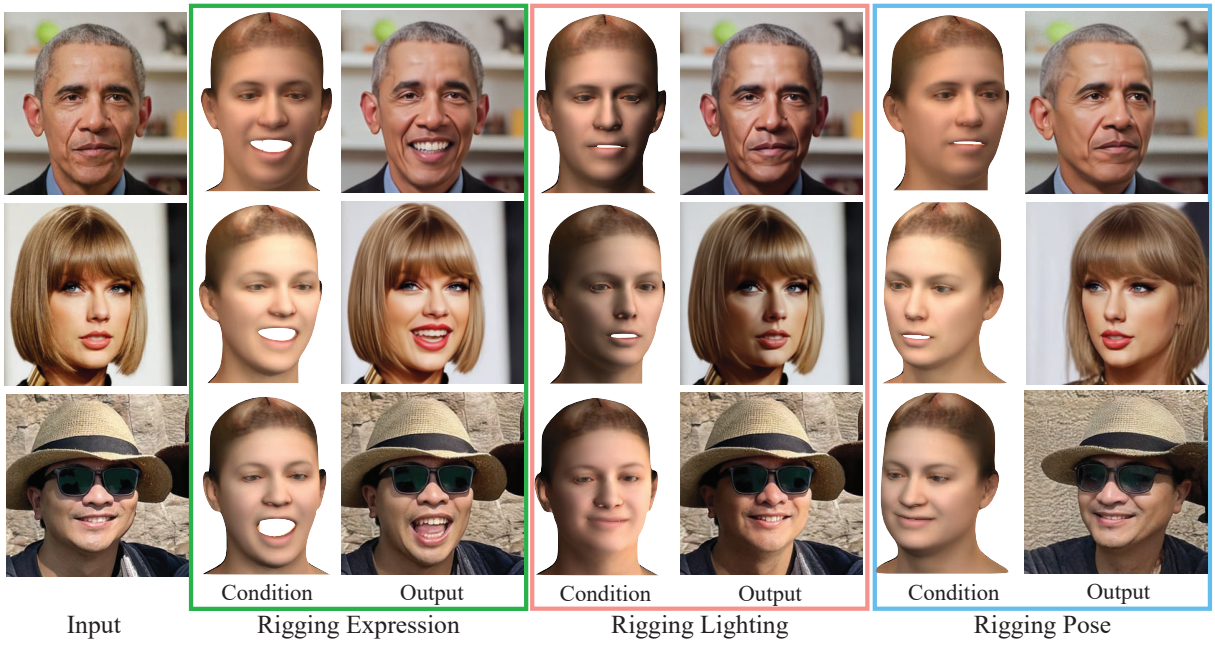


Figure 3.1. DiffusionRig takes in coarse physical rendering as the condition to "rig" the input image with learned personal priors. The edited images respect the rendering conditions, preserve the identity, and exhibit high-frequency facial details.

It is a longstanding problem in computer vision and graphics to photorealistically change the lighting, expression, head pose, etc. of a portrait photo while preserving the person's identity and high-frequency facial characteristics. The difficulty of this problem stems from its

fundamentally underconstrained nature, and prior work typically addresses this with zero-shot learning, where neural networks were trained on a large-scale dataset of different identities and tested on a new identity. These methods ignore the fact that such generic facial priors often fail to capture the test identity's high-frequency facial characteristics, and multiple photos of the same person are often readily available in the person's personal photo albums, e.g., on a mobile phone. In this work, we demonstrate that one can convincingly edit a person's facial appearance, such as lighting, expression, and head pose, while preserving their identity and other high-frequency facial details. Our key insight is that we can first learn generic facial priors from a large-scale face dataset [KLA19a] and then finetune these generic priors into personalized ones using around 20 photos capturing the test identity.

When it comes to facial appearance editing, the natural question is what representation one uses to change lighting, expression, head pose, hairstyle, accessories, etc. . Off-the-shelf 3D face estimators such as DECA [FFBB21] can already extract, from an in-the-wild image, a parametric 3D face model that comprises parameters for lighting (spherical harmonics), expression, and head pose. However, directly rendering these physical properties back into images yields CGI-looking results, as shown in the output columns of Figure 3.1. The reasons are at least three-fold: (a) The 3D face shape estimated is coarse, with mismatched face contours and misses high-frequency geometric details, (b) the assumptions on reflectance (Lambertian) and lighting (spherical harmonics) are restrictive and insufficient for reproducing the reality, and (c) 3D morphable models (3DMMs) simply cannot model all appearance aspects including hairstyle and accessories. Nonetheless, such 3DMMs provide us with a useful representation that is amenable to "appearance rigging" since we can modify the facial expression and head pose by simply changing the 3DMM parameters as well as lighting by varying the spherical harmonics (SH) coefficients.

On the other hand, diffusion models [HJA20a] have recently gained popularity as an alternative to Generative Adversarial Networks (GANs) [GPAM$^+$20] for image generation. Diff-AE [PCWS22a] further shows that when trained on the autoencoding task, diffusion models

can provide a latent space for appearance editing. In addition, diffusion models are able to map pixel-aligned features (such as noise maps in the vanilla diffusion model) to photorealistic images. Although Diff-AE is capable of interpolating from, e.g., smile to no smile, after semantic labels are used to find the direction to move towards, it is unable to perform edits that require 3D understanding and that cannot be expressed by simple binary semantic labels. Such 3D edits, including relighting and head pose change, are the focus of our work.

To combine the best of both worlds, we propose DiffusionRig, a model that allows us to edit or "rig" the appearance (such as lighting and head pose) of a 3DMM and then produce a photorealistic edited image conditioned on our 3D edits. Specifically, DiffusionRig first extracts rough physical properties from single portrait photos using an off-the-shelf method [FFBB21], performs desired 3D edits in the 3DMM space, and finally uses a diffusion model [HJA20a] to map the edited "physical buffers" (surface normals, albedo, and Lambertian rendering) to photorealistic images. Since the edited images should preserve the identity and high-frequency facial characteristics, we first train DiffusionRig on the CelebA dataset [LLWT15] to learn generic facial priors so that DiffusionRig knows how to map surface normals and the Lambertian rendering to a photorealistic image. Note that because the physical buffers are coarse and do not contain sufficient identity information, this "Stage 1 model" provides no guarantee for identity preservation. At the second stage, we finetune DiffusionRig on a tiny dataset of roughly 20 images of one person of interest, producing a person-specific diffusion model mapping physical buffers to photos of just this person. As discussed, there are appearance aspects not modeled by the 3DMM, including but not limited to hairstyle and accessories. To provide our model with this additional information, we add an encoder branch that encodes the input image into a global latent code ("global" in contrast to physical buffers that are pixel-aligned with the output image and hence "local"). This code is chosen to be low-dimensional in the hope of capturing just the aspects *not* modeled by the 3DMM, such as hairstyle and eyeglasses.

In summary, our contributions are:

- A deep learning model for 3D facial appearance editing (that modifies lighting, facial

Figure 3.2. Reconstruction with v.s. without personalized priors. Given the input image and its conditions (surface normals, albedo, and Lambertian rendering) automatically extracted using DECA, Stage 1 learns only generic face priors and fails to reconstruct the identity in both of the randomly sampled reconstructions. With Stage 2, DiffusionRig is able to faithfully reconstruct the input image using either of the two stochastically sampled noise maps.

expression, head pose, etc.) trained using just images with no 3D label,

- A method to drive portrait photo generation using diffusion models with 3D morphable face models, and
- A two-stage training strategy that learns personalized facial priors on top of generic face priors, enabling editing that preserves identity and high-frequency details.

3.2 Related Work

Our work is related to generative models, 3D Morphable Face Models (3DMMs), and personalized priors.

Generative Modeling. Since the proposal of early Generative Adversarial Networks (GANs) [GPAM$^+$20], researchers have made significant progress in generating photorealistic images of constrained classes, such as faces [Kar19, KLA$^+$20, KAL$^+$21]. Recently, denoising diffusion models [HJA20b], which learn to denoise random noise images into photorealistic images, have shown impressive synthesis results and gained popularity as an alternative to GANs. Different diffusion models are invented for faster sampling [SME20a] (used in this work), conditional generation [ND21, DN21a], and later pixel-aligned conditional generation [SHC$^+$22]. Similarly, we use pixel-aligned conditions, specifically surface normal, albedo, and Lambertian rendering images, as the condition that our diffusion model should satisfy. Closely

related to DiffusionRig are Diffusion Autoencoders (Diff-AE) that learn a latent space of facial attributes (e.g., +smiling v.s. −smiling) via the autoencoding task [PCWS22b]. Given binary labels of a certain attribute, the authors find the direction, along which the latent code should be pushed, to manipulate that attribute. 3D-aware generative models are a recent popular trend to combine 3D controllability with 2D image generation [GGU+20, TEB+20b, ZMG+22, TPF+22, CLC+22, CMK+21, HMBL21, TFM+22].

Facial Appearance Modeling. 3D Morphable Face Models or 3DMMs provide a valuable parameter space to describe (and in turn solve for) 3D facial characteristics [BV99]. The FLAME face model learned from 4D scans is a widely-used 3DMM that supports shape, pose, and expression change [LBB+17]. We refer the reader to a recent survey paper on Morphable Face Models [EST+20]. RingNet regresses FLAME parameters from 2D images [SBFB19]. Also a learning-based method, DECA additionally predicts albedo and lighting in spherical harmonics (SH) from a single face image [FFBB21]. An alternative to using 3DMMs for "face de-rendering" is directly predicting surface normals, albedo, and lighting in the image space, as in SfSNet [SKCJ18]. Although such approaches enjoy the benefit of being able to represent hair, accessories, etc. , image-space representations do not provide a physically meaningful parameter space for rigging like 3DMMs do. The geometry, albedo, and lighting from 3DMM are still extremely coarse and far from reality. The community has bridged the realism gap between 3DMM rendering and real photos through expensive hardware setups to capture fine-grained facial geometry [WCY+22, WZA+22] and reflectance fields [DHT+00]. Neural network-based, implicit appearance models have also been proposed to address the infeasibility of explicitly describing the appearance with precise reflectance and lighting [LSSS18, BLS+21, MHP+19, RBS+22, GPL+22, SZA+19, RTD+21, ZGSR21, NLML20, SBT+19, ZBT+20].

Personalized Priors. Learning personal priors has been more widely discussed in super-resolution, face restoration, and inpainting, by using examplar imagery [WOT20], personal supplemental attributes [YFHP18], an attention module with identity penalty [WZC+22], or facial component dictionaries [LCZ+20]. Conditional portrait image editing also shares the

objective of preserving the input identity [LSL+22, TEB+20a]. However, it remains a challenge how to compute an unbiased identity score, and these approaches do not explicitly learn personalized priors.

Closer to DiffusionRig that learns a personal prior from a set of personal album of the person, MyStyle [NAH+22] is a method to finetune a pre-trained StyleGAN model to achieve a generative model for a specific identity, while preserving the expressiveness of the latent space. However, it does not support precise 3D rigging to control the generation and requires a much larger personal dataset to obtain a smooth personalized latent space. DiffusionRig, on the other hand, focuses on controllable image editing and achieves the smooth editing naturally with the continuous physical space as conditions.

3.3 Method

To enable personalized appearance editing, our model, which we dub DiffusionRig, needs to (a) generate images based on different appearance conditions, such as novel lighting, and (b) learn personal priors so that the person's identity is not altered during editing.

To this end, we design a two-stage training pipeline as shown in Figure 5.2. At the first stage, the model learns generic face priors by being trained to reconstruct portrait images given their underlying "appearance conditions" represented as physical buffers automatically extracted using an off-the-shelf estimator. At the second stage, we finetune our model using portrait photos of just one person so that the model learns personalized priors, which are necessary to prevent identity shift during appearance editing.

3.3.1 Learning Generic Face Priors

Our first stage is designed to learn facial priors that enable photorealistic image synthesis conditioned on physical constraints like lighting. For the physical conditioning, we use DECA [FFBB21] to produce the physical parameters including the FLAME [LBB+17] parameters (shape β, expression ψ, and pose θ), albedo α, (orthographic) camera $\mathbf{c}$, and (spherical harmon-

ics) lighting $\mathbf{l}$ from the input portrait image. We then use the Lambertian reflectance to render these physical properties into three buffers: surface normals, albedo, and Lambertian rendering. Although these physical buffers provide pixel-aligned descriptions of the facial geometry, albedo, and lighting, they are rather coarse and nowhere close to photorealistic images (see the Lambertian rendering in Figures 3.1 and 3.2). Still, using these buffers, we can "rig" our generative model in a disentangled, physically meaningful way by changing the DECA parameters. For photorealistic image synthesis, we use a Denoising Diffusion Probabilistic Model (DDPM) as our generator because DDPMs can naturally take pixel-aligned conditions (more advantageous than latent code conditions as shown in Section 3.4.5) to drive the generation process.

Besides the pixel-aligned physical buffers, we keep the random noise images in DDPMs to explain the stochasticity during generation. In addition to the pixel-aligned buffers and noise map, we need another condition to encode *global* appearance information (as opposed to local information such as local surface normals) that is not modeled by the physical buffers, such as hair, hat, glasses, and the image background. Therefore, our diffusion model takes both physical buffers and a learned global latent code as conditions for image synthesis. Formally, our model can be described as $\hat{\varepsilon}_t = f_\theta\left([x_t, z], t, \phi_\theta\left(x_0\right)\right)$ where x_t is the noisy image at timestep t, z represent the physical buffers, x_0 is the original image, $\hat{\varepsilon}_t$ is the predicted noise, and f_θ and ϕ_θ are the denoising model and the global latent encoder, respectively.

It is theoretically possible that the global latent code also encodes local geometry, albedo, and/or illumination information, which could lead to the diffusion model ignoring the physical buffers entirely. Empirically, we find that the network learns to use the physical buffers for local information and does not rely on the global latent code, possibly because these buffers are pixel-aligned with the ground truth and thus more easily leveraged by the model.

3.3.2 Learning Personalized Priors

After learning the generic facial priors at the first stage, DiffusionRig is able to generate photorealistic images given coarse physical buffers. The next step is to learn personalized priors

for a given person to avoid identity shift during appearance editing. Personal priors are crucial to preserving identity and high-frequency facial characteristics, as shown in Figure 3.2. We achieve this by finetuning our denoising model on a specific person's photo album of around 20 images. During the finetuning stage, the denoising model learns the person's identity information. We fix the global encoder from the previous stage since it has learned to encode global image information not modeled by the physical buffers (which we want to retain). We show that this approach is simple and yet effective compared with GANs that need careful tuning, as mentioned in MyStyle [NAH+22].

For this small personalized dataset, we also extract the DECA parameters first. However, since DECA is a single-image estimator, its output is sensitive to extreme poses or expressions. Under the assumption that the general shape of a person's face does not change drastically within a reasonable period of time, we compute the mean of the shape parameters in FLAME over all the images in the album and use that mean shape when conditioning DiffusionRig.

3.3.3 Model Architecture

DiffusionRig consists of two trainable parts: a denoising model f_θ and a global encoder ϕ_θ. The architecture of our denoising model is based on ADM [DN21a] with modifications to reduce computational cost and take an additional global latent code as input. For the global code, we use the same method that ADM uses for their time embedding: We scale and shift the features in each layer using the global latent code. The encoder is simply a ResNet-18 [HZRS16] and we use the output features as the global latent codes.

Our loss function is a P2 weight loss [CLS+22] that computes distances between predicted and ground-truth noises: $\mathscr{L} = \lambda'_t ||\hat{\varepsilon}_t - \varepsilon_t||_2^2$, where λ'_t is a hyperparameter to control the loss weight at different timesteps. We empirically find that the P2 weight loss speeds up the training process and generates high-quality images compared with a constant loss weight.

3.3.4 Implementation Details

During the first stage, we train DiffusionRig on the FFHQ dataset [KLA19a], which contains 70,000 images. With Adam [KB14] as the optimizer with a learning rate of 10^{-4}, we train DiffusionRig for 50,000 iterations with a batch size of 256 (so the total number of samples seen by the model is 12,800,000). During the second stage, we use only 10–20 images of a single person. In the following, we show results for four celebrities (Obama, Biden, Swift, and Harris) and two non-celebrities. Please see the supplemental material and video for more results including more identities. We use 20 images for each person except for Harris, for whom we use only 10, and for the ablation study on the number of training images. We provide the personal photo album of two identities in the supplemental material. We finetune our model on each small dataset for 5,000 iterations with a batch size of 4 (so the total number of seen samples during finetuning is 20,000). We furthermore decrease the learning rate to 10^{-5} for the second stage. Training for the first stage takes around 15 hours using eight A100 GPUs, and the Stage 2 finetuning completes within 30 minutes on a single V100 GPU.

3.4 Experiments

We first show how to edit a person's appearance (e.g., facial expression, lighting, and head pose) by modifying the physical buffers that condition the model. We then demonstrate how to rig, with the global latent code, other aspects of a person's appearance not modeled by the physical buffers such as hairstyle and accessories. By swapping in the global latent code from another image, we can transfer portrait characteristics, such as hairstyle, accessories including glasses, and/or the image background, while preserving the physical properties (e.g., identity, pose, expression, and lighting) from the original image. Finally, we show the power of the learned personal priors by conditioning, for example, an Obama model on both the physical buffers and global latent code from a different person (to "Obama-fy" that person).

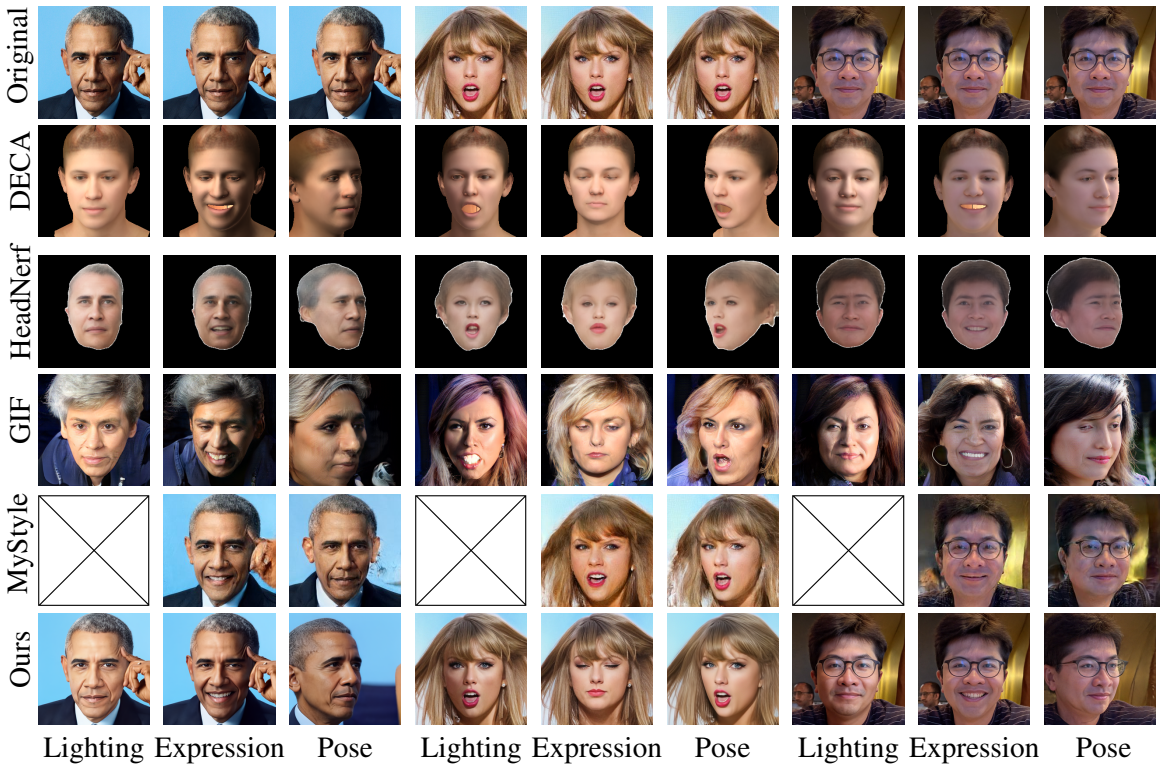


Figure 3.3. Appearance editing. DiffusionRig achieves convincing appearance edits while preserving the individual's identity using only 20 images per identity. GIF creates realistic-looking images but does not use personalized priors, leading to significant identity shifts. MyStyle is unable to make dramatic changes to the expression or pose without artifacts or minor identity shifts. In addition, MyStyle does not trivially support controllable relighting, so the corresponding fields have been left empty.

3.4.1 Rigging Appearance With Physical Buffers

In this section, we use our personalized model to rig the appearance with physical buffers. We show three different types of appearance rigging: relighting, expression change, and pose change. For relighting, we use different Spherical Harmonics (SH) parameters for producing the Lambertian rendering. To change the expression, we modify the expression and jaw rotation parameters of FLAME (the last three parameters of the pose vector). To vary the pose, we modify the head rotation parameters (the first three parameters of the pose vector). The 64-dimensional global latent code is produced by encoding the input image and remains unchanged when editing appearance.

Our results are displayed in Figure 3.3, where we depict three identities: two celebrities and one daily user. All the images have a resolution of 256×256. Additionally, 512×512 results can be found in the supplemental material. We compare our method against DECA [FFBB21], HeadNerf [HPX+22], GIF [GGU+20], and MyStyle [NAH+22], of which the first two are 3D face model estimation methods, and the latter two are GAN-based approaches. As Figure 3.3 shows, while GIF is capable of rigging the appearance by changing the expression and pose, it fails to preserve the individual's identity. DiffusionRig and MyStyle, on the other hand, are both personalized models that are able to preserve the identity. However, since our method is directly conditioned on physical buffers, we can rig the appearance in a physically-based manner, whereas MyStyle needs to search for and step into a certain direction within the latent space to produce the target appearance, limiting its controllability, interpretability, and capacity for dramatic appearance changes. We also observe more artifacts for MyStyle when doing appearance editing, which is likely due to the use of too few images during finetuning the StyleGAN model.

3.4.2 Rigging Appearance With Global Latent Code

By design, DiffusionRig finds it easier to learn what physical buffers can describe from the pixel-aligned buffers than from the global latent code. The latent code thus encodes what

physical buffers cannot describe including background, makeup, and hairstyle. In this part, we change the global latent code to show its effects on the generated images.

In Figure 3.4, we show a 2×3 matrix of generated images. Along the horizontal axis, we swap in the global latent code from another image of the same person while keeping the physical buffers identical (i.e., same physical buffers but different global codes). Along the vertical axis, we replace the physical buffers while keeping the same global latent code (i.e., same global code but different physical buffers). We can see that geometry information, such as head pose and expression, is preserved for each row, which shows that only the physical buffers (not the latent code) contain such information. This means that in DiffusionRig these physical properties are well disentangled from each other and from other appearance properties that physical buffers cannot describe. On the other hand, the information hard to model explicitly, including image background, glasses, and hair style/color, is encoded in the global latent code.

3.4.3 Identity Transfer With Learned Priors

In previous sections, we saw what information the physical buffers and the global latent code encode. Now, we demonstrate what information is encoded in the personalized diffusion models' weights. Here, we keep both physical buffers and global latent code the same but exchange the personalized model itself with another person's personalized model (i.e., model swapping without code or buffer swapping). The results of this experiment for four identities are shown in Figure 3.5. Each row uses the same physical buffers and latent code but another personalized model. Each column uses the same personalized model but different physical buffers and latent code. For example, the column "Obama-fy" shows four images that are generated by Obama's personal model but using the other celebrities' images as input. We see that across each row, while all inputs (physical buffers plus global latent code) are the same, the four different personalized models output different identities. These results further corroborate that our model is able to learn personalized priors from a small dataset.

Figure 3.4. Mix and match of physical buffers and global latent code. We mix the physical buffers from one image and the global latent code from another image to demonstrate how the two conditions encode disentangled information.

Figure 3.5. Swapping personalized models. We demonstrate the power of personalized priors by running one person's model on other identities. This creates the effect of "adding" one person's identity to another person. The images with green borders are "no-swap" results where the corresponding person's model is used.

Table 3.1. RMSE of DECA re-inference. All numbers are multiplies of 10^{-3}. We generate 1,000 images to compute the RMSE. For shape, expression, and pose, the RMSE is computed on rendered FLAME faces. For lighting, the RMSE is computed on re-inferred spherical harmonics directly. We only use our Stage 1 model since GIF is not a personalized model. Numbers for GIF and its vector-conditioned variant are cited from the original paper [GGU+20].

	Light ↓	Shape ↓	Exp. ↓	Pose ↓
GIF [GGU+20]	13.8	**3.0**	5.0	5.6
GIF, vector cond. [GGU+20]	–	3.4	23.1	29.7
DiffusionRig (Ours)	**11.2**	4.3	**2.8**	**4.2**
DiffusionRig, vector cond.	15.5	10.7	8.8	14.0
DiffusionRig, feature cond.	27.0	5.3	4.1	21.6

3.4.4 Baseline Comparisons & Evaluation Metrics

We evaluate our DiffusionRig quantitatively in three aspects: rigging quality, identity preservation, and photorealism, since these three qualities are the most important for our personalized appearance editing.

DECA Re-Inference Error. We follow the same setup as in GIF [GGU+20] to compute the DECA re-inference error. To evaluate relighting quality, we directly compute the RMSE on the re-inferred spherical harmonics. We show our results in Table 3.1. For our model, we also evaluate two ablated versions: "vector cond." and "feature cond." Instead of using pixel-aligned physical buffers as the condition, we use DECA's output parameters and features computed from physical buffers as conditions in our two ablated models. More details can be found in Section 3.4.5.

Face Re-Identification Error. An important metric for evaluating this work is whether DiffusionRig can preserve the identity after appearance editing, since identity shift is a notorious problem in generative model-based editing. To this end, we run a widely popular face re-identification network [Kin09] to automatically determine if the edited and original images are of the same person. As Table 3.2 shows, both MyStyle [NAH+22] and DiffusionRig preserve the identity in all 400 expression-edited images of Obama and another 400 of Swift. That said, for dramatic changes such as head pose change, DiffusionRig preserves the identity better than

Table 3.2. DiffusionRig vs. MyStyle [NAH+22] in expression and pose editing, as measured by an automatic face re-ID error [Kin09] (which has an obvious flaw; see text) as well as a user study on both realism and identity preservation.

	Auto. Face Re-ID ↑		User Study ↑			
	Obama and Swift		Obama		Swift	
	Expr.	Pose	Expr.	Pose	Expr.	Pose
MyStyle	**100%**	97.9%	79.4%	78.0%	64.5%	62.5%
Ours	**100%**	**99.3%**	**87.2%**	**86.5%**	**82.4%**	**80.2%**

MyStyle, as also demonstrated by Figure 3.3. One caveat of this error metric, though, is the obvious degenerate solution of not applying any edit at all, thereby achieving a perfect score. We refer the reader to Figure 3.3 and Table 3.1, which show that DiffusionRig avoids this degenerate solution.

User Study. To further evaluate both the photorealism and identity preservation of images from DiffusionRig against MyStyle, we conduct a user study involving Amazon Mechanical Turk. During the study, we show pairs of images, where the left image is an original image from the real image dataset, and the right image is a generated one. We occasionally include some real images on the right, too, for consistency check and quality control. We then ask the users whether the right image is a real image of the person on the left (so both photorealism and identity preservation are probed). We generate images that include either an expression or pose change for both DiffusionRig and MyStyle. We report our results in Table 3.2.

3.4.5 Ablation Study

We show several ablation studies to motivate the finetuning stage that injects the personalized prior and the choice of physical, pixel-aligned buffers to condition the model.

No Personalized Priors. We first show how DiffusionRig performs in the absence of personalized priors (i.e., trained on only the large dataset from Stage 1). Figure 3.2 shows that our model learns to use the physical buffers as conditions for pose, expression, and lighting, but it is incapable of preserving the person's identity during appearance editing.

Number of Images. Here we explore how the number of images used in Stage 2 affects

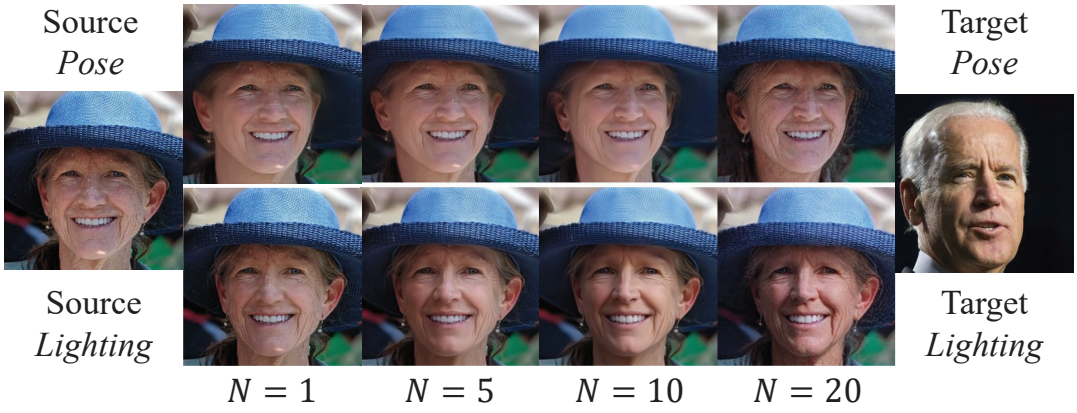


Figure 3.6. Quality w.r.t number of Stage 2 images. DiffusionRig achieves high-quality relighting and pose change with 20 images for Stage 2. Using fewer may yield blurry results and make them hard to rig with new conditions.

DiffusionRig's ability of learning personalized priors. We train three models of a non-celebrity with 1, 5, 10, and 20 images and test them on relighting, expression change, and pose change. As Figure 3.6 demonstrates, using just 1, 5, or 10 images yields worse results than using 20 images (unsurprisingly). With more images, DiffusionRig learns better-personalized priors that capture high-frequency face characteristics, such as the wrinkles in Figure 3.6.

Different Forms of Conditions. There are alternative ways to condition the image synthesis. We demonstrate that pixel-aligned physical buffers are the most effective form in accurately rigging the appearance. We explore the following two conditioning alternatives. **"Vector cond."** is when we directly concatenate DECA parameters, a 236-dimensional vector, to the global latent code without using pixel-aligned buffers. **"Feature cond."** means that we concatenate the physical buffers to the input image and pass them into the encoder to compute a global latent code, which is then used as a non-spatial feature condition.As shown in Figure 3.7, using pixel-aligned physical guidance is essential for accurate conditional image editing. Both vector and feature conditioning suffer from the generated images not following the desired physical guidance.

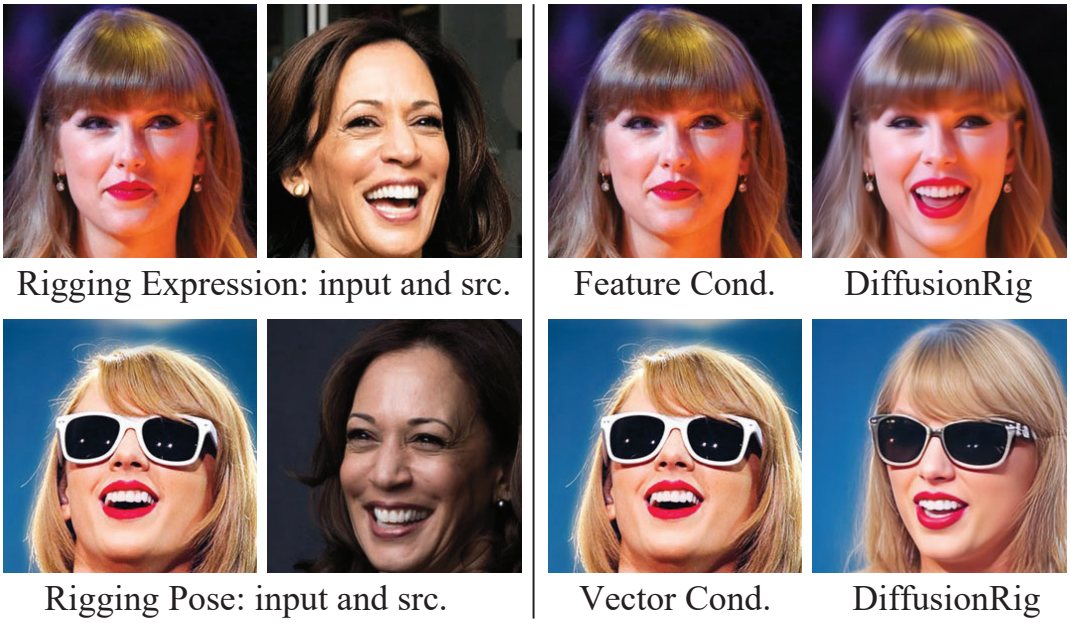


Figure 3.7. Ablation on the form of conditions. Neither feature conditioning nor vector conditioning is able to rig the input image to follow the physical properties of the target image.

3.5 Limitations & Conclusion

Although DiffusionRig achieves state-of-the-art facial appearance editing, it relies on a small portrait dataset to finetune, which limits its scalability for massive user adoption. Furthermore, when the edit involves dramatic head pose change, DiffusionRig may not stay faithful to the original background, since head pose change sometimes reveals what used to be occluded, therefore requiring background inpainting—a topic beyond the scope of this paper. Additionally, since DiffusionRig relies on DECA to get physical buffers, it will also be affected by DECA's limited estimation capability: for instance, extreme expressions usually cannot be well predicted. and the estimated lighting is sometimes coupled with the skin tone.

In this paper, we have presented DiffusionRig, a riggable diffusion model for identity-preserving, personalized editing of facial appearance. We introduced a two-stage method to first learn generic face priors and later personalized priors. Using both explicit conditioning via physical buffers and implicit conditioning via global latent code, we can drive and control our

model's facial image synthesis.

3.6 Acknowledgments

We thank Marc Levoy for the valuable feedback and everyone whose photos appear in this paper for their permission.

This chapter, in full, is a reprint of the material as it appears in "DiffusionRig: Learning Personalized Priors for Facial Appearance Editing". Ding, Zheng; Zhang, Cecilia; Xia, Zhihao; Jebe, Lars; Tu, Zhuowen; Zhang, Xiuming, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Patched Denoising Diffusion Models For High-Resolution Image Synthesis

4.1 Introduction

There have been explosive developments in generative adversarial learning [Tu07, GPAM$^+$14, RMC16, ACB17, KLA19a, DKD17], though many GAN models remain hard to train. VAE models [KW14] are easier to train, but the resulting image quality is often blurry. Diffusion generative models have lately gained tremendous popularity with generated images of superb quality [SDWMG15, HJA20a, SE20, SSDK$^+$20, CHIS23, RDN$^+$22]. Despite the excellent modeling capability of generative diffusion models, the current models still face challenges in both training and synthesis.

Due to direct optimization in the pixel space and multi-timestep training and inference, diffusion models are hard to scale up to high-resolution image generation. Therefore, current state-of-the-art models either use super-resolution methods to increase the generated images to higher resolutions [RDN$^+$22, SCS$^+$22], or optimize the latent space instead of the pixel space [RBL$^+$22]. However, both types of approaches still consist of high-resolution image generators that consume a large memory with a big model size.

To ameliorate the limitations in the current diffusion models, we propose a new method, Patch-DM, to generate high-resolution images with a newly-introduced feature collage strategy. The basic operating point for Patch-DM is a patch-level model that is relatively compact compared

to those modeling the entire image. Though it appears to have introduced compromises for a patch-based representation, Patch-DM can perform seamless full-size high-resolution image synthesis without artifacts of the boundary effects for pixels near the borders of the image patches. The effectiveness of Patch-DM in directly generating high-resolution images is enabled by a novel feature collage strategy. This strategy helps feature sharing by implementing a sliding-window based shifted image patch generation process, ensuring consistency across neighboring image patches; this is a key design in our proposed Patch-DM method to alleviate the boundary artifacts without requiring additional parameters. To summarize, the contributions of our work are listed as follows:

- We develop a new denoising diffusion model based on patches, Patch-DM, to generate images of high-resolutions. Patch-DM can perform direct high-resolution image synthesis without introducing boundary artifacts.
- We design a new feature collage strategy where each image patch to be synthesized obtains features partially from its shifted input patch. Through systematic window sliding, the entire image is being synthesized by forcing feature consistency across neighboring patches. This strategy, named feature collage, gives rise to a compact model of Patch-DM that is patch-based for high-resolution image generation.

Patch-DM points to a promising direction for generative diffusion modeling at a flexible patch-based representation level, which allows high-resolution image synthesis with lightweight models.

4.2 Related Work

Generative diffusion models.

Generative diffusion models [SDWMG15, HJA20a, SE20] which learn to denoise noisy images into real images have gained much attention lately due to its training stability and high image quality. Lots of progress has been made in diffusion models such as faster sampling[SME20a],

conditional generation[ND21, DN21a] or high-resolution image synthesis[RBL+22]. The traits of diffusion models have been amplified particularly by the success of DALL·E 2 [RDN+22] and Imagen [SCS+22] which generate high quality images from the given texts.

Patch-based image synthesis.

The practice of employing image patches of relatively small sizes to generate images of larger sizes has been a longstanding technique in computer vision and graphics, particularly in the context of exemplar-based texture synthesis [EL99]. While generative adversarial networks (GANs) have been utilized for expanding non-stationary textures[ZZB+18], image synthesis is still considered more challenging due to the complex structures present in images. To address this challenge, COCO-GAN[LCC+19] uses micro coordinates and latent vectors to synthesize large images by generating small patches first. InfinityGAN[LCL+22] further improves this by introducing Structure Synthesizer and Padding Free Generator to disentangle global structures and local textures and also generate consistent pixel values at the same spatial locations. ALIS[SSE21] proposes an alignment mechanism on latent and image space to generate larger images. Anyres-GAN[CGS+22], on the other hand, adopts a two-stage training method by first learning the global information from low-resolution downsampled images and then learning the detailed information from small patches. There are also some works share similar directions on patch-based diffusion models. [LL22] does a reshaping operation on the input image which pushes the dimensions of the height and width to the channels. The model still takes the whole image as input just with the shape changed. [WJZ+23] does a patch operation during the training stage by concatenating another position embedding layer to the input while still requires full-resolution during the inference stage. There are also other works applying patch-based diffusion models to specific applications like image restoration under weather conditions[ÖL23] and anomaly detection in brain MRI[BBK+24]. Both of them utilize a conditional diffusion mechanism by utilizing the weather-degraded images or images that miss some patches as conditions.

Our work, Patch-DM consists of a new design, feature collage, in which partial features of neighboring patches are cropped and combined for predicting a shifted patch. We borrow the term "collage" from the picture collage task [WQS+06] for the ease of understanding of our method, though our feature collage strategy only has a loose conceptual connection to picture collage [WQS+06]. Adopting positional embedding in Patch-DM also makes it easier to maintain spatial regularity. Although Patch-DM employs a shifted window strategy, its motivation and implementation are different from those of the widely-known Swin Transformers [LLC+21].

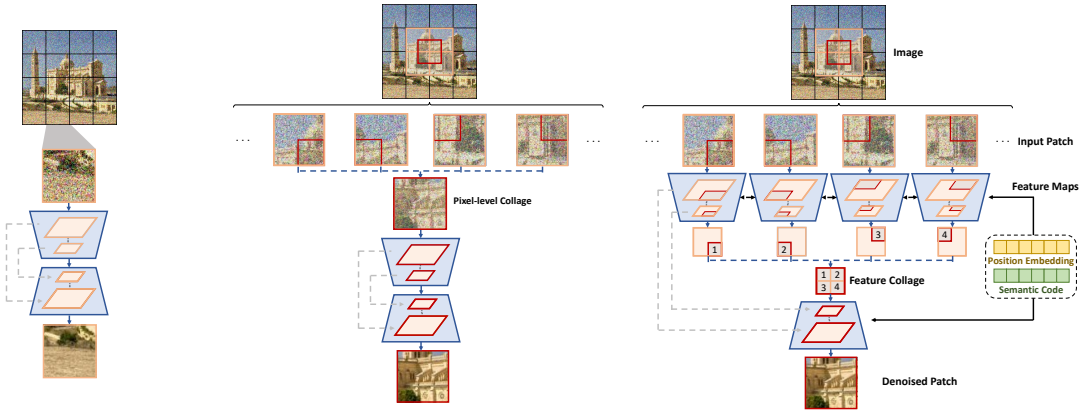
4.3 Background

Denoising diffusion models generate real images from randomly sampled noise images by learning a denoising function [HJA20a]. Instead of directly denoising the random noise image to a real image, denoising diffusion models learn to denoise the noise image through T steps. The forward process adds noise to the image x_0 gradually while the learned denoising function f_θ tries to reverse this process from the $x_T \sim \mathcal{N}(0, \mathbf{I})$. More formally, the forward process at time step $t (t = 1...T)$ can be defined as

$$x_t \sim \mathcal{N}(x_{t-1}; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}), \tag{4.1}$$

where β_t are hyperparameters that control the noise, making the noise level of x_t gradually larger through the timesteps. Note that x_t can be directly derived from the original image x_0 since Eq. 4.1 can be rewritten as

$$x_t \sim \mathcal{N}(x_0; \sqrt{\alpha_t} x_0, (1-\alpha_t)\mathbf{I}), \tag{4.2}$$



(a) No Collage on Patches (b) Patch Collage in Pixel Space (c) Patch Collage in Feature Space (Ours)

Figure 4.1. Patch Generation For Image Synthesis. (a) shows a very basic method of patch-wise image synthesis by simply splitting the images and generating patches independently. This method brings severe border artifacts. (b) alleviates the border artifacts by using shifted windows while generating images and doing patch collage in pixel space. (c) is our proposed method which collages the patches in the feature space. The features for neighboring features will be split and collaged for a new patch synthesis. We will show this method is a key design for us to generate high-quality images without border artifacts.

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. In order to generate the images from the noise input, the denoising model f_θ learns to reverse from x_t to x_{t-1}, which is defined as

$$\hat{\varepsilon}_t = f_\theta(x_t, t), \tag{4.3}$$

$$x_{t-1} \sim \mathcal{N}(x_t; \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\hat{\varepsilon}_t), \sigma_t \mathbf{I}), \tag{4.4}$$

where σ_t are hyperparameters that control the variance of the denoising process. The objective of the denoising model is $||\varepsilon_t - \hat{\varepsilon}_t||^2$. ε_t is ground truth noise added on image.

Therefore, after the denoising model is trained, the model can generate real images from random noise using Eq. 4.4. As can be seen, the whole generation process depends fully on the denoising process. Since the model denoises the image in the pixel space directly, the computation would be very expensive once the resolution gets higher.

4.4 Patched Denoising Diffusion Model

In this section, we describe our proposed Patched Denoising Diffusion Model (Patch-DM). Rather than using entire complete images for training, our model only takes patches for training and inference, and uses our proposed feature collage mechanism to systematically combine partial features of neighboring patches. Consequently, Patch-DM is capable of resolving the issue of high computational costs associated with generating high-resolution images, as it is resolution-agnostic.

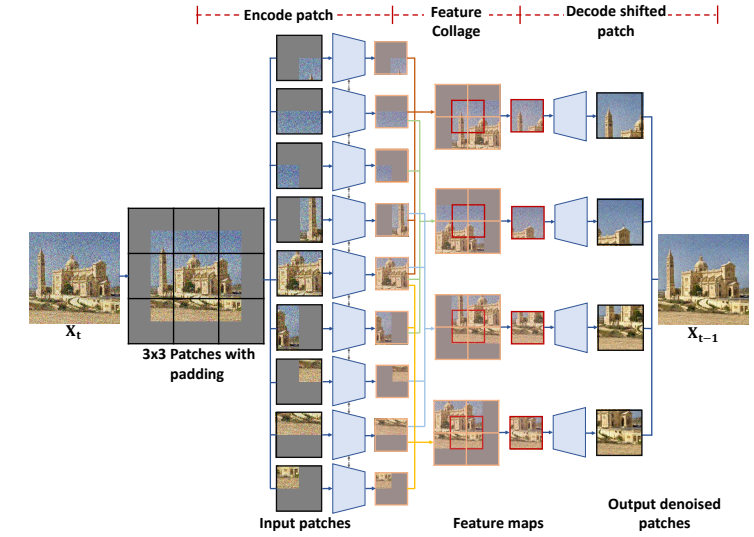
Before we dive into our model's training details, we first give an overview of the image generation process of our method. The training image from the dataset is $x_0 \in \mathbf{R}^{C\times H\times W}$, we split x_0 into $x_0^{(i,j)}$ where i, j is the row and column number of the patch, $x_0^{(i,j)} \in \mathbf{R}^{C\times h\times w}$. Instead of directly generating x_0 like most of methods do, our model only generates $x_0^{(i,j)}$ and concatenate them together to form a complete image.

A very basic way to do this is what we show in Figure 4.1(a) where the denoising model takes the noised image patch $x_t^{(i,j)}$ as input and output the corresponding noise $\hat{\varepsilon}_t{}^{(i,j)}$. However, since the patches do not interact with each other, there will be severe borderline artifacts.

A further way to do this is to shift image patches during each time step depicted in Figure 4.1(b). At different time steps, the model will take either the original split patch $x_t^{(i,j)}$ or the shifted split patch $x_t'^{(i,j)}$ so that the border artifacts can be alleviated which we call "Patch Collage in Pixel Space". However, in Section 4.6 we show that the border artifacts still exist.

To further improve this method, we propose a novel feature collage mechanism depicted in Figure 4.1(c). Instead of performing patch collage in the pixel space, we perform it in the feature space. The patch collage in the feature space is more in-depth and supports multi-level interaction. This allows the patches to be more cognizant of the adjacent features and prevent border artifacts from appearing while generating the complete images. More formally,

$$[z_1^{(i,j)}, z_2^{(i,j)}, ..., z_n^{(i,j)}] = f_\theta^E(x_t^{(i,j)}, t), \tag{4.5}$$



S

Figure 4.2. Detailed Inference process at each timestep.

where f_θ^E is the UNet encoder and $z_1^{(i,j)}, z_2^{(i,j)}, ..., z_n^{(i,j)}$ are the internal feature maps. We then split the feature maps and collage the split feature maps to generate shift patches

$$\begin{aligned} \hat{z}_k'^{(i,j)} = [&P_1(z_k^{(i,j)}), P_2(z_k^{(i,j+1)}), \\ &P_3(z_k^{(i+1,j)}), P_4(z_k^{(i+1,j+1)})], \end{aligned} \tag{4.6}$$

where $P1, P2, P3, P4$ are split functions as shown in Figure 4.1(c). Then we send these collaged shift features $\hat{z}_k'^{(i,j)}$ to the UNet decoder to get the predicted shift patch noise:

$$\varepsilon_t'^{(i,j)} = f_\theta^D([z_1'^{(i,j)}, z_2'^{(i,j)}, ..., z_n'^{(i,j)}], t). \tag{4.7}$$

In order to make the model generate more semantically consistent images, we also add position embedding and semantic embedding to the model so that f_θ will take another two inputs which are $\mathscr{P}(i,j)$ and $\mathscr{E}(x_0)$.

During inference time, we take a 3x3 example as illustrated in Figure 4.2, in order to generate the border patches, we first pad the images so that the feature collage can be done for each patch without information loss. At each time step t, image x_t is decomposed into patches

Table 4.1. Quantitative comparison with previous patch-based image generation methods. All models are trained on the natural images dataset (1024×512), standard benchmarks LHQ(1024×1024) and FFHQ(1024×1024). We use FID to measure the overall quality of generated images and sFID for the quality of high-level structures.

		Nature-21K(1024×512)		LHQ(1024×1024)		FFHQ(1024×1024)	
Method	Patch Size	FID	sFID	FID	sFID	FID	sFID
COCO-GAN[LCC+19]	64×64	70.980	74.208	35.693	88.988	80.059	209.683
InfinityGAN[LCL+22]	101×101	46.550	70.041	50.646	92.577	174.789	270.699
Anyres-GAN[CGS+22]	64×64	44.173	34.430	130.591	100.041	67.076	170.911
Patch-DM (Ours)	64×64	**20.369**	**34.405**	**23.777**	**37.217**	**19.696**	**36.512**

Figure 4.3. Generated 2048×1024 image. We double the number of patches so that the model can generate images with 2x resolution from 1024×512. The left image is a 2048×1024 image, and the right image is a zoom-in of the red bounding box, with a resolution of 256×256.

which are fed into the subsequent encoder. Before a feature map goes through the decoder, a split and collage operation is applied to it. Thus, the decoder outputs the predicted noise of the shifted patch. According to Eq. 4.4, we are able to obtain x_{t-1} and thus generate the final complete images.

4.5 Experiments

4.5.1 Implementation Details

Architecture. For the model architecture, we base our denoising U-Net model from [DN21a] with changes of taking global conditions and positional embeddings. We use two methods to obtain the global conditions. The first is to use a pretrained model to obtain the

image features and use the image features as the global conditions, while optimizing the features directly during training. In this case, we do not have to increase the model parameters and can scale to high-resolution images. However, when the number of images in the training dataset is too large, optimizing the pre-obtained image features requires more effort. We use this approach for global conditioning when training on datasets of 1024×512 images. The pretrained model we use for obtaining the image embeddings is CLIP[RKH$^+$21b]. We resize the images to 224×224 and send them to ViT-B/16 to obtain the features as global conditions; we then optimize these global conditions directly.

The second is jointly training an image encoder and using its output as the global conditions. Here, the jointly training image encoder may borrow the same architecture as in the denoising U-Net's encoder. It works particularly well when the training dataset is large. However, it requires another model, which would be a bottleneck in training on high-resolution datasets, since the computation would increase significantly as the resolution increases. We use this approach when training on large datasets of 256×256 images. We utilize global conditions with a dimension of 512 in both methods.

Classifier-free guidance. We also use the classifier-free guidance [HS22] to improve the training speed and quality. We use classifier-free guidance on both the global conditions and position embeddings. The dropout rate is 0.1 for the global conditions and 0.5 for position embeddings.

Patch size. We use a patch size of 64×64 in all our experiments. The denoising U-Net model's architecture is the same across all the datasets, as it is only related to the patch size regardless of the training images' resolution.

Inference. Once the denoising U-Net model has been trained, we train another latent diffusion model for unconditional image synthesis. The latent diffusion model's architecture is based on the one described in [PCWS22a]. The data we use for training the latent diffusion model is either from the output of the trained image encoder or the directly optimized image embeddings. To synthesize an image, we will first sample a latent code from the latent diffusion

FFHQ LSUN-Bedroom LSUN-Church

Figure 4.4. Generated images on FFHQ, LSUN-Bedroom, and LSUN-Church datasets using our proposed method. All the resolutions are 256×256.

model and then use this latent code to serve as the global conditions for sampling an image. During the sampling stage, we use the inference process proposed by DDIM [SME20a] and set the sampling step to 50.

Evaluation. We conduct both qualitative and quantitative evaluations on four datasets. For quantitative evaluation, we use FID, a popular metric in generative modeling[HRU$^+$17a] and sFID[NMDB21] for the quality of high-level structures. To compute FID, we follow the setting of [HRU$^+$17a] and generate 50K images to compute the metrics over the full dataset. We apply the same setting for sFID.

4.5.2 Results on 1k-Resolution Images

Setup. To show our model's capability on direct high-resolution image synthesis, we show our model's performance on three datasets: LHQ(1024× 1024), FFHQ(1024×1024) and a self-collected 21443 natural images from [Wal23](1024×512). We use a patch size of 64×64 across all datasets. Therefore, each image is split into 16×16 for 1024×1024-resolution and 16×8 for 1024×512-resolution.

We use the CLIP ViT-B/16 pretrained image visual encoder to obtain the image embeddings first. While downsampling to fit the CLIP model may result in the loss of some detailed image information in the embeddings, these details can still be "recovered" during the embedding

optimization in the training process, aided by the supervision of the original high-resolution image. Each image patch can be trained and sampled independently with feature collage assisting to be aware of the surrounding information. Since every image is segmented into smaller patches, the total number of model parameters is much smaller than other large diffusion models.

As most existing diffusion models merely can directly generate images of 1k resolution, and the general strategy for high-resolution synthesis is to sample hierarchically (generate relatively low-resolution images first and then perform super-resolution), our Patch-DM simplifies sample procedure using much more lightweight models, which is one of the main advantages.

Results. We compare our model with previous patch-based image generation methods in Table 4.1. From the table, we can see our method delivers the best overall quality of the generated images on both FID and sFID scores. Apart from the quantitative evaluation, we also present an image generated by our model in Figure 4.3. For more generated images, please refer to our supplementary material.

4.5.3 Results on 256×256 Images

Setup. To compare with other existing generative models, we also train our Patch-DM on three standard public datasets: FFHQ, LSUN-Bedroom, and LSUN-Church, and evaluate its sampling performance. All the resolution is 256×256. Thus, the number of patches is 4×4. Notice that the model architecture keeps the same; the only change here is the number of patches during training and inference. We use the same training setting across the three datasets.

Results. We report the quantitative results in Table 4.2 and qualitative results in Figure 4.4, respectively. In Table 4.2, we can see our model achieves competitive results while still outperforms previous patch-based methods. Figure 4.4 illustrates that despite producing small image patches, our denoising model exhibits minimal boundary artifacts and offers good visual quality. This demonstrates the effectiveness of our feature collage mechanism.

Model size comparison. Compared with other widely used diffusion models, our proposed method could achieve competitive performance using a smaller model with above

Table 4.2. Evaluation Metrics of unconditional image synthesis on three 256×256 datasets: FFHQ, LSUN-Bedroom, and LSUN-Church. For a fair comparison, results are reproduced in the same sampling steps as ours i.e. 50, using provided pretrained checkpoints of other diffusion models. We adopt a patch size of 64×64 for Patch-DM, Anyres-GAN, COCO-GAN and 101×101 for InfinityGAN. We bold the numbers to denote the best numbers in the same category (top: non-patch-based methods, bottom:patch-based methods).

	FFHQ		LSUN-Bedroom		LSUN-Church	
Method	FID	sFID	FID	sFID	FID	sFID
LDM [RBL+22]	8.76	**7.09**	3.40	7.53	4.23	11.44
UDM [KSS+22]	**5.54**	-	4.57	-	-	-
DiffAE [PCWS22a]	9.71	10.24	-	-	-	-
PGGAN [KALL18]	-	-	8.34	9.21	6.42	**10.48**
StyleGAN [KALL18]	-	-	**2.35**	**6.62**	**4.21**	-
COCO-GAN[LCC+19]	34.02	37.44	41.84	62.69	17.91	73.94
InfinityGAN [LCL+22]	28.87	127.92	10.71	19.28	7.08	33.58
Anyres-GAN [CGS+22]	24.48	55.77	15.65	56.24	17.09	80.66
Patch-DM (Ours)	**10.02**	**10.58**	**6.04**	**9.93**	**5.49**	**14.80**

mentioned indispensable components. Patch-DM is fully built upon the network on 64×64 patches regardless of the target image resolution and uses optimized global conditions to avoid the increase of model parameter amounts brought by higher input resolution. Comparison of model parameters with other classic diffusion models on 256×256 resolution is shown in Table 4.3. Notice that we use the same diffusion model architecture for the 1024×1024 and the 1024×512 resolutions.

4.5.4 Applications

We now demonstrate several applications of our Patch-DM. All of them are conducted without post-training.

Beyond patch generation.

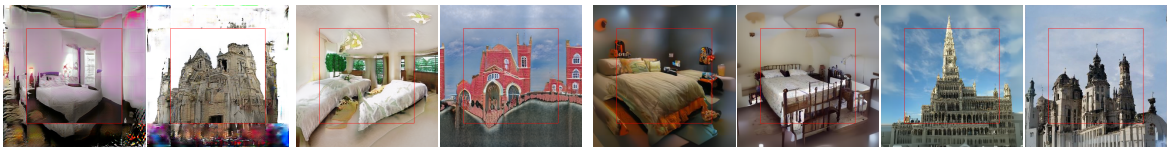
Since our method samples images using patches, we have the option to incorporate more patches during testing. This enables the model to produce images with higher resolutions compared to the ones in the training set without requiring further training. We adopt two ways to achieve this.

Table 4.3. Number of parameters comparison between different diffusion models on 256×256 resolution. SE means semantic encoder to extract global information. The size of our previously trained 1024×1024 and 1024×512 models is 70M+[size of optimized semantic embeddings] during training and [63M latent DPM] in inference.

Method	Model Size ↓
Base model + Super-resolution	
SR3 [SHC+22]	B[64]+625M
Direct generation	
ADM [DN21a]	552M
DiffAE [PCWS22a]	232M
LDM [RBL+22]	274M
Patch-DM (Ours, full model)	154M
Patch-DM (Ours, w/ SE, w/o latent DPM)	91M
Patch-DM (Ours, w/o SE, w/o latent DPM)	70M

The first one is to add more patches internally. We test this on our Nature-21K dataset. To generate 512×1024 images which has the same resolution as the training dataset, we need 8×16 patches (random gaussian noise maps) to start with. To generate a 2× resolution images, we can add another 8×16 patches internally so that the total patch number will become 16×32 that can generate images with a resolution of 1024×2048. The global conditions we use for these patches is the same as the original, while we interpolate the position embeddings in this setting. We provide a generated 2048×1024 image in Figure 4.3. As can be seen, our model can still generate consistent patches even though the newly added patches have never been used in the training process.

The second one is to add patches outside the original image. This way is similar to beyond-boundary generation in COCO-GAN[LCC+19] with a key difference that it needs a post-training process to improve the continuity among patches. We conduct an experiment on LSUN-Bedroom and LSUN-Church by adding more patches. The original resolution in the training data we use is 256×256, which is divided by 4×4 patches. We add more patches to the existing 4×4 patches so that the number grows to 6×6. Therefore, the model can generate an image with a resolution of 384×384. For the original 4×4 patches, we use the condition



COCO-GAN InfinityGAN Patch-DM(Ours)

Figure 4.5. Synthesized 384×384 images on LSUN-Bedroom (256×256) and LSUN-Church (256×256). The left half of each group is LSUN-Bedroom and the right half of each group is LSUN-Church. Despite only being trained on 256×256 images, our model can generate 384×384 images by adding more patches (outside the red bounding box). Extended images generated by our models are compared with COCO-GAN and InfinityGAN, which also have the ability to extend fields without further training.

generated from the latent diffusion model and the position embeddings as pre-defined. For the additional patches, we use the same semantic condition, while we don't use position embeddings for these added ones as in this case we're adding patches outside which the position embeddings could not be interpolated. Thus the model needs to synthesize the additional patches only according to the global conditions and the neighboring context information. We present our results in Figure 4.5 and compare it with COCO-GAN[LCC+19] without post-training and InfinityGAN[LCL+22] under the same setting.

Image outpainting. Another practical application would be image outpainting that only draws the outer part of the image while keeping the input image the same. To do this, we experiment using the model trained on LSUN-Church and LSUN-Bedroom. First, we send images from the LSUN-Church validation dataset and LSUN-Bedroom validation dataset to the image encoder to obtain the global conditions. Then, to keep the original image the same, we replace the inner predicted noised image patches with the ground truth noised images patches (adding corresponding noise to the input images) during each timestep while sampling. We present our results in Figure 4.6. It can be seen from the results that our model can "imagine" the surrounding areas reasonably and generate rather consistent outer parts of the image without obvious border effects.

Image inpainting. In this task, we infill the corrupted images with random masks, which requires the restored results to be consistent in context. We experiment on the LSUN-Church

Figure 4.6. Image outpainting on LSUN-Church and LSUN-Bedroom. The image inside the red bounding box is the input image from the validation dataset. We pad the image patches from 4×4 to 6×6 to enable the image outpainting. The image parts outside the red bounding box are the outpainting results.

Figure 4.7. Image Inpainting on LSUN-Church. Six pairs of images are presented here. For each pair, the left side is the masked image; the right one is the inpainted result by our model without further training on this task. The number of masked patches increasing from left to right.

validation set using already trained models without further tailored training. The original images are masked by different numbers of blocks ranging from 1 to 6, and the sampling process is only conditioned on local position embeddings w/o global conditions. The results are presented in Figure 4.7. From the figure, we can see that our model can infill the blocks consistently using surrounding patches, demonstrating that feature collage facilitates the model with the capability to be aware of adjacent information, enabling it to be naturally applied to inpainting tasks.

4.6 Ablation Study

Three indispensable components: semantic code, position embeddings and shift window strategy on feature levels, considerably eliminate border artifacts and improve our model performance. Here, we conduct ablation study to investigate the effects of these modules. We provide both qualitative results and quantitative results in Figure 4.8 and Table 4.4 respectively.

Global conditions.

We study the problem without global conditions; thus, the generation process will fully rely on the positional embedding and neighboring context information. We present our images

in Figure 4.8 (a). It's interesting to see how the model generates when no global conditions are given, which is a strong constraint for the model to generate semantic-related patches. From the given image, we can see that the model can still generate locally-consistent images; the image quality is however relatively low.

Position embeddings.

The last section shows that global conditions are necessary for our model to generate high-quality images. We then condition the model only on those to investigate the role of positional embeddings. The results are shown in Figure 4.8 (b). Without the position information, the model would generate distorted images with patchTheelonging to where they should be, although the whole image may follow a certain style. Hence, the positional embeddings are vital to our model.

Collage in the pixel space.

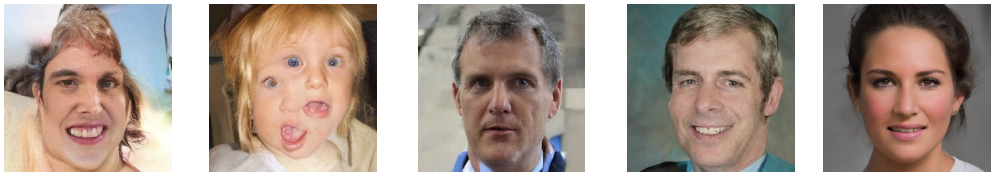
A straightforward idea is to perform collage in the pixel space as present in Figure 4.1(b); the images are decomposed by window-shifted patches from their original positions. To maintain patch size consistency, we add zero padding around the image. For the sampling procedure: In an odd-number step, original patches are generated independently, while in an even-number step, patches with shifted positions are sampled.

Under this scheme, we experiment in two different settings. The first is to take a fixed shift step (half patch size) along the height and width direction. The sample result is shown in Figure 4.8(c). There are still apparent artifacts along the border. This proves that even though the shift window on the image level could enable patches to be aware of surroundings during sampling, the awareness level is quite limited, and the final generation is similar to breaking the image into smaller patches.

The second setting is to shift the patch position with a randomly sampled step ranging from zero to patch size. The inference result is shown in Figure 4.8(d). The sample quality is much improved compared to the previous situation. However, the result is still not as photo-

Table 4.4. FID evaluation on 1,000 images of different ablation settings to investigate the importance of semantic condition, position embedding, and feature-level window shift.

Method	FID (1k) ↓
No global semantic condition	79.33
No position embedding	48.82
Pixel space fixed shift	49.80
Pixel space random shift	52.11
Patch-DM (Ours)	**37.99**



(a) No global conditions (b) No position embeddings (c) Pixel space fixed shift (d) Pixel space random shift (e) Ours

Figure 4.8. Ablation study on global conditions (a), position embeddings (b), and feature level shift (c, d).

realistic as Patch-DM. The reason is that although the random shift enables finer surrounding awareness, it lacks in-depth feature interaction as our model does. Therefore, the feature-level window shift and collage can significantly eliminate border artifacts and improve final inference quality.

4.7 Conclusion

We have presented a new algorithm, Patch-DM, a patch-based denoising diffusion model for generating high-resolution images. We introduce a feature collage strategy to combat the boundary effect for patch-based image synthesis. Patch-DM achieves a significant reduction in model size and training complexity compared to the standard diffusion models trained on the original size images. Competitive quantitative and qualitative results are obtained for Patch-DM when trained on several image datasets.

4.8 Acknowledgments

This work is supported by NSF Award IIS-2127544 and IIS-2211258.

This chapter, in full, is a reprint of the material as it appears in “Patched Denoising Diffusion Models For High-Resolution Image Synthesis”. Ding, Zheng; Zhang, Mengqi; Wu, Jiajun; Tu, Zhuowen, International Conference on Learning Representations (ICLR), 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Restoration by Generation with Constrained Priors

5.1 Introduction

Image restoration involves recovering a high-quality natural image x from its degraded observation $y = H(x)$ is a fundamental task in low-level vision. The challenge lies in finding a solution that 1) matches the observation through a set of degradation steps; and 2) aligns with the distribution of x. In scenarios where the degradation process H is unknown, the problem becomes a blind image restoration problem.

Discriminative learning approaches [GWX+22, ZCLL22, WLZS21, YRXZ21] aim to solve this inverse problem directly by training an inverse model $F(y)$, typically a neural network, using datasets of low- and high-quality image pairs (x, y). However, the trained model is limited to restoring images with degradations H present in the training set. This limitation places the burden of generalization on the construction of the training set. The effectiveness of these methods also heavily depends on the capacity of the inversion model and the characteristics of the loss function. Model-based optimization methods [ROF92, ZLZ+21, REM17, KEES22, CKM+22], on the other hand, assume that the degradation model is only known at inference time. They focus on learning the image prior $p(x)$, which can be represented as regularization terms [ROF92], denoising networks [ZZGZ17, REM17], or more recently pre-trained diffusion models [KEES22, CKM+22]. However, these methods generally assume that the degradation process

is known at inference time, limiting their practicality and often relegating them to synthetic evaluations.

In this paper, we adopt a markedly different approach to the image restoration problem. We observe that humans are able to recognize a degraded image (i.e., a 'bad photo') and envision a fix without knowing the imperfections in the image formation process. Such insights rely on our inherent understanding of what constitutes a high-quality image. Building on this observation, we propose to approach image restoration using the recent success of large generative models, which possess the capacity of forming high-quality imagery. Unlike prior works, we do not make any assumption on the degradation process. Our method solely relies on a well-trained denoising diffusion model.

The challenges then arise in how to project the input image into the generative process given the models are trained on mostly clean images. And once projected, how to constrain the generation to preserve the useful features in the input, e.g., the identity. We address the input projection by adding Gaussian noise to the low-quality image to be restored, matching the distribution of clean images added with noise. Once projected, we can then denoise the image as is normally done in the generation process of a diffusion model. To handle the second challenge of preserving useful signals in the input, we propose to constrain the generative space by finetuning the model with anchor images that share characteristic features with the low-quality input. When the anchor is given, such as from an album of other photos of the same identity, we can simply finetune the model with the provided images. When the anchor is missing, as in most single-image restoration scenarios, we propose to use a generative album as the anchor. The generative album is a set of clean images generated from the diffusion model with the low-quality input image imposing soft guidance, and thus closely resembles the input image.

Surprisingly, we find that our straightforward approach yields high-quality results on blind image restoration. Unlike previous methods, our approach does not rely on paired training data or assumptions about the degradation process. It thus generalizes well to real-world images with unknown degradation types, such as noise, motion blur, and low resolution. By

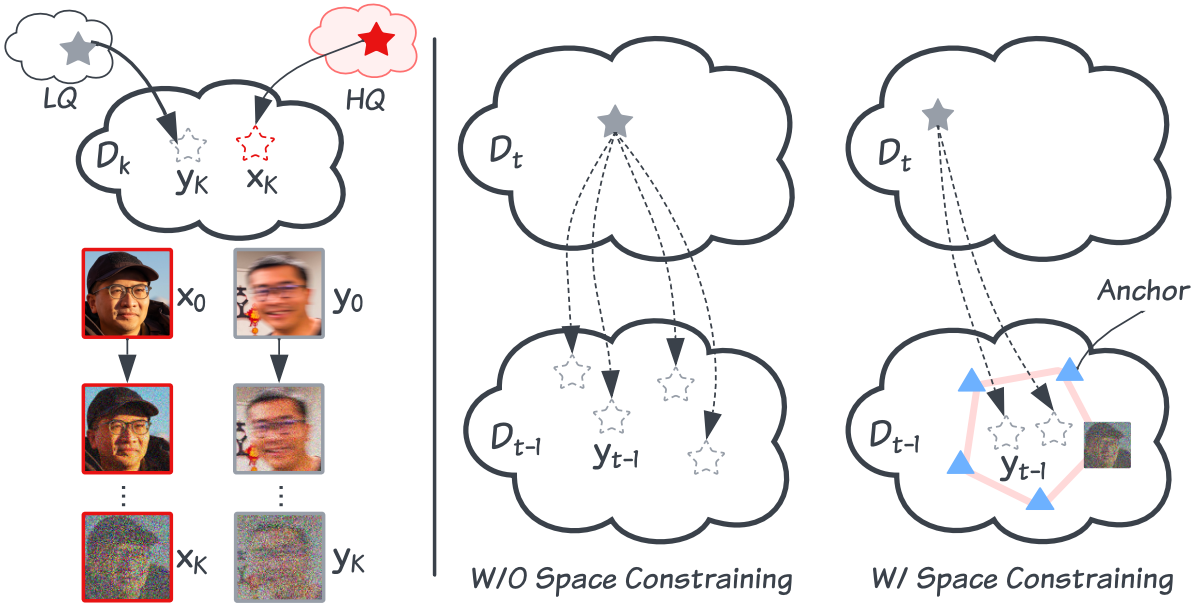


Figure 5.1. Left: Image projection. When sufficient Gaussian noise is added to the low- and high-quality image, we can bring them to the same distribution. The low-quality image can thus be denoised with a pre-trained diffusion model. **Right: With and without space constraining.** A regular diffusion step lands y_t in an arbitrary position in the generative space; with space constraining, the path of generation becomes more constrained towards the space defined by the anchor images.

effectively harnessing the generative capacity of a pre-trained diffusion model, our generation-based restoration approach produces high-quality and realistic images that are faithful to the input identity.

5.2 Related Works

Supervised Learning for Image Restoration.

The trend of leveraging advanced neural network architectures for image restoration has spanned from CNNs [ZZC+17, CJH+21, ZZZ18, TXL+20, ZLL+19] to GANs [LTH+17, KBM+18, KMWW19], and more recently, to transformers [ZAK+22, LCS+21, WCB+22] and diffusion models [SHC+22, WDT+22, SCC+22]. One aspect remains unchanged: these methods are trained on datasets comprising pairs of high-quality and low-quality images. Typically, these image pairs are synthetically generated, depicting a single type of degradation, leading to task-specific models for denoising [ZZC+17, ZZZ18, XC20, TXL+20, GYZ+19], deblurring [KBM+18, KMWW19, WDT+22, CJH+21], or super-resolution [LTH+17, WXDS21, SHC+22]. However, they fall short when applied to real-world low-quality images, which often suffer from diverse, unknown degradations.

In specific domains, particularly with facial images, numerous works have focused on training blind restoration models that simulate various degradation types during training. For instance, GFPGAN [WLZS21] and GPEN [YRXZ21] enhance pretrained GAN networks with modules to leverage generative priors for blind face restoration. Recent approaches like CodeFormer [ZCLL22], VQFR [GWX+22] and RestoreFormer[ZAK+22] exploit the low-dimensional space of facial images to achieve impressive results. Emerging works have also started building upon the success of diffusion models [HJA20c, SME20b, DN21b]. For example, IDM [ZHS+23] trains a conditional diffusion model for face image restoration by injecting low-quality images at different layers of the model. Conversely, DR2 [WZZ+23] combines the generative capabilities of pre-trained diffusion models with existing face restoration networks. Another line of works [LLY+18, LLR+20] seeks to enhance the results by incorporating addi-

tional information present in a guide image or photo album, which is often available in practice. Nevertheless, these methods rely on a synthetic data pipeline for training, which limits their generalizability. Diverging from these methodologies, our approach does not use paired data, synthetic or real, allowing it to generalize naturally to real data without succumbing to artifacts.

Model-based Image Restoration.

Unlike supervised learning methods, model-based methods often form a posterior of the underlying clean image given the degraded image, with a likelihood term from the degradation process and an image prior. [ZLZ+21, REM17] proposed using denoising networks as the image prior. These priors are integrated with the known degradation process during inference, and the Maximum A Posteriori (MAP) problem is addressed through approximate iterative optimization methods. DGP [PZD+21] proposes image restoration through GAN inversion, searching for a latent code that generates an image closely matching the input image after processing it through the known degradation. The recent success of pre-trained foundational diffusion models has inspired works [KS21, KVE21a, CKJ+21, KVE21b] to utilize diffusion models as such priors. Kawar et al. [KEES22] and Wang et al. [WYZ22] proposed an unsupervised posterior sampling method using a pre-trained denoising diffusion model to solve linear inverse problems. Chung et al. [CKM+22] extends diffusion solvers to general noise inverse problems. Despite these advancements, these methods generally assume that the degradation process is known at inference, limiting their practicality to synthetic evaluations. In contrast, our method does not assume any knowledge of the degradation model at training or inference.

Personalized Diffusion Models.

Personalization methods aim to adapt pre-trained diffusion models to specific subjects or concepts by leveraging data unique to the target case. In text-to-image synthesis, many works opt for customization by fine-tuning with personalized data, adapting token embeddings of visual concepts [GAA+22, GAA+23], the entire denoising network [RLJ+23], or a subset of the network [KZZ+23]. Recent studies [JZC+23, SXLJ23, XYF+23] propose bypassing per-object

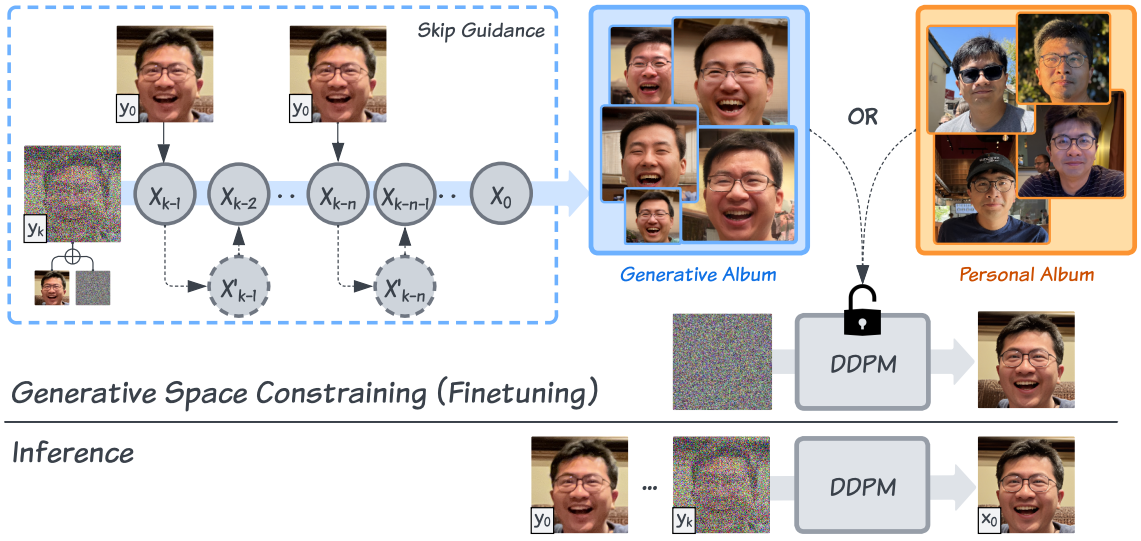


Figure 5.2. An illustration of our finetuning and inference stage. The core of our method is to constrain the generative space by fine-tuning a pre-trained diffusion model with either a generative album or a personal album. The generative album is generated from the input low-quality image with skip guidance to loosely follow the characteristics of the input. Once the generative space is constrained, at inference time, we can simply add noise to the input low-quality image and pass it through the diffusion model to do restoration.

optimization by training an encoder to extract embeddings of subject identity and injecting them into the diffusion model's sampling process. In other domains, DiffusionRig [DZX+23] learns personalized facial editing by fine-tuning a 3D-aware diffusion model on a personal album. In this work, we demonstrate that a personalized diffusion model represents a constrained generative space, directly usable for sampling high-quality images to restore images of a specific subject, without additional complexities. For *single-image restoration*, unlike previous instance-based personalization methods [JZC+23, SXLJ23, XYF+23], we generate an album of images close to the input and then constrain the diffusion model using this generative album. This approach enables restoration by directly sampling from the fine-tuned model, eliminating the need for guidance.

5.3 Method

5.3.1 Preliminaries

A diffusion model approximates its training image distribution $p_\theta(x_0)$ by learning a model θ that effectively reverses the process of adding noise. The commonly used Denoising Diffusion Probabilistic Models (DDPM) gradually introduce Gaussian noise into a clean image x_0:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\varepsilon, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{5.1}$$

The reverse generative process aims to progressively denoise x_t until it is free from noise. Once a diffusion model is trained, for any given time t and the corresponding noisy image x_t, it can iteratively denoise by sampling from $p(x_0|x_t)$ using the trained model.

The objective of image restoration, on the other hand, is to recover the latent high-quality image x_0 from a low-quality, partially observed image y_0. Contrary to previous methods that decompose the posterior distribution into the likelihood $p(y_0|x_0)$ and the prior $p(x_0)$ to solve a MAP problem, we propose to recover the complete observation by directly sampling from the posterior:

$$\hat{x} \sim p(x_0|y_0) \tag{5.2}$$

5.3.2 Restoration by Generation

We aim to maximally leverage the generative capacity of the diffusion model by using its iterative sampling process for restoration. A critical observation underlies this approach: when sufficient Gaussian noise is added to the degraded observation y_0, the resultant image y_t:

$$y_t = \sqrt{\alpha_t} y_0 + \sqrt{1-\alpha_t}\varepsilon, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{5.3}$$

becomes indistinguishable from the underlying clean image x_0 with the same noise. That is, there exists a large enough K such that

$$y_K \approx x_K \tag{5.4}$$

This phenomenon becomes apparent from Eq 5.1 and 5.3 as α decreases and when the same noise ε is sampled. It is also demonstrated in Fig 5.1, where adding noise to high-quality and low-quality images can progressively align their distributions, making them more similar over time, this suggests:

$$p(x_0|y_K) \approx p(x_0|x_K) \tag{5.5}$$

Based on this observation, we can sample a clean image x_0 from $p(x_0|y_K)$ using the same sampling process as from $p(x_0|x_K)$; in other words, we can denoise y_K iteratively directly with the pre-trained diffusion model. Since the sampling process remains unchanged, the resultant image should match the quality of the images generated from the original diffusion model.

We find it critical to select the optimal time K, which determines the amount of noise added to the low-quality input image to start the sampling process. If too little noise is added, the discrepancy between x_K and y_K becomes large, yielding low-quality samples as y_K does not align with the training distribution $p(x_K)$ of the diffusion model. On the other hand, with too excessive noise added, the original contents in the input y_K are hardly discernible. The generated sample, though with high quality, will not be faithful to the input. We aim to produce high-quality samples, while mitigating the information loss, and achieve so by constraining the generative space of the pre-trained diffusion model.

5.3.3 Generative Space Constraining

The loss of information is inherent in the diffusion process. Due to the stochasticity of the forward Markov chain, the clean image generated using the reverse process from x_t may not

Table 5.1. Quantitative comparison on real-world single-image blind face restoration on four datasets.

	Wider-Test		WebPhoto-Test		LFW-Test		Deblur-Test	
	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑
Input	183.03	15.68	161.82	20.26	131.68	27.51	169.43	27.53
GFPGAN[WLZS21]	59.38	56.48	114.15	55.13	64.10	60.46	178.40	58.03
CodeFormer[ZCLL22]	48.57	55.70	98.55	55.20	66.31	58.72	163.47	57.09
VQFR[GWX$^+$22]	52.64	54.23	105.94	52.44	63.73	57.52	168.36	54.45
DR2(+VQFR)[WZZ$^+$23]	69.40	53.62	143.96	51.92	67.70	57.42	173.33	55.34
Ours	**46.38**	**58.73**	**96.44**	**57.71**	**56.32**	**60.68**	**135.33**	**60.20**

match the original x_0. The larger t is, the larger the generative space $p(x_0|x_t)$ spans. The learned score functions guide x_t to the clean image space without constraining its content. This property is desirable for a generative model where the diversity of generation is valued. However, this is not ideal for image restoration where the input contents also need to be preserved. The goal is thus to constrain the generative space to a small subspace that tightly surrounds the underlying clean image.

We propose to use a set of anchor images to fine-tune the diffusion model, thus imposing the generative space. These anchor images can be given in the form of a *personal album*, or be generated as a *generative album* in the common scenario of single image restoration.

Personal Album as Additional Information.

In many real-world scenarios, additional information about the underlying clean image beyond a single degraded observation is available, such as an album of different clean images of the same subject. We personalize the pre-trained diffusion model in this case — fine-tuning it with the personal album. This approach naturally addresses the ill-posed nature of single-image restoration, producing results containing authentic high-frequency details absent in the degraded observation. This is demonstrated in identity preservation in face restoration tasks (Sec 5.4.2).

Generative Album from a Single Degraded Observation.

For single-image restoration, due to its ill-posed nature, we can only constrain the generative space to a subspace of high-quality realistic images close to the degraded observation.

To generate this album of high-quality images, we follow approaches similar to previous works on guided image generation [CKM+22, SZY+23, BCS+23]. Specifically, given a degraded image y_0, we first add noise ε_K to obtain y_K, then denoise it progressively with the pre-trained diffusion model. For the denoised image x_t, we apply a simple L_1 guidance that computes the distance between the input degraded image and the generated image:

$$x'_t = x_t - \lambda \nabla_{x_t} ||y_0 - \hat{x}_{0,t}||_2^2 \tag{5.6}$$

Unlike previous methods where the guidance needs to be strongly followed, our guidance, the low-quality input, is an approximation. Instead of applying the guidance at every step [SZY+23, CKM+22], we propose to apply this approximated guidance periodically at every n steps. The proposed *Skip Guidance* enforces the generated image to loosely follow the information in the degraded input while retaining the quality of images in the generative steps. We repeat this process multiple times to generate a set of images that form a generative album, which is used to fine-tune the diffusion model.

Once the diffusion model is fine-tuned with a personal or generative album, we restore a degraded image y_0 by adding noise ε_K. Then, we iteratively denoise y_K using the fine-tuned model for K steps, *without* further guidance. Notably, our approach does not rely on paired data for training and makes no assumptions about the degradation process at training or inference.

5.4 Experiments

With the core observation that generation can be directly applied for restoration, our method requires only a pre-trained unconditional diffusion model and is applicable to any image domain for which the diffusion model has been trained. We first show results of our restoration-by-generation approach on the standard task of single-image blind face restoration in Sec 5.4.1. In Sec 5.4.2, we extend our approach to personalized face restoration. Here, the objective is to restore a degraded image of a subject using other clean images of the same identity. Sec 5.4.3

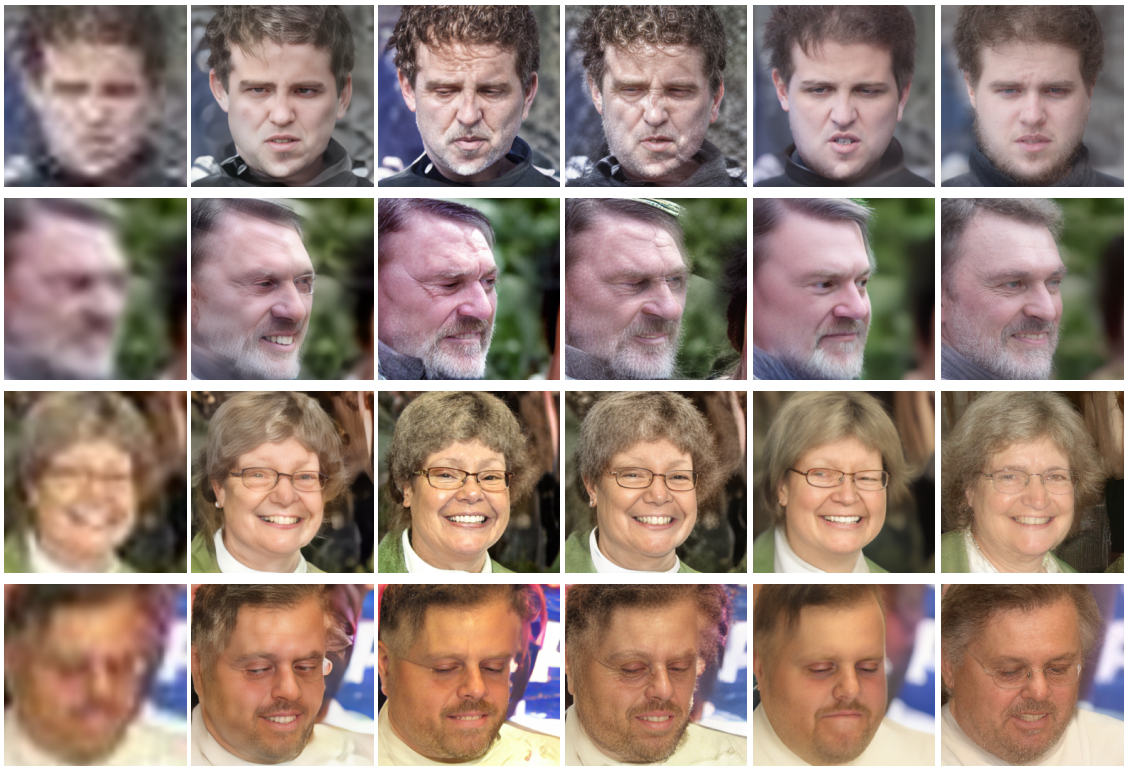


Figure 5.3. Qualitative comparison with baselines on Wider-Test. With strong generative capacity of the diffusion model, our method performs well on severely degraded images. We are able to produce high-quality and realistic images while prior works suffer from unrealistic artifacts.

presents the adaptation of our method to different image categories, such as dogs and cats, by simply swapping the pre-trained diffusion model. Notably, as our method does not presume any specific form of degradation, all our evaluations are conducted on real images with unknown degradation.

5.4.1 Blind Face Restoration with Generative Album

For the task of single-image blind face restoration, we utilize an unconditional diffusion model pretrained on the FFHQ dataset [KLA19b]. We first assess our approach on three widely-used real-world face benchmarks with degradation levels ranging from heavy to mild: Wider-Test (970 images) [ZCLL22], LFW-Test (1771 images) [WLZS21], and Webphoto-Test (407 images) [WLZS21]. These datasets are collections of in-the-wild images aligned using the method employed in FFHQ [KLA19b].

Our approach uses a *generative album* as the anchor for restoring these in-the-wild images. For each input low-quality image, we generate 16 images with skip guidance to form the album. We then fine-tune the diffusion model using this album to constrain the generative space. The process involves adding noise to the input low-quality image and denoising it for K steps with the fine-tuned model, where $K = 200$. The model is fine-tuned for 3,000 iterations with a batch size of 4 and a learning rate of 1e-5.

We benchmark our method against state-of-the-art supervised alternatives for blind face restoration, including the GAN-based GFPGAN [WLZS21], two codebook-based approaches (Codeformer [ZCLL22] and VQFR [GWX$^+$22]), and a diffusion-based approach DR2 [WZZ$^+$23]. Except for DR2, which combines a diffusion model with the pretrained supervised face restoration model VQFR [GWX$^+$22], all methods utilize supervised training with synthetic low-quality images from FFHQ.

Quantitative and qualitative results are provided. For the former, we use FID [HRU$^+$17b] and MUSIQ(Koniq) [KWW$^+$21] as metrics following CodeFormer [ZCLL22]. The quantitative scores are in Table 5.1. Previous methods, except for DR2 [WZZ$^+$23], are trained on FFHQ-

Figure 5.4. Comparison with previous methods on Deblur-Test. Previous methods do not include motion blur as part of the degradation simulation for training, and thus fail to restore the images. In contrast, our method does not make assumptions on the degradation types and generalizes more robustly.

512×512 for restoration. For a fair comparison, we downsize the outputs of these methods to 256×256 for metric calculation. Our results surpass all previous methods in terms of FID and MUSIQ across all datasets, despite not undergoing a supervised training approach for image restoration. Qualitative comparisons in Figure 5.3 illustrate that our method produces high-quality restoration results akin to those from an unconditional diffusion model, even with severely degraded input images.

Our method's agnosticism to the degradation process leads to superior generalization capabilities. To further demonstrate this, we constructed a motion blur dataset (Deblur-Test) by selecting 67 images from [LSC$^+$22] featuring moderate to severe real motion blur. The synthetic data pipeline in other supervised approaches does not model motion blur, resulting in poor performance on this out-of-distribution dataset. In contrast, our method consistently restores clean images from complex non-uniform motion blur, as seen in Figure 5.4, outperforming previous methods significantly, as shown in Table 5.1.

Figure 5.5. Qualitative Comparison on personalized face restoration. From top to bottom: Subject A, Obama and Hermione. With a personal album as anchor, we are able to restore images with faithful preservation of the input identity. Previous single-image methods alter the identity with lost details; previous reference-based methods fail to produce high-quality images and are prone to artifacts.

5.4.2 Personalized Face Restoration

We now evaluate our method on personalized restoration. Given a set of clean images of a subject, the goal is to restore any degraded image of the same subject using personalized features to preserve identity and recover high-frequency details that may have been lost in the degraded image. Our method naturally incorporates the personal album as the anchor. We use a personal album that contains around 20 images with diversity in pose, hairstyle, accessories, lighting, etc. We fine-tune on the personal album for 5,000 iterations. The model can then be used to restore any low-quality images of the same subject through direct sampling.

We compare our method against three single-image-based works: Codeformer [ZCLL22], VQFR [GWX+22], DR2 [WZZ+23], as well as an exemplar-based approach ASFFNet [LLR+20] which also incorporates a personal album for additional information. We evaluate our approach on three subjects: an elderly woman (Subject A), Obama and Hermione. We present the qualitative comparison in Figure 5.5. Single-image-based methods struggle to preserve identity – for example, wrinkles and other facial structures are often missing in the results of CodeFormer or DR2 for the elderly subject, altering their age and identity. By using a photo album as reference, ASFFNet preserves identity better, but fails to produce high-quality results. Our method, on the other hand, directly samples from the personalized generative space to do restoration, and thus produces faithful and high-quality results.

We also provide quantitative evaluation in Table 5.2 where we focus on the identity preservation. We use the identity score which uses the cosine similarity of the features given by a face recognition network ArcFace[DGXZ19]. For each subjects, we collect around 20 test images and compute their average identity scores. Table 5.2 shows that our method preserves the identity of the subject much better than both single-image-based methods and the exemplar-based approach ASFFNet.

Table 5.2. IDS comparison on three subjects. We use the cosine similarity of the features given by ArcFace[DGXZ19] to compute identity score.

	Subject A	Obama	Hermione
Input	0.721	0.502	0.483
CodeFormer[ZCLL22]	0.633	0.558	0.518
VQFR[GWX$^+$22]	0.560	0.527	0.483
DR2(+VQFR)[WZZ$^+$23]	0.384	0.400	0.392
ASFFNet[LLR$^+$20]	0.694	0.574	0.522
Ours	**0.731**	**0.716**	**0.664**

5.4.3 Beyond Face Restoration

Our model does not make any assumptions about the type of degradation or image contents, allowing it to be easily extended to other categories of data where a generative model is available. Specifically, we evaluate our approach's ability to generalize to restoring dog and cat images. We pre-train two diffusion models with the same architecture, one for dogs and one for cats, on the AFHQ Dog and Cat datasets [CUYH20]. Our testing involves three subjects: a gray cat, an English golden retriever, and an Australian shepherd. For each subject, we fine-tune the pre-trained diffusion model using an album of around 20 images. Once fine-tuned, given a low-quality image, we add noise to it and then denoise it using the fine-tuned model. Qualitative results in Figure 5.6 demonstrate that our method can effectively reconstruct high-frequency details such as fur, while preserving the identity.

5.5 Ablation Studies

Noise Step K. Our restoration-by-generation approach is predicated on the observation that sufficient noise added to a degraded image y_0 and subsequent denoising of the noisy image y_K with a pre-trained diffusion model yields a high-quality, realistic image. Here, we demonstrate this observation and analyze the effect of the choice of K, which determines the noise level added to initiate the sampling process. Figure 5.7 displays sampled images from y_K for varying K values. A smaller K leads to a y_K that falls outside the typical diffusion process's training

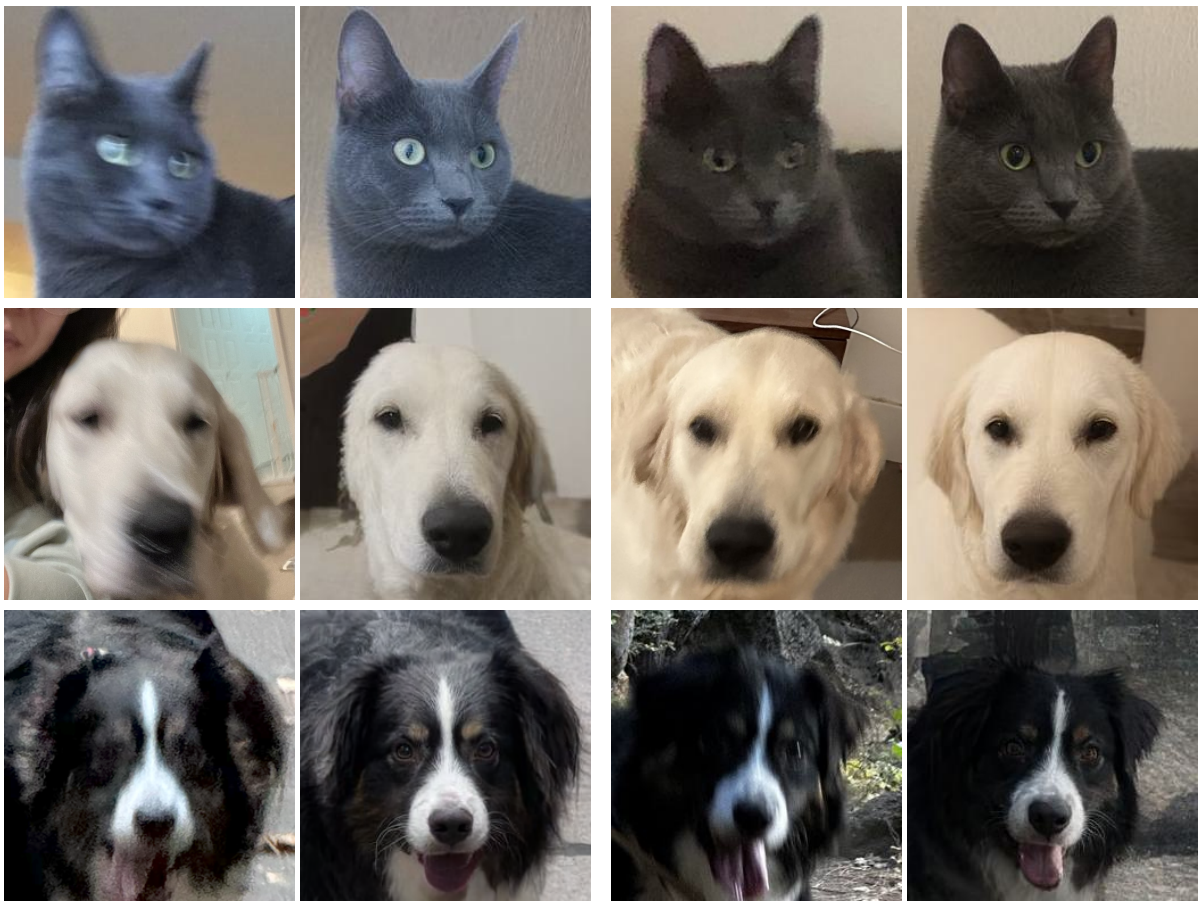


Figure 5.6. Results on real-world cat/dog restoration. Our method easily extends to other categories with corresponding pre-trained diffusion models. We show results on cats and dogs where we can reconstruct high-frequency details while preserving the identity.

trajectory, resulting in lower-quality sampled output. Conversely, while a larger K enhances sample quality as hypothesized, it may also produce outputs less faithful to the input.

Constraining Prior with Generative Album. In the same Figure 5.7, we illustrate the significance of prior constraining and the effectiveness of using a generative album. As shown, a generative space that is too diverse increases the difficulty of sampling high-quality images from a given input, especially when K is small. Conversely, for large K values, the sampled image can deviate significantly from the input. Constraining the generative space with an album close to the input ensures preservation of input information in the output for large K, while still allowing high-quality sampling from small K. Ablation on Skip Guidance is included in the supplementary.

Constraining Prior with Personal Album. When a personal album is available, we directly constrain the generative space with this album. This not only improves output quality and faithfulness, as with the generative album, but also aids in recovering information absent in the input. As demonstrated in Figure 5.8, compared to an unconstrained model (i.e., the pre-trained diffusion model), the personalized model produces higher-quality images that better preserve identity.

5.6 Conclusion

We propose a method for image restoration that involves simply adding noise to a degraded input and then denoising it with a diffusion model. The key to our approach is constraining the generative space with a set of anchor images. We demonstrate in single-image restoration tasks that this method yields high-quality restoration results, surpassing previous supervised approaches. Furthermore, we show that constraining the generative space with a personal album leads to a personalized restoration-by-generation model that is effective for any image of the same subject, producing results with high quality and faithful details.

Limitations and Future Work. Unlike the personalization case, for single-image

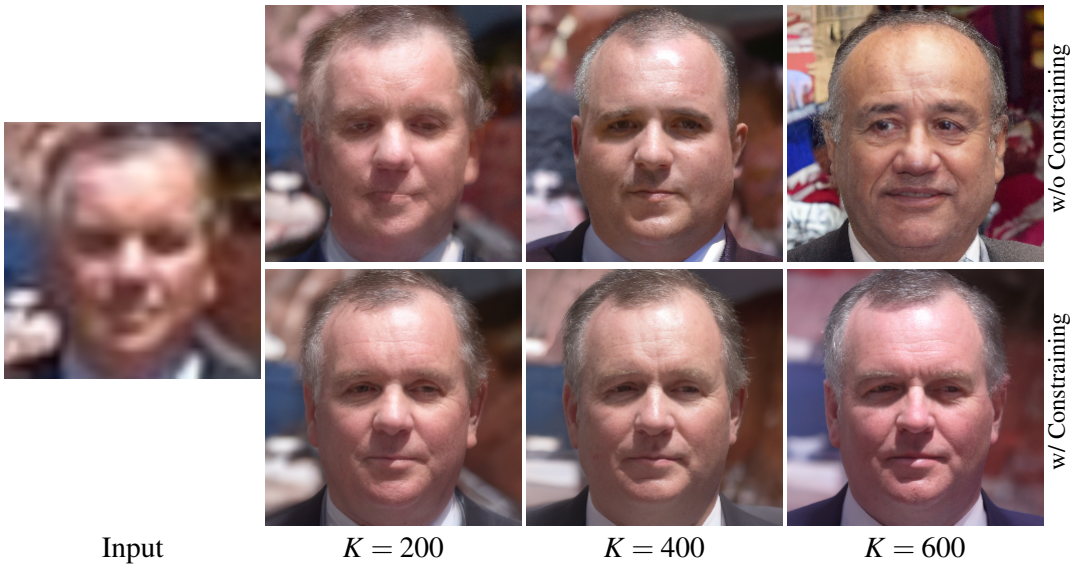


Figure 5.7. Ablation on Noise Step K and Constraining with Generative Album. As K increases, quality of images sampled from y_K improves, but alignment with the input reduces. Fine-tuning with a generative album notably enhances both image quality and input fidelity.

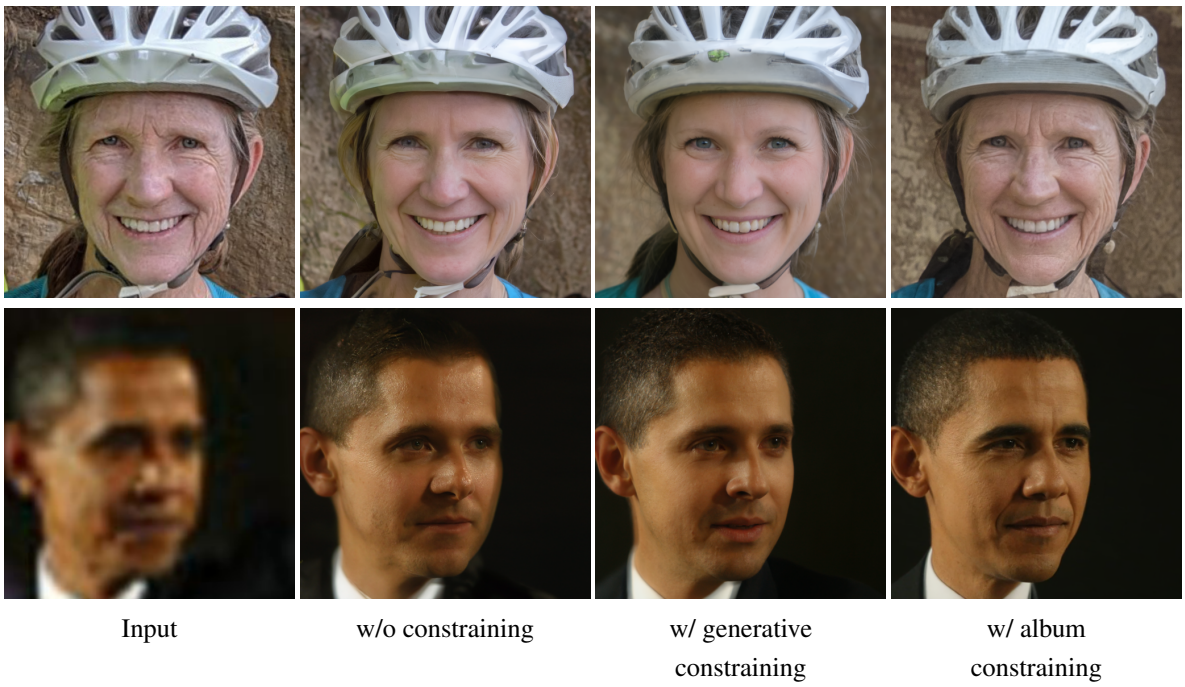


Figure 5.8. Constraining with personal album. Personalized model produces higher-quality images that better preserve identity compared to the model without constraining.

restoration, our approach requires fine-tuning for each input image. This is relatively slow compared to feed-forward approaches. Investigating methods to constrain the generative space without fine-tuning could be interesting. Furthermore, we have primarily validated our approach on class-specific image restoration tasks, largely due to the absence of a high-quality pre-trained diffusion model for natural images. Exploring whether our approach remains effective within a more diverse generative space would be intriguing. Such exploration could potentially address the challenge of blind restoration for general images.

5.7 Acknowledgments

We thank Marc Levoy for providing valuable feedback, and everyone whose photos appear in the paper, including our furry friends, Chuchu, Nobi and Panghu.

This chapter, in full, is a reprint of the material as it appears in "Restoration by Generation with Constrained Priors". Ding, Zheng; Zhang, Cecilia; Tu, Zhuowen; Xia, Zhihao, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Conclusion

This dissertation has presented a series of novel approaches that effectively address both controllability and efficiency challenges in visual generation as well as understanding. We began with the development of MaskCLIP, which harnesses the power of pretrained CLIP models for open-vocabulary universal image segmentation. By rethinking the fusion of semantic segmentation tasks with dense vision-language features, MaskCLIP establishes a flexible baseline for segmentation while significantly reducing the need for resource-intensive training procedures. This work not only advances open-world segmentation but also lays a foundation for further exploration into the integration of vision-language paradigms with segmentation tasks.

Building on the importance of controllability, we introduced DiffusionRig for personalized facial appearance editing. By leveraging a two-stage training methodology—initially capturing generic facial priors and subsequently refining these with a limited set of person-specific images—DiffusionRig achieves photorealistic face editing that robustly preserves identity and critical high-frequency details. This approach elegantly addresses the inherent trade-off between generalization and specificity, suggesting promising avenues for future personalized editing applications in both entertainment and professional settings.

Further emphasizing efficiency in generative modeling, Patch-DM was proposed as a novel patch-based denoising diffusion model that generates high-resolution imagery without succumbing to the common pitfalls of boundary artifacts. The introduction of a feature collage

strategy that enables smooth transitions between patches demonstrates that significant reductions in computational complexity are possible without compromising image quality. This result contributes an important perspective on designing lightweight architectures capable of supporting high-resolution generation, a critical requirement for real-time and resource-constrained applications.

Lastly, we tackled the challenge of personalized image restoration, formulating a method that repurposes pretrained diffusion models to restore degraded images while preserving essential characteristics. By strategically adding noise and constraining the generation process with anchor images or a "generative album," our approach effectively adapts to the nuances of the degradation process. This demonstrates the versatility of diffusion models beyond standard generative tasks and opens up opportunities for their application in robust image restoration.

In summary, the methods presented in this dissertation not only push the boundaries of what is achievable in controllable image generation and segmentation but also provide efficient computational strategies that are essential for practical deployment. Through MaskCLIP, DiffusionRig, Patch-DM, and Gen2Res framework, this dissertation has set forth a comprehensive research agenda that addresses critical challenges in visual generation. We believe that these contributions will inspire continued innovation in creating more accessible, controllable, and efficient computer vision systems, ultimately bridging the gap between algorithmic creativity and practical, human-centric applications.

Bibliography

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[BB15] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015.

[BBK+24] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain mri. In *Medical Imaging with Deep Learning*, pages 1019–1032. PMLR, 2024.

[BCS+23] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 843–852, June 2023.

[BLS+21] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (ToG)*, 40(4):1–15, 2021.

[BV99] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[CGS+22] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 170–188. Springer, 2022.

[CHIS23] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[CJH+21] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko.

Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021.

[CKJ+21] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[CKM+22] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[CKR+22] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. *arXiv preprint arXiv:2204.05626*, 2022.

[CLC+22] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[CLS+22] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022.

[CMK+21] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[CMS+22] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

[CSK21] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[CUYH20] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[DGXZ19] 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.

[DHT+00] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000.

[DKD17] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[DN21a] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[DN21b] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[DXXD22] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

[DZX+23] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023.

[EL99] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1033–1038, 1999.

[EST+20] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future. *arXiv:1909.01815 [cs]*, Apr 2020. arXiv: 1909.01815.

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn,

and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[FFBB21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[GAA+22] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[GAA+23] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

[GDG19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019.

[GGCL22] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.

[GGU+20] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020.

[GLKC22] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[GPAM+20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[GPL+22] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022.

[GWX+22] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and

Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022.

[GYZ+19] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[HJA20a] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[HJA20b] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, page 6840–6851. Curran Associates, Inc., 2020.

[HJA20c] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[HKL+22] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *CVPR*, 2022.

[HMBL21] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. *arXiv:2104.07659 [cs]*, Apr 2021. arXiv: 2104.07659.

[HPX+22] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[HRU+17a] 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.

[HRU+17b] 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.

[HS18] Ryota Hinami and Shin'ichi Satoh. Discriminative learning of open-vocabulary

object retrieval and localization by negative phrase augmentation. In *EMNLP*, 2018.

[HS22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JYX+21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[JZC+23] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.

[KAL+21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

[KALL18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[Kar19] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KBM+18] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[KEES22] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[KGHD19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature

pyramid networks. In *CVPR*, 2019.

[KHG+19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[Kin09] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[KLA19a] 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.

[KLA19b] 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.

[KLA+20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[KMWW19] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019.

[KS21] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021.

[KSL+21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *CVPR*, pages 1780–1790, 2021.

[KSS+22] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pages 11201–11228. PMLR, 2022.

[KVE21a] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

[KVE21b] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 1866–1875, 2021.

[KW14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[KWW+21] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[KZZ+23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[LBB+17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[LCC+19] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: generation by parts via conditional coordinating. In *ICCV*, pages 4512–4521, 2019.

[LCL+22] Chieh Hubert Lin, Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-pixel image synthesis. In *ICLR*, 2022.

[LCS+21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[LCZ+19] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019.

[LCZ+20] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European conference on computer vision*, pages 399–415. Springer, 2020.

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[LJZ+21] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[LL22] Troy Luhman and Eric Luhman. Improving diffusion model efficiency through patching. *arXiv preprint arXiv:2207.04316*, 2022.

[LLC+21] 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*, 2021.

[LLR+20] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020.

[LLST20] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. In *CVPR*, 2020.

[LLWT15] 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.

[LLY+18] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[LMH+21] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

[LSC+22] Wei-Sheng Lai, Yichang Shih, Lun-Cheng Chu, Xiaotong Wu, Sung-Fang Tsai, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Face deblurring using dual camera fusion on mobile phones. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

[LSL+22] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 3d-fm gan: Towards 3d-controllable face manipulation. In *European Conference on Computer Vision*, pages 107–125. Springer, 2022.

[LSSS18] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4):1–13, Jul 2018. arXiv: 1808.00362.

[LTH+17] 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.

[LWB+22] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

[LZZ+22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, June 2022.

[MCL+14] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[MHP+19] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[NAH+22] 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. *arXiv preprint arXiv:2203.17272*, 2022.

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[NLML20] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[NMDB21] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

[ÖL23] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[PCWS22a] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[PCWS22b] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. *arXiv:2111.15640 [cs]*, Mar 2022. arXiv: 2111.15640.

[PWS+21] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, pages 2085–2094, 2021.

[PZD+21] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021.

[QKW+22] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[RBS+22] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[REM17] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

[RKH+21a] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[RKH+21b] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[RLJ+23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[RMC16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[ROF92] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[RTD+21] Mallikarjun B. R, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, and Christian Theobalt. Photoapp: Photorealistic appearance editing of head portraits. *arXiv:2103.07658 [cs]*, Mar 2021. arXiv: 2103.07658.

[RZC+22] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.

[SBFB19] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[SBT+19] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019.

[SCC+22] 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.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models

with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

[SE20] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[SHC+22] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[SKCJ18] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. *arXiv:1712.01261 [cs]*, Apr 2018. arXiv: 1712.01261.

[SLT+22] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *ICLR*, 2022.

[SME20a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[SME20b] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[SSE21] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14144–14153, 2021.

[SXLJ23] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

[SZA+19] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 2387–2397, 2019.

[SZY+23] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32483–32498. PMLR, 23–29 Jul 2023.

[TCYZ05] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005.

[TEB+20a] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

[TEB+20b] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[TFM+22] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[TPF+22] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022.

[Tu07] Zhuowen Tu. Learning generative models via discriminative approaches. In *CVPR*, 2007.

[Tu08] Zhuowen Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008.

[TXL+20] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, and Hong Liu. Attention-guided cnn for image denoising. *Neural Networks*, 124:117–129, 2020.

[Wal23] Wallpaperscraft. Wallpaperscraft. https://wallpaperscraft.com/, 2013-2023.

[WCB+22] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[WCC+22] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022.

[WCY+22] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022.

[WDT+22] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022.

[WJZ+23] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

[WLZS21] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021.

[WOT20] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[WQS+06] Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. In *CVPR*, pages 347–354, 2006.

[WXDS21] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[WYZ22] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference*

on Learning Representations, 2022.

[WZA+22] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022.

[WZC+22] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022.

[WZZ+23] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023.

[XC20] Zhihao Xia and Ayan Chakrabarti. Identifying recurring patterns with deep neural networks for natural image denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2426–2434, 2020.

[XLZ+19] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019.

[XYF+23] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[XZH+21] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.

[XZW+22] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022.

[YFHP18] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[YRXZ21] 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.

[ZAK+22] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[ZBT+20] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020.

[ZCLL22] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

[ZGSR21] Mona Zehni, Shaona Ghosh, Krishna Sridhar, and Sethu Raman. Joint learning of portrait intrinsic decomposition and relighting. *arXiv:2106.15305 [cs]*, Jun 2021. arXiv: 2106.15305.

[ZHS+23] Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, and Matthias Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7312–7322, 2023.

[ZLD22] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022.

[ZLL+19] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[ZLZ+21] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

[ZLZ+22] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv:2203.11876*, 2022.

[ZMG+22] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European Conference on Computer Vision*, pages 18–35. Springer, 2022.

[ZRHC21] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-

vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021.

[ZYZ+22] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, June 2022.

[ZZB+18] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[ZZC+17] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[ZZGZ17] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017.

[ZZP+17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[ZZP+19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[ZZZ18] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.