This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SODA: Bottleneck Diffusion Models for Representation Learning

Drew A. Hudson* Daniel Zoran Mateusz Malinowski Andrew K. Lampinen
Andrew Jaegle James L. McClelland Loic Matthey Felix Hill Alexander Lerchner
Google DeepMind

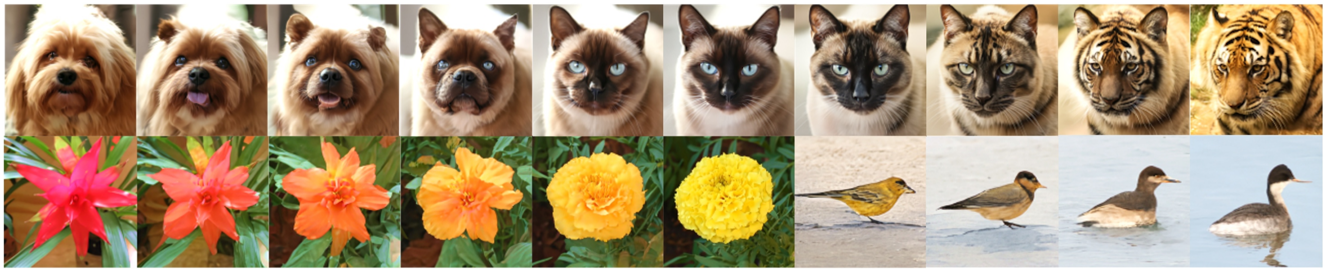

Figure 1. **Latent Interpolations.** SODA learns to encode images into compact latent representations. By traversing its latent space, we can interpolate between images, morphing from one image category to another and smoothly transitioning between semantic attributes.

Abstract

We introduce SODA, a self-supervised diffusion model, designed for representation learning. The model incorporates an image encoder, which distills a source view into a compact representation, that, in turn, guides the generation of related novel views. We show that by imposing a tight bottleneck between the encoder and a denoising decoder, and leveraging novel view synthesis as a self-supervised objective, we can turn diffusion models into strong representation learners, capable of capturing visual semantics in an unsupervised manner. To the best of our knowledge, SODA is the first diffusion model to succeed at ImageNet linear-probe classification, and, at the same time, it accomplishes reconstruction, editing and synthesis tasks across a wide range of datasets. Further investigation reveals the disentangled nature of its emergent latent space, that serves as an effective interface to control and manipulate the produced images. All in all, we aim to shed light on the exciting and promising potential of diffusion models, not only for image generation, but also for learning rich and robust representations. See our website at soda-diffusion.github.io.

1. Introduction

> What I cannot create, I do not understand.
>
> – *Richard P. Feynman*

Synthesis, the ability to create, is considered among the highest manifestations of learning [1, 2]. As opposed to passive analysis of a text or an image, conceiving them out of thin air involves profound understanding of the underlying factors and intricate generative processes that give rise to the final product [3]. Indeed, learning to write in a new language is often more challenging than reading it. Figuring out the solution to a math problem is fundamentally harder than verifying it [4]. And just as the chef learns more about the culinary arts than the diner to prepare a tasty meal, and the novelist knows more about narrative structures than the reader to tell a good story, the artist better grasps perspective and composition to craft a breathtaking masterpiece.

Analogously, in AI, the recent years have witnessed remarkable progress at the generative domain, with large-scale diffusion modeling proving to be a powerful and flexible technique that can create vivid imagery of astonishing realism and incredible detail. And yet, while the vast majority of research harnesses these models for the straightforward goal of synthesis or editing alone [5–17], only little attention has been given to their representational capacity [18–20], leaving this promising direction rather unexplored. Surely, models that can weave from scratch such rich depictions of high fidelity, likely learn much along the way about the underlying properties, processes, and components that make up the resulting pictures. How then can we leverage this untapped potential of diffusion models for the purpose of representation learning, and extract the knowledge they acquire for the benefit of downstream tasks?

Motivated to achieve this aim, we present SODA, a self-supervised diffusion model, designed for both perception and synthesis. It couples an image encoder with the classic diffusion decoder [5], both trained in tandem for novel view generation [21] – a task we choose to employ here, not only for its own sake, but as a self-supervised objective. The encoder transforms an input view into a concise latent representation, which then guides the denoising of an output view, by modulating the decoder's activations.

This setup introduces a desirable information bottleneck

[*] Main contributor. The paper presents the findings of Drew's internship project at Google DeepMind, hosted by Felix Hill and Alexander Lerchner. Contact: dorarad@google.com; lerchner@google.com.

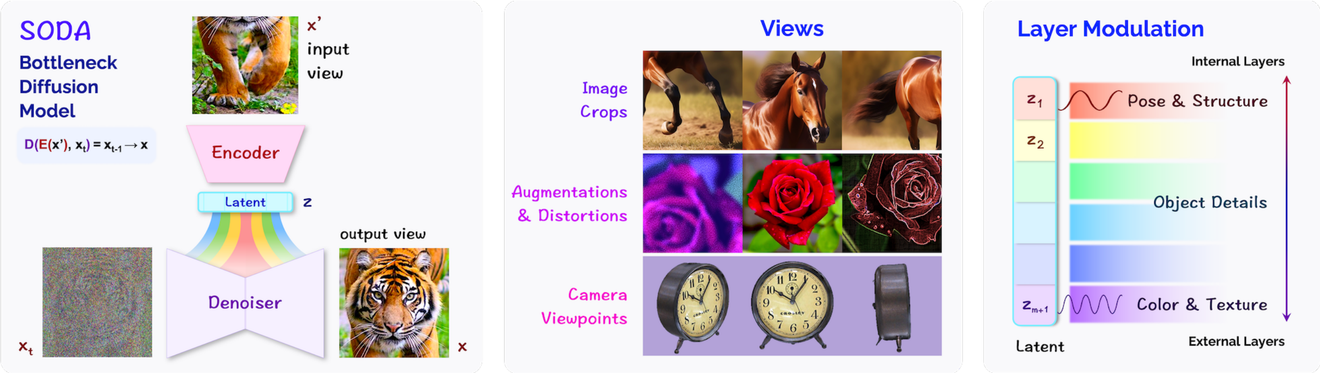

Figure 2. **(Left) Model Overview.** SODA consists of two modules: an *encoder* and a *denoising decoder*. The encoder transforms a source view into a *compact latent representation*, which guides the denoising of a target view, by having a dedicated latent subvector modulating each of the decoder's layers **(Right, Layer Modulation)**. The model is useful both for representation learning, by applying the latents to downstream perception tasks, as well as for synthesis, by iteratively generating novel views. **(Center) Views.** We consider views as any set of images that relate visually or semantically: They can be crops, augmentations, or images captured from different camera perspectives.

between the encoder and the decoder [22], that in contrast to the typical diffusion framework, equips our model with an explicit and interpretable visual latent space. As our experiments confirm, its advantages are twofold: it both encourages the emergence of disentangled and informative **representations** that capture image key properties and semantics, which thus can be applied to downstream tasks, and further provides effective means to control and manipulate the produced outputs, for the gain of image editing and **synthesis**. We further devise and integrate multiple new ideas into the network architecture and training procedure: layer modulation, modified classifier-free guidance, and an inverted noise schedule, so to maximize its representation skills.

We demonstrate our model's strengths and versatility by evaluating it along a series of classification, reconstruction and synthesis tasks, spanning an extensive collection of datasets that covers both the simulated and real-world kinds. SODA possesses strong representation skills, attaining high performance in linear-probing experiments over the ImageNet dataset among others. Moreover, it excels at the task of few-shot novel view generation, and can flexibly synthesize images either conditionally or unconditionally, as indicated by metrics of fidelity, consistency and diversity. Finally, we inspect the model's emergent latent space and discover its disentangled nature, which offers controllability over the semantic traits of the images it produces, as validated both qualitatively and quantitatively.

Overall, SODA integrates together three research ideas that we seek to establish and promote: First, diffusion models are not only adept at image generation, but are also capable of learning strong representations. Second, novel view synthesis can serve as a powerful self-supervised objective for model pre-training. And third, the compactness of the latent space, which could be reached by constricting the bottleneck between the encoder and the denoiser, plays a pivotal role in enhancing the latent representations' quality, informativeness and interpretability.

2. Related Work

Diffusion. The advent of diffusion models has lately marked a breakthrough in the field of visual synthesis. Originally inspired by theories of thermodynamics [23], it approaches generative modeling by following a reversible and iterative denoising process, the forward direction of which slowly erodes the structure within the data distribution, while the backward direction is gradually restoring it. Since its early inception back in 2015 [24], tremendous strides have been made in the quality and diversity of the created outputs, thanks to innovations of the framework's training and sampling techniques [25–28]. Consequently, diffusion models have been widely adopted for numerous tasks and modalities [29–33], synthesizing images, videos, audio and text [34–38], and even advancing planning [39] and drug discovery [40], effectively becoming one of the leading paradigms for generative modeling nowadays.

But while most literature highlights its generative feats, only a handful of works have studied diffusion modeling's representational capacity, mainly repurposing pre-trained text-to-image models for classification [41], segmentation [42], or multimodal reasoning [43]. The reliance on such models makes it unclear whether the downstream capabilities arise from the diffusion approach itself, or are actually attributable to the exceptionally large scales, long training and voluminous captioned data, which, essentially, provides rich and textual semantic supervision. To address this shortcoming, we focus here instead on the fully-unsupervised regime, and train our model from scratch on standardized benchmarks, seeking to asses the value and potential of diffusion-based representations derived from images alone.

Visual Encoding. Closer to our work is DRL [20], that extends early research on denoising auto-encoders [44–47], and conditions a denoiser on an encoded clean version of its own target. It is mainly explored from a theoretical perspective, along with preliminary results on MNIST and CIFAR-10. DiffAE [48] follows up, integrating style modulation into the encoder [49–51], while InfoDiffusion [52] regularizes it with mutual-information loss. Our approach builds upon this line of research, but instead of auto encoding the same image, we generate novel views. Notably, we discover that this, in turn, remarkably enhances the model's representation skills, as evidenced by substantial gains in downstream performance. We further couple this idea with multiple technical innovations, pertaining both architecture and optimization, geared to realize the representational capabilities of diffusion models to their fullest. And in contrast to prior works, we provide an extensive empirical study of diffusion-based representation learning, encompassing a broad suite of datasets over multiple different tasks.

Hybrid Models. A couple of partially related works are unCLIP [8] and Latent Diffusion Models [10], both of which utilize a frozen pre-trained encoder (CLIP and VQ-GAN respectively) to cast images onto a compressed latent space over which a diffusion model can operate. Consequently, we note that the latent representations used in both these approaches are in fact not derived by diffusion itself, but rather through either contrastive or adversarial pre-training. As such, they differ fundamentally from our study, which aims to explore the effectiveness of diffusion-based pre-training as a means for representation learning.

Downstream Tasks. For each of the tasks we explore – classification, disentanglement, reconstruction, and novel view synthesis – we compare SODA to the leading prior works. These include models such as SimCLR, DINO, and MAE for linear-probe classification [53–58], NeRF-based approaches for novel view generation [59, 60], and classic variational models for the task of disentanglement [61–63]. Whereas these techniques are designed for particular objectives or depend on domain-specific assumptions, SODA exhibits a greater degree of versatility, as it tackles representational and generative tasks alike.

3. Approach

SODA is a self-supervised diffusion model that learns a bidirectional mapping between images and latents. It consists of an image encoder $\mathcal{E}(\boldsymbol{x}') = \boldsymbol{z}$ that casts an *input view* $\boldsymbol{x}'$ into a low-dimensional latent $\boldsymbol{z}$, which is then used to guide the synthesis of a novel *output view* $\boldsymbol{x}$, that relates to the input $\boldsymbol{x}'$ (Figure 2). Concretely, $\boldsymbol{x}$ is produced through a diffusion process that is conditioned on the encoding $\boldsymbol{z}$ via *feature modulation* [49]. This design equips SODA with an explicit and compact latent space, which not only offers ample control over the generative process, but can also be leveraged for downstream perception tasks (Section 4)[1].

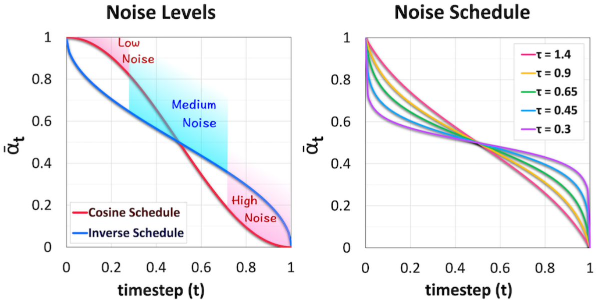

Figure 3. **Inverted Noise Schedule** used for SODA's training. It prioritizes medium noise levels, which aid representation learning.

We first present an overview of the model (**Section 3.1, Figure 2**), followed by an in-depth discussion of each of its core components: the encoder's architectural design (**Section 3.2**), the mechanisms involved in the synthesis of novel views (**Section 3.3**), and the optimization techniques we develop to cultivate strong and meaningful representations (**Section 3.4**).

3.1. Model Overview

As a denoising diffusion model [5, 25], SODA is formally defined by a pair of forward and backward Markov chains, iteratively transforming a sample $\boldsymbol{x}_T$ from the normal distribution into the target one ($\boldsymbol{x}_0$) and vice versa. Each forward step t erodes $\boldsymbol{x}_t$ by adding low Gaussian noise $\boldsymbol{\epsilon}_t$ according to a fixed variance schedule α_t. Meanwhile, the respective backward step performs image denoising, and aims to estimate $\boldsymbol{\epsilon}_t$ in order to recover $\boldsymbol{x}_{t-1}$ from its successor $\boldsymbol{x}_t$. It is carried out by a decoder $\mathcal{D}$, implemented as a convolutional UNet [64] (with $2m + 1$ activation layers h_i).

To tackle the denoising challenge, we assist the decoder by conditioning it on a latent vector $\boldsymbol{z}$, which guides its operation by modulating the activations $\boldsymbol{h}$ at each of its layers through adaptive group normalization that controls their scale and bias: $\boldsymbol{z}_s\text{GroupNorm}(\boldsymbol{h}) + \boldsymbol{z}_b$ where $(\boldsymbol{z}_s, \boldsymbol{z}_b)$ are linear projections of $\boldsymbol{z}$, applied evenly across the activation grid. See Appendix B for the architectural details of the decoder, as well as closed-form equations of the diffusion forward and backward steps. The latent $\boldsymbol{z}$ is created by the newly introduced image encoder $\mathcal{E}$, discussed next.

3.2. The Encoder

The latent representation $\boldsymbol{z}$ is at the model's core, serving as the communication channel between the encoder $\mathcal{E}$ and the denoising decoder $\mathcal{D}$, while guiding the latter through the diffusion process. It is derived by a ResNet encoder $\mathcal{E}(\boldsymbol{x}') = \boldsymbol{z}$, representing a clean source view $\boldsymbol{x}'$ that semantically or visually relates to the target view $\boldsymbol{x}$ (Section 3.3).

The driving motivation behind this idea stems from the denoising task's inherent *under-determination*: Seeking to fulfill it to the best of its ability, the decoder will leverage

[1] Our model is named after the soda drink. Indeed, the fizzing in soda bottles is an everyday example of the diffusion phenomena.

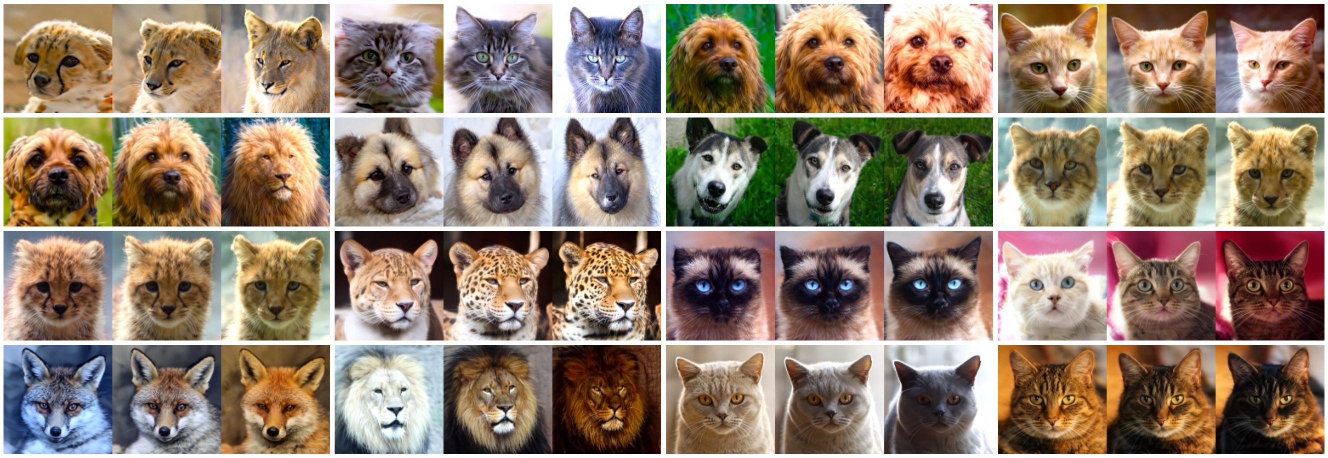

Figure 4. **Latent Controllability (Unsupervised).** We explore SODA's latent space and discover directions that correspond to semantic attributes like face size and structure, lighting and viewpoint, maturity, expressions, fur length and color, texture, brightness and tone. SODA is trained fully unsupervised over images only. It successfully disentangles and encodes meaningful aspects into its learned representations.

any pertinent knowledge or useful clue that could inform it of the missing content to fill in. This in turn incentivizes the encoder to distill into $\boldsymbol{z}$ the most striking and prominent commonalities between the source and target views. Since we further constrict the encoding $\boldsymbol{z}$ into a **low-dimensional space**, and use it to guide the decoder via **global feature modulation**, it yields a fruitful combination that encourages $\boldsymbol{z}$ to specifically capture the image's *high-level semantics*, while delegating the reconstruction of *localized and high-frequency details* to the denoiser itself.

From that perspective, learning a latent $\boldsymbol{z}$ that supports image **denoising**, rather than pure reconstruction from scratch, as in most auto-encoders [65–67], liberates our encoder from the need to compress *all* information about the image into the representation, and let it instead focus on the image's most distinctive and descriptive qualities.

3.2.1 Layer Modulation & Masking

To enhance the model's latent space disentanglement, we introduce two intertwined mechanisms of modulation and masking. In **layer modulation**, we partition the latent vector $\boldsymbol{z}$ into $m+1$ sections – half the number of layers in the decoder $\mathcal{D}$. We use each $\boldsymbol{z}_i$ to modulate the respective pair or layers (h_i, h_{2m-i}), thereby promoting specialization among the latent sub-vectors. Just like light rays refracted through a prism, they are encouraged to capture visual traits at different levels of granularity, from coarser to finer, so to guide the decoder's operation through the layers.

To improve the localization and reduce the correlations among the m sub-vectors, we present **layer masking** – a layer-wise generalization of classifier-free guidance [27]. During training, we zero out a random susbset of $\boldsymbol{z}_{1:m}$, effectively performing layer-wise guidance dropout, that mitigates the decoder's reliance on sub-vector dependencies, allowing them to decouple and specialize independently. At sampling, we extrapolate the model's output in the conditional direction: $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t|\boldsymbol{z}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t|\boldsymbol{0})$. This endows SODA with finer control over the generative process, and opens the door for image editing and style mixing [49], as we can selectively condition the decoder on some levels of granularity, like structural or positional aspects, while giving it free rein to unconditionally vary other ones, such as lighting, texture, or color palette (see supplementary figures).

3.3. Novel View Generation

We loosely consider *views* to be any set of images that hold some relation among each other, such as visual or semantic (Figure 2): they can be various augmentations or distortions of an original image, as is commonly explored in the contrastive learning literature [58], they can show a 3D object from different poses and perspectives [21], or they can simply share the same semantic category with one another.

Cardinality. We permit the trivial singular case where all views are identical, which then turns the model into an auto-encoder. Conversely, we can extend the conditioning and create a novel view based on a *set of k input views*, instead of just a single one. We map each input x^i to its latent with a shared encoder $\mathcal{E}$, and aggregate the resulting latents $\boldsymbol{z}^{1:k}$ into a single vector $\boldsymbol{z}$, either by taking their mean, or by processing them through a shallow transformer.

Perspective. The model can incorporate richer forms of conditional information, such as the camera perspective associated with each view: Specifically, for experiments over 3D datasets like ShapeNet (Section 4.2.2), we concatenate a grid of ray positions and directions $\boldsymbol{r} = (\boldsymbol{o}, \boldsymbol{d})$, embedded with sinusoidal positional encoding [68], to the linearly-mapped RGB channels of the source and denoised views, $\boldsymbol{x'}$ and $\boldsymbol{x}_t$. This allows us to conditionally generate novel views that match the requested pose and orientation. See supplementary for illustrations and implementation details.

Guidance. We extend classifier-free guidance for novel view synthesis, and instead of masking just the latent $\boldsymbol{z}$, we randomly and independently mask either the latent or the pose information $\boldsymbol{r}$. As our empirical findings suggest, this idea not only hones the model's generative skills, but further enables conditioning on partial information, allowing SODA to either unconditionally conceive *novel objects* at a requested pose, or alternatively, generate arbitrary novel views of given objects at the absence of source or target positional information, based on an image only (Appendix G).

Cross Attention. The technique of layer modulation offers the encoding $\boldsymbol{z}$ with *global control* over each of the decoder's layers, by evenly setting their scales and biases across the grid. We further study alternative mechanisms, and explore the integration of cross attention, so to support *spatial modulation*. Instead of layer-wise modulation (which we use by default), we partition the latent $\boldsymbol{z}$ into n sub-vectors, and perform cross attention between the sub-vectors set and the decoders' activations, akin to the word-based attention common in text-to-image generation. We find that cross attention aids the model at 3D novel view synthesis, while layer modulation performs better for image editing, reconstruction, and representation learning.

3.4. Training & Sampling

Noise Schedule. We train the model with the standard MSE objective [5], but introduce a new noise schedule to better fit the representation learning task: Indeed, diffusion models commonly set the variance of the additive noise term ϵ to follow either a cosine [69], sigmoid [70] or linear [5] decay schedules, prioritizing noise levels that are close to the margins, either of the high or low ends (Figure 3). Those schedules have been found useful for image synthesis.

However, from a representation learning perspective, denoising images with overly high or low noise levels fails to provide effective training signal for the model to learn from: Too little noise does not present the denoiser with a challenging enough task, thereby diminishing the encoder's necessity. Meanwhile, too heavy noise puts excess pressure on the latent $\boldsymbol{z}$ to fully capture every pixel-level detail of the image, turning denoising into mere reconstruction. We thus incorporate a new inverted noise schedule, that promotes medium noise levels in lieu of the extremes, which proves highly conducive to representation quality (Section 4.1).

Additional Settings. Two modifications we find beneficial for representation learning are: (1) adding low Gaussian noise to the *encoder's input images*; (2) optionally setting the encoder to have a higher learning rate than the decoder, so to positively impact their learning dynamics by allowing the encoder to adapt faster as it guides the decoder in the denoising task. While the model is robust to the selection of the learning-rate ratio, tuning it could improve downstream results. Once trained, we use DDPM [5] for sampling.

Table 1. **Linear-Probe Classification on ImageNet**, evaluating discriminative and generative approaches of comparable model sizes. *Top1 (Crop+Flip)* presents scores with light augmentations only, revealing the greater robustness of generative approaches in general and SODA in particular to data augmentations. ($*$) denotes cropping only, and (†) denotes no augmentation. ($\star$) indicates that the decoder is not used for the downstream linear probing.

Method	Arch.	#θ	Top1	Top5	Top1 Crop+Flip
Discriminative Approaches					
Supervised [71]	RN50×2	94	79.9	95.0	-
SimCLR [58]	RN50×2	94	74.2	92.0	46.7
BYOL [72]	RN50×2	94	77.4	**93.6**	63.8
SwAV [57]	RN50×2	94	73.5	-	54.2
DINO [56]	ViT-B/16	86	74.9	-	61.1
SwAV + multi-crop	RN50×2	94	77.3	-	58.7
DINO + multi-crop	ViT-B/16	86	**78.2**	-	**65.3**
Generative Approaches					
Vanilla EncDec	RN50×2	118	8.5	17.6	10.2
Vanilla AutoEnc	RN50×2	118	14.3	28.9	-
iGPT [55]	GPT-2	76	41.9	-	41.9^{*}
iGPT-L [55]	GPT-2	1386	65.2	-	65.2^{*}
BEiT [54]	ViT-B/16	86	56.7	-	-
BigBiGAN [73]	RV50×4	86	61.3	81.9	$61.3^{\dagger}$
MAE [53]	ViT-B/16	86	68.0	-	68.0
Diffusion-based Approaches (Generative)					
Palette [30]	UNet	118	11.4	22.3	8.7
Unconditional Diffusion	UNet	118	24.5	44.4	28.3
SODA w/o bottleneck	RN50×2	$94{+}35^{\star}$	34.2	52.9	29.7
SODA w/o modulation	RN50×2	$94{+}32^{\star}$	56.7	78.5	51.1
SODA w/o novel views	RN50×2	$94{+}36^{\star}$	55.1	75.5	48.2
SODA w/o noise sched.	RN50×2	$94{+}36^{\star}$	62.0	82.9	56.8
SODA (*ours*)	RN50×2	$94{+}36^{\star}$	**72.2**	**90.5**	**69.1**

Table 2. **Image Reconstruction on ImageNet,** comparing discrete [74, 75] and continuous [8] approaches. The suffix of discrete methods refers to their codebook cardinality.

Method	Latent Dim	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
DALL-E2	1024	9.0	0.11	16.53	0.66
DALL-E,8K	512×32×32	22.8	0.73	32.01	1.95
VQGAN,1K	256×16×16	19.4	0.50	7.94	1.98
VQGAN,16K	256×16×16	19.9	0.51	4.98	1.83
VQGAN,8K	256×32×32	22.2	0.65	**1.49**	1.17
SODA (*ours*)	2048	**23.6**	**0.93**	2.77	**0.19**

4. Experiments

We evaluate SODA through a suite of quantitative and qualitative experiments, demonstrating its strong representation skills and generative capabilities over 12 different datasets grouped into 4 tasks: We begin with linear-probe classification (**Section 4.1**), showing the model's utility for downstream perception. We proceed to image reconstruction and few-shot novel view synthesis (**Section 4.2**), illustrating its ability to envision 3D objects from new unseen perspectives. We then explore the model's disentanglement & controllability (**Section 4.3**), as substantiated by comparative analysis and latent-space interpolations.

In the supplementary and our website website (*soda-diffusion.github.io*), we provide additional samples, visualizations and animations, and provide further details about our evaluation procedures, discussing the datasets, metrics, baselines, and implementation details (**Appendices C-F**). We conclude with ablation studies (**Appendix G**), that em-

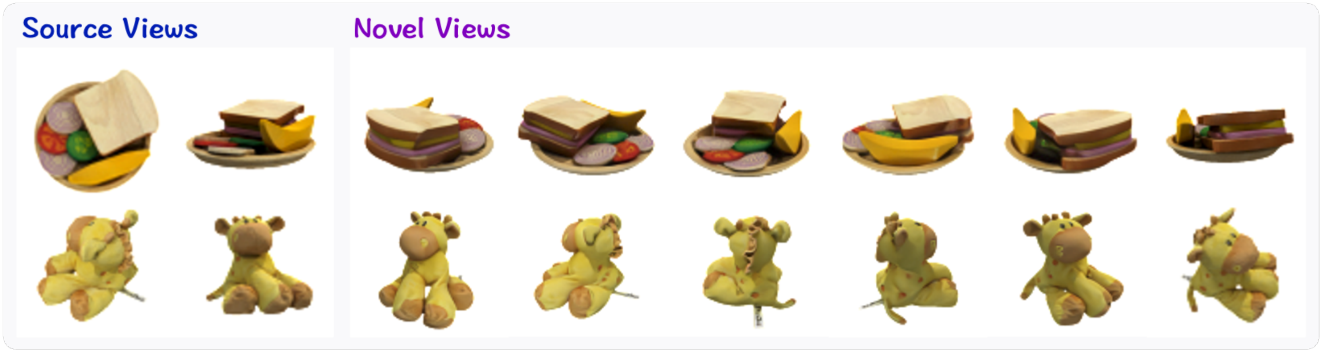

Figure 5. **Novel View Synthesis**, where given two source views of GSO objects, SODA generates their images from new perspectives.

pirically validate the contribution of the model's components and design choices. Taken altogether, the evaluation offers solid evidence for the efficacy, robustness and versatility of our approach.

4.1. Linear-Probe Classification

We assess the quality of SODA's learned representations through linear-probe analysis [72, 76–78], over ImageNet and CelebA, which complement one another: the former calls for fine-grained clustering into 1000 possible *categories*, while the latter involves rich identification of diverse semantic *attributes*. We train our model in a self-supervised fashion: using RandAugment [79] for ImageNet, and Gaussian data augmentation for CelebA. We then fit a linear classifier over the latent vectors z that predicts the category or attributes, and measure the resulting performance. We note that training diffusion models for representation learning is computationally efficient, since iterative sampling is necessary for generative purposes only.

As shown in Table 1, SODA reaches **72.24%** accuracy (**top1**) on the ImageNet1K linear-probe classification task, outshining competing generative approaches such as MAE, BEIT and iGPT [53–55], and significantly reducing the gap with discriminative and constrastive approaches like DINO, SwAV, and SimCLR [56–58]. Meanwhile, for CelebA, our model attains the strongest results (72.7% F1) compared to competing approaches (Supp Table 6), eclipsing even the language-supervised CLIP embeddings (71.1% F1).

SODA proves remarkably **robust to the choice of data augmentation**, as it performs strongly regardless of the selected strategy, seeing only a minor decrease of 3.1% when switching the heavier RandAugment for a lighter **crop+flip** augmentation. This stands in stark contrast to the high sensitivity of contrastive methods to data augmentations, with e.g. BYOL and SimCLR suffering from major drops of 13.6% and 27.5% respectively when light augmentation is applied (crop+flip), and other approaches relying on particular schemes such as MultiCrop [57] among others.

The newly introduced **image encoder** plays an instrumental role in the model's downstream performance, enabling a 3x boost over features obtained from an *unconditional diffusion* model [19][2], which for ImageNet scores 24.49% only. A significant improvement further arises from the use of **novel view synthesis as a self-supervised representation learning objective**: Indeed, maintaining a distinction between the model's source and target views yields a 17.12% increase for ImageNet, compared to when they match (i.e. auto-encoding), even though the same data augmentation is applied in both cases (Figure 6)[3].

Other contributors include the **compact bottleneck** and **feature modulation**, which respectively raise accuracy by 38.04% and 15.51% for ImageNet, and 12.74% and 11.25% F1 for CelebA (Supp Table 11). Designing a new inverted **noise schedule** that favors medium noise levels in lieu of the extremes likewise strengthen the model's representation capacity, eliciting a 10.25% increase over ImageNet.

4.2. Visual Synthesis

Next, we analyze the model's generative skills, evaluating its ability to faithfully reconstruct an image x from its latent encoding z (**Section 4.2.1**)[4], and generate novel views of 3D objects from requested camera perspectives, given one or more conditional source views (**Section 4.2.2**). We evaluate the targets and predictions' similarity along multiple dimensions: pixel-wise (PSNR [81]), structural (SSIM [82]), perceptual (FID [83]) which accounts for sharpness and realism, and semantic (LPIPS [84]) (Appendix E.2).

4.2.1 Image Reconstruction

As indicated by Table 2, SODA produces excellent reconstructions, surpassing competing approaches like VQGAN [75], StyleGAN2 [85], DALL-E [74] and unCLIP (DALL-E2) [8], especially in terms of structural and semantic similarity (SSIM & LPIPS). Visually speaking, our samples

[2] The baseline obtains image encodings by pooling the features of the best-performing denoiser layer (a middle one) over a lightly-noised image.

[3] For auto-encoding, we sample a new augmentation at every training step, but use it both as the source and as the target.

[4] To evaluate our model at the task of reconstruction, we keep the source the and the target views the same during training $x' = x_0$.

Table 3. **Novel View Synthesis**, from a single source view, on real and synthetic objects, comparing Geometry-free and -aware approaches. *PSNR*, *SSIM*, and *LPIPS* respectively express pixel-wise, structural and semantic similarity, while *FID* captures sharpness and realism. For these experiments, we integrate cross attention with feature modulation (Section 3.3). ($\star$) denotes short sampling, with 20 steps only.

	GSO				ShapeNet				NMR			
Method	**PSNR ↑**	**SSIM ↑**	**FID ↓**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**FID ↓**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**FID ↓**	**LPIPS ↓**
Geometry-Aware Approaches												
PixelNeRF [60]	24.93	0.919	48.72	0.086	26.58	0.940	25.34	0.073	28.19	0.932	34.54	0.082
NeRF-VAE [59]	22.20	0.882	74.83	0.113	24.60	0.915	45.79	0.101	25.56	0.893	71.67	0.134
Geometry-Free Approaches												
Vanilla Auto-Encoder	24.46	0.941	75.96	0.129	24.71	0.943	42.23	0.096	24.22	0.948	51.42	0.116
SRT [80]	21.97	0.877	38.64	0.110	26.31	0.934	17.96	0.073	25.69	0.898	7.90	0.090
Diffusion-based Approaches (Geometry-Free)												
Palette [30]	13.42	0.672	8.49	0.199	14.44	0.582	6.93	0.177	14.10	0.609	6.03	0.212
DALL-E2 [8]	15.68	0.793	8.32	0.147	18.75	0.823	6.54	0.101	20.32	0.899	3.61	0.087
SODA w/o bottleneck	20.97	0.926	4.39	0.071	24.31	**0.949**	2.83	0.051	23.31	0.948	**0.75**	0.053
SODA w/o modulation	21.02	0.929	4.10	0.069	25.02	0.944	2.97	0.048	25.34	0.949	0.77	0.051
SODA (shorter sampling)*	24.38	0.930	2.35	0.065	26.71	0.946	1.31	0.046	27.13	0.936	1.10	0.063
SODA (*ours*)	**24.97**	**0.945**	**1.51**	**0.054**	**27.42**	0.947	**0.74**	**0.039**	**28.71**	**0.952**	0.81	**0.048**

are sharper and crispier than DALL-E's, while being more accurate than StyleGAN2 and unCLIP inversions, perhaps due to their lack of a trainable encoder (see supplementary examples). The results are significant given the order-of-magnitude lower dimensionality of the latents z from which we restore the images: 2K for SODA versus 65-524K for VQGAN and DALL-E[5], illustrating here an advantage of continuous representations over discrete codebooks.

4.2.2 Novel View Generation

For the task of few-shot novel view synthesis, we focus on the 3D regime and look into 3 datasets that span both synthetic object renderings and real-world scans of household items (Google Scanned Objects [86], custom ShapeNet [87], and NMR [88]). We compare our model to geometry-free and -aware approaches such as PixelNeRF [60], Scene Representation Transformer (SRT) [80], and Palette [30]. We condition the models on 1-9 source views, and test them on held-out validation objects that do not appear in training.

SODA consistently beats the competing approaches across the 3 datasets and for different numbers of source views, as indicated by FID, SSIM and LPIPS (Table 3 and Supplementary Figure 7). It reaches the largest gains along LPIPS and FID, producing significantly sharper images that better match the source views both structurally and semantically. We observe that settings of 1-3 source views benefit the most from our model, where for the single-source case, it improves FID scores by an order-of-magnitude and often almost halves the LPIPS scores. For the GSO dataset, as we increase the source views number, we score a little lower on the pixel-wise PSNR than PixelNeRF, perhaps due to the probabilistic nature of our approach. Yet, in terms of computational efficiency, contrary to the slow and heavy rendering of geometry-aware methods, SODA maintains strong performance with as little as 20 sampling steps.

Figure 5 and the supplementary animations feature objects synthesized from various perspectives, showcasing the viewpoint consistency SODA achieves. For multiple sources, we find that our proposed transformer-based view aggregation (Section 3.3) surpasses the stochastic conditioning technique of 3DiM [89] (Figure 6). Our approach further outperforms the denoiser-only Palette diffusion model, which fits translational tasks that closely follow the source layout, like colorization or super-resolution, but struggles at structural transformations, corroborating the need for our dedicated image encoder.

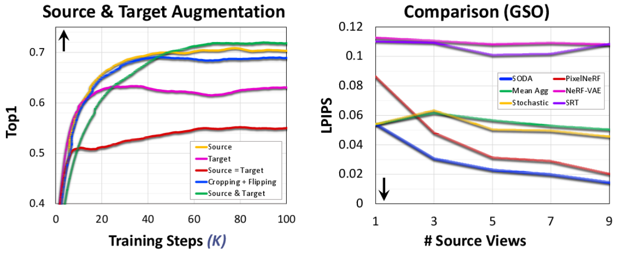

Figure 6. **(Left):** Contribution of the Novel View Synthesis objective to downstream classification on ImageNet. **(Right):** Generative quality comparison with baseline approaches and aggregation methods as we increase the number of source views.

4.3. Disentanglement & Controllability

The concept of disentanglement has been a recurring theme in representation learning research over the years [90–94]. While formal definitions may vary [95–99], a common aim lies in the discovery of abstract and meaningful latent representations that linearly align with the natural axes of variation. Disentanglement could enhance the encodings' interpretability, and, in the context of generative modeling, support greater controllability. In the following, we inspect the model's latent space, and analyze it quantitatively and qualitatively along the dimensions of disentanglement, controllability and informativeness.

[5] The overall latent dimension is $16{\times}16{\times}256$ – $32{\times}32{\times}512$.

4.3.1 Qualitative Evaluation

Latent Interpolations. We begin by visualizing latent interpolations for our model, linearly traversing the latent space from one vector z_1 to another z_2 (Figure 1 and the supplementary figures & animations). We observe smooth variations over traits of texture and structure. Notable in particular are image categories that seamlessly morph from one to another (e.g. from a tiger to a cat to a dog to a wolf), while for CelebA, we see gradual transformations ranging from broad shifts of pose and orientation to finer transitions of hair, facial features and expressions.

Attribute Manipulation. We go beyond interpolations and identify meaningful latent directions that correspond to individual axes of variation. To infer them, we explore two techniques: **supervised**, by normalizing the linear-probe's weight matrix, and **unsupervised**, through PCA decomposition [100] (details at Appendix E.3). Figure 4 and the supplementary figures show perturbations along the discovered directions, which influence various semantic properties: from age and gender in CelebA, to tone, clarity and lighting conditions in LSUN, to object's dimensions, thickness and material in ShapeNet and GSO. Indeed, we see that these manipulations are disentangled: one attribute is altered while the others are mostly kept intact, demonstrating the well-behaved nature of SODA's emergent latent space and the quality and strength of its learned representations. Indeed, we emphasize that both SODA's training and the following PCA-based discovery of interpretable latent directions are fully-unsupervised, derived from images only.

Layer Modulation. We investigate the effect of layer modulation and masking (Section 3.2.1), blocking the encoder's guidance from select decoder's layers and inspecting the impact on the produced outputs. As illustrated in the supplementary, it allows for selective modification of input images at different levels of granularity, so to preserve certain factors while unconditionally regenerating other ones. We can thus resample a new color palette while retaining shape and structure, or change the background while preserving the subject's identity. We find that layer masking improves the model's robustness to these forms of partial conditioning. Meanwhile, the role played by the initial Gaussian **noise map** x_T is closely linked to the chosen augmentation scheme: it controls fine stochastic subtleties like fur or freckles when SODA is trained to reconstruct the source image, and could conversely shape the underlying layout when heavier data augmentations are applied.

4.3.2 Quantitative Evaluation

To quantitatively bolster the findings above, we analyze our approach with DCI [94], which measures representations along Disentanglement, Completeness and Informativeness by assessing the degree of 1-to-1 correspondence between latent and ground-truth factors of variation (Appendix E.3). We evaluate models over multiple semantically-annotated datasets, ranging from the diagnostic SmallNORB and MPI3D to the realistic CUB and CelebA [101–105].

Table 4. **Disentanglement Analysis (DCI)** for various datasets. SODA achieves improvements of 27.2-58.3% in Disentanglement, 5.0-23.8% in Completeness, and comparable Informativeness to variational approaches. We report here the 3 metrics' average. StyleGAN2 achieves 68.71% (W) and 69.07% (W+) for CelebA. See Supplementary Table 7 for the full comparison.

Method	**CelebA**	**CUB**	**MPI3D**	**3DShapes**	**SmallNORB**
AnnealedVAE [106]	53.09	46.94	30.95	72.11	33.14
FactorVAE [63]	54.76	47.62	43.99	91.26	50.04
DIP-VAE (I) [107]	58.63	46.39	55.94	94.11	47.16
DIP-VAE (II) [107]	57.02	47.22	44.40	94.07	51.20
β-VAE [67]	56.96	47.39	47.77	**94.45**	50.40
β-TCVAE [108]	56.61	48.59	50.33	88.51	50.56
SODA w/o layer mod.	65.61	54.75	70.71	80.09	59.10
SODA (*ours*)	**74.67**	**56.98**	**73.41**	94.08	**64.78**

As Table 4 and supplementary Tables 6 and 7 show, SODA outshines both variational and adversarial approaches, improving Disentanglement by 27.2-58.3% and Completeness by 5.0-23.8% across 4 different datasets, with the sole exception of the synthetic 3DShapes, for which both SODA and most variational methods attain excellent scores. For Informativeness, results are mostly comparable, with SODA taking the lead for some datasets, while StyleGAN or DIP-VAE [107] improving scores for others. Our experiments further validate the contribution of layer modulation and masking, respectively yielding 3.2-13.5% and 2.5-8.9% increases in latent-space Disentanglement, and 3.8% and 1.7% mean increase in Completeness. Visually, SODA's samples are significantly sharper than the variational ones, and it achieves remarkable boosts in realism (FID) and semantic similarity (LPIPS).

5. Conclusion

We introduced SODA, a self-supervised diffusion model, designed for both perception and synthesis. It re-purposes the task of novel view generation as a training objective for representation learning. By conditioning a denoiser on an image encoder, and imposing an information bottleneck between the two, SODA learns strong semantic representations that enable downstream classification, as well as reconstruction, editing and synthesis. While we focused on single-object images, as in LSUN, ShapeNet, or ImageNet, we believe that exploring the applicability of our approach to dynamic compositional scenes is a promising direction for future research. We hope our work will help bridging the gap between novel view synthesis and self-supervised learning, two flourishing topics that are often pursued independently, and bring us one step closer to unlocking the potential of generative models in general and diffusion models in particular to advance the representational frontier.

References

[1] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002. 1

[2] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 1

[3] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 1

[4] Christos H Papadimitriou. Computational complexity. In *Encyclopedia of computer science*, pages 260–265. 2003. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 5, 14

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 14

[7] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 14, 21

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3, 5, 6, 7, 16, 20

[9] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[10] 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. 3

[11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[13] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[14] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022.

[15] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[16] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[17] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1

[18] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 1

[19] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. *arXiv preprint arXiv:2303.09769*, 2023. 6, 20

[20] Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021. 1, 2

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 4

[22] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2

[23] Sybren Ruurds De Groot and Peter Mazur. *Non-equilibrium thermodynamics*. Courier Corporation, 2013. 2

[24] 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. PMLR, 2015. 2

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[26] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4, 14

[28] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[29] 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*, 45(4):4713–4726, 2022. 2

[30] 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. 5, 7, 20, 23

[31] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022.

[32] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv e-prints*, pages arXiv–2303, 2023.

[33] Allan Jabri, Sjoerd van Steenkiste, Emiel Hoogeboom, Mehdi SM Sajjadi, and Thomas Kipf. DORsal: Diffusion for object-centric representations of scenes et al. *arXiv e-prints*, pages arXiv–2306, 2023. 2

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[35] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[36] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[37] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

[38] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022. 2

[39] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022. 2

[40] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022. 2

[41] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 2

[42] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 2

[43] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? *arXiv preprint arXiv:2305.16397*, 2023. 2

[44] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[45] Krzysztof J Geras and Charles Sutton. Scheduled denoising autoencoders. *arXiv preprint arXiv:1406.3269*, 2014.

[46] B Chandra and Rajesh Kumar Sharma. Adaptive noise schedule for denoising autoencoder. In *Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part I 21*, pages 535–542. Springer, 2014.

[47] Qianjun Zhang and Lei Zhang. Convolutional adaptive denoising autoencoders for hierarchical feature extraction. *Frontiers of Computer Science*, 12:1140–1148, 2018. 2

[48] 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. 3, 16, 22

[49] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 4, 14, 21

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 14

[51] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 23

[52] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. *arXiv preprint arXiv:2306.08757*, 2023. 3

[53] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3, 5, 6, 16, 17, 21

[54] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 5, 21

[55] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 5, 6, 21

[56] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5, 6, 17, 21

[57] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 5, 6, 21

[58] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4, 5, 6, 16, 17, 21

[59] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, pages 5742–5752. PMLR, 2021. 3, 7, 15, 22

[60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3, 7, 15, 22, 23

[61] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 3, 19, 22

[62] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[63] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 3, 8, 16, 18, 22

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[65] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 4

[66] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 22

[67] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016. 4, 8, 16, 18, 22

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 14, 24

[69] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 5, 14

[70] Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. *arXiv preprint arXiv:2212.11972*, 2022. 5

[71] 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. 5, 14

[72] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 5, 6, 21

[73] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019. 5

[74] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 5, 6, 21

[75] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5, 6, 16, 21

[76] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 6

[77] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[78] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[79] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6, 17, 23

[80] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 7, 15, 22, 23

[81] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6, 19

[82] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6, 19

[83] 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. 6, 19

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 19

[85] 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. 6, 16

[86] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 7, 16, 21

[87] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7, 16, 21

[88] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 7, 16, 21

[89] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2022. 7, 15, 22

[90] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 7

[91] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[92] Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992.

[93] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[94] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018. 7, 8, 19

[95] Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992. 7

[96] Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

[97] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[98] Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 125:126–127, 2017.

[99] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 7

[100] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 8, 19

[101] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004. 8, 16, 21

[102] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019. 16, 21

[103] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 16

[104] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 16, 21

[105] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 8, 16, 21

[106] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $beta$-vae. *arXiv e-prints*, pages arXiv–1804, 2018. 8, 16, 18, 22

[107] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 8, 16, 18, 22

[108] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 8, 16, 18, 22

[109] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv e-prints*, pages arXiv–1412, 2014. 14

[110] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the

impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 14

[111] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 14

[112] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. In *International Conference on Learning Representations*, 2016. 14, 23

[113] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 14

[114] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 14

[115] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. 2018. 14

[116] Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1308–1316, 2019. 15

[117] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 16, 21

[118] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 16, 21

[119] 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. 16, 21

[120] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 16, 21

[121] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 16

[122] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022. 16

[123] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 16

[124] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 17

[125] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015. 19

[126] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 19

[127] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 19

[128] 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. PMLR, 2021. 20

[129] Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv e-prints*, pages arXiv–1901, 2019. 21, 23

[130] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 21

[131] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021. 22