Diffusion Model as Representation Learner

Xingyi Yang   Xinchao Wang*
National University of Singapore
xyang@u.nus.edu, xinchao@nus.edu.sg

Abstract

*Diffusion Probabilistic Models (DPMs) have recently demonstrated impressive results on various generative tasks. Despite its promises, the learned representations of pre-trained DPMs, however, have not been fully understood. In this paper, we conduct an in-depth investigation of the representation power of DPMs, and propose a novel knowledge transfer method that leverages the knowledge acquired by generative DPMs for recognition tasks. Our study begins by examining the feature space of DPMs, revealing that DPMs are inherently denoising autoencoders that balance the representation learning with regularizing model capacity. To this end, we introduce a novel knowledge transfer paradigm named **RepFusion**. Our paradigm extracts representations at different time steps from off-the-shelf DPMs and dynamically employs them as supervision for student networks, in which the optimal time is determined through reinforcement learning. We evaluate our approach on several image classification, semantic segmentation, and landmark detection benchmarks, and demonstrate that it outperforms state-of-the-art methods. Our results uncover the potential of DPMs as a powerful tool for representation learning and provide insights into the usefulness of generative models beyond sample generation. The code is available at* `https://github.com/Adamdad/Repfusion`.

1. Introduction

In the ever-evolving landscape of machine learning, generative models have emerged as a captivating approach to tackle the intricacies of data distributions. Among these marvels, Diffusion Probabilistic Models (DPMs) stand tall, boasting a remarkable prowess in producing realistic and diverse samples. Powered by the elegent design of diffusion, these models elegantly transform a humble noise into target data distribution, unfurling a breathtaking array of variations and unyielding fidelity in their artistic creations.

---

*Corresponding Author.
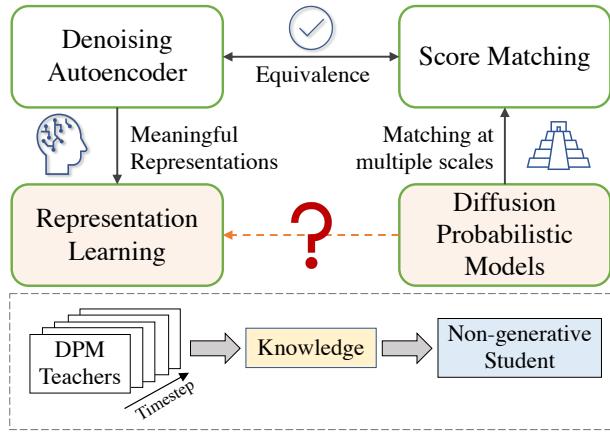
Figure 1. Illustration of our core idea on establishing a link between representation learning and diffusion models. Building upon this connection, we propose to utiliz the knowledge acquired from pre-trained DPMs. This knowledge is selectively distilled into a student network, empowering it to carry out recognition tasks with enhanced proficiency.

Although the generative ability of DPMs has been extensively studied, their potential for representation learning has not been fully explored. Recent research has demonstrated that the diffusion models already have a semantic latent space[32], and could be extended to tasks like controllable image generation [51], representation learning [1, 46] and image segmentation [5, 76], albeit through complicated model modifications. Nevertheless, the usefulness of the learned feature space of DPMs remain unclear. In this work, we intend to close the gap between generative DPM and its capability in representation learning by answering the following question: *Can the representation learned by DPMs be effectively reused for recognition tasks?*

We answer this question by first analyzing the inherent relationship between the diffusion model and the standard auto-encoder, as shown in Figure 1. Conceptually, DPMs are designed to predict noise from perturbed input distributions. Essentially, DPMs can be viewed as *denoising autoencoders* (DAEs) [68, 67] with varying denoising scales. DAEs have been well-established as a powerful technique for self-supervised learning, which captures the underlying

structures of data. Its latent features are useful for downstream tasks. On another note, DPMs are commonly known as score-matching models [62], which model the gradient of the log probability density with a sequence of intermediate latent variables. By connecting score-matching with DAEs [66], DPMs should also be considered an extension of DAE. It is, therefore, intuitive that DPM produces meaningful encoding from its input.

Although the representational power of the diffusion model has well-established theoretical foundations, leveraging the off-the-self DPM for non-generative tasks poses significant challenges. Primarily, these models are parameterized as time-conditioned Unet [55, 14, 63], a specialized structure unsuited for tasks such as classification and object detection. Secondly, existing DPMs are computationally heavy [78], making it difficult to use the original model for discriminative tasks without substantial modifications. Another obstacle arises from the fact that a single DPM can be perceived as a composition of networks, each indexed by its input timestep. Determining the suitable time-step remains non-trivial. Consequently, the representations learned by DPMs do not readily benefit other non-generative tasks.

In this paper, our primary objective is to reuse the knowledge encoded in the DPM for recognition tasks [13]. To achieve this, we look into the mathematical formulation of DPM's latent space, and show that DPMs can be seen as DAEs that strike a balance between learning meaningful features and regularizing the model capacity. Building on this insight, we propose a novel knowledge transfer approach, termed **RepFusion**. Our approach utilizes the knowledge distillation techniques to transfer the representation learning capability of the trained generative models to improve discriminative tasks. Specifically, we dynamically extract intermediate representations at different time steps and use them as auxiliary supervision for student networks. To determine the optimal time selection, we measure the informativeness of a given representation and optimize it through the REINFORCED algorithm [72]. Moreover, the reinforced time-steps selection is aligned with the derived property of DPMs, thereby providing a mechanism for adapting to different downstream tasks and increasing the flexibility and generalizability of our approach.

Our experiments demonstrate that **RepFusion** consistently improves performance on several image classification, semantic segmentation, and landmark detection benchmarks, indicating the powerful representation learning capability of DPMs. These findings shed light on the potential utility of generative models beyond their traditional use in sample generation and highlight the opportunities for exploring pre-trained models in representation learning.

To sum up, our contributions can be divided into three main parts:

1. We investigate the potential of repurposing diffusion

models for representation learning, an area that has been relatively unexplored in prior research.

2. By establishing the relationship between DPMs and denoising auto-encoders, we verify the statistical and empirical properties of features extracted from DPMs.

3. We introduce a novel knowledge distillation approach called **RepFusion**, which utilizes pre-trained DPMs to enhance recognition tasks. Extensive evaluations on image classification, segmentation, and landmark detection benchmarks demonstrate the effectiveness of DPMs as powerful tools for representation learning.

2. Related Work

Denoising Autoencoders. Autoencoders are a type of neural network that can learn to reconstruct their input at the output. Denoising autoencoder (DAE) is a variant of autoencoder that recover clean output from corrupted input. With the denoising objective, DAE has been one of the dominant approaches for unsupervised representations learning [68, 67, 18, 36]. Several studies have demonstrated that adjusting the noise scale [17, 11, 86] improves the features. Apart from learning representation, DAE has been recognized as a generative model by matching the score function [66, 62, 7, 30]. This paper reframes modern diffusion models as auto-encoders to make their latent features more suitable for recognition tasks.

Diffusion Probabilistic Models. DPMs [24] have emerged as a cutting-edge approach for generating high-quality samples, with applications like conditioned image generation [85, 52, 57], 3D generation [50], video synthesis [25] and domain generalization [82]. DPMs are essentially score-based models [66, 63] match the score functions over multiple scales [62]. The objective of this study is to investigate the potential of diffusion models in learning representation, which has not received much attention before.

Learn Representation from Generative Models Generative models, such as diffusion-based models have been found to learn meaningful semantics [51, 32], with applications on semantic segmentation [76], correspondence [43, 64] and image editing [61]. Others endeavors to learn from model generated data [4, 39, 74, 21]. However, current approaches [5] require complex modifications of the model. In contrast, our paper takes a novel approach by distilling well-trained DPMs for general representation learning.

Model Reuse and Knowledge Distillation. The popularity of pre-trained models creates a growing demand for reusing model [71, 29, 28, 77, 79, 53] and data [40, 41] to enhance performance and minimize costs [70, 15, 16, 44]. Knowledge distillation (KD) [22, 80, 65, 38] emerges as a promising solution, facilitating knowledge transfer from teacher models to students, benefiting DPMs with reduced sampling time [45, 58, 45] and improved model efficiency [78]. Our

research pioneers the exploration of knowledge distillation from diffusion-based models for recognition tasks.

3. Diffusion Models are Auto-Encoders

In this section, we review DPM's formulation and demonstrate its connection to DAEs in terms of representation learning. Our main insight is that DPM behaves similarly to a regularized auto-encoder, with its representation characteristics being controlled by noise scales.

3.1. Recap on Diffusion Probabilistic Models

The diffusion model is a probabilistic generative model designed for denoising by systematically reversing a progressive noising process. It initiates with clean data denoted as $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and iteratively applies Gaussian noise with zero mean and variance β_t for a total of T steps, progressively generating a noisy version at each step,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

where $t \in [1, T]$ and $0 < \beta_{1:T} < 1$ denote the noise scale scheduling. This process converges to a Gaussian white noise distribution, i.e., $\mathbf{x}_T \to \mathcal{N}(0, \mathbf{I})$. Sampling from the noise-perturbed distribution $q(\mathbf{x}_t) = \int q(\mathbf{x}_{1:t}|\mathbf{x}_0)d\mathbf{x}_{1:t-1}$ requires numerical integration over the steps, while the use of Gaussian parametrization enables the generation of an arbitrary time-step $\mathbf{x}_t$ through a closed-form formulation

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The reverse process is parameterized as a time-conditioned denoising neural network $\mathbf{s}(\mathbf{x}, t; \boldsymbol{\theta})$ with $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t + \beta_t\mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})), \beta_t\mathbf{I})$. This neural network is trained to minimize re-weighted evidence lower bound (ELBO) that fits the noise

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \mathbf{s}(\mathbf{x}_t, t; \boldsymbol{\theta})||_2^2\right] \quad (3)$$

With a trained denoiser $\mathbf{s}(\mathbf{x}, t; \boldsymbol{\theta}^*)$, we are able to generate the data by solving a reverse process.

Connecting DAE with Diffusion Model. The relation between diffusion model and DAE lies at the core of our paper. In essence, diffusion models are created through a gradual denoising procedure, akin to a cascade of DAEs, with the only distinction on the input t and shared parameters.

A compelling demonstration of this connection can be traced back to the advent of *score-based* generative model [26, 66, 6, 3, 7, 2]. In that sense, the denoising autoencoders involves fitting the derivative of the log-likelihood of the data, which is known as *score matching*. The connection is further reinforced by a recent study [63] that diffusion models learn to generate data by matching scores.

This profound insight provides a deeper understanding of the intrinsic relationship between these two powerful models, paving the way for our later discussion.

3.2. Probing the Representations in DPMs

Despite the connection between DPM and DAE, we typically ignore internal latent representation of DPM. This section examines the features learned by DPM, particularly the characteristics of the latent space at various time steps.

An autoencoder (AE) consists of an encoder $E : \mathbb{R}^L \to \mathbb{R}^d$ and a decoder $D : \mathbb{R}^d \to \mathbb{R}^L$, where $d < L$. The encoder maps the noisy input $\mathbf{x}' = \mathbf{x} + \sigma^2\boldsymbol{\epsilon}$ to a hidden space $\mathbf{h}$, and the decoder maps it back $\mathbf{x}'$ to make the prediction $\mathbf{y}$. We assume a linear AE with skip connection, which mimics the Unet [55] used in DPMs. This encoder-decoder interaction is mathematically expressed as:

$$\mathbf{h} = W_E\mathbf{x}'; \quad \mathbf{y} = W_D\mathbf{h} + W_S\mathbf{x}' \quad (4)$$

Here, $W_E \in \mathbb{R}^{d \times L}$, $W_D \in \mathbb{R}^{L \times d}$ and $W_S \in \mathbb{R}^{L \times L}$ are the encoding, decoding, and skip-connection matrices, respectively. This simplification enables us to examine the latent space properties of DPM in Proposition 1.

Proposition 1.[1] *Assume a linear DPM with skip connection. For $\mathbf{x}_t = \sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}$ and $\mathbb{E}[\mathbf{x}_0] = 0$, minimizing Eq.3 is equivalent to minimize*

$$\mathcal{L}(W_E, W_D, W_S) = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - (W_DW_E\mathbf{x}_t + W_S\mathbf{x}_t)||_2^2] \quad (5)$$
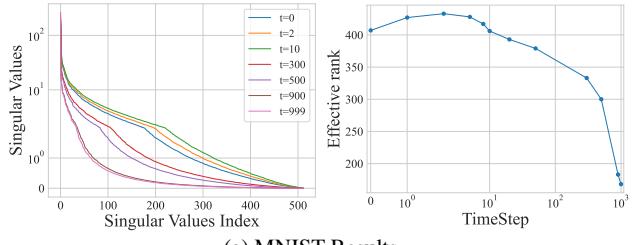
$$= \underbrace{\bar{\alpha}_t P^\top \Sigma_{xx} P}_{\text{Representation Learning}} + \underbrace{||I - \sqrt{1-\bar{\alpha}_t}P||_2^2}_{\text{Regularization}} \quad (6)$$

Here, $P = W_DW_E + W_S$, and Σ_{xx} is the covariance matrix of $\mathbf{x}_0$. The expression consists of two terms: the first term is related to representation learning, while the second term is related to regularization.
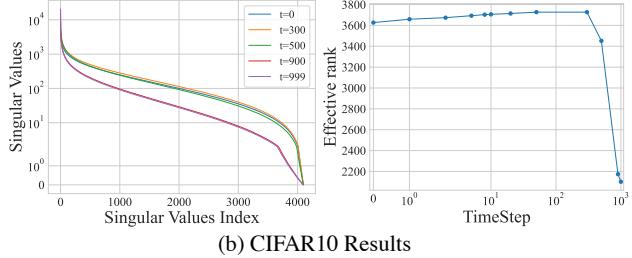
(i) The first term ensures a meaningful latent space $\mathbf{h}$. For example, when $W_S = -I$, we get back to linear AE that encodes the data into principal components space [27, 8].

(ii) The second term enforces spectral regularization [47], driving the diagonal values of the weight matrix $W_DW_E + W_S$ towards $\frac{1}{\sqrt{1-\bar{\alpha}_t}}$. This regularization encourages a compact latent space as more noise is added, aligning with the manifold hypothesis [10].

Trade-off in DPMs. Eq. 6 indicates a trade-off between learning a meaningful representation and the penalize the spectrum of the optimal parameters. As t gets larger and $\bar{\alpha}_t$ decreases, the model prioritizes the regularization term over the representation learning term. Consequently, the singular values of the learned representations tend to become smaller. if the model becomes overly regularized, its ability

---

[1]Due to the page limit, full derivation is in the supplementary material.

(a) MNIST Results

(b) CIFAR10 Results

Figure 2. Comparison of (Left) the singular values and (Right) effective rank of learned features at different timestep on MNIST and CIFAR-10 datasets.

Figure 3. T-SNE feature visualization on for unconditional DDPM trained on MNIST and CIFAR-10 datasets.

Figure 4. Averaged self-attention map visualization on different feature layers with $T \in \{0, 100, 999\}$ on CIFAR-10.

to learn informative features may become limited. Ideally, an optimal representation should exhibit uniform singular values close to 1, with a larger effective rank number [56] that prevents dimensional collapse [60].

Singular Values and Rank Number. To validate the trade-off in DPMs, we analyzed the singular values and effective rank number [56] of embeddings extracted from well-trained DPMs[2] on the CIFAR10 [31] and MNIST [34].

As illustrated in Figure 2, the representations show a slight improvement when t is small, but they quickly deteriorate as t increases due to a rapid reduction in singular values and the effective rank number. It indicates that the feature quality grows earlier in the reverse process, while gradually becoming uninformative in the later stages.

Latent Embedding Visualization. Visualizing the features using T-SNE for unconditional DDPMs, we observed a time-varying pattern. Figure 3 displays the data embeddings obtained through T-SNE for unconditional DDPMs. Our results reveal that, even without supervision, DPMs are able to group input samples at properly selected timestep, particularly $t = 300$ on MNIST and $t = 100$ on CIFAR-10. However, when t is too small or large, the class-wise representations become inseparable, leading to less distinct clustering structures and blurred decision boundaries.

Attention Patterns. We also examined the attention patterns of DPMs on the CIFAR-10 dataset for different noise magnitudes corresponding to different scales of feature extraction. We calculate the self-attention map at $T \in \{0, 100, 999\}$ by averaging across 128 samples and all at-

tention heads. The pixel value at the i-th column and j-th row refers to the attention weights between the i-th and the j-th token. As shown in Figure 4, higher noise levels led to the network attending to more image tokens, capturing long-range dependencies. Conversely, smaller t corresponded to local attention, capturing fine-grained patterns.

To sum up, DPMs encompass a latent space that evolves over time. Notably, the features at a moderately small timestep are found to possess the best geometric structure, potentially leading to optimal performance in the recognition task. Nevertheless, this optimal time step varies across various tasks and datasets, thereby posing significant challenges for applying trained DPM in non-generative tasks.

4. RepFusion

Given a pre-trained denoising network $s(\cdot, \cdot; \boldsymbol{\theta}^*)$, our goal is to reuse its knowledge for recognition tasks. This can be achieved by distilling a recognition encoder $f(\cdot; \boldsymbol{\theta}_f)$ from a DPM teacher. However, as discussed in Section 3, determining the ideal time step for knowledge transfer is non-trivial. To address this issue, we introduce a reinforcement learning approach to solve the selection problem. Our pipeline is illustrated in Figure 5.

4.1. Distill Knowledge from Diffusion Teacher

We intend to distill the intermediate representation from a pre-trained diffusion model to a recognition student. Given an input $\mathbf{x}$ and its label y, we extract the feature pair from both the diffusion model at timestep t and stu-

---

[2]We feed clean images and timestep t into the encoder and compute the statistics of the embedding vectors at the MID-BLOCK layer.
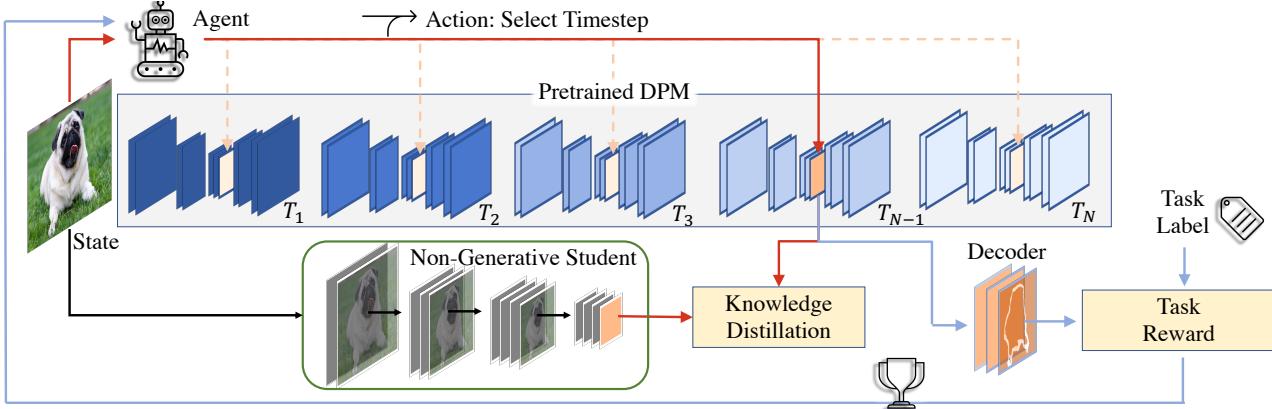
Figure 5. Overall pipeline for the **RepFusion**. Given an input sample, we adopt REINFORRCE algorithm [72] to determine the optimal time-step for knowledge distillation.

dent model respectively, e.g. $\mathbf{z}^{(t)} \leftarrow \boldsymbol{s}(\mathbf{x}, t; \boldsymbol{\theta}^*)$ and $\mathbf{z} \leftarrow f(\mathbf{x}; \boldsymbol{\theta}_f)$. We minimize the distance between $\mathbf{z}^{(t)}$ and $\mathbf{z}$ with respect to a loss function $\mathcal{L}_{\mathrm{kd}}$

$$\min_{\boldsymbol{\theta}_f} \mathbb{E}[\mathcal{L}_{\mathrm{kd}}(\mathbf{z}^{(t)}, \mathbf{z})] \qquad (7)$$

In the experiment, we show that arbitrary distillation function $\mathcal{L}_{\mathrm{kd}}$ can effectively boost the model's performance. Our implementation employs L2 distance [54], attention transfer [84], and relational knowledge distillation [49].

After the distillation phase, the student is reapplied as a feature extractor and fine-tuned on with the task label

$$\min_{\boldsymbol{\theta}_f} \mathbb{E}[\mathcal{L}_{\mathrm{task}}(y, \hat{y})], \quad \text{where } \hat{y} = f(\mathbf{x}; \boldsymbol{\theta}_f) \qquad (8)$$

Here, $\hat{y}$ represents the student model's prediction and $\mathcal{L}_{\mathrm{task}}$ is the task loss function. In this way, the distilled model using Eq. 7 is utilized to initialize the task training. As discussed, the selection of an optimal value for t is a critical aspect of our pipeline. We provide a detailed explanation of our selection strategy for t in Section 4.2.

4.2. Reinforced Time Selection for Distillation

Applying DPM for knowledge distillation presents a noteworthy challenge due to the mismatch between a group of time-indexed teachers and a single student network. This section outlines our criteria for time-step selection and explains how we optimize it through reinforcement learning.

Intuitively, we seek to extract the most informative feature, akin to a "golden standard", that can guide the student. To achieve this, we identify the optimal t^* for each sample $\mathbf{x}$, such that the DPM's feature $\mathbf{z}^{(t^*)}$ is most predictive for y, thereby facilitating knowledge distillation. The realization of this objective leads to a nested optimization problem:

$$t^* = \operatorname*{argmin}_{t \in [0, T]} \left\{ \inf_{\boldsymbol{\theta}_g} \mathcal{L}_{\mathrm{task}}(y, g(\mathbf{z}^{(t)}; \boldsymbol{\theta}_g)) \right\} \qquad (9)$$

Here, $g(\cdot; \boldsymbol{\theta}_g)$ refers to an additional decoder that maps the feature of the DPM to the label space. While its generally impractical to train g on all $(t, \mathbf{x})$ combinations[3] exhaustively, we have developed an efficient solution for determining the optimal distillation strategy using the REINFORCED algorithm [72].

Specifically, we define the state space as the input sample $\mathbf{x}$, and the action space as timestep t. We set the reward function $\mathcal{R}_{\mathbf{x}}^t$ to be the negative task loss $-\mathcal{L}_{\mathrm{task}}(y, g(\mathbf{z}^{(t)}; \boldsymbol{\theta}_g))$. A policy network $\boldsymbol{\pi}_{\boldsymbol{\theta}_\pi}(t|\mathbf{x})$ is trained to determine, on a per-sample basis, which timestep should be evaluated. The network parameters $\boldsymbol{\theta}_\pi$ are optimized to maximize the reward function

$$\max_{\boldsymbol{\theta}_\pi} J(\boldsymbol{\theta}_\pi) = \mathbb{E}[\sum_t \boldsymbol{\pi}_{\boldsymbol{\theta}_\pi}(t|\mathbf{x}) \mathcal{R}_{\mathbf{x}}^t] + \lambda_H H(t) \qquad (10)$$

where $H(t)$ [48] is an entropy term to promote action diversity, and $\lambda_H = 0.1$ is its weighting factor.

Besides, the decoder g is trained jointly with the policy network to minimize task loss and improve the prediction accuracy of the ground-truth label. As such, we are optimizing Eq. 9 equivalently.

By assembling all the components, we present the **RepFusion** algorithm in Figure 5. In each training step, we begin by sampling the time from the action distribution $t \sim \text{Categorical}_T(\boldsymbol{\pi}_{\boldsymbol{\theta}_\pi}(t|\mathbf{x}))$ and subsequently extract the corresponding feature map $\mathbf{z}^{(t)}$ for each input sample. The student network is optimized to replicate the teacher's representation using Eq. 7. Meanwhile, the policy network and teacher decoder are updated according to Eq. 9 and Eq. 10. Despite its simplicity, we observe that this pipeline yields impressive results across various tasks.

---

[3]Gradient-based optimizations are not feasible due to the discrete nature of t.

5. Experiment

This section demonstrates that our **RepFusion** can be effectively utilized in a variety of recognition tasks with competitive performance, despite being learned from generative DPMs. The distilled encoders are evaluated on a range of downstream applications, including semantic segmentation, keypoint detection, and image classification. More experiments can be found in the supplementary materials.

5.1. Experimental Setting

Datesets and Pre-trained DPMs. We evaluate the proposed method on 3 tasks across 4 different visual datasets. Those datasets include CelebAMask-HQ for semantic segmentation, WFLW for face landmark detection and CIFAR-10 and Tiny-ImageNet for image classification. The **CelebAMask-HQ** [35] dataset is a collection of 30,000 high-resolution face images. Each image comes with its segmentation mask of 19 facial attributes. The **WFLW** dataset [73] consists of 10,000 faces that have been annotated with 98 landmarks, with 7,500 for training and 2,500 for testing. **CIFAR-10** [31] is an image dataset containing 32×32 images from 10 classes, and is split into 50,000 images for training and 10,000 for validation. **Tiny-ImageNet** [33] comprises images from 200 classes downsized to 64×64. Specifically, each class contains 500 training, 50 validation, and 50 test images.

We have gathered open-source DPMs models on diffusers[4] and guided diffusion[5]. To perform knowledge distillation, we utilize ADM [14] model that was trained on 64×64 ImageNet, as the teacher model for TinyImageNet. For face parsing, we use DDPM [24] trained on CelebA-HQ, as the teacher model. We do not apply distillation on WFLW, but direct transfer the weight learned from the CelebA-HQ segmentation as initialization. Furthermore, we reuse the DDPM that was trained on unconditional CIFAR-10 as the teacher model for the same dataset.

Distillation. We utilize the pipeline described in Section 4 to train the networks f, g, π for 200 epochs with a mini-batch size of 128 across all tasks. The networks are optimized using SGD with an initial learning rate of 0.1, a momentum term of 0.9, and a weight decay of 1e-4. We apply learning rate decay at the 100th and 150th epochs. During training, the only augmentation applied to training images is random horizontal flipping, and no pre-trained weights are loaded by default. Feature distillation is performed using the embeddings extracted from the MID-BLOCK of the DPM and the last layer of the student network. A hint loss is used for feature distillation by default, and this selection will be verified in the ablation study. Furthermore, we define g as a linear layer for classification and FCN [42] for

---

Table 1. Semantic segmentation results on CelebAMask-HQ [35], compared with hint-based knowledge distillation [54].

Teacher	Student	mIOU	aAcc	mAcc
-	ResNet-18	61.22	94.65	70.29
ResNet-18	ResNet-18	62.73	94.46	71.88
ResNet-50	ResNet-18	63.12	94.72	72.44
RepFusion(ours)	ResNet-18	**67.37**	**94.75**	**76.37**
-	ResNet-50	65.81	94.83	75.31
ResNet-50	ResNet-50	68.21	94.96	77.23
RepFusion(ours)	ResNet-50	**70.63**	**95.24**	**80.08**

Table 2. Semantic segmentation results on CelebAMask-HQ [35], compared with self-supervised learning approaches.

Method	mIOU	aAcc	mAcc
ResNet-18	61.22	94.65	70.29
+ MoCov2 [12]	66.38	**94.84**	74.23
+ SwAV [9]	67.14	94.66	75.92
+ DeepClusterv2 [9]	66.92	94.51	76.11
+ **RepFusion**(ours)	**67.37**	94.70	**76.37**
ResNet-50	65.81	94.83	75.31
+ MoCov2 [12]	69.81	95.32	79.55
+ SwAV [9]	70.41	95.26	79.40
+ DeepClusterv2 [9]	68.83	95.27	79.43
+ **RepFusion**(ours)	**70.89**	**95.47**	**80.14**

segmentation. The policy network π is a 3-layer convolutional net followed by 3 linear layers.

Segmentation. We use BiSeNetv1 [81] with ResNet-18 and ResNet-50 backbones as the baseline for semantic segmentation. The networks are distilled on CelebA-HQ at 256×256 resolution, then trained for 160k iterations on segmentation labels using SGD optimizer with batch size 16, initial learning rate 0.01, momentum 0.9, and polynomial learning rate scheduler. Random cropping of size 448×448 is applied during training. Finally, we evaluate the performance of the model on the test set at a scale of 512×512 and report its performance based on the mean Intersection over Union (mIOU), mean accuracy of each class (mAcc), and all-pixel accuracy (aAcc).

Landmark Detection. The distilled segmentation student on CelebA-HQ is reused for Top-down heatmap-based keypoint estimation [75] with HRNetv2-w18 [69] and ResNet-50. The model was trained for 60 epochs with a batch size of 128 using the Adam optimizer with an initial learning rate of 2e-3. Images were resized to 256 during both training and testing. Performance was evaluated on using the normalized mean error (NME) metric on 7 subsets.

Classification. On CIFAR-10 and Tiny-ImageNet, we follow the conventional setting and apply ResNet-

Figure 6. **Face parsing** results on CelebAMask-HQ dataset. **RepFusion** excels on segmenting rare patterns and minority categories.

Figure 7. **Face keypoint detection** results on WFLW [73] dataset. **RepFusion** can successfully handle the hard cases (large pose and low image quality).

18 [20], WideResNet (WRN) 28-2 [83] and MobileNetv2 (MBNv2) [59] as the student models.

Baselines. We compared our method to several self-supervised learning approaches, including MoCov2 [12], SwAV [9], and DeepCluster-v2 [9]. To ensure a fair comparison, we used the official implementation and pre-trained the models directly on CelebA-HQ without large-scale pre-training, keeping all hyper-parameters the same as in their respective papers. However, we did not compare with DatasetGAN [87] and DatasetDDPM [5] as they are primarily designed for few-shot segmentation.

Besides, we assessed our performance by comparing it with the other distillation techniques. Specifically, we initially trained a teacher model from scratch and distilled a student model by mimicking the teacher's features.

5.2. Evaluation on Recognition Tasks

5.2.1 Semantic Segmentation

To evaluate the effectiveness of our method, we conducted a comparative analysis on CelebAMask-HQ by comparing it with two distinct lines of research, namely (1) knowledge distillation from a supervised teacher and (2) self-supervised representation learning models.

Quantitative Results. We show the numerical results for face parsing with distilled and self-supervised models in Table 1 and Table 2. First, our approach significantly outperforms the knowledge distillation counterparts. Precisely, while raw ResNet-18 achieved an mIOU score of 61.22, our method achieved an mIOU of 67.37. This represents an improvement of 6.15 over the baseline model and a 4.25 improvement over the distilled student from ResNet-50. Our **RepFusion** also matches the performance with up-to-date self-supervised representation learner. These findings underscore the notable improvement in the segmentation task through knowledge distillation from off-the-shelf DPMs.

Qualitative Results. Additionally, we visualize the segmented facial attribute with ResNet-50 model in Figure 6,
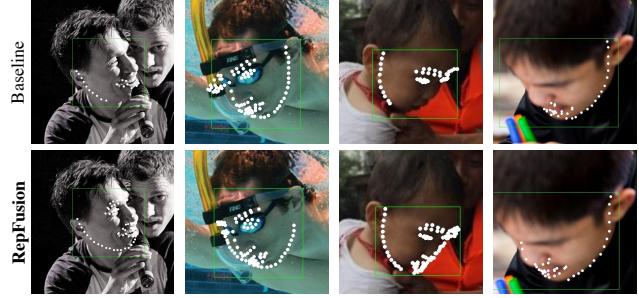
where we highlight the errors made by the baseline model using white rectangles. It is evident that the baseline model struggles to accurately segment rare patterns and minority categories such as `white hair`, `hat` and distinguish between the `left eyebrows` and `right eyebrows`. By distilling from DPMs, our **RepFusion** enables the model to capture fine-grained details to better discern between these features, resulting in more accurate segmentation.

5.2.2 Keypoint Detection

In this study, we compare our distill image encoder on CelebA-HQ to other self-supervised learning approaches for facial keypoint detection.

Quantative Results. We present the results in Table 3. All of the self-supervised methods yielded performance improvements. However, our proposed approach outperformed the other methods, achieving the smallest mean errors for both ResNet-50 and HRNetv2w18.

It is noteworthy that the distillation process is proved to be particularly beneficial in challenging scenarios characterized by large pose variations, heavy occlusions, and extreme illumination. In such scenarios, our proposed method achieved significant improvements of 0.39/0.18 on NME_{pose}, 0.25/0.15 on $NME_{\text{occlusion}}$, and 0.24/0.09 on $NME_{\text{illumination}}$ scores. These findings highlight the effectiveness of our proposed distillation approach in enhancing the model's performance in challenging scenarios.

Qualitative Results. Figure 7 showcases the landmark detection visualization obtained by ResNet-50. The outcomes expose that the baseline model, lacking distillation, encounters challenges in recognizing faces with extensive head movements or off-angle orientations. In contrast, our proposed technique successfully identifies landmarks that are otherwise unobservable. These findings validate the efficacy of our approach in augmenting the model's recall in detecting facial landmarks, especially in complex scenarios characterized by notable pose variations, occlusions, and

Table 3. Normalized Mean Error (NME) on WFLW for head keypoint detection. A smaller value indicates better performance.

Method	NME_{test} ↓	NME_{pose} ↓	$NME_{illumination}$ ↓	$NME_{occlusion}$ ↓	NME_{blur} ↓	NME_{makeup} ↓	$NME_{expression}$ ↓
HRNetv2w18	4.33	7.81	4.42	5.35	4.92	4.36	4.55
+ MoCov2 [12]	4.27	7.64	4.25	5.22	4.86	4.22	4.49
+ SwAV [9]	4.26	7.54	4.19	5.23	4.82	4.19	4.49
+ DeepClusterv2 [9]	4.31	7.58	4.22	5.19	4.84	4.23	4.48
+ **RepFusion**	**4.23**	**7.42**	**4.18**	**5.10**	**4.80**	**4.15**	**4.48**
ResNet-50	4.58	8.15	4.58	5.63	5.23	4.54	5.01
+ MoCov2 [12]	4.49	7.99	4.52	5.49	5.14	4.42	4.85
+ SwAV [9]	4.47	7.97	4.53	5.49	5.12	4.40	4.87
+ DeepClusterv2 [9]	4.51	8.00	4.55	5.51	5.18	4.39	4.88
+ **RepFusion**	**4.47**	**7.97**	**4.49**	**5.48**	**5.10**	**4.39**	**4.84**

Table 4. Test Accuracy (%) comparison on CIFAR-10.

Teacher:Acc	Student:Acc	Hint [54]	AT [84]	RKD [49]	Avg Imp.
ResNet-18:94.47	ResNet-18:94.47	95.06	94.66	94.90	+0.40
WRN28-2:94.15		94.87	94.63	94.57	+0.22
WRN28-10:95.18		94.95	94.35	95.04	+0.31
RepFusion(Ours)		95.01	95.09	94.58	**+0.42**
ResNet-18:94.47	WRN28-2:94.15	94.39	94.42	94.57	+0.31
WRN28-2:94.15		94.53	94.12	94.07	+0.09
WRN28-10:95.18		94.35	94.05	94.61	+0.18
RepFusion(Ours)		94.52	94.42	94.66	**+0.38**

Table 5. Test Top-1 Accuracy (%) comparison on Tiny-ImageNet.

Teacher:Acc	Student:Acc	Hint [54]	AT [84]	RK [49]	Avg Imp.
ResNet-18:60.37	ResNet-18:60.37	61.48	61.93	61.06	+1.12
WRN28-2:59.30		61.91	61.11	61.31	+1.07
WRN28-10:65.73		61.65	61.12	62.08	+1.25
RepFusion(Ours)		62.16	62.05	62.38	**+1.83**
ResNet-18:60.37	MBNv2:61.69	61.99	61.84	62.40	+0.42
WRN28-2:59.30		61.97	61.20	62.37	+0.20
WRN28-10:65.73		61.81	61.63	61.63	+0.04
RepFusion(Ours)		61.93	62.01	62.27	**+0.42**

suboptimal image quality.

5.2.3 Image Classification

In this section, we aim to compare the effectiveness of learned representations obtained through distillation from DPMs by assessing the classification performance [23, 19] on widely used benchmark datasets, namely CIFAR-10 and TinyImageNet. We compare our approach with typical distillation schemes on top of Hint-based [54], Attention Transfer-(AT) [84], and Relational Knowledge Distillation (RKD) [49] losses.

Results. Table 4 and Table 5 present a comparison of the test accuracy of various models on the CIFAR-10 and Tiny-ImageNet datasets. On CIFAR-10, our proposed method performs comparably to other teachers in terms of the average increase in test accuracy, exhibiting a gain of +0.42/+0.38 compared to its student. On Tiny-ImageNet, our proposed technique achieves the highest average accuracy gain, with an improvement of +1.83 over ResNet-18 and +0.42 on MBNv2. These results highlight the effectiveness of our approach in improving the performance of the student models on recognition tasks.

Furthermore, all distillation losses (Hint, AT, and RKD) lead to a boost in student model accuracy, with hint-based

distillation showing the best improvement. Notably, our model outperforms student models that were distilled from supervised teachers, establishing its efficacy in acquiring valuable representations for classification tasks.

Comparison Among Tasks. In our experiments, we found that when distilling from the DPM teacher, there was a notable performance improvement in segmentation compared to a more modest enhancement in classification. This suggests that the features learned by DPMs are more effective for tasks that prioritize *local information* over *global semantics*. This conclusion is supported by findings from other recent studies [37, 88].

5.3. Ablation Study and Analysis on Time Selection

This section validates the necessity and effectiveness of our time-step selection approach. We specifically investigate its importance and the achieved outcomes.

Do We Need Time-step Selection? To assess the necessity of the reinforced strategy in selecting time steps, we conducted experiments using alternative selection methods, namely *random* and *manual* selection. For the *random* strategy, a time-step is sampled uniformly from the set of time-steps $\{0, \ldots, T\}$ for each data point. In contrast, the *manual* selection method involves selecting a fixed time-step t for all data points, with $t \sim$
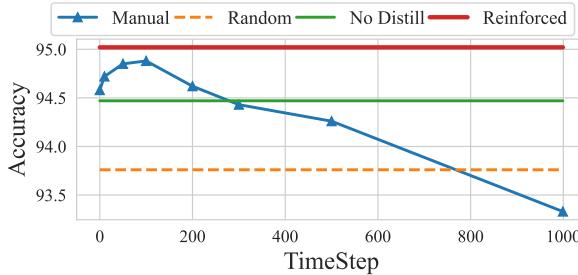
Figure 8. Ablative study on the time-step selection.

$\{0, 10, 50, 100, 200, 300, 500, 999\}$. We validate the results on CIFAR-10 with the ResNet18 model.

Figure 8 presents the performance with different time selection strategies. The results indicate that distillation from randomly sampled or large fixed timesteps leads to a significant degradation in performance, even compared to the no-distill baseline. This finding aligns with the derivation in Section 3 that large timestep features are unlikely to provide discriminative information, leading to a negative impact on performance. While a fixed small t approach can slightly improve performance, it is not universally effective. In contrast, our reinforced strategy that identifies the optimal t on a per-sample basis outperforms the fixed t approach.

What Time Has Been Selected? One natural question is what timestep is selected across various tasks. Our core observation reveals that different tasks give slightly varied outcomes. Initially, during training, the time step t is randomly selected, but gradually it tends to converge to a narrow range of $0 - 200$. We have included the full results in the supplementary, owing to page limitations.

6. Conclusion

In this work, we demonstrate the efficacy of Diffusion Probabilistic Models (DPMs) in enhancing visual representations for recognition tasks. By establishing a connection between DPMs and denoising auto-encoders, we empirically validate the statistical properties of DPM-extracted features. However, utilizing these features for non-generative tasks poses challenges. To address this, we propose a novel knowledge distillation approach called **RepFusion**, which leverages a reinforcement learning framework to determine the optimal timing for representation distillation. Our method demonstrates consistent improvements across various recognition benchmarks, highlighting the remarkable potential of DPMs in learning effective representations and facilitating transfer to downstream tasks.

Acknowledgment

References

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

[2] Guillaume Alain and Yoshua Bengio. What regularized autoencoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[3] Guillaume Alain, Yoshua Bengio, and Salah Rifai. Regularized auto-encoders estimate local statistics. In *ICLR*, 2013.

[4] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *CoRR*, abs/1711.04340, 2017.

[5] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[6] Yoshua Bengio, Guillaume Alain, and Salah Rifai. Implicit density estimation by local moment matching to sample from auto-encoders. *arXiv preprint arXiv:1207.0057*, 2012.

[7] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[8] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[9] 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.

[10] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[11] 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.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[13] 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.

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

[15] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[16] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *arXiv preprint arXiv:2305.10924*, 2023.

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

[18] 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.

[19] 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.

[20] 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.

[21] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[22] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[23] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

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

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

[26] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[27] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[28] Yongcheng Jing, Yining Mao, Yiding Yang, Yibing Zhan, Mingli Song, Xinchao Wang, and Dacheng Tao. Learning graph neural networks for image style transfer. In *European Conference on Computer Vision*, 2022.

[29] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Amalgamating knowledge from heterogeneous graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[30] Durk P Kingma and Yann Cun. Regularized estimation of image statistics by score matching. *Advances in neural information processing systems*, 23, 2010.

[31] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[32] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

[33] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[34] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[35] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[37] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models, 2023.

[38] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[39] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *CoRR*, abs/2305.12954, 2023.

[40] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *Advances in neural information processing systems*, 2022.

[41] Songhua Liu, Jingwen Ye, Runpeng Yu, and Xinchao Wang. Slimmable dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[42] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[43] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv*, 2023.

[44] Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the Structural Pruning of Large Language Models. *arXiv preprint arXiv:2305.11627*, 2023.

[45] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14297–14306, June 2023.

[46] Sarthak Mittal, Guillaume Lajoie, Stefan Bauer, and Arash Mehrjou. From points to functions: Infinite-dimensional representations in diffusion models. *arXiv preprint arXiv:2210.13774*, 2022.

[47] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[48] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[49] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[50] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023.

[51] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[52] 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.

[53] Sucheng Ren, Xingyi Yang, Songhua Liu, and Xinchao Wang. SG-Former: Self-guided Transformer with Evolving Token Reallocation. In *IEEE International Conference on Computer Vision*, 2023.

[54] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[55] 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.

[56] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[58] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[59] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[60] Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[61] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023.

[62] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[63] 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*, 2021.

[64] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *ArXiv*, abs/2306.03881, 2023.

[65] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[66] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[67] 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.

[68] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[69] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[70] Peihao Wang, Rameswar Panda, Lucas Torroba Hennigen, Philip Greengard, Leonid Karlinsky, Rogerio Feris, David Daniel Cox, Zhangyang Wang, and Yoon Kim. Learning to grow pretrained models for efficient transformer training. In *The Eleventh International Conference on Learning Representations*, 2023.

[71] Peihao Wang, Rameswar Panda, and Zhangyang Wang. Data efficient neural scaling law via model reusing. 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 36193–36204. PMLR, 23–29 Jul 2023.

[72] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32, 1992.

[73] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018.

[74] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2XP-ASG: generating adversarial scenes for vehicle-to-everything perception. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 3584–3591. IEEE, 2023.

[75] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[76] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023.

[77] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, 2022.

[78] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[79] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. *Advances in neural information processing systems*, 35:25739–25753, 2022.

[80] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[81] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[82] Runpeng Yu, Songhua Liu, Xingyi Yang, and Xinchao Wang. Distribution shift inversion for out-of-distribution prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[83] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[84] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. 2017.

[85] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

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

[87] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[88] 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.