Free-ATM: Harnessing Free Attention Masks for Representation Learning on Diffusion-Generated Images

David Junhao Zhang[1*], Mutian Xu[2,4†], Jay Zhangjie Wu[1], Chuhui Xue[3], Wenqing Zhang[3], Xiaoguang Han[2,4], Song Bai[3], and Mike Zheng Shou[1‡]

[1] Show Lab, National Univeristy of Singapore
[2] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[3] ByteDance
[4] The Future Network of Intelligence Institute, CUHK-Shenzhen

Abstract. This paper studies visual representation learning with diffusion-generated synthetic images. We start by uncovering that diffusion models' cross-attention layers inherently provide *annotation-free attention masks* aligned with corresponding text inputs on generated images. We then investigate the problems of three prevalent representation learning methods (*i.e.,* contrastive learning, masked modeling, and vision-language pretraining) on diffusion-generated synthetic data and introduce customized solutions by fully exploiting the aforementioned free attention masks, namely Free-ATM. Comprehensive experiments demonstrate Free-ATM's ability to enhance the performance of various representation learning frameworks when utilizing synthetic data. This improvement is consistent across diverse downstream tasks including image classification, detection, segmentation and image-text retrieval. Meanwhile, by utilizing Free-ATM, we can accelerate the pretraining on synthetic images significantly and close the performance gap between representation learning on synthetic data and real-world scenarios.

Keywords: Synthetic Data · Diffusion · Representation Learning

1 Introduction

Representation learning is a type of machine learning where models learn to identify patterns or structures in data. In the past few years, several representation learning techniques have emerged, including contrastive learning [9, 11, 20, 22], masked modeling [21,73], and vision-language pretraining [36,42,53], *etc.* Although these advancements have led to significant progress in visual representation learning, the majority of them rely on pretraining on large-scale datasets, such as ImageNet [13] which contains millions of images. However, manually building

* Work is partially done during an internship at ByteDance.
† Project lead.
‡ Corresponding author.

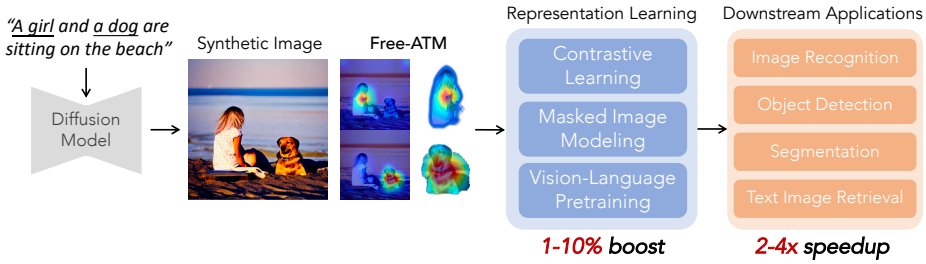

Fig. 1: ***Free-ATM****:* Free attention masks can be naturally adapted into different representation learning frameworks for diffusion-generated data, boosting performance and speeding up pretraining on these synthetic data.

a sizable dataset with decent richness and diversity is often time-consuming and costly. Moreover, present-day concerns about data privacy and usage rights further complicate the acquisition of massive data [52], creating additional obstacles to the development of representation learning.

To overcome these challenges, using synthetic data for representation learning presents itself as a logical option, given its advantages such as cost-effectiveness, virtually limitless scalability, enhanced control over data distribution, and improved data privacy and security.

In the computer vision area, there have been some attempts that leverage synthetic data for image recognition tasks. Besnier *et al.* [6] and Zhao *et al.* [77] both employ BigGAN [7] to produce informative images for training image classifiers. DatasetGAN [76] and BigDatasetGAN [41] adopt StyleGAN [35] and BigGAN [7] for generating images with pixel-wise labels for segmentation tasks. In addition to using GANs, He *et al.* [27] find that the revolutionary text-to-image diffusion models such as GLIDE [49] can generate not only high-quality but also diverse images in a customized label space for benefiting image recognition. This recent study is noteworthy for firstly demonstrating promising results of image understanding using **diffusion-generated images**. Albeit promising, there has been a lack of in-depth exploration focusing on representation learning on diffusion-generated data. We attempt to remedy this defect from the perspective of both ***i)*** diffusion-generated data and ***ii)*** representation learning frameworks.

i) In contrast to class-specific GANs that can only generate images of individual objects, text-to-image diffusion models have the capability to produce diverse images featuring *multiple* objects by utilizing different text tokens. More importantly, we find that the cross-attention layers of diffusion models naturally provide semantic attention masks aligned with corresponding text inputs on generated images *without* any manual annotations (also indicated in [31, 78]), which helps to locate each foreground object as shown in Fig. 1. We can get these attention masks freely when using diffusion model to generate images, requiring no additional cost. ***ii)*** In view of representation learning, there are some issues when directly applying standard representation learning approaches to diffusion-generated synthetic data. To tackle these challenges, we suggest

employing free attention masks (Free-ATM) with diffusion-generated images to bolster representation learning on synthetic data. In the following part, we analyze these problems using three renowned representation learning frameworks and demonstrate how Free-ATM can be effectively used to overcome them.

Contrastive Learning (CL) methods [9–11, 20, 22] usually treat an image with a single object as a complete entity and conduct random crop and augmentations to get positive pairs at the image level. Yet, such a paradigm is not suitable for diffusion-generated images containing multiple objects. As discussed earlier, the diversity of diffusion-generated images, which often include multiple instances, presents a significant challenge for traditional random crops in contrastive learning. This is due to the high risk of positive pairs being originated from distinct instances, resulting in ambiguity in model training as the discriminate features of each instance are pulled in (see top row of Fig. 2 (a)). To mitigate this issue, we leverage the Free-ATM to ensure that each positive pair comes from the same object. Meanwhile, negative pairs are formed by selecting different instances of images based on their corresponding masks (Fig. 2 (a)).

Masked Image Modeling (MIM) [5, 21, 73] achieves success in various vision challenges by reconstructing randomly masked visual patches. The diffusion model might produce images with plain or empty backgrounds as shown in Fig. 2(b). If random masks are used, a large portion will fall into areas with little to no semantic or informational content, making the reconstruction of these regions not particularly meaningful for the model. Differently, we employ the Free-ATM to increase the masking ratio of foreground object patches. Our strategy enables the MIM model to learn both universal and targeted representations.

Vision-and-Language Pretraining (VLP) [36, 42, 53] predominantly relies on position features, such as those belonging to the objects of interest in an image, to gain a better understanding of the relationships between words and objects. Previous methods usually locate these features through the use of bounding boxes detector. However, many objects in synthetic images might be out of domain and not be recognized by an object detector trained on specific real-world data. Additionally, the detection process is time-intensive. Fortunately, the Free-ATM can naturally align each text prompt with its corresponding object position. As shown in Fig. 2 (c), we apply attention masks to supply position information *without* requiring the extra step of object detection. This not only brings greater efficiency to VLP models, but also enhances their overall effectiveness.

Notably, we do not aim at designing specific modules to enhance a single representation learning framework. Instead, our work explores how to better utilize synthetic images for general visual representation learning by proposing simple yet effective Free-ATM adaptions. Extensive experiments show that our Free-ATM enhances the results of different representation learning frameworks on synthetic images, which is validated by consistent improvements on different benchmarks (PASCAL VOC [17], COCO [45], Cityscapes [12], ADE20K [79], and ImageNet [13]) across various downstream tasks including image classification, detection, segmentation, and image-text retrieval. Meanwhile, Free-ATM can significantly accelerate the pretraining process on synthetic images. In addition,

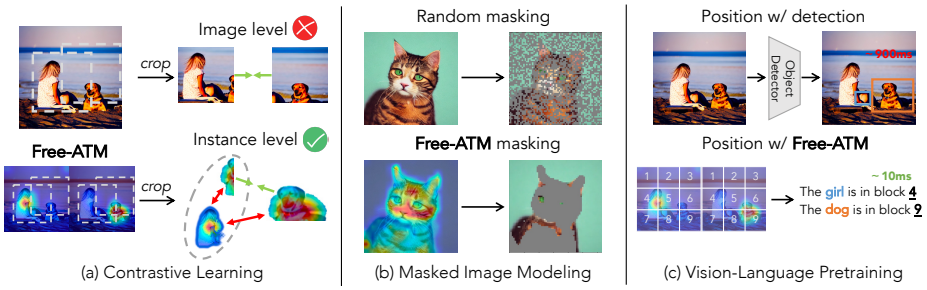

Fig. 2: Existing problems in current representation learning frameworks on synthetic images and our Free-ATM solutions.

by leveraging Free-ATM, the performance gap between representation learning on synthetic data and real-world scenarios can possibly be closed. Moreover, mixing synthetic data with real data for pretraining can further boost performance.

2 Related Work

Text-to-Image Diffusion Models. Diffusion models [14, 32, 33, 61, 62] have been making waves in the generative model's landscape, primarily due to their unique approach to generating new data samples. These models commence with a straightforward random noise and progressively denoise it through numerous steps with learned transformations until it mirrors a sample from the desired data distribution. Recently, large-scale text-to-image diffusion models such as Stable Diffusion [56], Imagen [58], and GLIDE [49] have made considerable strides and produced striking visual outcomes. Rombach *et al.* propose an approach where the diffusion process takes place in the latent space, utilizing a UNet [57] to predict noise and a VAE [37] decoder to convert the latent feature into pixel space. This approach streamlines the text-to-image diffusion process and accelerates it, becoming a popular choice for diffusion models. Building on the latent diffusion model, Hertz *et al.* [31] discover that the cross-attention map, which is employed for text-visual interaction in UNet, can accurately represent the foreground object when given the appropriate prompt. They leverage this characteristic to manipulate images as per requirements by directly modifying the attention map. Similarly, Zhao *et al.* [78] use the UNet of the diffusion models as the core structure and extract the cross-attention map to boost the zero-shot segmentation performance.

In our study, we use the latent diffusion model to generate images in line with the text prompts and to extract the cross-attention map. Using human annotation-free sources, we investigate how attention maps, as an additional resource, can enhance representation learning on synthetic images.

Synthetic Data from Generative Models. Generative Adversarial Networks (GANs) [19] have the capability to produce highly realistic and superior quality images. There are numerous studies that utilize GANs to synthesize datasets akin

to ImageNet [13]. Li *et al.* [41] present BigdatasetGAN, a method that generates a vast amount of images corresponding to ImageNet classes with pixel-level labels, but it necessitates additional human annotation. Jahanian *et al.* [34] propose a unique method that employs a GAN to generate multiple views of an image, which are then used as positive pairs for contrastive learning. Recently, diffusion models have been gaining precedence in the field of image generation. He *et al.* [27] demonstrate that innovative text-to-image diffusion models, such as GLIDE [49], can produce data in a custom label space for image recognition. Furthermore, works by Azizi *et al.* [2] and Mert Bulent *et al.* [59] employ Imagen [58] and Stable Diffusion [56] to generate images with class labels, thereby improving supervised classification performance. Adding to this, Brandon *et al.* [64] utilize diffusion model inversion to augment images for classifying small datasets.

Rather than focusing on task-specific synthetic data [69, 70, 74], we delve into the broader realm of representation learning on synthetic images. Our work is concurrent with StableRep [63], which uses a single caption to generate multiple images, forming multiple samples in a positive pair. Differently, we leverage the overlooked attention masks, which can offer pixel-level labels without any need for human annotations, benefiting representation learning on synthetic images.

Representation Learning. *(i) Contrastive Learning* is built upon the foundational idea of drawing positive pairs nearer while distancing negative pairs in the representational space. This method has proven to be effective in learning visual representations without the need for labeled data [3, 30, 48, 50, 63, 67, 71, 75]. A significant advancement in this area is the SimCLR framework [9], which has substantially improved the quality of the learned representations via a non-linear transformation head. MoCo [22], another impactful work, maintains a memory bank for a vast array of negative samples, and employs a momentum-based method for gentle updates, ensuring improved consistency during learning. *(ii) Masked Modeling* [4, 5, 21, 68, 80], such as MAE [21], SimMIM [73], and iBOT [80], uses masked patch reconstruction in combination with basic data augmentation to effectively learn robust representations. *(iii) Vision-Language Pre-training (VLP)*, exemplified by models like CLIP [53], BLIP [42], and ViLT [36], seeks to boost performance on downstream vision and language tasks by pretraining the model on large-scale image-text pairs to align visual and text features without any task-specific supervision.

In this work, we harness the Free-ATM from diffusion generators to enhance the above-mentioned three representation learning frameworks on synthetic data.

3 Method

3.1 Attention Mask from T2I Diffusion Model

Latent Text-to-Image Diffusion model [56] is a generative model that uses a text prompt to create high-quality images through a controlled diffusion process. It first encodes the text prompt into a latent space representation, then uses a diffusion process to gradually transform a noise input into the final image, guided

by the encoded text. The model represents the joint understanding of textual and visual data via cross-attention interaction in UNet. Specifically, in the single diffusion step and layer, the text embedding is projected into key as $K \in \mathbb{R}^{L \times C}$ and the visual noise is projected into query as $Q \in \mathbb{R}^{H \times W \times C}$, where L is the text sequence length, H, W are height and width of visual feature and C is the feature dimension. The cross-attention mask is achieved by the multiplication of Q and K, resulting in

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{1}$$

where $A \in \mathbb{R}^{H \times W \times L}$ is the attention mask, illustrating the relationship between textual and visual elements. The attention mask $a \in \mathbb{R}^{H \times W}$, corresponding to specific nouns such as 'dog' in an L length sentence, is selected. We empirically explore how to extract attention masks from the UNet as shown in Fig. 5 and decide to compile the attention masks derived from every layer and timestep within the diffusion models. These masks are then resized and averaged to form a new mask.

It is worth noting that synthetic data generation and representation learning pretraining on synthetic data are two separate stages. We obtain attention masks from the synthetic data generation stage, where we use a conditioned diffusion model to generate images. Diffusion models inherently produce these attention masks during image generation, eliminating any additional computational cost.

3.2 Prompts Generation

We use ImageNet [13]'s label-space as prompt bases for generating synthetic images, aiming for a broad variety to aid in learning universal representations. To diversify these images, we transform labels into sentences using a large language model, avoiding unrealistic prompts that could harm downstream task performance. For more realistic and varied images, we employ GPT-3.5-turbo for prompt augmentation. Our templates are adapted to ImageNet's class hierarchy, with formats like *"[Class (with other class)] is/are [somewhere]" or "[Class] with [other class] is/are [doing something] [somewhere]"*.

3.3 Free-ATM for Representation Learning on Synthetic Images

Upon generating diverse images and complimentary attention masks via augmented prompts used as input to diffusion models, we suggest modifications to three common representation learning frameworks on synthetic data: Contrastive Learning, Masked Image Modeling, and Vision-Language Pretraining. These proposed adjustments fully exploit the attention masks to address the problem of representation learning on synthetic data.

1) Contrastive Learning. We adopt two widely used contrastive learning methods - SimCLR [9] and MoCo-v2 [10], to incorporate the use of free attention masks. For ease of explanation, we use SimCLR as a representative example, as depicted in Fig. 3 (a). As previously discussed in the introduction, using

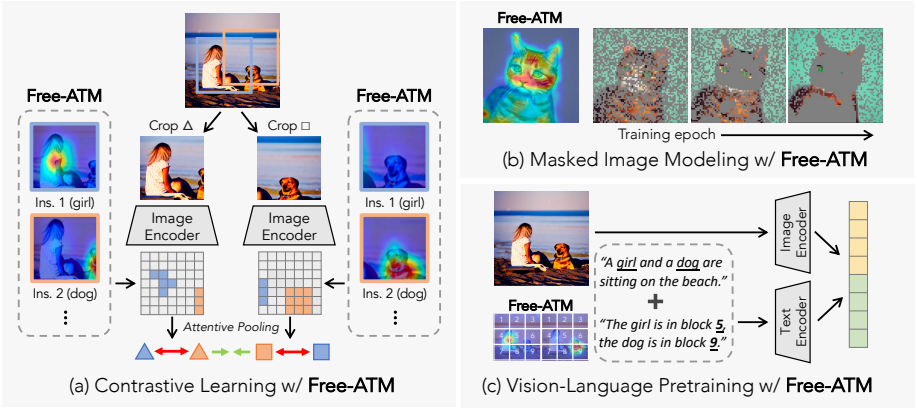

Fig. 3: Different representation learning frameworks with Free-ATM. Free attention masks, obtained during the synthesizing stages with synthetic images, can easily be adapted for use in various representation learning methods on synthetic images.

image-level features can be problematic when a synthetic image contains multiple instances, such as a girl and a dog. The augmented positive pairs, after random cropping, may contain different instances. This could negatively impact network training, as the distinct instance features that should be differentiated are instead conflated. To mitigate this, we use instance features based on the attention mask in place of image features.

Specifically, given an image, we apply a series of augmentations and a random crop, resulting in two cropped images, denoted as x and x'. Concurrently, our instance attention masks are cropped in line with the image operation, yielding two sets of masks $a^1, \cdots, a^N$, $a'^1, \cdots, a'^N$, with each set containing N instances. These two image crops are then input into the encoder (e.g., ResNet50 [25]), resulting in outputs $z = f(x)$, $z' = f(x')$. At this stage, we utilize the features $z \in \mathbb{R}^{h \times w \times c}, z' \in \mathbb{R}^{h \times w \times c}$, which are derived prior to the last average pooling layer. Next, the $m-th$ instance attention masks a^m, a'^m, resized to match the spatial resolution of the encoded features, are mapped to the features z and z', thereby applying attentive pooling. This process results in:

$$z^m = \frac{1}{\sum_{i,j} a^m_{i,j}} \sum_{i,j}^{h,w} a^m_{i,j} z_{i,j}, \tag{2}$$

and $z'^m \in \mathbb{R}^C$. Following the application of attention pooling, the features z^m and z'^m are transitioned from the image level to the instance level. Subsequently, a straightforward Multilayer Perceptron (MLP) layer is applied to these features. For instance-level features, we redefine the contrastive loss for z^m as

$$l = -\log \frac{\exp([z^m, z'^m])}{\exp([z^m, z'^m]) + \sum_{n=1[m \neq n]}^{N} \exp([z^m, z^n])}, \tag{3}$$

where $[z^m, z'^m] = \frac{z^{m\top} \cdot z'^m}{|z^m| \cdot |z'^m|}$. In this loss computation, we consider the features of the same instance from the two crops as the positive pair, and the features of different instances as negative pairs. In Equation (3), N signifies the total number of instances in the image. When extended to the batch dimension, N can represent the total number of instances in the batch.

For MoCo-v2, the encoders for the two image crops are distinct. One encoder is updated by the Exponential Moving Average (EMA) [22] of the other encoder. Furthermore, instance-level features, as opposed to image-level, are updated and stored in the memory bank. In addition to addressing the issue where positive pairs may comprise different instances, this strategy enables every image to provide a wealth of information. This greatly aids the network and allows for the learning of a more diverse representation, given the fact that each image typically has multiple instances.

Differences from object prior methods [28, 29]. They need to use non-trivial process to gradually get the object priors during the pre-training stage. For example, [28] uses a K-Means to determine the K instances masks in each training step, which needs multiple iterations and is quite time-consuming. In contrast, our approach leverages the Free-ATM of the diffusion model, offering a direct and effective object localization without added steps in pre-training.
2) Masked Image Modeling. Diffusion model might generate images with plain or empty background. When pretraining on synthetic images, network will learn useless information by reconstructing randomly masked patches in empty background. Our Free-ATM naturally embodies the importance score of the foreground object mask, thus removing the blank background parts. The scores of the attention mask, ranging from low to high, correspond to the probability of foreground patches.

An intuitive approach would be to mask the patches with the highest attention mask scores and then use pretraining from scratch to reconstruct the masked images. Nonetheless, despite the presence of synthetic images with empty backgrounds, a significant number of them still feature meaningful backgrounds. Therefore, focusing solely on reconstructing foreground patches might hinder the development of a more comprehensive representation of the entire set of synthetic images. To mitigate this and make a balance, during the initial stages of training, we continue to mask the patches randomly. As the training epochs increase (Fig. 3 (b)), we gradually raise the ratio of masked patches determined by the highest importance scores, and reduce the ratio of randomly selected masked patches. This balanced approach enables the network to learn both universal and targeted representations simultaneously.

Differences from HPM [66]. HPM introduces an additional teacher-student network to predict foreground patches. Meanwhile, the teacher network requires an extra learnable decoder with 8 transformer blocks for mask prediction during pre-training, increasing GPU memory and training duration. Instead, we do not need an additional teacher network and a heavy decoder. Our attention masks are directly from the diffusion model and can naturally fit in MAE [21] framework during pre-training w/o added training cost.

3) Vision-Language Pretraining.

Position grounding in vision-language models [43, 46] is crucial for cross-modality tasks. Our Free-ATM inherently includes the bounding box of the object (nouns). This is made possible as we can readily transform the attention mask into a binary mask, with pixels marked as '1' representing the foreground region, thereby allowing us to obtain the bounding box. Rather than extracting regions using bounding boxes as inputs for the visual encoder, we employ position-aware prompts, inspired by [65], which does not impose additional parameters or computational demands on the vision-language model. As illustrated in Fig. 3 (c), we initially divide the image into N blocks. After determining the bounding box of the object, we can identify the block in which the object's center is located. Using this information, we generate position-aware prompts following the template: "*The [O] is in block [P].*"

Subsequently, we concatenate these prompts for all objects in the images with the original prompt. This forms the input for the text encoder, enhancing the position-aware ability of the vision-language model. The model in focus, BLIP, is adapted to our needs for Vision-Language Pretraining (VLP). We then conduct end-to-end training using conventional objectives. Consistent with the methods outlined in [36, 42, 54], the training process involves the use of Language Modeling (LM) loss, Image-Text Matching (ITM) loss, and Image-Text Contrastive (ITC) loss. Note that the object's positional information is only required during the pre-training stage. For downstream tasks, we evaluate the model using standard end-to-end methods, without the need for object information.

Differences from PTP [65]. PTP obtains the bounding box by an offline detection model like Fast-RCNN [18], which is slower. On the contrary, Free-ATM does not need such an object detector. Moreover, the offline detector of PTP is trained on close-domain real data so it might fail to detect objects in out-of-domain synthetic images. In contrast, Free-ATM is extracted from synthetic data generation process and each object has a correlated attention mask for location.

4 Experiments

4.1 Implementation Details of Settings

The following outlines some setting details of our experiments. **1)** In pretraining for CL (Tab. 1) and MIM (Tab. 2), we generate 1.2 million images using augmented prompts, a quantity that matches precisely the original ImageNet-1K [13] for fair comparisons. For pretraining in the VL task (Tab. 3) , we select a subset from CC3M [60], which encompasses 0.3 million image-text pairs, and employ the original captions to generate images. For scalability experiments in Fig. 6, we use 14M prompt in text-image pairs from COCO [45], Visual Genome [40], Conceptual 12M [8] and SBU caption [51], following BLIP [42]. **2)** CL and MIM experiments directly employ the attention masks as described in Section 3.1, without any further processing. For VLP, we convert the attention masks into binary masks using a fixed threshold of 0.45. This creates binary masks that

Table 1: SimCLR [9] and MoCo-v2 [10] downstream results. SimCLR and MoCo-v2 on pure synthetic images are our pretraining baselines. With Free-ATM(w/ ours), pretraining on pure synthetic images(DeepIF) achieves consistent improvements over baselines, even leading to better results than pretraining on real images.

Images		PASVOC	COCO		Cityscapes	ImageNet
		AP^b_{50}	AP^b	AP^m	$mIoU$	Lin. acc
Real	random ini	59.0	31.4	28.5	65.2	–
	supervised	81.6	39.2	35.5	74.2	–
	SimCLR	80.4	37.7	34.2	74.8	66.8
	MoCo-V2	82.4	39.8	36.9	75.0	67.3
Synthetic	Stable Diffusion 1.5					
	SimCLR	79.1	36.4	33.1	73.7	64.4
	w/ ours	80.0 (↑0.9)	37.8 (↑1.4)	34.2 (↑1.1)	74.7 (↑1.0)	65.4 (↑1.0)
	MoCo-v2	81.6	37.9	35.2	74.1	65.1
	w/ ours	82.2 (↑0.6)	39.6 (↑1.7)	36.4 (↑1.2)	74.9 (↑0.8)	66.2 (↑1.1)
	DeepIF					
	SimCLR	79.5	37.1	33.5	74.2	65.6
	w/ ours	80.4 (↑0.9)	38.4 (↑1.3)	34.6 (↑1.1)	75.0 (↑0.8)	66.5 (↑0.9)
	MoCo-v2	81.9	38.8	35.5	74.6	65.9
	w/ ours	**82.5** (↑0.6)	**40.6** (↑1.8)	**37.1** (↑1.6)	**75.2** (↑0.6)	**67.4** (↑1.5)
Mixed	Stable Diffusion 1.5					
	SimCLR	79.8	37.5	33.7	74.1	66.3
	w/ ours	80.6 (↑0.8)	38.7 (↑1.2)	34.8 (↑1.1)	75.0 (↑0.9)	67.2 (↑0.9)
	MoCo-v2	82.2	39.2	36.0	74.6	66.8
	w/ ours	82.8 (↑0.6)	40.8 (↑1.6)	37.2 (↑1.2)	75.3 (↑0.7)	67.8 (↑1.1)
	DeepIF					
	SimCLR	80.9	38.0	34.0	74.4	66.5
	w/ ours	81.8 (↑0.9)	39.5 (↑1.5)	35.3 (↑1.3)	75.3 (↑0.9)	67.4 (↑0.9)
	MoCo-v2	82.4	39.8	36.4	75.1	67.1
	w/ ours	**83.1** (↑0.7)	**41.2** (↑1.4)	**37.6** (↑1.2)	**75.9** (↑0.8)	**68.0** (↑0.9)

Table 2: MAE [21]downstream results.

		COCO		ImageNet	ADE20K
		AP^b	AP^m	acc	$mIoU$
Real	supervised	47.9	42.9	81.0	47.4
	MAE	50.3	44.9	83.6	48.1
Synthetic	Stable Diffusion 1.5				
	MAE	48.0	42.7	82.7	47.6
	w/ ours	48.9 (↑0.9)	43.7 (↑1.0)	83.4 (↑0.7)	48.4 (↑0.8)
	DeepIF				
	MAE	49.5	44.1	83.0	47.9
	w/ ours	**50.5** (↑1.0)	**44.9** (↑0.8)	**83.7** (↑0.7)	**48.6** (↑0.7)
Mixed	Stable Diffusion 1.5				
	MAE	49.9	44.7	83.3	48.2
	w/ ours	50.8 (↑0.9)	45.4 (↑0.7)	83.9 (↑0.6)	48.7 (↑0.5)
	DeepIF				
	MAE	50.2	44.6	83.5	47.9
	w/ ours	**51.0** (↑0.8)	**45.6** (↑1.0)	**84.2** (↑0.7)	**48.8** (↑0.9)

Table 3: BLIP [42] retrieval results. Incorporating Free-ATM (w/ ours) in pretraining leads to a marked improvement compared to the BLIP baseline.

	MS-COCO	finetune		zero-shot	
		tr@1	ir@1	tr@1	ir@1
Real	BLIP	58.1	44.2	42.1	30.4
Synthetic	Stable Diffusion 1.5				
	BLIP	52.3	40.9	23.2	20.9
	w/ ours	54.9 (↑2.6)	43.8 (↑2.9)	31.8 (↑8.7)	28.2 (↑7.3)
	DeepIF				
	BLIP	56.8	42.7	35.6	24.2
	w/ ours	**59.0** (↑2.2)	**44.8** (↑2.1)	**43.1** (↑7.5)	**31.1** (↑6.9)
Mixed	Stable Diffusion 1.5				
	BLIP	54.9	42.1	34.2	25.3
	w/ ours	60.8 (↑5.9)	46.2 (↑4.1)	42.3 (↑8.1)	30.6 (↑5.3)
	DeepIF				
	BLIP	57.5	43.8	38.9	27.5
	w/ ours	**61.4** (↑3.9)	**47.1** (↑2.3)	**43.7** (↑4.8)	**32.0** (↑4.5)

identify the image blocks where the object prompt is likely located. **3)** For Tab. 1-3, the 'Real', 'Synthetic', and 'Mixed' sections refer to the use of exclusively real data, exclusively synthetic data, and a 50:50 combination of both, respectively, **with equal image counts in each for fair comparisons**. Under 'Mixed' section, we follow [27] to do pretraining on real images first and then on synthetic images with our Free-ATM(w/ours). **4)** 'Random init' means starting without pretraining. 'SimCLR, MoCo-v2, MAE, BLIP' indicates regular representation learning on synthetic data as baselines. 'w/ ours' signifies our adapted method using synthetic images with Free-ATM. **5)** All results including baselines about synthetic images use augmented prompts so the improvement is solely brought by Free-ATM.

4.2 Contrastive Learning

SimCLR and MoCo-v2 Pretraining. We employ ResNet50 [26] as the encoder for our pretraining. To maintain fairness in our comparisons, we adhere to the same 200 pretraining epochs across all settings, and all hyperparameters for training are aligned with those outlined in the original SimCLR [9] and MoCo-v2 [10] papers. As shown in Tab. 1, SimCLR and MoCo-v2 serve as our baselines for synthetic or mixed images. Free-ATM (w/ ours) can be seamlessly integrated into these frameworks without altering training strategies.

Object Detection, Segmentation and Linear Probing. The pretrained network is utilized to initialize our feature extractor. For object detection in PASVOC [17], we adhere to the standard protocol of fine-tuning a Faster R-CNN detector (with a C4 backbone) on the VOC trainval07+12 set, as per the 2x schedule mentioned in [67]. The evaluation is then carried out on the VOC test2007 set. For object detection and instance segmentation in COCO [45], we, in accordance with [22], modify the Mask-RCNN [23] to be equipped with FPN [44]. The complete model is fine-tuned on the training dataset, following a standard 1x schedule (12 epochs), and we report the results as bounding box AP (AP^b) and instance mask AP(AP^m). In the case of semantic segmentation for Cityscapes [12], we fine-tune the model over a span of 160 epochs, reporting the results in terms of $mIoU$ (mean Intersection over Union). For linear classification, we follow the MoCo [22] to freeze the features and train a supervised linear classifier (a fully-connected layer followed by softmax) for 100 epochs.
Results. As shown in Tab. 1, contrastive pretraining using purely synthetic data(generated by Stable Diffusion) leads to a noticeable performance discrepancy on the downstream tasks when compared to the use of purely real data. However, by incorporating Free-ATM (w/ ours), we observe significant performance gains e.g., SimCLR and MoCo-v2 see improvements of 1.4%/1.1% and 1.7%/1.2% on the COCO dataset, respectively. This effectively bridges the performance gap between synthetic and real-data training. Notably, with higher-quality synthetic images (using DeepIF [1]), Free-ATM eliminates this performance gap between real and synthetic images entirely. These results highlight the power of Free-ATM for instance-level contrastive learning on synthetic images.

Furthermore, combining real and synthetic data with Free-ATM leads to even greater performance gains, outperforming models trained on real images alone. Importantly, this is achieved without the need for costly human annotation or data collection, if a model for Text-to-Image diffusion is provided.

4.3 Masked Image Modeling

MAE Pretraining. We employ MAE [21], a representative masked modeling method. We follow its original pretraining protocol for the ViT-B [15], spanning a total of 1600 epochs. The overall mask ratio is set at 75%. Additionally, we adopt a strategy to incrementally increase the ratio β of masked patches according to the highest attention score, with a ceiling of 0.8, following a linear progression as epochs increasing. Thus, a portion of the masked patches, specifically $(\beta \times 0.75)$ of them, are determined based on the attention score. The remaining masked patches, which account for $((1-\beta) \times 0.75)$ of the total, are selected at random.
Image Classification, Object Detection and Semantic Segmentation. For classification tasks, we carry out fine-tuning for 100 epochs on the ImageNet-1K training set and subsequently report the top-1 accuracy on the validation set. For detection tasks, we fine-tune Mask R-CNN [24] end-to-end on COCO. The ViT-B backbone is adapted for use with FPN [44]. For segmentation tasks, we use UperNet [72] as the segmentation head. This is then subjected to end-to-end fine-tuning on the ADE20K [79] dataset for a total of 160k iterations.

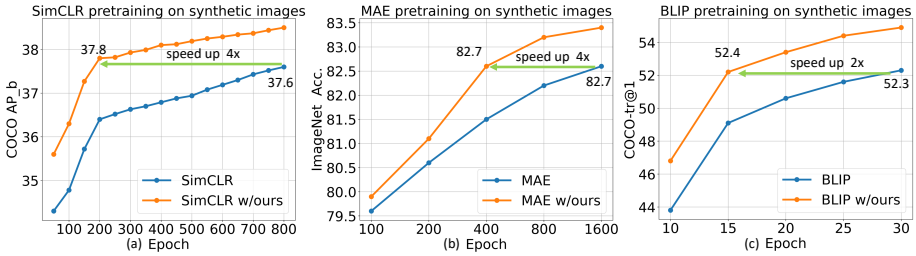

Fig. 4: Pretraining Acceleration on synthetic images. (a) shows that by using Free-ATM, SimCLR pretraining on synthetic images for 200 epochs can yield same outcomes to what would otherwise necessitate 800 epochs of training on synthetic images without Free-ATM, achieving 4 times speed up. (b) and (c) indicate a 4× and 2× acceleration in MAE and BLIP pretraining, respectively.

Results. Tab. 2 demonstrates that Free-ATM (w/ ours) consistently enhances MAE(baseline) performance on purely synthetic images from Stable Diffusion and Deep-IF across various downstream tasks. This aligns with the trend observed when comparing MAE [21] to contrastive learning. Our findings suggest that Free-ATM also benefits masked image modeling on synthetic images and helps bridge the performance gap between synthetic and real-world data.

4.4 Vision Language Pretraining

Pretraining. We train the vision-language model following BLIP [42] on 0.3 M text-image pairs until the loss converges.

COCO-Retrieval. We evaluate BLIP on the COCO datasets, focusing on image-to-text retrieval (TR) and text-to-image retrieval (IR). The pre-trained model is fine-tuned using both ITC and ITM loss functions. Additionally, we implement a re-ranking strategy to further refine the retrieval results.

Results. Tab. 3 illustrates that with free attention mask, which provides position-aware prompts, enables the model to attain a significant improvement over baseline results obtained using purely synthetic data, especially for zero-shot testing with 8.7% and 7.3% improvements. Furthermore, combining synthetic data with the mask and real data considerably enhances the results. Different from other experiments pretraining on synthetic ImageNet, which use simple prompts to synthesize images, BLIP experiments were conducted on images synthesized by complex captions from CC3M, which proved challenging for Stable Diffusion [56] to create well-text-aligned images. This image misalignment leads to synthetic images underperforming zero-shot text-image retrieval significantly compared to real ones unless it is corrected through fine-tuning. However, as shown in Tab. 3, if we use a more text-aligned image generator like DeepIF [1], BLIP can enhance zero-shot performance with synthetic images to match that of real data. This proves our method's generalization ability to unseen classes and its potential for improved results with more advanced text-to-image models.

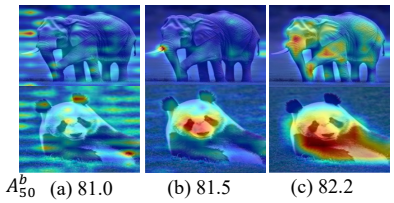

Fig. 5: Different attention mask extractions from (a) single diffusion step and single unet block, (b) average of all diffusion steps and down-blocks in UNet and (c) average of all diffusion steps and all blocks in UNet. Evaluation is conducted with MoCo-V2 VOC detection.

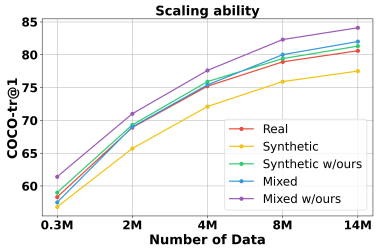

Fig. 6: Scaling ability of our Free-ATM. As the number of synthetic images for BLIP pretraining rises, Fee-ATM consistently enhances performance.

4.5 Pretraining Acceleration

Our method not only improves the results but also accelerates the pre-training process on synthetic data significantly. Specifically, as shown in Fig. 4(a) our method boosts contrastive pre-training (SimCLR) speed by 4 times. With our method, synthetic data pretraining hits 37.8 in just 200 epochs, compared to 800 epochs needed to reach 37.6 without Free-ATM. Similarly, as shown in Fig. 4(b)(c), Free-ATM can accelerate the mask image modeling and vision language pretraining process by 4 and 2 times, respectively.

4.6 Ablation Study

Attention Masks Selections. We use attention masks from different layers and timesteps of UNet to study mask quality's effect on methods. Differences in attention mask quality can be seen in Fig. 5, which also shows that better masks lead to better pretraining results. We find the attention masks from the average of all diffusion steps and all blocks in UNet has the best visual quality and use them as the default setting.

Attention Masks Quality. To further evaluate the quantitative quality of the attention masks, we provide the m-IOU results in the Tab. 4. First, we use DenseCRF [39] to transform our attention masks into binary masks. Second, we use the SAM [38] to extract the instance mask as the ground truth and then calculate the m-IOU. For quick implementation, we select synthetic images which contain one instance, including 23 animal classes and around 10000 images. As shown in Tab.4, our attention mask achieves 38.8 m-IOU. This result is very similar to the unsupervised results of slot attention [47] (specifically designed to extract foreground object masks). Therefore, these results validate that the attention masks aren't random and their quality is good enough to allocate the foreground object. Furthermore, as shown in Tab. 5, Free-ATM outperforms PTP [65] with object detector, which also indicates Free-ATM's good location quality and its enhanced capability over object detectors trained on closed domain data, particularly for synthetic data.

Table 4: Attention masks quality of our method.

	mIOU
Slot Attention [47]	38.6
Ours	**38.9**

Table 5: Comparisons with PTP [65] (object detector).

	finetune		zero-shot	
BLIP [42]	*tr*@1	*ir*@1	*tr*@1	*ir*@1
PTP [65]	53.8	42.5	28.2	26.5
Ours	**54.9**	**43.8**	**31.8**	**28.2**

Table 6: Ablation of prompt designs.

		base	augment
Moco-V2 [10]	AP^b_{50}	81.0	**81.6**
MAE [21]	*Acc.*	81.8	**82.7**

Table 7: Ablation of different text-to-image models.

	finetune		zero-shot	
BLIP [42]	*tr*@1	*ir*@1	*tr*@1	*ir*@1
VQGAN [16]	35.6	32.0	17.1	15.4
DALL-E2 [55]	44.5	38.6	18.5	15.6
Stable [56]	52.3	40.9	23.2	20.9
DeepIF [1]	**56.8**	**42.7**	**35.6**	**24.2**

Image Quality. We assess the quality impact of synthetic images produced by various text-to-image models. DeepIF [1] excels in generating images of superior visual quality and more accurate text-image alignment compared to Stable Diffusion [56], VQGAN [16], and DALL-E2-LAION [55]. We conduct BLIP [42] pretraining on images synthesized by these four generators. Tab. 7 reveals that higher image quality and better text-image alignment correlate with improved results. This suggests that as more powerful generators emerge in the future, the prospect of pretraining on synthetic data becomes increasingly promising.

Prompts Design. We investigate the significance of augmented prompts (Section 3.2) in our study, simply employing a basic prompt, "a photo of [class]," for comparison purposes. As indicated in Tab. 6, augmented prompts are conducive to the generation of a broader range of images, enhancing the pretraining process. Therefore, we use the augmented prompts to generate more diverse images, which benefits the learning of visual representations.

Data Scalability. We assess the effects of increasing number of diffusion-generated images using Free-ATM. As shown in Fig. 6, from a scale of 0.3M to 14M, Free-ATM consistently elevates the performance of BLIP pretraining on synthetic data. The findings demonstrate performance enhancements of Free-ATM as the volume of synthetic images grows, highlighting its scalability comparable to real images. Moreover, mixing synthetic and real data with Free-ATM is able to achieve scalable and superior performance compared to using purely real data.

5 Conclusion

We have proposed Free-ATM, a technique that utilizes the readily available attention masks from diffusion generators, to improve representation learning on diffusion-generated images. We anticipate that our approach will provide new directions for future research on leveraging synthetic images to tackle computer vision problems. We anticipate that our approach will provide new directions and insights for future research on leveraging synthetic images to tackle computer vision problems.

Acknowledgement

David Junhao Zhang and Mike Zheng Shou are supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. David Junhao Zhang is also supported by NUS IDS-ISEP schoalrship. Mutian Xu and Xiaoguang Han are supported in part by NSFC-62172348, Guangdong Provincial Outstanding Youth Fund (No. 2023B1515020055), NSFC-61931024, and Shenzhen Science and Technology Program No. JCYJ20220530143604010.

References

1. https://github.com/deep-floyd/IF 11, 12, 14
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023) 5
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: NeurIPS (2019) 5
4. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022) 5
5. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: ICLR (2022) 3, 5
6. Besnier, V., Jain, H., Bursuc, A., Cord, M., Pérez, P.: This dataset does not exist: Training models from generated images. In: ICASSP (2020) 2
7. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019) 2
8. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 9
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 1, 3, 5, 6, 10
10. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3, 6, 10, 14
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021) 1, 3
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 3, 11
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 1, 3, 5, 6, 9
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021) 4
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 11
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 14
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010) 3, 11

18. Girshick, R.: Fast r-cnn. In: ICCV (2015) 9
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM (2020) 4
20. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: NeurIPS (2020) 1, 3
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022) 1, 3, 5, 8, 10, 11, 12, 14
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 1, 3, 5, 8, 11
23. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 11
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 11
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 7
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10
27. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., QI, X.: IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION? In: ICLR (2023) 2, 5, 10
28. Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: ICCV (2021) 8
29. Hénaff, O.J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J., Arandjelović, R.: Object discovery and representation networks. In: ECCV (2022) 8
30. Hénaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S., Oord, A.v.d.: Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272 (2019) 5
31. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 2, 4
32. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 4
33. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 4
34. Jahanian, A., Puig, X., Tian, Y., Isola, P.: Generative models as a data source for multiview representation learning. In: ICLR (2022) 5
35. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. TPAMI (2021) 2
36. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML (2021) 1, 3, 5, 9
37. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4
38. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. ICCV (2023) 13
39. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NeurIPS (2011) 13
40. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language

and vision using crowdsourced dense image annotations. International journal of computer vision (2017) 9
41. Li, D., Ling, H., Kim, S.W., Kreis, K., Barriuso, A., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: CVPR (2022) 2, 5
42. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022) 1, 3, 5, 9, 10, 12, 14
43. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: CVPR (2022) 9
44. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 11
45. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 3, 9, 11
46. Liu, Z., Stent, S., Li, J., Gideon, J., Han, S.: Loctex: Learning data-efficient visual representations from localized textual supervision. In: ICCV (2021) 9
47. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. NeurIPS (2020) 13, 14
48. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR (2020) 5
49. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022) 2, 4, 5
50. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 5
51. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS (2011) 9
52. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: ICCV (2017) 2
53. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 3, 5
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 9
55. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021) 14
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 4, 5, 12, 14
57. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 4
58. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) 4, 5
59. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR (2023) 5

60. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 9
61. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435 (2023) 4
62. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 4
63. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: ECCV (2019) 5
64. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023) 5
65. Wang, A.J., Zhou, P., Shou, M.Z., Yan, S.: Position-guided text prompt for vision-language pre-training. In: CVPR (2023) 9, 13, 14
66. Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., Zhang, Z.: Hard patches mining for masked image modeling. In: CVPR (2023) 8
67. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: CVPR (2021) 5, 11
68. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR (2022) 5
69. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. NeurIPS (2023) 5
70. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. arXiv preprint arXiv:2303.11681 (2023) 5
71. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018) 5
72. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (2018) 11
73. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR (2022) 1, 3, 5
74. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. In: NeurIPS (2023) 5
75. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: CVPR (2019) 5
76. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: CVPR (2021) 2
77. Zhao, B., Bilen, H.: Synthesizing informative training samples with gan. NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (2022) 2
78. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. arXiv preprint arXiv:2303.02153 (2023) 2, 4
79. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019) 3, 11
80. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: ICLR (2022) 5