Visual Prompt Tuning for Generative Transfer Learning

Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania,
Han Zhang, Yuan Hao, Irfan Essa, Lu Jiang
Google Research

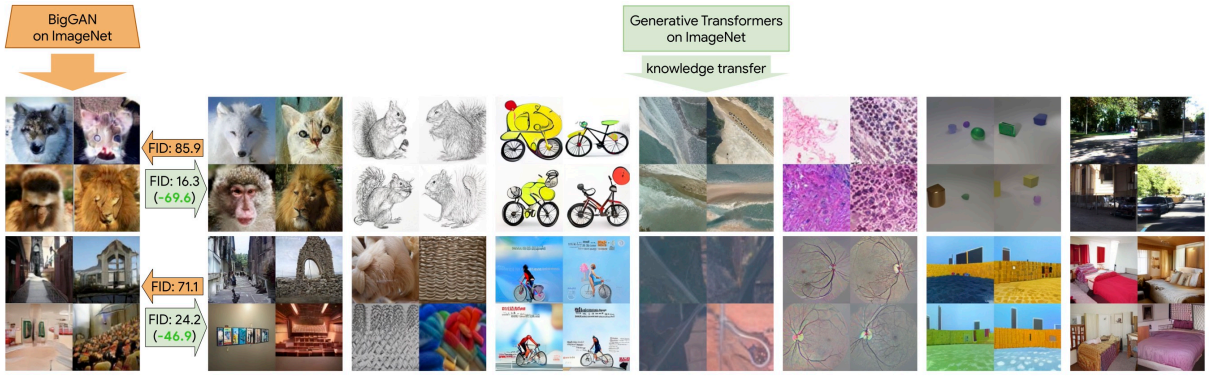


Figure 1. Image synthesis by knowledge transfer. Unlike previous works using GANs as base model and test transfer on relatively narrow visual domains, we transfer knowledge of generative vision transformers [7, 15] to a wide range of visual domains, including natural (*e.g.*, scene, flower), specialized (*e.g.*, satellite, medical), and structured (*e.g.*, road scenes, infograph, sketch) with a few training images. Notably, the prompt tuning significantly improves the prior best FID on two benchmarks ImageNet (85.9→16.3) and Places (71.3→24.2).

Abstract

Learning generative image models from various domains efficiently needs transferring knowledge from an image synthesis model trained on a large dataset. We present a recipe for learning vision transformers by generative knowledge transfer. We base our framework on generative vision transformers representing an image as a sequence of visual tokens with the autoregressive or non-autoregressive transformers. To adapt to a new domain, we employ prompt tuning, which prepends learnable tokens called prompts to the image token sequence and introduces a new prompt design for our task. We study on a variety of visual domains with varying amounts of training images. We show the effectiveness of knowledge transfer and a significantly better image generation quality.[1]

1. Introduction

Image synthesis has witnessed tremendous progress recently with the advancement of deep generative models [2, 12, 20, 67, 69]. An ideal image synthesis system generates diverse, plausible, and novel scenes capturing the appearance of objects and depicting their interactions. The success of image synthesis does heavily rely on the availability of a large amount of diverse training data [73].

Transfer learning, a cornerstone invention in deep learning, has proven indispensable in an array of computer vision tasks, including classification [35], object detection [18,19], image segmentation [23, 24], *etc*. However, transfer learning is not widely used for image synthesis. While recent efforts have shown success in transferring knowledge from pre-trained Generative Adversarial Network (GAN) models [46, 60, 71, 76], their demonstrations are limited to narrow visual domains, *e.g.*, faces or cars [46, 76], as in Fig. 1, or requiring a non-trivial amount of training data [60, 71] to transfer to out-of-distribution domains.

In this work, we approach transfer learning for image synthesis using generative vision transformers, an emerging class of image synthesis models, such as DALL·E [53], Taming Transformer [15], MaskGIT [7], CogView [13], NÜWA [75], Parti [79], among others, which excel in im-

[1] https://github.com/google-research/generative_transfer

age synthesis tasks. We closely follow the recipe of transfer learning for image classification [35], in which a source model is first trained on a large dataset (*e.g.*, ImageNet) and then transferred to a diverse collection of downstream tasks. Except, in our setting, the input and output are reversed and the model generates images from a class label.

We present a transfer learning framework using *prompt tuning* [38,40]. While the technique has been used for transfer learning of discriminative models for vision tasks [1,29], we appear to be the first to *adopt prompt tuning for transfer learning of image synthesis*. To this end, we propose a parameter-efficient design of a prompt token generator that admits condition variables (*e.g.*, class), a key for controllable image synthesis neglected in prompt tuning for discriminative transfer [29, 38]. We also introduce a marquee header prompt that engineers learned prompts to enhance generation diversity while retaining the generation quality.

We conduct a large-scale study to understand the mechanics of transfer learning for generative vision transformers. Two types of generative transformers – *AutoRegressive (AR)* and *Non-AutoRegressive (NAR)* – are examined. AR transformers (*e.g.*, DALL·E [53], Taming Transformer [15], Parti [79]) generate image tokens sequentially with an autoregressive language model. NAR transformers (*e.g.*, MaskGIT [7], MUSE [6]) or diffusion models (*e.g.*, Imagen [58], Latent Diffusion [57]) decompose image synthesis as a series of refinement or denoising steps. In this work, we study transfer learning of class-conditional AR [15] and NAR [7] transformer models trained on ImageNet to comply with existing transfer learning settings [60, 71]. In addition to investigating proposed prompt tuning, we also conduct an analysis of two other transfer learning methods, *i.e.* full fine-tuning and adapter tuning, in the context of generative transfer learning using vision transformers. We compare their strengths and weaknesses in Sec. 4.1.

Our study shows that generative vision transformers with prompt tuning outperform state-of-the-art methods using GANs [60, 71] by a vast margin, which is verified on 19 tasks of diverse visual distributions and drastically different amounts of training data in VTAB [81]. Fig. 1 compares domains, showing the great expansion of downstream domains to what is achieved by previous works. On the on-manifold domains on which previous studies have focused, our method slashes the prior state-of-the-art in FID from 71 to 24 on Places [85] and 86 to 16 on Animal Face [61] datasets. Moreover, our method shows highly-competitive data efficiency, generating diverse images following the target distribution when trained from a few images per class.

In summary, our contributions are as follows:

- We present a generative visual transfer learning framework for vision transformers with prompt tuning [38], proposing a new prompt token generator design.
- We conduct a large-scale empirical study for generative transfer learning to validate our proposed prompt tuning and relevant transfer learning methods (*e.g.*, full fine-tuning, adapter tuning) on several visual domains (*e.g.*, VTAB) and scenarios (*e.g.*, few-shot). We show state-of-the-art image synthesis performance.
- To our knowledge, we are first to propose the use of prompt tuning for transfer learning of generative transformers. Importantly, we provide the quantitative evidence on the necessity of generative knowledge transfer on VTAB [81], the common and challenging transfer learning benchmark.

2. Preliminary

2.1. Generative Vision Transformers

This paper uses generative vision transformers to denote vision transformers for image synthesis. Broadly, there are two types of generative transformers, *AutoRegressive (AR)* and *Non-AutoRegressive (NAR)* transformers, both consisting of two stages – image quantization and decoding. The two models share the same first stage: image quantization by a Vector-Quantized (VQ) auto-encoder [15, 54, 67, 78]. The VQ encoder converts image patches into indices (or tokens) in a codebook. The 2D image is then flattened into a 1D sequence to which a special token indicating its class label is prepended.

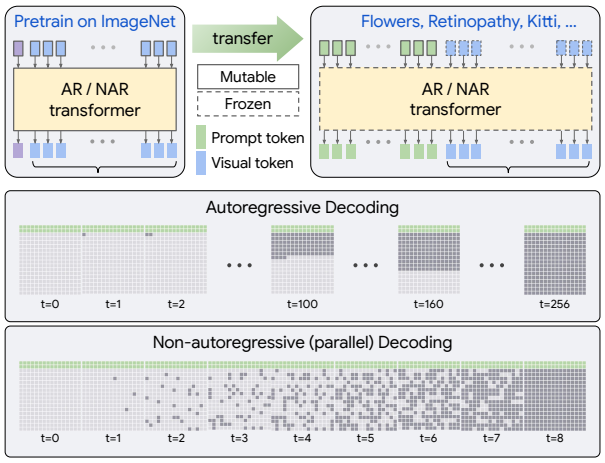
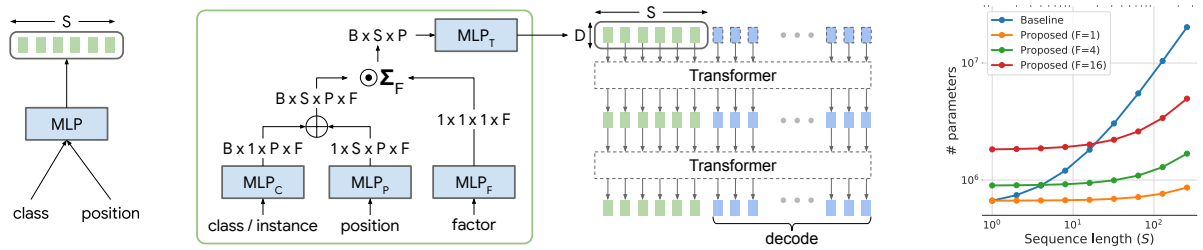


Figure 2. Our method transfers knowledge from generative vision transformers (*e.g.*, autoregressive [15] or non-autoregressive [7]) trained on a large dataset to various visual domains by prepending learnable prompt tokens (green) to visual tokens (blue).

AR and NAR transformers differ in the second stage. AR transformers [8, 13, 15, 53, 75, 79], such as DALL·E [53], Taming Transformer [15], learn an AR decoder on the flattened token sequence to generate image tokens sequentially from previously generated tokens. As in Fig. 2, the generation follows a raster scan ordering, generating tokens from left to right, line-by-line. Finally, the generated tokens are mapped to the pixel space using the VQ decoder.



(a) Baseline prompt token generators of length S conditioned on class.

(b) The proposed parameter efficient prompt token generator via factorization of class / instance and position. $\oplus$ is an element-wise sum, $\odot$ is an element-wise product, Σ_F is a sum over F dimension. S: sequence length, B: batch size, P: feature dimension, D: token dimension.

(c) Number of parameters with respect to the sequence length and different number of factors F.

Figure 3. Prompt token generators and their use in transformer. (a) a straightforward extension of baseline prompt token generators [29, 38, 40] with a class condition. When using an MLP with a single dense layer of P units, the number of trainable parameters is $P{\cdot}(C{\cdot}S{+}D)$. (b) The proposed parameter efficient prompt token generators that factorizes data dependent conditions (e.g., class, instance) and token position. Under a similar design choice as baseline models, the number of trainable parameters is $P{\cdot}(F{\cdot}(C{+}S){+}D)$, which could be significantly fewer when $F \ll \min(C, S)$. (c) Number of parameters for prompt token generators with respect to the sequence length (S), while setting $P = 768$, $D = 768$, and $C = 100$ with different number of factors F.

NAR or diffusion models, including DALL·E 2 [52], MaskGIT [7], Latent Diffusion [57], or Imagen [58], decompose image synthesis as a series of refinement or denoising steps. For prompt tuning, we need a NAR model with the transformer backbone [7, 17, 21, 36, 37, 39, 83], and use a leading NAR image transformer called MaskGIT [7].

NAR transformers are trained on the masked modeling proxy task [11]. For inference, the model adopts a non-autoregressive decoding method to synthesize an image in a few steps [7, 21, 36, 39]. As in Fig. 2, the NAR transformer starts from a blank canvas with all tokens masked, and generates an image in 8 steps or so. In each step, it predicts all tokens in parallel and retains the ones with the highest prediction scores. The remaining tokens are masked out and predicted in the next iteration. NAR transformers [7, 39] have shown faster inference than AR transformers.

2.2. Prompt Tuning

Prompt tuning [38, 40] is introduced recently in natural language processing as a way of efficiently adapting pretrained large language models to downstream tasks. Here, prompt is a sequence of additional tokens prepended to a token sequence. In prompt engineering [3], their values are often chosen by heuristic. On the other hand, in prompt tuning [38, 40], tokens are parameterized by learnable parameters and their parameters are updated via gradient descent to adapt transformers to the downstream tasks. Due to its simplicity and as transformers' central role in language foundation models, prompt tuning has been applied to some vision tasks for knowledge transfer, *e.g.*, image classification [1, 29], detection and segmentation [45], but not yet for image synthesis.

3. Visual Prompt for Generative Transfer

Fig. 2 overviews the proposed generative transfer learning framework. We aim at transferring a generative prior, parameterized by generative vision transformers, while utilizing the same VQ encoder and decoder trained from the large source dataset. We use prompt tuning to adapt to the target distributions while leaving the transformer parameters frozen. We discuss how to learn visual prompts (Sec. 3.1), a new prompt generator for conditional image synthesis (Sec. 3.2), and a prompt design for generating visually diverse images (Sec. 3.3).

3.1. Learning Visual Prompt

A sequence of prompt tokens is prepended to the visual tokens to guide the pretrained transformer models to the target distribution. Prompt tuning, learning the parameters of the token generator, is optimized by gradient descent with respective loss functions, while fixing the parameters of the pretrained transformers. To be specific, let $\mathcal{Z} = \{z_i\}_{i=1}^{H \times W}$ be a sequence of visual tokens (*i.e.*, an output of VQ encoder followed by the vectorization) and $\mathcal{P}_\phi = \{p_{s;\phi}\}_{s=1}^{S}$ be a sequence of prompt tokens. For the AR transformer, the loss is given as follows:

$$\mathcal{L}_{\text{AR}} = \mathbb{E}_{x \sim P_{\mathcal{X}}} \big[-\log P_\theta(\mathcal{Z}|\mathcal{P}_\phi) \big] \tag{1}$$

$$P_\theta(\mathcal{Z}|\mathcal{P}_\phi) = \prod\nolimits_{i=1}^{H \times W} P_\theta(z_i|z_{<i}, \mathcal{P}_\phi) \tag{2}$$

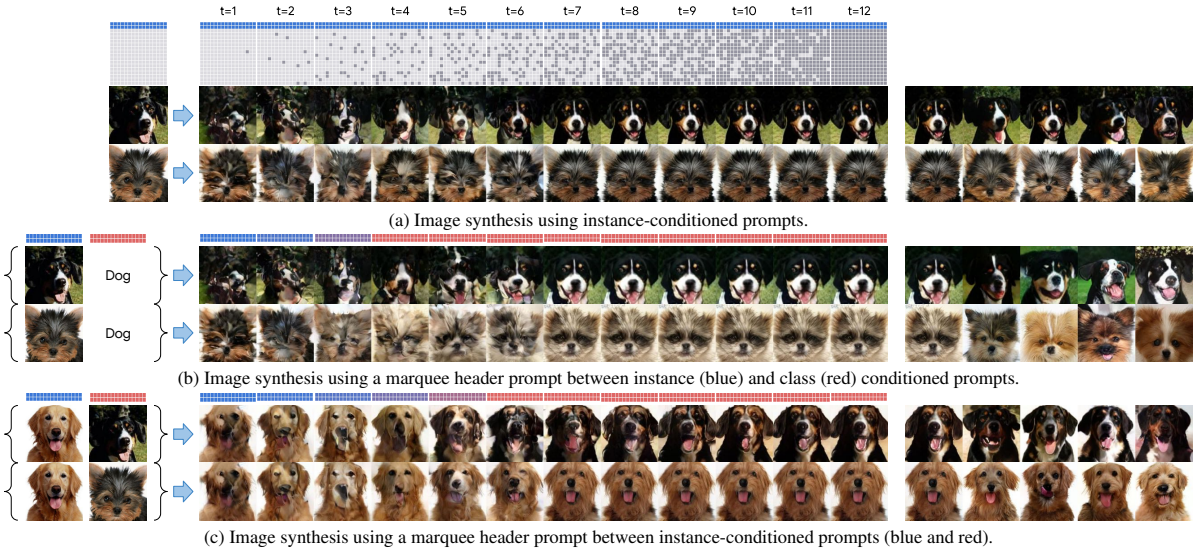
For the NAR transformer, we follow that of MaskGIT [7]:

$$\mathcal{L}_{\text{NAR}} = \mathbb{E}_{x \sim P_{\mathcal{X}}, M \sim P_{\mathcal{M}}} \big[-\log P_\theta(\mathcal{Z}_M|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) \big] \tag{3}$$

$$P_\theta(\mathcal{Z}_M|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) = \prod\nolimits_{i \in M} P_\theta(z_i|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) \tag{4}$$

where $M \subset \{1, ..., H{\times}W\}$ is a set of visual token indices sampled from a masking schedule distribution $P_{\mathcal{M}}$, $\overline{M}$ is its complement, and $\mathcal{Z}_M = \{z_i\}_{i \in M}$. Prompt tuning proceeds by minimizing the respective loss with respect to the prompt parameters ϕ while fixing the transformer parameters θ:

$$\phi^* = \arg\min_{\phi} \mathcal{L}_{\text{AR/NAR}} \tag{5}$$



(a) Image synthesis using instance-conditioned prompts.

(b) Image synthesis using a marquee header prompt between instance (blue) and class (red) conditioned prompts.

(c) Image synthesis using a marquee header prompt between instance-conditioned prompts (blue and red).

Figure 4. Iterative decoding of NAR transformers. (4a) instance prompts generate images of high-fidelity but with low diversity. Marquee header prompts enhance generation diversity by interpolating (4b) from instance to class prompts or (4c) between instance prompts.

While we focus on the prompt tuning due to the virtue of effectiveness and compute-efficiency for large source transformers, we note that the proposed learning framework is amenable with other methods, such as adapter [28] or fine-tuning [35], with learnable prompts. See a detailed comparison in Appendix B.4.

After prompt tuning, we generate visual tokens for image synthesis by iterative decoding. For AR transformer,

for $i \leftarrow 1$ to $H \times W$ **do**
  $\hat{z}_i \sim P_\theta(z_i|\hat{z}_{<i}, \mathcal{P}_\phi)$
end for

For the NAR model, parallel decoding [7] is used:

Require: $\overline{M} = \{\}, T, \{n_1, ..., n_T\}, \sum_{t=1}^{T} n_t = H \times W$
for $t \leftarrow 1$ to T **do**
  $\hat{z}_i \sim P_\theta(z_i|\widehat{\mathcal{Z}}_{\overline{M}}, \mathcal{P}_\phi), \forall i \in M$
  $\overline{M} \leftarrow \overline{M} \cup \{\arg\text{topk}_{i \in M}\big(P_\theta(\hat{z}_i|\widehat{\mathcal{Z}}_{\overline{M}}, \mathcal{P}_\phi), k = n_t\big)\}$
end for

where $\{n_1, ..., n_T\}$ is a masking schedule that decides the number of tokens to decode at each step. We refer to [7] for details on decoding for NAR transformers. Illustrations of decoding steps for both models are in Fig. 2.

3.2. Prompt Token Generator Design

For transfer learning of discriminative tasks, prompts are designed without condition variables [29]. For generative tasks, it is beneficial to have condition variables (*e.g.*, class, attribute) for better control in generation. We achieve this with a simple design of treating class conditions as another prompt, as in Fig. 3a.

One critical issue is that the number of learnable parameters increases as the product of three factors: the number of classes C, the prompt sequence length S and the feature dimension P. For example, when using a prompt of length S=128, hidden P=768 and embedding dimension D=768, the token generator would introduce 10.4M parameters for C=100 class conditions, as in Fig. 3c. The bottleneck occurs at the 3d weight tensor of size $C \times S \times P$.

To make it parameter efficient, we propose a factorized token generator (Fig. 3b). We encode class and sequence position index via MLP_C and MLP_P with F factors, respectively. The MLP outputs are element-wise summed, multiplied by a 1d factor vector from MLP_F, and reduced along the factor dimension. The output is then fed to MLP_T to produce a prompt of length S. As in Fig. 3c, the number of parameters of the proposed architecture is greatly reduced, requiring only 0.76M parameters, down from 10.4M, for a prompt of length 128 when $F = 1$.[2] We empirically find that $F = 1$ is sufficient for NAR transformers. For AR transformers, extra capacity is needed by setting $F = 16$.

Moreover, we build a new type of prompt tokens conditioned on individual data instances, inspired by the instance-conditioned GAN [5]. We assign each data a unique index and map it into a distinct embedding via MLP_C. When both class label and instance index are used, instance index is simply treated as an extra class, indexed from C. To train the model, we sample between class label and instance index. As we explain below in Sec. 3.3, instance-conditioned prompts add more fine-grained control on generation.

3.3. Engineering Learned Prompts

Given the wealth of learned prompts conditioned on the class and instance proposed in Sec. 3.2, we propose a new

[2]The proposed factorization can be extended to incorporate the "depth" position of deep visual prompt [29] to reduce the number of parameters.

Model		(# tr params)	Mean	Mean (≤10K)	C101	Flowers	Pet	DTD	Kitti	SUN	EuroSAT	Resisc
MineGAN [71]		(88M)	151.5	114.0	102.4	132.1	130.1	87.4	117.9	77.5	111.5	81.0
cGANTransfer [60]		(105M)	85.1	63.8	89.6	61.6	48.6	70.3	48.9	31.1	45.6	50.3
Non-Autoregressive	Prompt ($S=1$)	(0.67M)	53.7	19.7	13.5	13.8	11.9	**25.8**	32.3	**7.3**	45.9	28.5
	Prompt ($S=16$)	(0.68M)	39.9	**18.6**	12.7	**13.2**	11.1	26.0	30.0	7.4	**35.8**	24.9
	Prompt ($S=128$)	(0.76M)	**36.4**	**18.6**	**12.9**	13.4	**10.9**	**25.9**	**29.9**	7.7	38.4	**24.8**
	Scratch	(172M)	42.7	60.0	72.7	57.2	70.3	66.1	33.8	9.2	39.5	32.0
Autoregressive	Prompt ($S=1$)	(0.86M)	73.2	44.1	45.4	28.9	42.2	37.1	66.8	18.8	37.3	35.1
	Prompt ($S=16$)	(0.88M)	47.4	34.5	41.4	19.6	36.6	33.4	41.4	16.4	32.6	28.8
	Prompt ($S=256$)	(1.06M)	39.0	32.3	39.6	17.3	34.9	32.5	37.1	15.0	29.6	26.7
	Prompt ($S=256$, $F=16$)	(5.16M)	**36.9**	**26.6**	**27.2**	**14.1**	**27.2**	**30.0**	34.6	**12.8**	26.4	**22.2**
	Scratch	(306M)	39.6	61.8	76.0	56.1	52.5	92.7	**31.6**	13.5	**19.4**	29.5

Table 1. FIDs (lower the better) on VTAB tasks. The number of trainable parameters (second column) are computed assuming 100 classes. The mean FID over 19 VTAB tasks (third column), over small-scale datasets (≤10K, fourth column) and those with a small to mid-scale training data are reported. Complete results are in Appendix C.1.3. The **best** and the second best results are highlighted in each column.

prompt engineering strategy, a "Marquee Header" prompt, tailored to the non-autoregressive transformer decoding, for enhancing generation diversity.

We interpolate the learned prompt representations (*e.g.*, outputs of MLP_C). To account for the iterative decoding, the interpolation between prompts is carried out over multiple decoding steps. This is shown in Fig. 4b, where we start the decoding process using instance-conditioned prompts (blue header) but gradually transition to a class-conditioned prompt (red header) over decoding steps. Unlike the generation in Fig. 4a where the instance-conditioned prompts are used all along, the marquee header prompt generates diverse images while maintaining the generation quality and following characteristics of reference instances (*e.g.*, pose, color pattern, hairiness). Fig. 4c shows a consistent trend when applying the prompt between two image instances.

The marquee header prompt is formulated as follows:

$$\text{PMT}(t) = (1-w_t)\text{PMT}_1 + w_t\text{PMT}_2 \quad (6)$$

$$w_t = \min\left\{\left(\frac{t-1}{T_\text{cutoff}-1}\right)^2, 1\right\} \quad (7)$$

where $t=1,...,T$ is a decoding step, $T_\text{cutoff} \leq T$ is a cutoff step, and PMT_i is a prompt representation (*e.g.*, an output of MLP_C). The schedule in Eq. (7) makes a smooth transition of prompts from PMT_1 to PMT_2. We keep Eq. (7)'s formulation as simple as possible and note that there could be various other prompt formulations, which we leave their investigations as our future work.

4. Experiments

We conduct extensive experiments of generative transfer learning by prompt tuning. Sec. 4.1 evaluates the efficacy on diverse visual domains on the VTAB benchmark [81]. Sec. 4.2 assess the task of few-shot transfer learning on six common benchmarks. Sec. 4.3 presents more discussions.

4.1. Generative Transfer on VTAB

Dataset. We employ the visual task adaptation benchmark (VTAB) [81] – a suite of 19 visual recognition tasks based on 16 datasets. VTAB covers diverse image domains (*e.g.*, natural, structured, and specialized such as medical or satellite imagery) and tasks (*e.g.*, object and scene recognition, distance classification, and counting). while VTAB serves as a standard yet challenging benchmark for transferring representation, this work provides the first study of generative transfer learning on the VTAB benchmark.

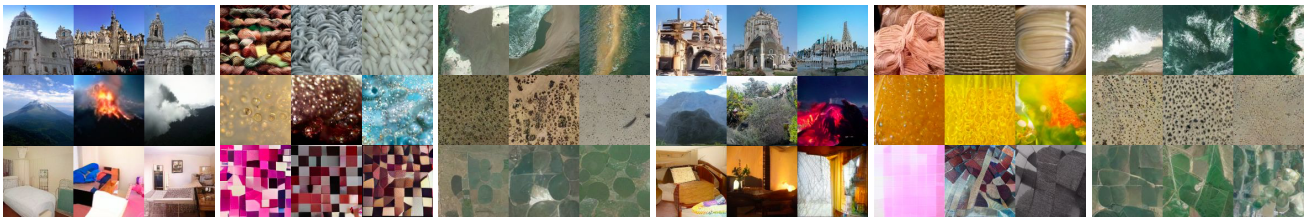
Setting. We train class-conditional image generation models on the VTAB (full) tasks, where the class-conditional prompts are trained on the "train" split, using the same hyperparameters across tasks. We investigate the generative transfer of AR [15] and NAR transformers [7] trained on 256×256 images of the ImageNet dataset as source models. Both models contain 24 transformer layers, comprised of 306M and 172M model parameters, respectively. See more implementation details in Appendix C.1.2.

Baselines. We compare our method against state-of-the-art GAN-based transfer learning methods, including MineGAN [71] and cGANTransfer [60]. Both models use BigGAN [2] trained on ImageNet as the source. BigGan's FID on the ImageNet validation is 7.4 which is better than our pretrained AR transformer (18.7) and almost on par with that of NAR transformer (6.2).

In addition, we compare generative transformers trained from scratch on VTAB with a comparable number of training epochs. We provide an analysis under different compute budgets in Appendix B.4.

Evaluation. We use Frechet Inception Distance (FID) [27]. FID is computed using $20k$ generated images and $20k$ real images randomly sampled from a respective dataset.

Results. We report mean FIDs over 3 runs in Tab. 1. As shown in Tab. 1, prompt tuning is effective for both AR and NAR generative transformers, especially when the number of training images is small (*e.g.*, $\leq 10k$). We find that the NAR model transfers better than the AR model. Nevertheless, both models with class-conditional prompt tuning show significant gains in performance over GAN-based baselines. These comparisons validate the superiority of prompt tuning over the prior state-of-the-arts. The result



(a) SUN397 (NAR) (b) DTD (NAR) (c) Resisc (NAR) (d) SUN397 (AR) (e) DTD (AR) (f) Resisc (AR)

Figure 5. Class conditional generation using (a – c) NAR (S=128) and (d – f) AR (S=256, F=16) transformers with prompt tuning.

Method	# params	10 epoch	200 epoch	800/1600 epoch
Prompt ($S=128$)	0.76M	27.6	18.5	17.7
Prompt + Adapter	5.43M	20.1	15.7	15.1
Prompt + Fine-tune	172M	**19.5**	**15.0**	**14.2**
Scratch	172M	–	60.0	22.7

Table 2. FID vs the number of train epochs for different learning methods of NAR transformers on VTAB small-scale datasets ($\leq$10k). Each number of trainable parameters is provided in the second column. Complete results are in Appendix B.1.

also provides the first quantitative evidence on the necessity of generative knowledge transfer on the VTAB benchmark.

In Fig. 5, we show generated images using 128 prompt tokens for NAR transformers and 256 prompt tokens (with $F=16$) for AR transformers on a few VTAB tasks. Due to limited space, we report complete results on all 19 tasks in Appendix C.1.3 and generated images in Appendix C.1.4.

Tab. 1 also shows that prompt tuning of generative transformers benefits greatly from a long prompt, reducing mean FID from 53.7 to 36.4 by increasing the length from 1 to 128. This is achieved by only adding extra 0.1M parameters (0.76M overall), thanks to our parameter-efficient design of the prompt token generator. Besides, AR transformers generally require prompts with more learnable parameters, which needs increasing the number of factors. The performance is still on par with that achievable with the baseline prompt, while using significantly less number of parameters (5.6M instead of 20.5M), as shown in Appendix B.3. The above results verify the design of our parameter-efficient prompt token generator.

Transfer learning settings. We compare prompt tuning with other transfer learning settings including *1)* full fine-tuning, *2)* adapter tuning [28], and *3)* learning from scratch on the target domain. To adapt these methods for generative transfer learning, we integrate prompt tuning with them to introduce class conditioning for image synthesis.

The results are given in Tab. 2, with detailed results available in Appendix B.4. Our findings indicate that prompt tuning is the most efficient approach, making it likely the only feasible option for transferring from large transformers. However, prompt tuning may not be the most expressive method for transfer learning, as its generation quality is often outperformed by adapter tuning or full fine-tuning, which have more tunable parameters. Nevertheless, our results consistently show the necessity of generative knowledge transfer when learning from limited training data.

4.2. Few-shot Generative Transfer

After validation on VTAB, we examine few-shot transfer learning, where the number of training images is further reduced. We focus on studying the transfer of the NAR transformer, *i.e.*, MaskGIT [7], and provide more comparisons to existing few-shot image generation models, either with [60, 71] or without [63, 84] knowledge transfer.

Dataset. We study few-shot generative transfer learning on three broadly-used benchmarks: Places [85], ImageNet [10], and Animal Face [61]. Following [60, 71], for Places and ImageNet, we select 5 classes[3] and use 500 images per class for training. For Animal Face, we consider two scenarios – following [60], we use 100 images per class for training from 20 classes (denoted as "Animal Face" in Tab. 3); alternatively, following [63, 84], we use all images of dog (389) and cat (160) classes (denoted as "dog face" and "cat face" in Tab. 3) for training.

Moreover, we test on three challenging off-manifold domains, *i.e.* DomainNet Infograph, Clipart (345 classes) [49], and ImageNet sketch (1000 classes) [70] where only two training images per class are used for transfer.

Setting. We study the class-and-instance conditional generative transfer as in Sec. 3.2 that is particularly suitable for few-shot transfer scenarios

Baselines. In addition to the transfer learning baselines, *i.e.*, MineGAN [71] and cGANTransfer [60], we compare to competitive models specially design for few-shot learning, *e.g.*, DiffAug [84] and LeCam GAN [63].

Evaluation. We report FIDs using $10k$ generated images, except for experiments on dog and cat faces, where we generate $5k$ images following [84]. For Places, ImageNet, and Animal Face, we use the entire training data (*i.e.*, 2500 for Places and ImageNet, 2000 for Animal Face, 389 and 160 for dog and cat faces, respectively) for the reference distribution. We sample $10k$ images for the reference distribution to compute FID for DomainNet and ImageNet sketch.

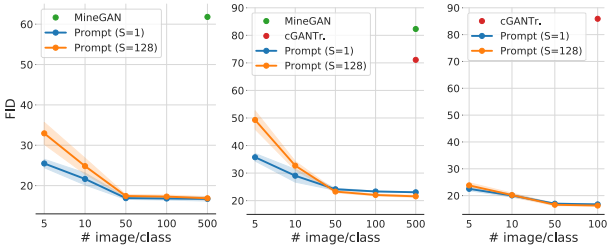
[3] Cock, Tape player, Broccoli, Fire engine, Harvester for ImageNet, and Alley, Arch, Art gallery, Auditorium, Ballroom for Places.

(a) DomainNet Clipart (2 shot; FID=22.4) (b) DomainNet Infograph (2 shot; FID=20.6) (c) ImageNet Sketch (2 shot; FID=14.4)

Figure 6. Class conditional generation of few-shot transfer models. Images in red boxes are two training images of each class.

Dataset (shot)	ImageNet (500)	Places (500)	Animal Face (100)	Dog Face (389)	Cat Face (160)
MineGAN [71]	61.8†	82.3†	–	93.0*	54.5*
cGANTransfer [60]	–	71.1‡	85.9‡	–	–
DiffAug [84]	–	–	–	58.5*	42.4*
LeCam GAN [63]	–	–	–	54.9*	34.2*
Ours (class)	**16.9**	24.2	16.3	65.4	40.2
Ours (instance)	19.6	**19.5**	**13.3**	**26.0**	**31.2**

Table 3. FIDs of image generation models on few-shot benchmark. Numbers with †, ‡, ∗ are from [71], [60], [63], respectively.



(a) ImageNet (b) Places (c) Animal Face

Figure 7. FIDs for models trained with varying numbers of images per class for class-conditional few-shot generative transfer.

		# params	Small	Medium	Large	Natural	Struct.	Spec.
S=16	baseline	1.81M	18.6	34.6	89.1	23.8	50.9	41.7
	F=1	0.68M	18.6	36.1	89.5	25.2	51.9	41.5
	F=4	0.95M	18.6	35.5	88.4	24.4	51.5	41.4
	F=16	2.02M	18.5	35.0	86.8	24.3	50.8	40.4
S=128	baseline	10.4M	18.2	30.8	86.4	22.0	46.9	39.9
	F=1	0.76M	18.5	30.6	88.9	22.5	47.1	40.5
	F=4	1.30M	18.1	31.5	88.0	23.3	48.2	38.0
	F=16	3.39M	17.9	30.8	86.5	22.6	47.4	37.7

Table 4. Ablation on prompt token generators for NAR transformers on VTAB. We report FIDs averaged by different categorizations of tasks.

Results. In Tab. 3, we report FIDs of our method using prompts of $S = 128$. When conditioned on the class, our method improves FIDs upon existing generative transfer learning methods. When comparing with few-shot generation methods on dog and cat face datasets, our method with a class condition slightly under-performs, likely due to that dataset having one class. When conditioned on instances, our models outperform highly-competitive few-shot generation models such as DiffAug, cGANTransfe, and LeCam GAN. We provide visualizations in Appendix C.2.1.

We visualize generated images conditioned on the class by our models in Fig. 6, which shows the two images used in transfer training for each class in red boxes. We observe reasonable generalization, achieved by two training images, to target domains that are visually distinct from the source ImageNet dataset.

Data Efficiency. We conduct experiments to investigate data efficiency. We train models on 5, 10, 50, and 100 training images per class for ImageNet, Places, and Animal Face datasets. The same number of images is used for the reference set to make FIDs comparable across settings.

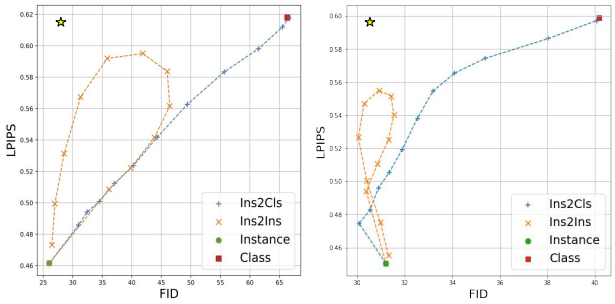
Results are in Fig. 7. Our method shows superior data efficiency, achieving substantially lower FIDs with only 5 training images per class, to MineGAN [71] or cGANTransfer [60] based on GANs trained with 20 or 100 times more images per class. We find that using long prompts is not favorable when the number of training images is too small as models start to overfit to the small train set. We discuss how the prompt length affects the adaptation-diversity trade-off in Appendix B.2. The above results substantiate the efficacy of our method on the few-shot image synthesis task.

Enhancing Generation Diversity via Prompt Engineering. As in Sec. 3.3 and Figs. 4b and 4c, our model offers a way to enhance generation diversity by composing prompts. We report quantitative metrics to support our claim.

We conduct experiments on the dog and cat faces dataset using marquee header prompts with different T_{cutoff} values. For the fidelity metric, we compute the FID. To measure the diversity, we follow [46] and report the intra-cluster pairwise LPIPS distance, where we generate $5k$ samples and map them to one of the training images.[4]

Results are in Fig. 8. Ideally, we expect a model with low FID and high intra-cluster LPIPS scores (yellow star at top-left corner). When generating samples using the class-condition prompt (red square), we generate diverse images, but with poorer fidelity. When conditioned on data instances (green dot), the FID is improved but at the cost of reduced diversity. Instance to class Marquee header prompts (blue) control the generation diversity and fidelity. Moreover, instance to instance Marquee header prompts, which interpolate the prompts between two instances, shows an improved trade-off between fidelity and diversity.

[4]We use a pixel-wise L2 distance for computation efficiency instead of LPIPS distance in [46].



(a) Dog Faces (b) Cat Faces

Figure 8. Marquee header prompt shows clear tradeoff between fidelity (FID) and diversity (LPIPS) when interpolating from instance to class (blue). It shows a better tradeoff when interpolating between instances (orange), achieving low FID and high LPIPS.

4.3. Analysis and Discussion

Parameter-efficiency. Tab. 4 provides results of using different prompt token generators, where the baseline indicates the non-factorized prompt tuning method. As shown, FIDs of prompt tuning with the proposed factorization reasonably match those of the baseline. while achieving comparable or better FIDs than the baseline using 70% fewer parameters.

Adaptation-Diversity Trade-Off. We study the instance-conditioned prompts with various lengths. Fig. 9 shows the generated images with $S=1$ (top) and $S=128$ (bottom) where more results can be found in Appendix. With a longer prompt, the synthesized images follow, more faithfully, the conditioned image, but seem less diverse. With a short prompt, on the other hand, the model still captures more dominant characters of the conditioned image (*e.g.*, color, class), but lacking fine details. The results suggest that the adaptation and diversity could be controlled with the prompt length.

5. Related Work

Transfer learning [47,62,74,87] improves the performance of downstream tasks using knowledge from the source domain. It is particularly effective when the amount of training data is limited for downstream tasks. Knowledge transfer of deep neural networks has been realized in various forms, such as linear probing [9,26], side-tuning [82], bias-tuning [4, 80], fine-tuning [35, 51], or adapter [28, 55, 56]. Recently, prompt tuning [38,40–42] has emerged as a powerful tool for transfer learning of transformer-based large language models in NLP. It has also been applied to vision-language models [16, 30, 50, 77, 86] that are limited to the input of text encoders. Since the introduction of Vision Transformer [14], prompt tuning has been studied for vision tasks where the pre-trained model is an image encoder [1, 29]. While previous works have shown the effectiveness of prompt tuning for discriminative tasks (e.g., classification [1,29]), this paper proposes an effective visual prompt tuning approach for image synthesis.

Figure 9. Instance-conditioned generation with (top) $S=1$ and (bottom) $S=128$. Images in red are the conditioned instances.

Generative models have been extensively studied for image synthesis, including variational autoencoder [34,64,66], diffusion [12, 57] and autoregressive [48, 65, 69] models. A large volume of progress has been made around the generative adversarial network (GAN) [20] thanks to its ability at synthesizing high-fidelity images [2, 31, 32, 59]. As such, generative knowledge transfer has been studied to transfer knowledge of pretrained GAN models. TransferGAN [72], following a usual practice of fine-tuning on the target dataset, has demonstrated that transferring knowledge from pretraining improves the performance when training with limited data. Freezing a few layers of the discriminator [44] further improves, while stabilizing the training process. MineGAN [71] introduces a miner, which projects random noise into the embedding space of the pretrained generator, and trains it with discriminator while fixing generator parameters. cGANTransfer [60] makes explicit transfer of knowledge on classes of the source dataset to new classes. Albeit showing improvement, these methods still require careful training (e.g., early stopping) and have evaluated on a few datasets. In our work, we extensively test methods on a wide variety of visual domains (e.g., VTAB) and show improvement by a large margin over existing GAN-based generative transfer methods.

6. Conclusion

We present a method for learning image generation models from diverse data distributions and varying amount of training data via knowledge transfer from the source model trained on a large dataset. A simple modification on prompt token designs allows to learn a parameter and compute efficient class and instance conditional image generation models of autoregressive and non-autoregressive vision transformers. We provide comprehensive experimental results of image synthesis across diverse visual domains, tasks, and the number of training images. In addition, we show how to apply learned prompts for novel image synthesis in the form of marquee header prompts using just a few images.

Acknowledgment. We thank Brian Lester for helpful discussion on prompt tuning, Boqing Gong and David Salesin for their feedback on the manuscript.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 3, 8

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 1, 5, 8

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 3

[4] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33:11285–11297, 2020. 8

[5] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned GAN. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 4

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 1, 2, 3, 4, 5, 6, 15

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 2

[9] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2018. 8

[10] 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. 6

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 8

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1, 2

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

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2, 5

[16] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 8

[17] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP-IJCNLP*, 2019. 3

[18] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 1, 8

[21] Jiatao Gu and Xiang Kong. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of ACL-IJCNLP*, 2021. 3

[22] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 15

[23] 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. 1

[24] 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. 1, 13

[25] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2020. 13, 16

[26] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 8

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[28] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona

Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 4, 6, 8, 15, 16

[29] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 2, 3, 4, 8

[30] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 8

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 8

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 8

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 16

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 8

[35] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020. 1, 2, 4, 8, 16

[36] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation. *arXiv preprint arXiv:2112.05112*, 2021. 3

[37] Xiang Kong, Zhisong Zhang, and Eduard Hovy. Incorporating a local translation mechanism into non-autoregressive translation. *arXiv preprint arXiv:2011.06132*, 2020. 3

[38] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 2, 3, 8

[39] Jose Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *ECCV*, 2022. 3

[40] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 3, 8

[41] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 8

[42] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 8

[43] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 16

[44] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 8

[45] Xing Nie, Bolin Ni, Jianlong Chang, Gaomeng Meng, Chunlei Huo, Zhaoxiang Zhang, Shiming Xiang, Qi Tian, and Chunhong Pan. Pro-tuning: Unified prompt tuning for vision tasks. *arXiv preprint arXiv:2207.14381*, 2022. 3

[46] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1, 7

[47] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 8

[48] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 8

[49] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 8

[51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 8

[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. 3

[53] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *ICML*, 2021. 1, 2

[54] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[55] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 8

[56] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018. 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 8

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[59] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 1, 2022. 8

[60] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 5, 6, 7, 8, 16

[61] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on pattern analysis and machine intelligence*, 34(7):1354–1367, 2011. 2, 6

[62] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018. 8

[63] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 6, 7

[64] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 8

[65] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 8

[66] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 8

[67] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurIPS*, 2017. 1, 2

[68] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 13

[69] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 1, 8

[70] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[71] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 5, 6, 7, 8, 16

[72] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. 8

[73] Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *CVPR*, 2019. 1

[74] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. 8

[75] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021. 1, 2

[76] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-shot generative domain adaptation. *arXiv preprint arXiv:2111.09876*, 2021. 1

[77] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 8

[78] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. *arXiv preprint arXiv:2110.04627*, 2021. 2

[79] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[80] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 8

[81] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 2, 5

[82] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision*, pages 698–714. Springer, 2020. 8

[83] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. M6-ufc: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021. 3

[84] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 6, 7

[85] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 2, 6

[86] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 8

[87] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 8