Few-Shot Incremental Learning for Label-to-Image Translation

Pei Chen[1], Yangkang Zhang[1], Zejian Li[1]*, Lingyun Sun[1,2]
[1]Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Zhejiang University
[2]Zhejiang-Singapore Innovation and AI Joint Research Lab
{chenpei, yangkz, zejianlee, sunly}@zju.edu.cn

Abstract

Label-to-image translation models generate images from semantic label maps. Existing models depend on large volumes of pixel-level annotated samples. When given new training samples annotated with novel semantic classes, the models should be trained from scratch with both learned and new classes. This hinders their practical applications and motivates us to introduce an incremental learning strategy to the label-to-image translation scenario. In this paper, we introduce a few-shot incremental learning method for label-to-image translation. It learns new classes one by one from a few samples of each class. We propose to adopt semantically-adaptive convolution filters and normalization. When incrementally trained on a novel semantic class, the model only learns a few extra parameters of class-specific modulation. Such design avoids catastrophic forgetting of already-learned semantic classes and enables label-to-image translation of scenes with increasingly rich content. Furthermore, to facilitate few-shot learning, we propose a modulation transfer strategy for better initialization. Extensive experiments show that our method outperforms existing related methods in most cases and achieves zero forgetting.

1. Introduction

In this work, we consider the task of generating images from semantic label maps. Semantic maps depict the layouts and semantic classes of images, which can be regarded as human doodles. Existing works have made great progress in generation spatial alignment [36], diversity [45], fine details [47], and style controls [62]. Nevertheless, these approaches still suffer from two issues. Firstly, they require a vast quantity of labeled data in training. However, manually labeling data is a costly and complicated process, and thus semantically fine-annotated data is often expensive to acquire. Secondly, existing label-to-image translation models require that training samples of all classes are prepared beforehand and learned at once. However, in practice, a trained translation model is often expected to perform new image generation tasks by learning novel semantic classes. A naive approach to achieve this is to retrain the model on all the old and new data, which is both time-consuming and computationally expensive. Therefore, it is necessary to equip a label-to-image translation model with the ability to learn new semantic classes flexibly without retraining.

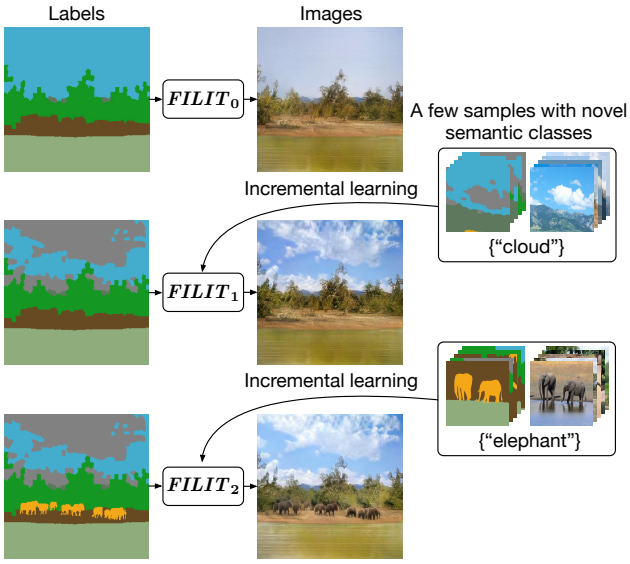

Figure 1. **An illustration of our proposed FILIT.** FILIT is capable of continually learning novel semantic classes from a few samples, without forgetting old semantic classes. In this example, the model starts with a version that can depict images consisting of the sky, mountain, tree, dirt, and river (FILIT$_0$). When given a few unseen images labeled with a new semantic class “cloud”, the model incrementally learns to depict clouds without forgetting learned classes (FILIT$_1$). Similarly, given other new images labeled with “elephant”, the model incrementally learns to generate elephants by the river (FILIT$_2$). Therefore, as incrementally learning on new labeled images, FILIT can generate scenes with increasingly rich content. Best viewed magnified on the screen.

*Corresponding author

Humans can incrementally learn a large number of different tasks without forgetting already-acquired knowledge [31]. To imitate the human learning process, incremental learning [35] has been proposed. It aims at continuously updating a trained model with samples from new tasks. A long-standing problem in incremental learning is how to prevent "catastrophic forgetting" [10, 30]. To address this problem, methods based on regularization [1, 24, 43, 54], rehearsal [4, 27, 38] and expansion [23, 53] are proposed. To alleviate the dependency on the amount of data for new tasks, advanced methods are proposed to solve few-shot incremental learning problems [11, 21, 29, 61]. Notably, recent efforts have demonstrated that generative models could also incrementally learn a sequence of datasets to generate different images. The method of memory replay [50] treats generated data from previous tasks as parts of training samples in new tasks. LifelongGAN [57] employs knowledge distillation for conditional image generation. PiggybackGAN [55] factorizes previously learned filters into a set of piggyback filters to perform new tasks.

Inspired by the above works, we propose a Few-shot Incremental learning method for Label-to-Image Translation (FILIT). It enables a pre-trained translation model to learn novel semantic classes from a few samples incrementally (Fig. 1). To achieve this, we adopt semantically-adaptive normalization and convolution filters in the generator (Sec. 3.2), which customizes convolution filters and normalization for each pixel of input feature according to the pixel's semantic class. When learning a new task, the model only learns a few modulation parameters for base convolution and normalization (Sec. 3.3). In addition, we propose a modulation transfer strategy to accelerate the convergence of modulation parameters for a new class (Sec. 3.4).

Experimental results show FILIT effectively learns new classes and successfully achieves zero forgetting of learned classes (Sec. 5.1). Ablation studies exemplify the efficacy of the semantically-adaptive design and transfer strategy (Sec. 5.2), and the required number of the extra parameters is low (Sec. 5.3). Further experiments show that a trained FILIT model can even continually learn semantic classes from datasets in other domains (Sec. 5.4).

Real label-to-image applications require massive classes to train the model, but it is infeasible to collect data of all classes at once. With FILIT, we provide a set of classes for basic creation with the pre-trained model and allow users to incrementally add new classes with a few labeled images. Such design places a low data annotation burden on users. Our contributions are three-fold: 1) We present a few-shot incremental learning method for label-to-image translation. It enables the flexible addition of novel classes. To the best of our knowledge, we are the first to target this problem. 2) We propose to adopt semantically-adaptive filters and normalization in the model. The model learns new classes with only a few extra parameters and avoids forgetting. 3) We propose a modulation transfer strategy to accelerate the convergence of incremental learning.

2. Related Work

Incremental Learning of Generative Models. Incremental learning generally involves training a model on a sequence of classes (tasks). Its main challenge is how to avoid catastrophic forgetting when learning novel tasks [30]. To address the problem, reply-based methods [2, 4, 28, 38] retain old knowledge by rehearsing on previous training data. Regularization-based methods [1, 5, 20, 54] constrain the changes of models by penalizing the drifting of important parameters or using distillation losses [9, 24, 32, 43]. Expansion-based methods [23, 52, 53] add additional task-specific components as new tasks arrive.

Pioneering works on incremental generative models are relatively less. Seff *et al*. [41] incorporate the idea of Elastic Weight Consolidation (EWC) into the training of GANs. Wu *et al*. [50] explore the idea of memory replay in label-conditioned image generation. GANmemory [8] proposes to train sequential modulation parameters for full connected layers and convolution layers to form sequential targeted generative models. LifelongGAN [57] uses distillation losses to retain learned knowledge. PiggybackGAN [55] maintains a filter bank trained on different tasks. Hyper-LifelongGAN [56] utilizes hypernetworks to generate dynamic base filters for new tasks.

Few-Shot Incremental Learning. Challenges of incremental learning become tougher when novel tasks contain very few training samples. TOPIC [48] first proposes to use a neural gas network to achieve few-shot class-incremental learning (FSCIL) in classification tasks. Following the FSCIL setting, later works propose strategies such as vector quantization [6] and calibrated classifiers [21] to improve classification performance. ONCE [37] and iMTFA [11] introduce the idea of few-shot incremental learning to object detection and instance segmentation.

Label-to-image Translation. Generating images conditioned on semantic maps, where each pixel is assigned a category label, is called label-to-image translation. Pix2pix [16] first adopts conditional GANs [33] to achieve universal image translation tasks. Pix2pixHD [49] extends it with a coarse-to-fine generator and a multi-scale discriminator to generate images with high-resolution. Additionally, spatially-adaptive normalization [36], spatially-varying conditional convolution kernels [25] and class-adaptive normalization [46] are proposed to avoid semantic information being washed away. Above label-to-image translation models depend on large-scale data for training, and do not support continually learning novel classes from new datasets. To this end, we introduce the few-shot incremental learning scheme into label-to-image translation.

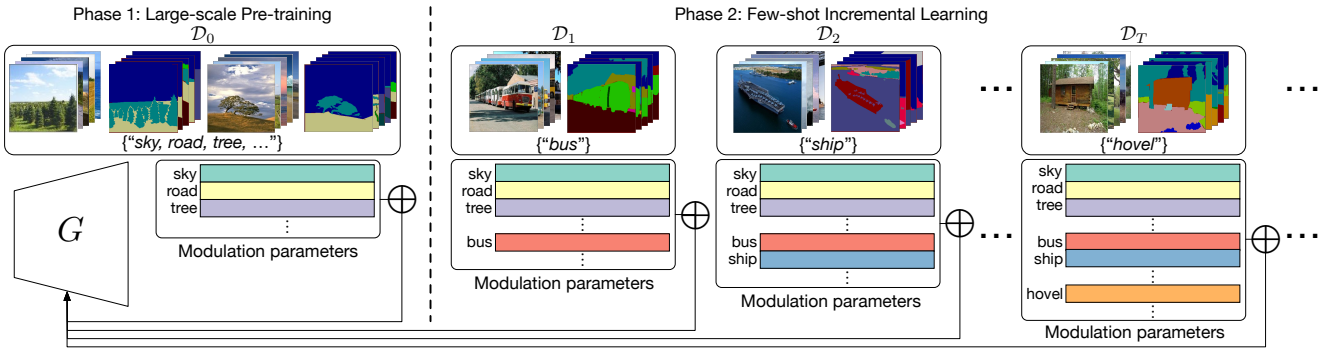

Figure 2. **The overall scheme of our proposed FILIT**. It contains two training phases: the large-scale training phase and the few-shot incremental learning phase. In the incremental learning phase, FILIT only learns class-specific modulation parameters for normalization and convolution without changing other parameters of the pre-trained generator.

3. Method

3.1. Problem Formulation

We consider the problem that a label-to-image translation model continually learns new semantic mappings from a sequence of labeled training sets $\{\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_T\}$. Each training set contains a set of realistic images and their semantic label maps. Each dataset $\mathcal{D}_t$ contains a novel semantic class which is unseen in previous datasets $\{\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_{t-1}\}$ for $t \geq 1$. When given $\mathcal{D}_t$, the model is expected to learn the new semantic mapping while maintain the translation abilities learned from previous datasets.

Incremental learning in label-to-image translation is different from incremental learning among individual tasks [8]. To learn among individual tasks, a model is equipped with multiple task-specific modules updated in the training. After training, the model loads the associated module to compose a target generative model for one specific task [8]. Namely, the model only performs a solitary task at once, such as learning to generate flowers or cats. However, in label-to-image translation, more than one semantic class is generated jointly [16, 36]. The model performs multiple tasks (classes) simultaneously, such as learning to depict cats playing before flowers.

Therefore, there are three main requirements of the defined task. Firstly, label-to-image translation requires the model to generate multiple already-learned semantic classes in a scene, instead of switching between individual tasks. Secondly, incremental learning requires the classes to be learned one by one incrementally rather than being jointly trained or trained from scratch and the model should not forget learned classes in the process. Thirdly, in the few-shot incremental learning, the number of samples to train a new class is limited, for example, only 20 samples.

3.2. Method Overview

We propose a few-shot incremental learning method for label-to-image translation (FILIT). We design a semantically-adaptive generator on top of CLADE [46]. The generator stacks ResNet blocks [13] with normalization, convolution, and LeakyReLU. Particularly, the normalization and convolution layers are semantically-adaptive to introduce semantic information. Both normalization and convolution have class-specific modulation parameters for each semantic class to facilitate incremental learning.

The generator is trained adversarially with a U-Net discriminator [39, 40]. It is an encoder-decoder network and performs a pixel-wise $(N+1)$-class classification task. Specifically, $(N+1)$ is the number of output channel in the last layer of the discriminator. N is the number of semantic classes needs to be predicted in real images. Pixels in generated images are always classified as the extra one class (fake pixels). The discriminator provides spatial and semantical information back to the generator. The pixel-wise classification is suitable for incremental learning. When learning a novel class, the discriminator extends one more output channel in the last layer to discriminate the new class.

We adopt a two-phase pipeline scheme to train the model (Fig. 2). The scheme consists of a large-scale pre-training phase and a few-shot incremental learning phase. When initially trained on $\mathcal{D}_0$, FILIT jointly learns all the base semantic classes with a large number of training samples. When continually trained on $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T\}$, FILIT incrementally learns the novel semantic classes with a few samples, while keeping the memory of previous classes.

3.3. Semantically-Adaptive Normalization and Convolution

In this part, we introduce the design of the normalization and convolution layer in a basic block of the generator. We maintain a modulation parameter

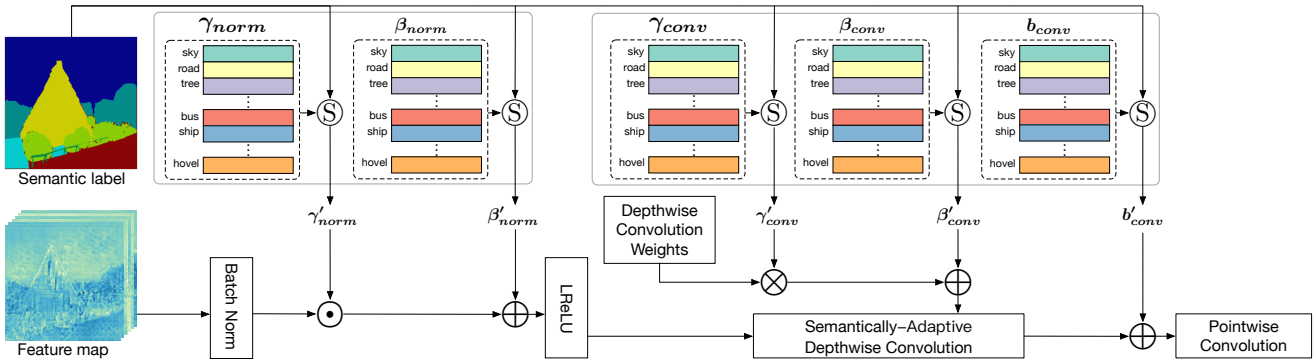

Figure 3. **The basic block of the proposed FILIT.** It consists of semantically-adaptive normalization and convolution. We maintain a modulation parameter bank $\{\gamma_{norm}, \beta_{norm}, \gamma_{conv}, \beta_{conv}, b_{conv}\}$ to modulate normalization and convolution for different semantic classes. γ_{norm} and β_{norm} are class-specific scale and shift parameters of all classes to modulate the normalization layer. γ_{conv}, β_{conv} and b_{conv} are class-specific scale, shift and bias parameters to modulate the convolution layer. Ⓢ is the Guided Sampling operation. $\odot$ is the Hadamard product and $\otimes$ is the outer product operation.

bank $\{\gamma_{norm}, \beta_{norm}, \gamma_{conv}, \beta_{conv}, b_{conv}\}$ in each block to produce semantically-adaptive filters and normalization. Specifically, γ_{norm} and β_{norm} are class-adaptive scale and shift parameters of all classes to modulate the normalization layer. Similarly, γ_{conv}, β_{conv} and b_{conv} are class-adaptive scale, shift and bias parameters to modulate the convolution layer (Fig. 3). In particular, γ_{norm} consists of N vectors of size C_{in} to modulate N semantic classes, where C_{in} denotes the number of channels in the layer. Other modulation parameters share the same structure.

Let $f \in \mathbb{R}^{C_{in} \times H \times W}$ be the input feature of a normalization layer, and $y \in \mathbb{L}^{H \times W}$ be the semantic map of f. H and W are the height and width of the feature map. $\mathbb{L}$ is the set of semantic classes. The feature f is firstly normalized and then modulated with class-specific scale and shift parameters. The scale and shift are obtained by Guided Sampling [46]. Specifically, we sample the class-specific parameters from γ_{norm} and β_{norm} for each pixel $f_{i,j}$ $(0 \leq i < H, 0 \leq j < W)$ according to its semantic class $y_{i,j}$. Thus, we obtain the dense modulation parameters $\gamma'_{norm}, \beta'_{norm} \in \mathbb{R}^{C_{in} \times H \times W}$ for f. The whole normalization can be formulated as

$$\hat{f}_{c,i,j} = (\gamma'_{norm})_{c,i,j} \frac{f_{c,i,j} - \mu_c}{\sigma_c} + (\beta'_{norm})_{c,i,j} \quad (1)$$

$\hat{f}_{c,i,j}$ is the normalized pixel. μ_c and σ_c are the mean and standard deviation of f in the channel c for $0 \leq c < C$.

Next, $\hat{f}$ is fed into the convolution layer. To reduce computation cost, we use depthwise separable convolution [15], which contains a depthwise convolution and a pointwise convolution. The semantically-adaptive modulation is operated on the depthwise convolution. The filter in the depthwise convolution layer is denoted as $\mathcal{F} \in \mathbb{R}^{C_{in} \times s_w \times s_h}$, and the bias as $b \in \mathbb{R}^{C_{in}}$. s_w and s_h are the width and height of the filter. First, we also adopt Guided Sampling operation to get dense modulation parameters $\gamma'_{conv}, \beta'_{conv}, b'_{conv} \in \mathbb{R}^{C_{in} \times H \times W}$. Then, we modulate the depthwise filters as

$$\begin{aligned} \hat{\mathcal{F}} &= \gamma'_{conv} \otimes \frac{\mathcal{F} - \mu(\mathcal{F})}{\sigma(\mathcal{F})} + \beta'_{conv} \\ \hat{b} &= b + b'_{conv} \end{aligned} \quad (2)$$

Here $\hat{\mathcal{F}} \in \mathbb{R}^{C_{in} \times H \times W \times s_w \times s_h}$ is the spatial class-specific filter, and $\hat{b} \in \mathbb{R}^{C_{in} \times H \times W}$ is the class-specific bias. $\mu(\mathcal{F}) \in \mathbb{R}^{C_{in}}$ and $\sigma(\mathcal{F}) \in \mathbb{R}^{C_{in}}$ denote the channel-wise mean and standard deviation of $\mathcal{F}$, respectively. $\otimes$ is the outer product operation. The filter is normalized along the spatial footprint [8] to remove base style information to facilitate learning unseen semantic classes. To equip the filter with the target styles of pixels' classes, our modulation applies a scale γ'_{conv} and a shift β'_{conv} to the base filter. The scale and shift introduce more flexibilities to the adapted filter compared with scaling the filter only [58]. Finally, the bias in the convolution is also modulated spatially in our design. Thus, we customize a filter for each pixel in the feature map determined by the pixel's semantic class.

To implement convolution, we use $\hat{\mathcal{F}}$ to execute Multiply-Add operations on patches of the input feature map $\hat{f} \in \mathbb{R}^{C_{in} \times H \times W}$. The convolution operation on each patch can be formulated as

$$\hat{\hat{f}}_{c,i,j} = \sum_{u=0}^{s_w} \sum_{v=0}^{s_h} \hat{f}_{c,i+u,j+v} \hat{\mathcal{F}}_{c,i,j,u,v} + \hat{b}_{c,i,j} \quad (3)$$

Why should we design semantically-adaptive architecture to perform pixel-wise modulation? Such design is to meet discussed requirements at the end of Sec. 3.1. Firstly, our design can perform pixel-wise modulation on the feature according to its semantic map without switching and thus can generate images with multiple classes simultane-

ously. Secondly, the modulation bank to parameterize semantic classes is separated from the base network. As a result, the model only learns new modulation parameters for an unseen class in incremental learning. At the same time, because the base network and the learned part of the modulation bank are fixed, the model does not forget learned classes. Thirdly, the design facilitates few-shot incremental learning. The number of new modulation parameters to learn is low, while the base network contains universal visual information in the trained data.

3.4. Training

Training phase 1: Large-scale Pre-training. This pre-training phase is similar to conventional methods, which are trained on a large-scale dataset [40]. The pre-training dataset $\mathcal{D}_0$ consists of base semantic classes with a reasonably large amount of training samples. Images of all classes are merged for pre-training. G_0 and D_0 denote the generator and discriminator in the pre-training phase, and they are optimized by:

$$\arg\min_G \max_D \mathcal{L}_{full} \equiv \mathcal{L}_{GAN} + \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{con}\mathcal{L}_{con} \quad (4)$$

$\mathcal{L}_{GAN}$, $\mathcal{L}_{vgg}$ and $\mathcal{L}_{con}$ denote adversarial loss [40], perceptual loss [17] and consistency loss [40], respectively. The consistency loss is used to encourage the discriminator to focus on differences in characteristics between classes in generated and real images.

Training phase 2: Few-shot Incremental Learning. In this step, we conduct incremental learning on datasets $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T\}$ to progressively learn a sequence of new semantic classes unseen in the pre-training phase. The generator and the discriminator for incremental learning is initialized by the trained G_0 and D_0. When learning a new class, the trained G_0 only learns the new class's modulation parameters and other parameters of the generator are fixed to avoid forgetting. The number of parameters required to learn for a new class is only $C_{in} \times 5$ for each block.

To accelerate the convergence of incremental learning, we propose a modulation transfer technique. It initializes the modulation parameters of novel classes by transferring modulations of learned similar classes. The assumption behind is that semantic classes with perceptually similar visual appearances have modulation parameters close to each other [42]. To seek similar learned classes for a new class, we utilize the encoder of the pre-trained discriminator to extract hidden features for both the new and learned classes. Recall that our discriminator is a U-Net [39], and its hidden features summarize patch-wise semantic information.

Formally, we divide pixels in the hidden features according to their spatial labels. Such extraction is performed over images in the pre-trained dataset $\mathcal{D}_0$ and the new dataset $\mathcal{D}_t$. Thus, we obtain a set of features for each semantic class. Assuming each set of features subjects to a Gaussian distribution, we use L_2 Wasserstein distance [12] to estimate the similarity. The modulation parameters of the new class is initialized as a linear combination of those modulations of the three closest classes. Finally, the model is also trained with Eq. (4) in the incremental learning.

Training details. We use the ADAM optimizer [19] with $\beta_1 = 0, \beta_2 = 0.999$. The learning rates are set to 0.0001 for the generator and 0.0004 for the discriminator, respectively. We set the loss weight $\lambda_{vgg} = 10$ and $\lambda_{con} = 10$. We resize images to 256×256 for training. In the incremental learning, the encoder of discriminator is fixed. Both the generator and the discriminator are trained for 100 epochs for each novel class.

4. Experimental Setup

4.1. Dataset

We conduct experiments on the following datasets.

ADE20K [59] consists of 20,210 training and 2,000 validation images annotated with 151 semantic classes. We split the whole dataset into 21 sub-datasets with $\mathcal{D}_0$ for pre-training and $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{20}\}$ for incremental learning. $\mathcal{D}_0$ contains 20,712 images with 131 semantic classes. Each dataset $\mathcal{D}_t (1 \leq t \leq 20)$ for incremental learning contains an unseen class in $\mathcal{D}_{t-1}$. $\mathcal{D}_t$ is further divided into a support dataset and a query dataset: the support dataset is used for learning the new task with 20 images, and the query dataset is used for testing. There are 1098 testing images for 20 tasks in total.

COCO-Stuff [3] consists of 123,287 images annotated with 172 semantic classes, and is also divided into 21 sub-datasets. The pre-training dataset $\mathcal{D}_0$ contains 107,928 images with 152 classes. The support dataset of D_t contains 50 images. There are 14,359 testing images for 20 tasks.

The 20 semantic classes for incremental training in ADE20K and COCO-Stuff are those having the least number of samples in the original dataset. Models are trained on the support datasets of Task 1 to 20 sequentially. After Task 20, models are tested on all the query datasets jointly.

4.2. Baselines

We compare FILIT with incremental learning models for conditional image generation and FILIT's variants.

- *LifelongGAN* [57] introduces knowledge distillation losses into BicycleGAN [60] to perform conditional image generation in the incremental learning setting. We reproduce LifelongGAN for our experiments.
- *PiggybackGAN* [55] factorizes filters trained on previous tasks to learn new tasks, and maintains a task-specific filter bank to memorize learned tasks. We use the code from [18] for our experiments.

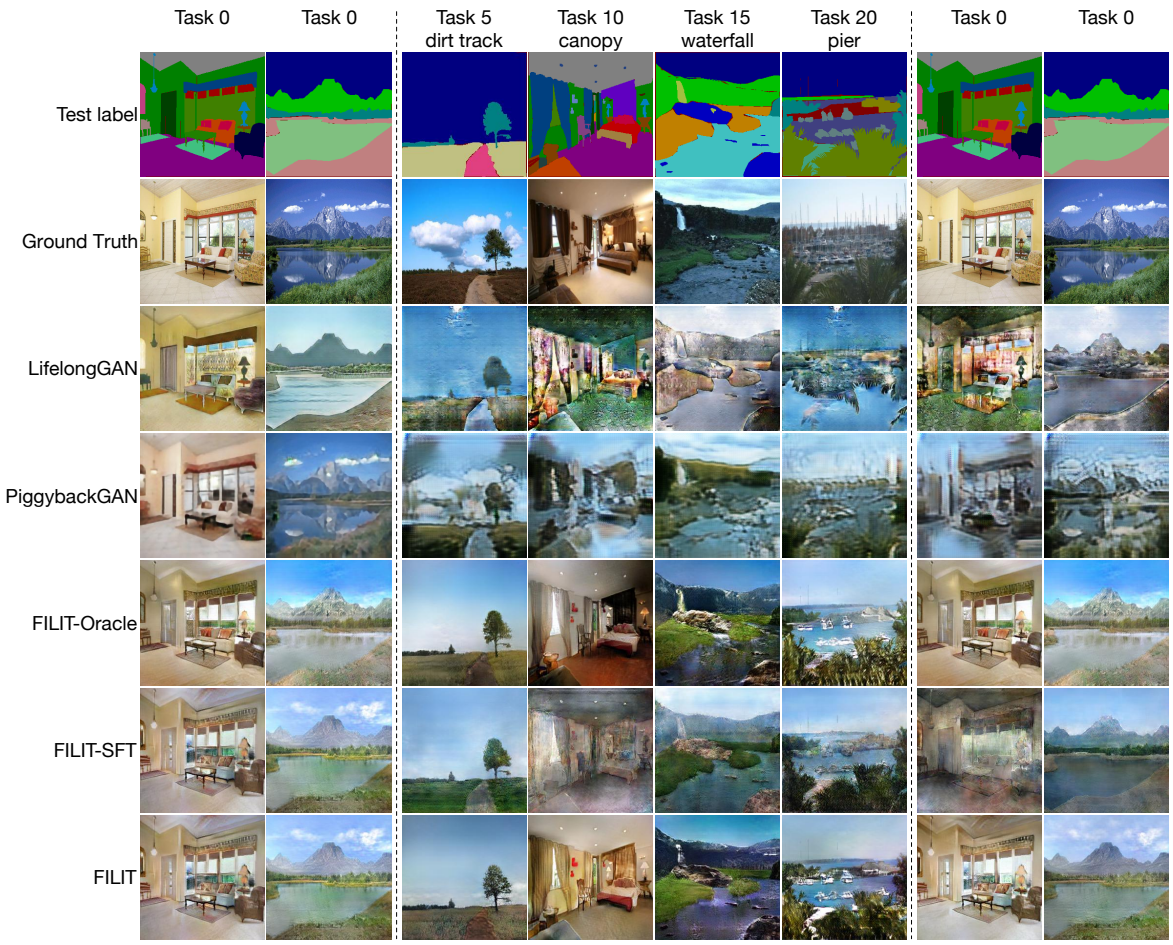

Figure 4. Qualitative results on ADE20K dataset. The first two columns show the generative results on Task 0 after pre-training. The central four columns show the incremental learning results on Task 5, 10, 15 and 20. The last two columns on the right display results of recalling Task 0 after the incremental learning, and they are to examine catastrophic forgetting.

- *FILIT-Oracle* is a variant of FILIT trained with $\{\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_T\}$ jointly without incremental learning. This model marks the performance upper bound that FILIT can achieve in the incremental learning.

- *FILIT-SFT (Sequential Fine-Tuning)* is another variant of FILIT. It is also pre-trained on the Task 0, and then it is fine-tuned in a sequential manner. In the fine-tuning process, the whole generator is used to learn new tasks, and the decoder of discriminator is fixed [34].

LifelongGAN and PiggybackGAN are trained with our two-phase pipeline.

4.3. Metrics

We adopt the evaluation metics from previous works [25, 36, 40, 46] including mean Intersection-over-Union (mIoU), pixel accuracy (accu) and Fréchet Inception Distance (FID) [14]. Particularly, mIoU and accu measure segmentation accuracy. FID measures the distance between the distributions of generated images and the distributions of real images. We use the segmentation model UperNet101 [51] for ADE20K dataset and DeepLabV2 [7] for COCO-Stuff dataset. To evaluate the extent of catastrophic forgetting, we define the forgetting rate of mIoU, accu and FID, denoted as FmIoU, Facc and FFID, respectively. Specifically, after pre-trained on Task 0, we test on the $\mathcal{D}_0$ to obtain $(\text{mIoU})_0$, $(\text{accu})_0$ and $(\text{FID})_0$. Then, after training on Task 1 to Task 20, we test the model on $\mathcal{D}_0$ again to obtain $(\text{mIoU})_{20}$, $(\text{accu})_{20}$ and $(\text{FID})_{20}$. The FmIoU is defined as $\frac{|(mIoU)_0 - (mIoU)_{20}|}{(mIoU)_0}$, and Faccu and FFID are similarly defined. In addition, we use SceneFID [44] to measure the generative performance of novel classes learned in the incremental learning process. Specifically, given images containing objects of novel classes, we use a ROI operator to crop these objects out. All cropped objects in real and generated images are resized and used to calculate SceneFID.

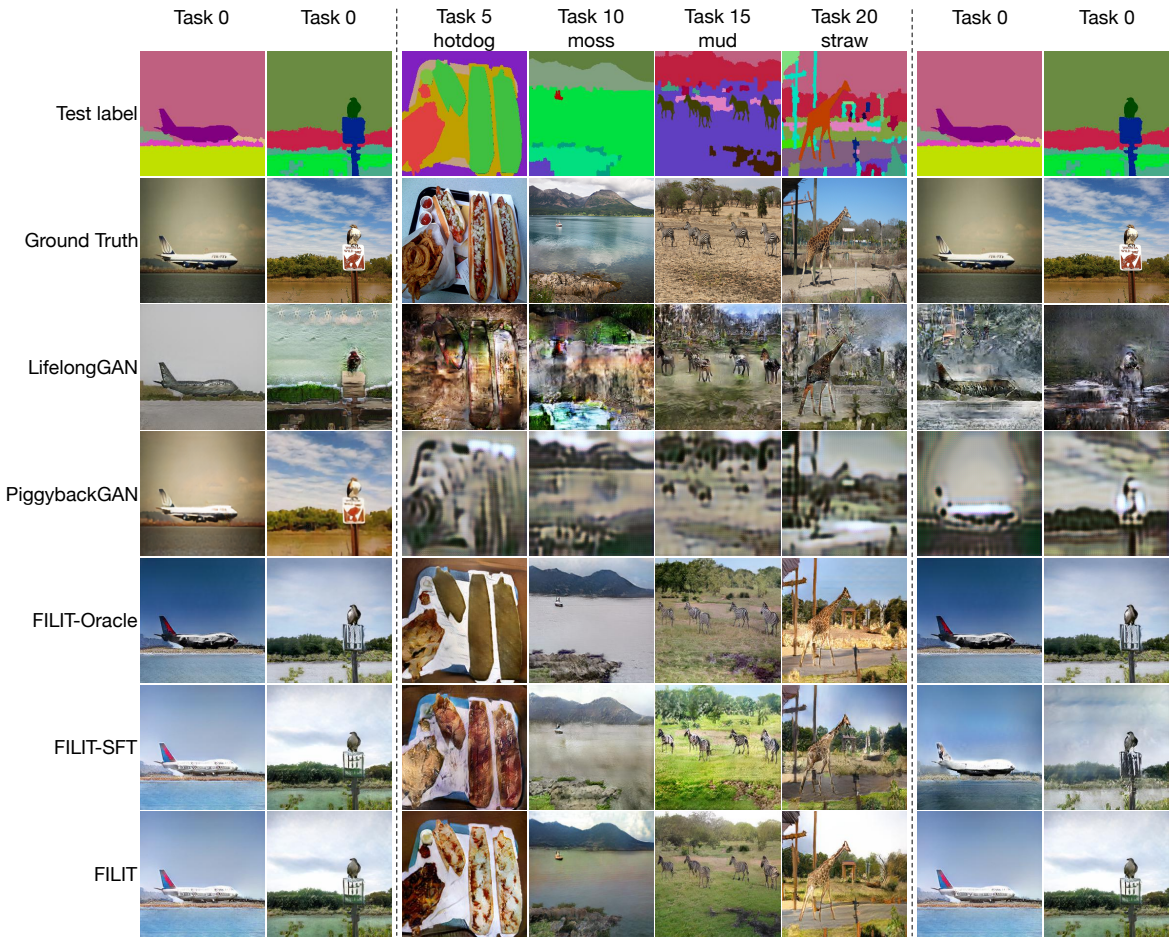

Figure 5. Qualitative results on COCO-Stuff dataset. The first two columns show the generative results on Task 0 after pre-training. The central four columns show the incremental learning results on Task 5, 10, 15 and 20. The last two columns on the right display results of recalling Task 0 after the incremental learning, and they are to examine catastrophic forgetting.

Method	ADE20K							COCO-Stuff						
	mIoU↑	accu↑	FID↓	FmIoU↓	Faccu↓	FFID↓	SceneFID↓	mIoU↑	accu↑	FID ↓	FmIoU↓	Faccu↓	FFID↓	SceneFID↓
LifelongGAN	10.0	25.2	170.5	37.7%	32.6%	>50%	112.3	2.2	4.8	160.0	>50%	>50%	>50%	57.3
PiggybackGAN	5.2	15.4	253.9	>50%	>50%	>50%	153.3	2.6	7.7	208.4	>50%	>50%	>50%	116.2
FILIT-Oracle	**23.5**	**50.4**	**74.4**	-	-	-	**77.8**	**18.0**	**38.0**	**32.0**	-	-	-	28.8
FILIT-SFT	16.8	40.7	170.2	35.6%	24.4%	>50%	147.8	15.0	32.4	64.4	16.3%	13.5%	>50%	37.6
FILIT	23.2	49.6	77.1	**0%**	**0%**	**0%**	80.7	**18.0**	37.9	32.7	**0%**	**0%**	**0%**	**22.3**

Table 1. Quantitative results on ADE20K and COCO-Stuff. ↑ means a higher value is better, and vice versa. The best and second best performances are highlighted by using bold and underline, separately.

5. Results

5.1. Qualitative and Quantitative Results

Figs. 4 and 5 present generated images on ADE20K and COCO-Stuff. Our model generates images as visually appealing as the fully-trained model. Especially, objects of new semantic classes are harmonious with those of learned classes, and the learned classes are not forgotten in the incremental learning process. Tab. 1 reports quantitative results. The reported mIoU, accu and FID reflect generative performance on incremental learning (Task 1 to Task 20), and FmIoU, Faccu, and FFID reflect the forgetting rate of Task 0 after incremental learning. For consistency of metrics, we fixed the random seed in testing. In this case, FILIT achieves 0% forgetting rates. In summary, the results show that FILIT achieves comparable performance with the full-trained model and outperforms other compared models.

Method	mIoU↑	accu↑	FID↓	SceneFID↓
FILIT	**23.2**	**49.6**	**77.1**	**80.7**
w/o adaptive convolution filters	23.0	48.8	81.6	84.0
w/o adaptive normalization	22.5	48.5	83.7	88.4
w/o modulation transfer	23.0	49.4	84.1	87.2

Table 2. Results in the ablation study for 20-task incremental learning on ADE20K dataset.

Method	Task 0	Task 1	Task 20	Additional↓
PiggybackGAN †	7.84M	12.22M	95.44M	4.38M
PiggybackGAN ○	13.16M	13.64M	22.32M	0.48M
FILIT	18.99M	19.05M	20.15M	**0.06M**

Table 3. The numbers of additional parameters when learning a new semantic class of ADE20K. † means the numbers of Task 0 and Task 1 are from Zhai et al. [56] and the number of Task 20 is calculated accordingly. ○ means the numbers are measured in the implementation of [18].

5.2. Ablation Study

We conduct ablation studies on variants of our proposed model on ADE20K dataset. Variants include replacing semantically-adaptive convolution filters with fixed filters, replacing normalization layers, and no modulation transfer. Tab. 2 shows semantically-adaptive convolution, normalization, and the modulation transfer strategy all contribute to the improvement of FID and SceneFID, and they have minor influence on the segmentation performance.

5.3. Model Expansion

FILIT adds class-specific modulation parameters to the original model when learning a new task, and it belongs to the expansion-based method. We compare the model expansion of FILIT with PiggybackGAN [55], which is also an expansion-based method. As shown in Tab. 3, the additional parameters for each subsequent class of FILIT are only 0.06M. In summary, FILIT requires little amount of expansion while achieving compelling performance.

5.4. Experiments on Cross-Domain Tasks

To further investigate the generalization ability of FILIT, we use the model trained on all tasks of ADE20K to continually learn cross-domain samples from DeepFashion [26] and CelebAMask-HQ [22] datasets. Specifically, we pick ten samples from DeepFashion as $\mathcal{D}_{21}$, which contains nine novel semantic classes. Similarly, ten samples containing sixteen classes from CelebAMask-HQ are selected to construct $\mathcal{D}_{22}$. As images of DeepFashion and CelebAMask-HQ are dissimilar from those in ADE20K, we do not use the modulation transfer strategy in this experiment, and we train 30,000 epochs for each task. FILIT generates vivid images of persons and faces (Fig. 6). In particular, after training on Task 21 and Task 22, FILIT still generates high-quality images of Task 0 in ADE20K without forgetting.

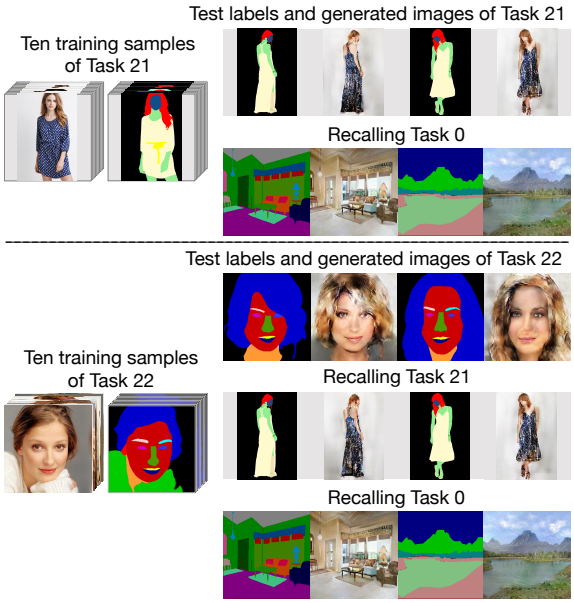

Figure 6. Incremental learning results from DeepFashion and CelebAMask-HQ dataset in a ten-shot setting. After training on ten samples from DeepFashion, we recall Task 0 from ADE20K dataset. After training on another ten samples from CelebAMask-HQ, we recall Task 0 and Task 21 again. Both recalling experiments show FILIT does not forget learned tasks.

6. Conclusion

In this paper, we propose a novel few-shot incremental learning method for label-to-image translation (FILIT). It continually learns new semantic classes with a few samples without training from scratch. To achieve this, we adopt semantically-adaptive convolution filters and normalization. It separates class-specific modulation parameters from the base network to support adding novel classes and avoid forgetting. To accelerate the convergence of incremental learning, we propose a modulation transfer strategy. Experimental results demonstrate that with a few additional parameters, FILIT effectively learns new classes from datasets in the same and other domains while achieves zero forgetting. FILIT presents new possibilities of practical applications with label-to-image translation models.

Acknowledgements

This paper is funded by National Key R&D Program of China (2018AAA0100703), the National Natural Science Foundation of China (No. 62006208 and No. 62107035) and the Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD IDEA Grant.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 2

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11817–11826. Curran Associates, Inc., 2019. 2

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018. 5

[4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 2

[5] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 2

[6] Kuilin Chen and Chi-Guhn Lee. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 6

[8] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. GAN memory with no forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 16481–16494. Curran Associates, Inc., 2020. 2, 3, 4

[9] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5138–5146, 2019. 2

[10] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 2

[11] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1185–1194, 2021. 2

[12] Clark R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31:231–240, 1984. 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–646. Springer, 2016. 3

[14] 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 (NeurIPS)*, 30:6626–6637, 2017. 6

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. 4

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 2, 3

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 5

[18] Kaushilk. PiggybackGAN. https://github.com/kaushik333/Piggyback-GAN-Pytorch. Last accessed on Nov 17, 2021. 5, 8

[19] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2

[21] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[23] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: a continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning (ICML)*, pages 3925–3934. PMLR, 2019. 2

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2935–2947, 2017. 2

[25] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 570–580. Curran Associates, Inc., 2019. 2, 6

[26] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1096–1104, 2016. 8

[27] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. volume 30, pages 6467–6476. Curran Associates, Inc., 2017. 2

[28] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[29] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. 35(3):2337–2345, 2021. 2

[30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: the sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[31] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013. 2

[32] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 205:103167, 2021. 2

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[34] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020. 6

[35] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: a review. *Neural Networks*, 113:54–71, 2019. 2

[36] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019. 1, 2, 3, 6

[37] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13846–13855, 2020. 2

[38] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 350–360. Curran Associates, Inc., 2019. 2

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3, 5

[40] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 5, 6

[41] Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets. *arXiv preprint 1705.08395*, 2017. 2

[42] 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 (CVPR)*, pages 12167–12176, 2021. 5

[43] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3420–3429, Oct 2017. 2

[44] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. volume 35, pages 2647–2655, May 2021. 6

[45] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7962–7971, 2021. 1

[46] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 2, 3, 4, 6

[47] Hao Tang, Song Bai, and Nicu Sebe. Dual attention GANs for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 1994–2002, 2020. 1

[48] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12183–12192, 2020. 2

[49] TingChun Wang, MingYu Liu, JunYan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 2

[50] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost Van de Weijer, and Bogdan Raducanu. Memory replay GANs: learning to generate images from new categories without forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5966–5976. Curran Associates, Inc., 2018. 2

[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 6

[52] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[53] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[54] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, pages 3987–3995. PMLR, 2017. 2

[55] Mengyao Zhai, Lei Chen, Jiawei He, Megha Nawhal, Frederick Tung, and Greg Mori. Piggyback GAN: Efficient lifelong learning for image conditioned generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 397–413. Springer, 2020. 2, 5, 8

[56] Mengyao Zhai, Lei Chen, and Greg Mori. Hyper-LifelongGAN: scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2246–2255, 2021. 2, 8

[57] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong GAN: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2759–2768, 2019. 2, 5

[58] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained GANs for generation with limited data. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11340–11351. PMLR, 2020. 4

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 5

[60] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476. Curran Associates, Inc., 2017. 5

[61] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6801–6810, 2021. 2

[62] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5104–5113, 2020. 1