Exploring Incompatible Knowledge Transfer in Few-shot Image Generation

Supplementary Material

Overview

This supplementary material provides the additional experiments and results to further support our main findings and proposed method for few-shot image generation (FSIG). These were not included in the main paper due to the space limitations. The supplementary material is organized as follows:

Contents

Algorithm 1 Proposed Knowledge Truncation for Few-shot Image Generation

```
Input: target samples $\mathbf{X}$, number of total iterations $\mathbf{N}$, number of warmup iterations $\mathbf{N}_w$, interval for filter importance estimation $\mathbf{T}$, Conv filters of the model $\mathbf{W}$, FI of the filter $\mathcal{F}(\mathbf{W})$, quantile of filter importance $q(\mathbf{W})$, quantile threshold for filter that is selected for pruning $t_l$, threshold for filter that is selected for preserving $t_h$, total pruning rate $p\%$.
Initialize with pretrained GAN: $G_t \leftarrow G_s$ and $D_t \leftarrow D_s$.
for $iteration = 1; iteration \leq \mathbf{N}_w; iteration++$ do
    Update the discriminator $D_t$ given $\mathbf{X}$, via Eqn. 1   # warmup for $D_t$
end for
for $iteration = \mathbf{N}_w + 1; iteration \leq \mathbf{N}; iteration++$ do
    If $iteration \% \mathbf{T} = 0$ :
        Estimate $\mathcal{F}(\mathbf{W})$ via Eqn. 2;
        Preserve $\mathbf{W}$ if $t_h \leq q(\mathbf{W})$;
        Fine-tune $\mathbf{W}$ if $t_l \leq q(\mathbf{W}) \leq t_h$;
        Prune $\mathbf{W}$ if $q(\mathbf{W}) \leq t_l$;
        Update the operation of $\mathbf{W}$ in Memory Bank. # size of the Memory Bank $\sim$ 5,000
    ELSE
        Extract the operation of $\mathbf{W}$ in Memory Bank;
        Update model via Eqn. 1.
end for
Output: The adapted GAN
```

A. Pesudo-code of our proposed method

Here, we describe the proposed method in algorithmic format. We use the generator G for example, but the same operations are also applied to the discriminator D during adaptation. As mentioned in Sec. 5.1 in the main paper, we use Fisher Information (FI) as the measurement for importance estimation in main experiments, and we use the same notation in the algorithm. We summarize the proposed method in Algorithm 1.

B. Pseudo-code for implementing Intra-LPIPS

Here we describe the implementation details of Intra-LPIPS, the diversity evaluation method proposed in [14]. We follow prior works [14, 25, 26] to apply Intra-LPIPS as an additional metric to evaluate the diversity of generated images by different adapted GAN models. To make it easy to understand, we summarize the pseudo-code of Intra-LPIPS in Algorithm 2. A small value of Intra-LPIPS indicates that the generated images are more close to few-shot target samples.

C. A brief description of Importance Probing (IP)

In the main paper, we apply IP, a modulation based method that estimates the filter importance for target adaptation proposed in AdAM [25], to determine the filters importance for target adaptation. We discover that the filters with least importance for adaptation are highly correlated with the knowledge that is incompatible to the target domain.

In this section, we briefly discuss the implementation of IP. We refer the readers to [25] for more details. In IP, for each filter in the generator $\mathbf{W}$, they generate a new filter $\mathbf{M}$ with the same dimension as $\mathbf{W}$ to modulate $\mathbf{W}$. Then, the modulated filter $\mathbf{W}^{'}$ can be written as follows:

$$\mathbf{W}^{'} = \mathbf{W} \odot (\mathbf{J} + \mathbf{M}), \tag{1}$$

where $\mathbf{J}$ is the all-one matrix, $\odot$ indicates the Hadamard product. Then, they adapt the modulation matrix $\mathbf{W}^{'}$ to target domain for 500 iterations (called “probing”) and only update $\mathbf{M}$ during the probing stage. After that, they use the importance (estimated) of modulation matrix $\mathbf{M}$ as approximation of importance for the original filters $\mathbf{W}$.

In contrast to IP, we directly estimate the filter importance for adaptation by leveraging on-the-fly gradient information during training (therefore no probing stage). The results in Table 2 in the main paper shows we can achieve comparable performance even without our proposed knowledge truncation method.

D. Additional experiment details

Here we provide additional details to help reproduce our results.

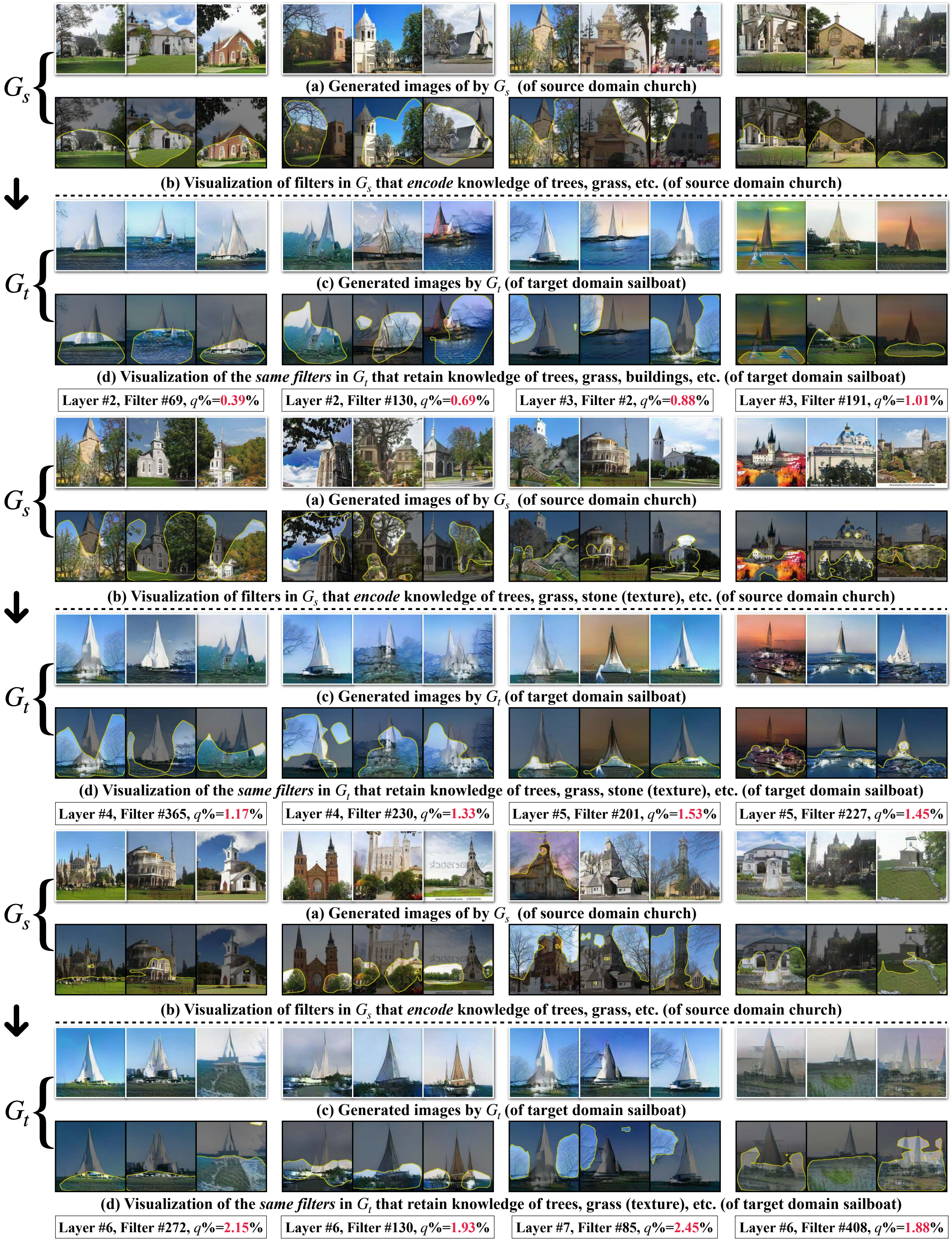
Algorithm 2 Pseudo-code of Intra-LPIPS

```
# Input: 1. Generated images X=[x1, ..., xn];
#        2. Suppose we have 2-shot target samples with cluster center: c0, c1;
#        3. Cluster_0, Cluster_1 = [], []
# Output: Avg Intra-LPIPS over 2 clusters
# -------------------------------------------- #
# Step 0. Define the LPIPS distance function (Zhang et al.)
import lpips

lpips_fn = lpips.LPIPS(net=’vgg’)  # default setup

# Step 1. Assign generated images to the cluster with smallest LPIPS distance to the cluster center
for X[i] in X:
    dist0 = lpips_fn(X[i], c0)
    dist1 = lpips_fn(X[i], c1)
    if dist0 < dist1:
        Cluster_0.append(X[i])
    else:
        Cluster_1.append(X[i])
# -------------------------------------------- #

# Step 2. Compute Intra-LPIPS
lpips_dist = []
While not done:  # randomly sample image pairs
    for img_i, img_j in Cluster_0:
        lpips_dist.append(lpips_fn(img_i, img_j))
    for img_i, img_j in Cluster_1:
        lpips_dist.append(lpips_fn(img_i, img_j))
return lpips_dist.mean()
# -------------------------------------------- #
```

Code Implementation. For GAN dissection experiments and analysis, we follow their implementation and use the official repo. In main experiments, following prior SOTA methods [14, 25, 26], we use StyleGAN-V2 as GAN architecture for a fair comparison. The implementation is from this repo. We additionally use the ProgressiveGAN [8] to validate our proposed analysis and method, using the code base from Karras *et al*. [8]. For performance evaluation, we use the tool (*e.g*., to compute FID) used in [4]. During adaptation, we strictly follow the default hyperparameters as prior works [14, 26]. Before the main adaptation, there is a warmup stage where we only update D_t for a few iterations (*e.g*., 250 in our experiments) to let the D_t aware of the target domain information to have a better estimation of the filter importance for the target adaptation. We evaluate the filter importance for adaptation every 50 iterations. Nevertheless, we found our proposed method is not very sensitive to the choice of these hyperparameters. We use batch size 4 and an initial learning rate 0.002 (with Cosine Scheduler in PyTorch [15]) in all experiments, similar to prior works [14, 25, 26]. We use images with resolution 256 x 256 for adaptation.

Datasets. In main paper, we follow prior works [13, 14, 25, 26] in the choice of source and target datasets for adaptation. To make a fair and comprehensive comparison, we use FFHQ [10] as source domain. In this supplementary material, we additionally use LSUN Church [23] as source domain for visualization. We also include more target datasets for adaptation, *e.g*., face paintings [5, 18, 22], Haunted Houses or Palace [6]. We perform 10-shot adaptation in most setups. Nevertheless, we also evaluate our method given different number of target training samples, see results in Table S4.

Implementation details of pruning EWC and AdAM. In literature, EWC [13] and AdAM [25] proposed different criteria to preserve source knowledge by identifying parameter importance. EWC (Li *et al*.) applied Fisher Information (FI) to measure parameter importance to the source domain. They then penalize the change of parameters during adaptation, weighted by the importance of the parameters: if a parameter is deemed to be important, it will be penalized more given the same change. AdAM proposed a modulation based method to evaluate the parameter importance for adaptation, see details in Sec. C. Therefore, in Table 1 in the main paper, we propose to prune least important filters using the importance measurement proposed in their original works, and empirically we observed the improved FID score of the generated images, which implies the effectiveness and generalizability of our proposed method.



(a) Generated images of by G_s (of source domain church)

(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass, buildings, etc. (of target domain sailboat)

Layer #2, Filter #69, q%=0.39% | Layer #2, Filter #130, q%=0.69% | Layer #3, Filter #2, q%=0.88% | Layer #3, Filter #191, q%=1.01%

(a) Generated images of by G_s (of source domain church)

(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, stone (texture), etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass, stone (texture), etc. (of target domain sailboat)

Layer #4, Filter #365, q%=1.17% | Layer #4, Filter #230, q%=1.33% | Layer #5, Filter #201, q%=1.53% | Layer #5, Filter #227, q%=1.45%

(a) Generated images of by G_s (of source domain church)

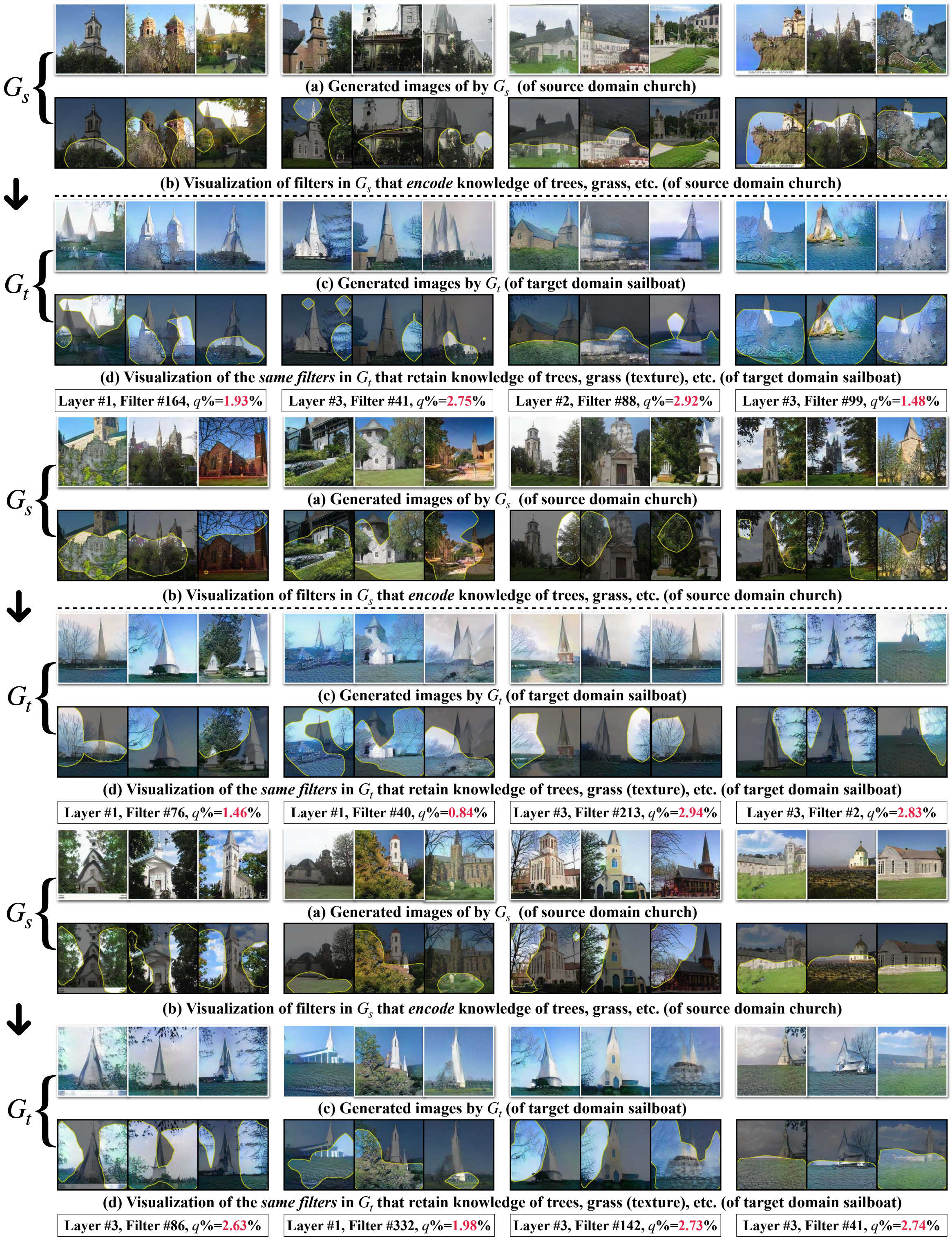
(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass (texture), etc. (of target domain sailboat)

Layer #6, Filter #272, q%=2.15% | Layer #6, Filter #130, q%=1.93% | Layer #7, Filter #85, q%=2.45% | Layer #6, Filter #408, q%=1.88%

Figure S1. Additional GAN dissection results using AdAM. Note that in AdAM fine-tuning is applied during adaptation from G_s to G_t to update these low importance filters. However, we observe that incompatible knowledge (tree, building, grass) remains in the same filters in G_t after fine-tuning.


G_s
G_t
G_s
G_t
G_s
G_t

(a) Generated images of by G_s (of source domain church)

(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass (texture), etc. (of target domain sailboat)

Layer #1, Filter #164, q%=1.93%
Layer #3, Filter #41, q%=2.75%
Layer #2, Filter #88, q%=2.92%
Layer #3, Filter #99, q%=1.48%

(a) Generated images of by G_s (of source domain church)

(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass (texture), etc. (of target domain sailboat)

Layer #1, Filter #76, q%=1.46%
Layer #1, Filter #40, q%=0.84%
Layer #3, Filter #213, q%=2.94%
Layer #3, Filter #2, q%=2.83%

(a) Generated images of by G_s (of source domain church)

(b) Visualization of filters in G_s that *encode* knowledge of trees, grass, etc. (of source domain church)

(c) Generated images by G_t (of target domain sailboat)

(d) Visualization of the *same filters* in G_t that retain knowledge of trees, grass (texture), etc. (of target domain sailboat)

Layer #3, Filter #86, q%=2.63%
Layer #1, Filter #332, q%=1.98%
Layer #3, Filter #142, q%=2.73%
Layer #3, Filter #41, q%=2.74%

Figure S2. Additional GAN dissection results using EWC. Similar to AdAM, in EWC fine-tuning is applied during adaptation from G_s to G_t to update these low importance filters. However, we observe that incompatible knowledge (tree, building, grass) remains in the same filters in G_t after fine-tuning.

E. Additional GAN dissection results

Additional GAN dissection results for AdAM. In Figure 2, we use IP to estimate the filter importance, therefore we use AdAM as the adaptation method for analysis. In this section, we provide additional GAN dissection results as supplement to Figure 2 in the main paper.

Additional GAN dissection results for EWC. Since EWC [13] also proposed parameter importance estimation method for FSIG adaptation, it is naturally to validate our findings in Sec. 3 with EWC method as well. In this section, we additionally include EWC for GAN dissection, and we note that we have similar observations as Sec. 3 in the main paper.

The parameter (filter) importance estimation criteria of AdAM and EWC can be found in Sec.D. The detailed additional visualization results can be found in Figure S1 (AdAM) and Figure S2 (EWC).

F. Ablation Study: The impact of high-importance filters

In the main paper, we emphasize our contributions of investigating the incompatible knowledge transfer, its relationship with the least important filters, and the proposed method to address this unnoticed issue for FSIG. Besides knowledge truncation, following prior works, we also preserve useful source knowledge for adaptation. Specifically, we preserve filters that are deemed important for target adaptation by freezing them. We select the high-importance filters by using a quantile (t_h, *e.g.* 75%) as a threshold. See Algorithm 1 for details. In this section, we conduct a study to show the effectiveness and impact of preserving different amount of filters that are deemed to be most relevant for target adaptation, the results are in Table. Note that we do not prune any filters in this experiment.

Table S1. We preserve different amount of filters during adaptation and evaluate the performance of adapted generators. We do not prune any filters (*i.e.*, $t_l = 0\%$) in this experiment. The experiment setups are the same as Table 1 in the main paper. t_h is the quantile that we start to preserve filters. *e.g.*, if $t_h = 90\%$, we only preserve 10% filters (see Algorithm 1 for details).

$\boldsymbol{t_h}$	30%	40%	50%	60%	70%	80%	90%
Babies	47.55	47.42	**46.54**	49.19	51.09	52.00	67.33
AFHQ-Cat	90.13	69.64	69.13	65.92	**57.02**	57.27	60.56

As shown in Table S1, varying amounts of filters for preservation do in fact increase performance in different ways. In practice, we select $t_h = 50\%$ for FFHQ $\mapsto$ Babies and $t_h = 70\%$ for FFHQ $\mapsto$ AFHQ-Cat, and this choice is intuitive: for target domains that are semantically closer to the source, preserving more source knowledge might improve the performance.

G. Ablation Study: Effectiveness of dynamic filter importance estimation

In the main paper, our proposed method includes a dynamic filter importance estimation scheme (denote as "dynamic"). In contrast to prior work [13, 25] that evaluate the parameter importance only once before the adaptation stage (denote as "static"), we regularly estimate the filter importance for target adaptation every certain iterations (*e.g.* 50 iterations in our experiments). In this section, we conduct a study to show the effectiveness of the proposed dynamic importance estimation scheme (compared to the "static" estimation scheme) and the results are shown in Table S2.

H. Ablation Study: Additional importance measurement

Evaluating the importance of weights in generative tasks is still underexplored. In the main paper, we follow some prior works [1, 13] to use the Fisher Information (FI) as the measurement for importance estimation and obtain competitive performance across different datasets (See Table 1 in the main paper). Nevertheless, there could be different ways to evaluate how well the obtained weights given the adaptation task. In literature, Class Salience [17] (CS) is used as a tool to estimate which area/pixels of a given input image stand out for a specific classification decision, and it is similar to FI that leverages the gradient information. Therefore, we note that CS could have a connection with FI as they both use the knowledge encoded in the gradients for knowledge importance estimation.

We perform an experiment to replace FI with CS in filter importance measurement and compare with prior works. Note that, in [17], CS is computed w.r.t. input image pixels. To make CS suitable for our problem, we modify it and compute CS w.r.t. different filters by averaging the importance of all parameters within a filter to calculate the importance of that filter. After that, the same as main paper, we preserve the filters that are deemed to be salient for target adaptation by freezing them, prune the filters with least CS value and fine-tune the rest of filters, and we regularly evaluate the filter importance.

Table S2. In this experiment, we study the effectiveness of the proposed dynamic framework for knowledge truncation and preservation. In AdAM, they propose a modulation based method (see details in [25]) which is static, i.e., they conduct one-time filter importance identification before the main adaptation stage. To show the effectiveness of the proposed dynamic framework for FSIG is better than its static counterpart proposed in [25], we conduct a study and show results in below. Note that we modify the original AdAM such that it freezes the important parameters rather than modulating them, hence we have a more direct comparison. We use the same experiment setup as Table 1 in the main paper for fair comparison.

Method	**FFHQ → Babies**		**FFHQ → Cat**	
	FID (↓)	Intra-LPIPS (↑)	FID (↓)	Intra-LPIPS (↑)
TGAN [19]	101.58	0.517	64.68	0.490
EWC [13]	79.93	0.521	74.61	0.587 {static}
AdAM [25] {static, modulation, w/o prune}	48.83	0.590	58.07	0.557
AdAM {static, modulation, w/ prune} **(Ours)**	**43.12**	-	**53.94**	-
AdAM {static, freezing, w/o prune}	50.81	0.581	61.60	0.559
AdAM {static, freezing, w/ prune} **(Ours)**	46.87	-	57.56	-
Ours {dynamic, freezing, w/prune}	**39.39**	**0.608**	**53.27**	0.569

The results in Table S3 are obtained with our proposed method using FI and CS as the measurements for filter importance estimation.

Table S3. In this experiment, we apply different measurements (*i.e.*, FI and CS) for filter importance estimation and compare with state-of-the-art methods. The experiment setup is the same as Table 1 in the main paper. We evaluate the performance under different source → target adaptation setups where source and target domains have different proximity.

Method	**FFHQ → Babies**		**FFHQ → Cat**	
	FID (↓)	Intra-LPIPS (↑)	FID (↓)	Intra-LPIPS (↑)
TGAN [19]	101.58	0.517	64.68	0.490
EWC [13]	79.93	0.521	74.61	0.587
AdAM [25]	48.83	0.590	58.07	0.557
Ours w/ Class Salience	39.95	0.607	56.46	0.574
Ours w/ Fisher Information	**39.39**	**0.608**	**53.27**	0.569

As can be observed in Table S3, our proposed method can be applied with different measurements for importance estimation. Meanwhile, we note that a good FID score does not necessarily infer to a good diversity of generated images from the obtained generator [20]. Compared to prior works, our proposed method can consistently achieve a **good trade-off** of performance (between FID and intra-LPIPS) of the adapted models across different adaptation setups.

I. Additional qualitative results

Few-shot adaptation with additional setups. We provide qualitative results with additional adaptation setups to show the generalizability of our proposed method. Specifically, we visualize the generated images before and after adaptation. We use the same experiment setups as in the main paper and 10-shot target samples for adaptation. As results below, since our proposed method reliably removes the filters that are deemed to be incompatible with the target domains, **there is not much incompatible knowledge transferred to the generated images**. Meanwhile, we show that the generated images are diverse and high fidelity, as we also preserve knowledge important for target adaptation:

- **FFHQ ↦ Sketches:** Figure S3
- **FFHQ ↦ Sunglasses:** Figure S4
- **FFHQ ↦ Metfaces:** Figure S5
- **FFHQ ↦ Raphael's Paintings:** Figure S6
- **FFHQ ↦ Amedeo-Modigliani's Paintings:** Figure S7

- **LSUN Church ↦ Haunted House:** Figure S8

- **LSUN Church ↦ Palace:** Figure S9

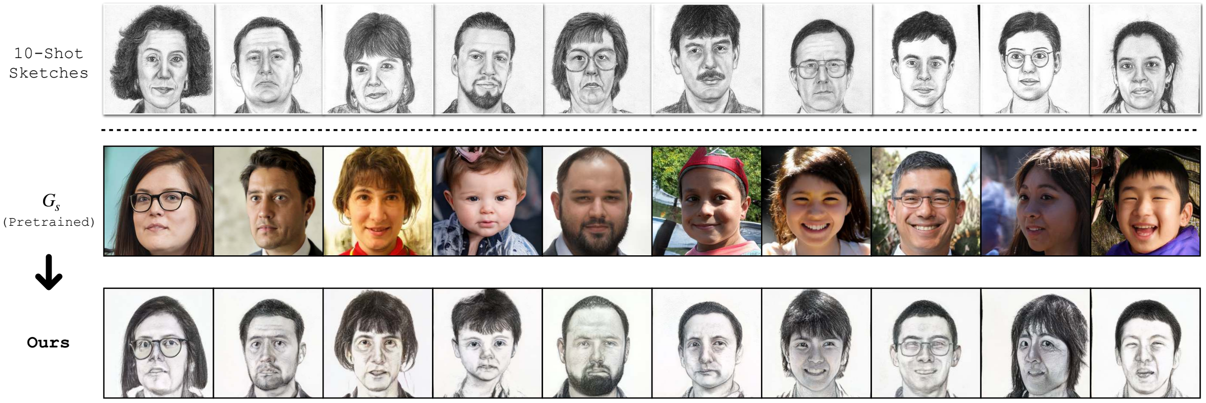


Figure S3. Additional FSIG results with FFHQ ↦ Sketches.

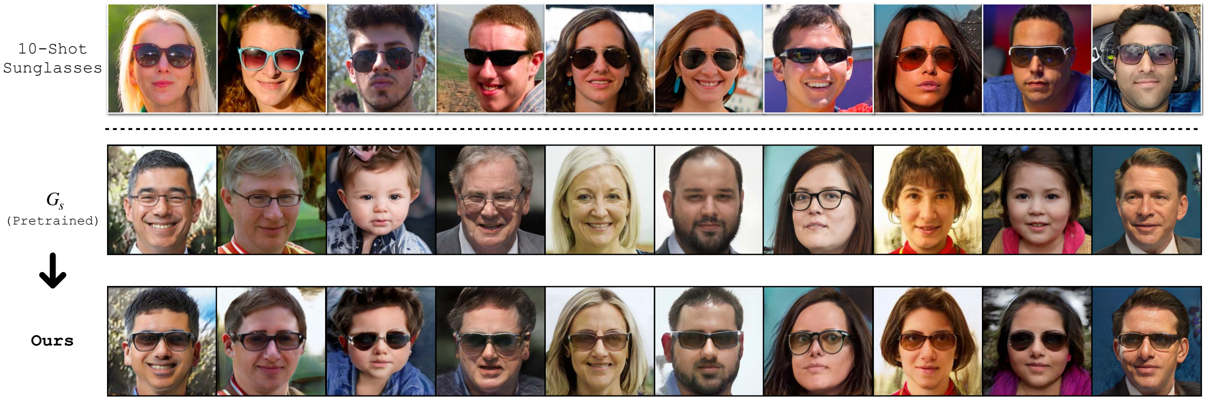


Figure S4. Additional FSIG results with FFHQ ↦ Sunglasses.

Interpolation in the latent space. In our methods, the MLP layers in the generator are not tuned. In Fig. S10 , we show the semantic manipulation results of our target model that it facilitates editing applications. We follow GenDA [21] to linearly interpolate between two different latent codes after few-shot adaptation and visualize the intermediate generated images, which are still of high quality.

Comparison with SOTA on additional target domains. The discussion and results are in Figure S11.

J. Additional quantitative results

FSIG with different number of shots. In the main paper, we mainly focus on the adaptation setup where we have only 10-shot samples for adaptation. In this section, we comprehensively evaluate the proposed method given different number of target training samples. The results are in Table S4.

Evaluation on additional dataset. We note that we have evaluate our method and compare with SOTA methods comprehensively in the main paper, see Table 1 and Figure 4. Here, we include additional results of FFHQ ↦ Sketches [18]. The results are included in Table S5.

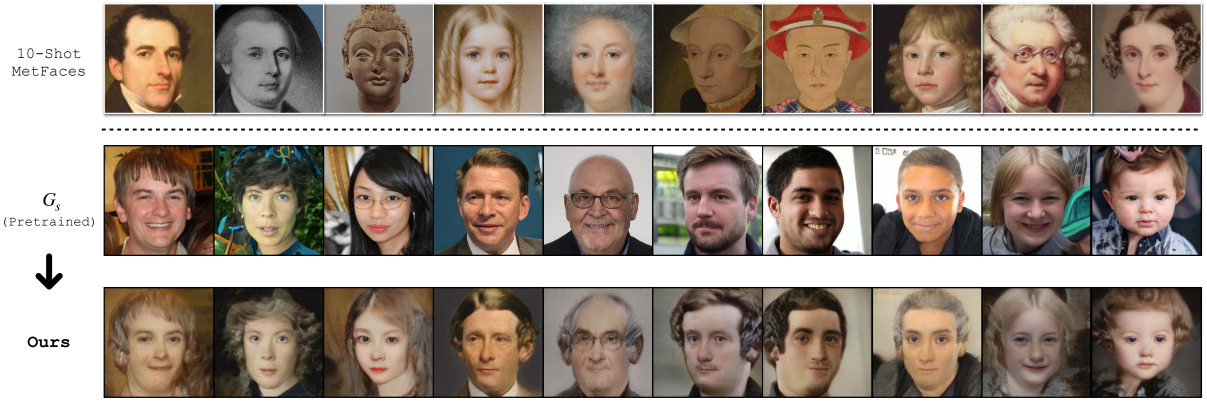


Figure S5. Additional FSIG results with FFHQ ↦ MetFaces.

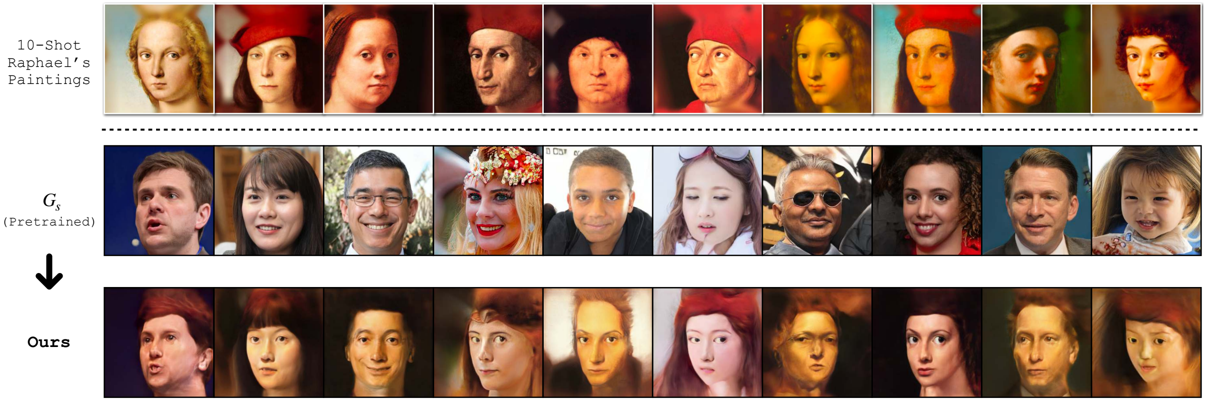


Figure S6. Additional FSIG results with FFHQ ↦ Raphael's paintings.

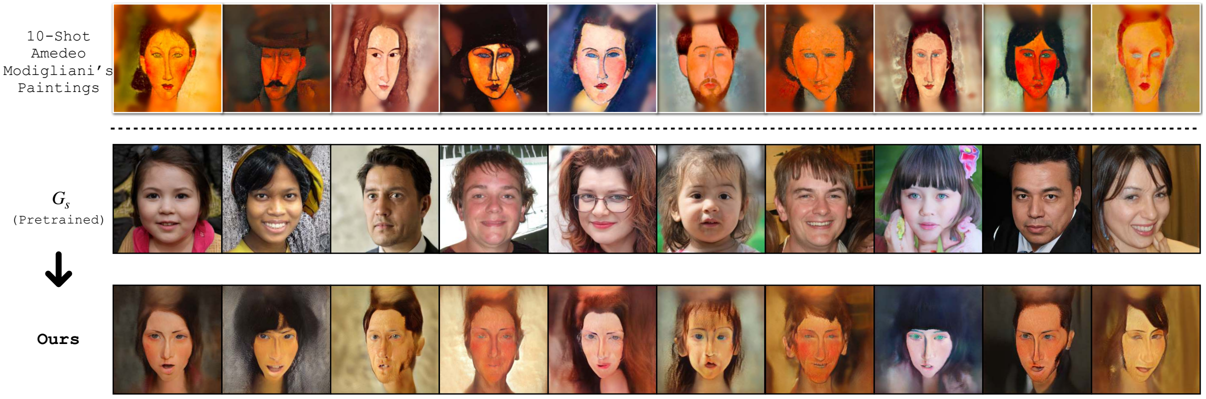


Figure S7. Additional FSIG results with FFHQ ↦ Amedeo-Modigliani 's paintings.

Additional evaluation metric. We note that we evaluate our proposed method via different quantiative metrics (FID, Intra-LPIPS) across six different datasets, and achieve competitive performance compared to prior works. Here we additionally provide the results of Kernel Inception Score (KID [2]), which is supplement of FID in our work, to evaluate the difference between abundant generated images and the entire target domain. The results are in Table S7. We show that our proposed method can achieve comparable KID score with other SOTA FSIG methods.

Additional Ablation Studies. Table S6 1) & 3): We conduct experiments and show that our proposed dynamic selection is indeed *important and compatible* with pruning. Table S6 5) & 6): We randomly re-initialize the filters with the lowest importance, and we show that those filters may not learn knowledge properly due to less iterations in few-shot

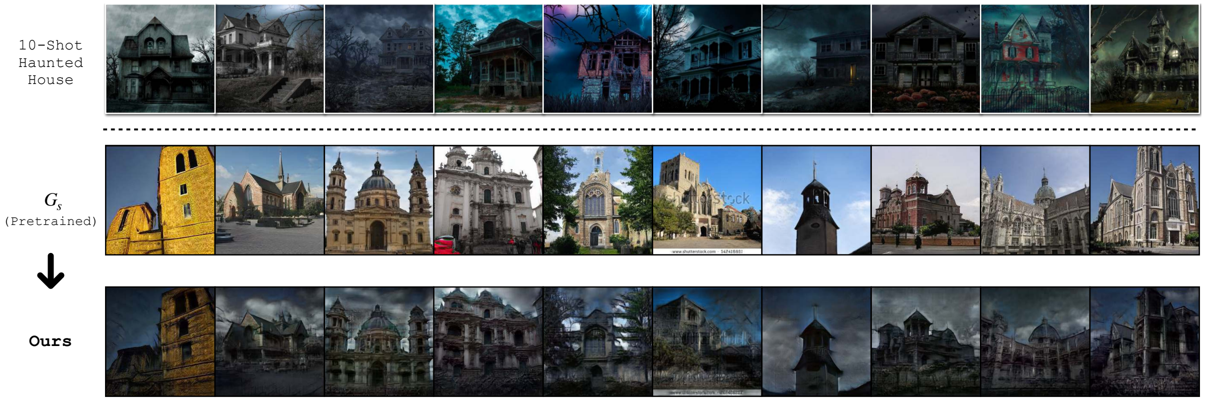


Figure S8. Additional FSIG results with Church ↦ Haunted Houses.

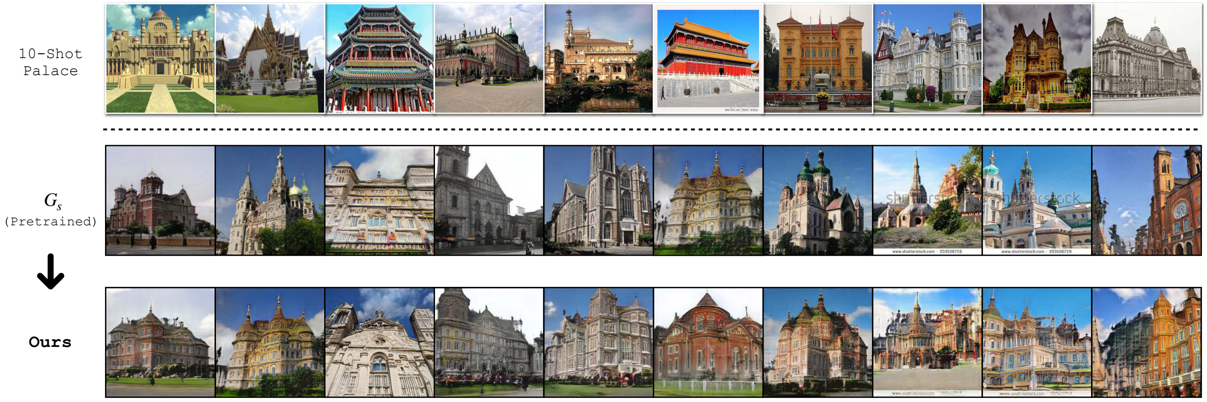


Figure S9. Additional FSIG results with Church ↦ Palace

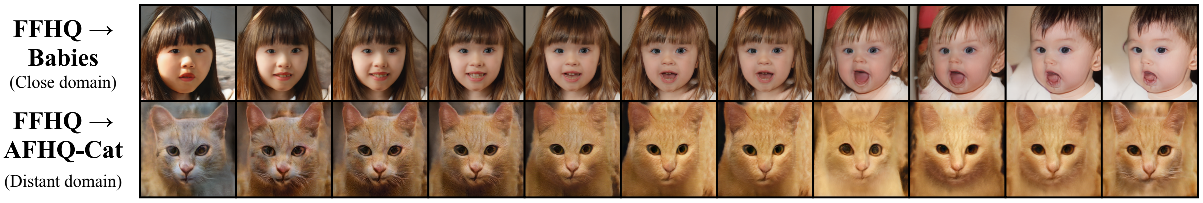


Figure S10. Visualization of the generated images using linear interpolation between two latent codes after adaptation.

adaptation. Therefore, we propose to prune lowest important filters. **Table S6 7)**: We show the results of dynamic scheme with modulation method [25], the performance is comparable with the freezing scheme. In our paper, we adopt freezing which is simple to implement.

K. Training longer leads to severe overfitting of existing FSIG methods

In this section, we conduct a study to show that issue of incompatible knowledge *cannot be addressed by fine-tuning for more iterations* by Eqn. 1 in the main paper (*i.e.*, GAN loss), as with more iterations in fine-tuning would lead to severe overfitting as also pointed out in previous work [26]. Specifically, we follow the experiment setup in the main paper (*e.g*. Figure 1 and Figure 2 as we analyze the incompatible knowledge transfer in that section) and evaluate the diversity of generated images at different adaptation steps. The results are in Table S8 and it clearly shows the diversity of generated images gradually collapse to the few-shot target samples. We conjecture that this is due to $\mathcal{L}_{adv}$ where the generatore G_t is encouraged to replicate few-shot target samples to fool the discriminator. Therefore, based on the above observations, it is important to remove the knowledge incompatible to the target domain before mode collapse becomes severe.

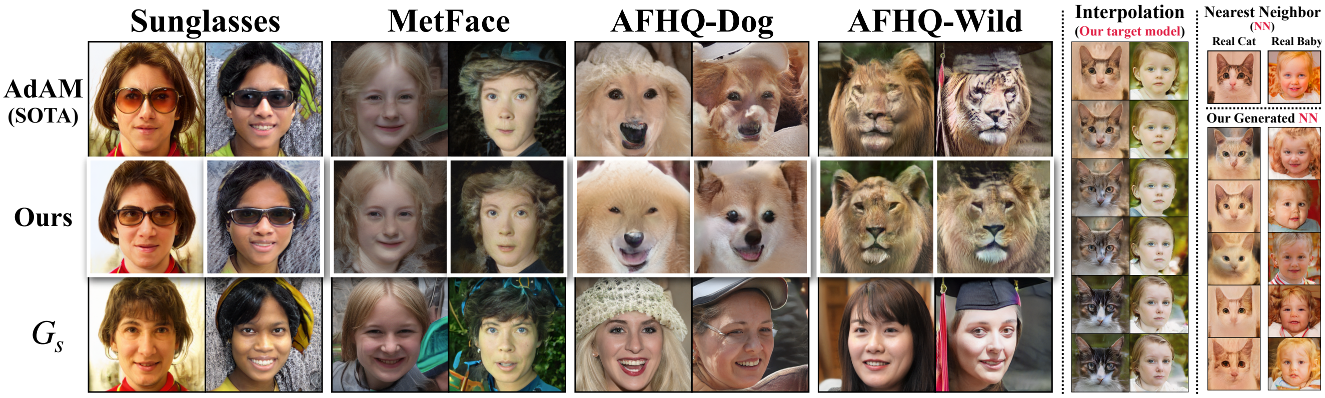


Figure S11. Additional visualization and comparison results. **Left:** Comparison with SOTA on additional target domains (Sunglasses, MetFaces, etc.), FFHQ is the source domain. **Mid:** Similar to Figure S10, we provide additional visualized images by using linearly interpolated latent codes. **Right:** For a selected target image, we visualized the nearest neighbour of that target image. To this end, we generate 5,000 images after adaptation, and assign images with smallest LPIPS distance [24] to each of the target image. **Best viewed in color and zooming in.**

Table S4. FID (↓) with respect to the number of shots for adaptation. We use the same experiment setups as Table 1 in the main paper.

Number of Samples	**Domain**	TGAN [19]	TGAN+ADA [9]	EWC [13]	AdAM [25]	**Ours**
1-shot	FFHQ ↦ Babies	172.49	188.84	104.5	77.71	**74.34**
5-shot		108.65	105.19	88.51	52.85	**43.53**
10-shot		101.58	102.98	79.93	48.83	**39.39**
25-shot		54.83	56.66	44.67	27.77	**25.13**
50-shot		48.39	52.94	39.32	24.69	**22.46**
100-shot		39.04	45.71	34.49	19.63	**18.85**
200-shot		33.65	38.84	32.65	17.06	**15.71**
500-shot		27.21	26.31	28.11	16.17	**13.67**
All		25.63	25.47	24.57	13.59	**12.34**
Number of Samples	**Domain**	TGAN [19]	TGAN+ADA [9]	EWC [13]	AdAM [25]	**Ours**
1-shot	FFHQ ↦ AFHQ-Cat	125.52	125.81	139.11	118.25	**117.69**
5-shot		90.24	86.94	136.65	79.53	**74.47**
10-shot		64.68	80.16	74.61	58.07	**53.27**
25-shot		40.52	48.61	56.23	32.38	**31.85**
50-shot		33.87	35.76	43.58	26.43	**24.67**
100-shot		27.78	28.16	36.93	21.50	**19.60**
200-shot		24.73	26.78	33.43	19.79	**18.02**
500-shot		20.25	19.01	32.73	16.80	**13.56**
All		10.52	9.56	18.76	6.52	**6.22**

Table S5. We additionally evaluate the proposed method in FFHQ ↦ Sketches setup. We use the same experiment setup as Table 1 in the main paper. Compared to prior state-of-the-art methods [11, 13], our proposed method can achieve better FID by a large margin with comparable diversity. Meanwhile, we also note that the entire Sketch domain [18] contains only ∼300 images, which could be not very stable for evaluation and we only include the results in Supplementary for reference.

Metric	**Domain**	TGAN [19]	TGAN+ADA [9]	EWC [13]	CDC [14]	AdAM [25]	**Ours**
FID (↓)	FFHQ ↦ Sketches	53.42	66.99	71.25	45.67	45.03	**39.29**
Intra-LPIPS (↑)		0.394	0.414	0.421	0.453	0.459	0.442

Table S6. FID (↓) with the same hyper-parameters as Tab. S2. We follow Tab.S2 to evaluate the filter importance for target adaptation: only once for *static* and periodically for *dynamic*.

Method (the same experiment setup as Tab. S2)	**FFHQ → Babies**	**FFHQ → Cat**
1) Ours {static, no freezing, w/ prune}	56.62	64.25
2) Ours {static, w/ freezing, w/ prune} (Tab.S2)	46.87	57.56
3) Ours {dynamic, no freezing, w/ prune}	51.08	59.71
4) Ours {dynamic, w/ freezing, w/ prune} (Tab. S2 & Tab. 1)	**39.39**	**53.27**
5) Ours {static, w/ freezing, w/ random re-initialize}	50.54	59.93
6) Ours {dynamic, w/ freezing, w/ random re-initialize}	41.67	55.21
7) Ours {dynamic, w/ modulation, w/ prune}	39.97	53.81

Table S7. Besides the FID and Intra-LPIPS results in the main paper, we additionally include KID (↓) score of different methods. We use the same experiment setup as Table 1 in the main paper. Following prior works [3,9], we scale the KID values with 10^3.

Domain	TGAN [19]	TGAN+ADA [9]	EWC [13]	CDC [26]	AdAM [25]	**Ours**
FFHQ ↦ Babies	60.91	64.84	54.94	47.45	32.74	**29.43**
FFHQ ↦ AFHQ-Cat	46.03	64.54	57.79	192.38	35.53	**35.16**

Table S8. **Intra-LPIPS of generated images during adaptation** (see details of Intra-LPIPS in Algorithm 2). The issue of incompatible knowledge cannot be addressed by training longer iterations. In particular, we empirically observe that training longer iterations will lead to severe overfitting for different existing FSIG methods, including early baseline method [19] and recent state-of-the-art methods [13,25]. Here we evaluate the diversity of generated images of different methods using Intra-LPIPS over 10-shot target samples. Lower value indicates that the generated images are less diverse and collapsing to few-shot target sample. The experiment setups are the same as Table 1 in the main paper.

Method	**Domain**	iter-0	iter-500	iter-750	iter-1000	iter-1250
TGAN [19]	Church ↦ Sailboat	0.678	0.563	0.507	0.436	0.394
EWC [13]		0.679	0.641	0.611	0.599	0.583
AdAM [25]		0.693	0.611	0.546	0.512	0.484
Ours		0.683	0.605	0.559	0.528	0.509
Method	**Domain**	iter-0	iter-500	iter-750	iter-1000	iter-1250
TGAN [19]	FFHQ ↦ AFHQ-Cat	0.664	0.620	0.563	0.545	0.517
EWC [13]		0.665	0.656	0.638	0.622	0.618
AdAM [25]		0.670	0.658	0.615	0.585	0.557
Ours		0.667	0.655	0.635	0.605	0.592

L. Future works

We review current SOTA approaches for FSIG in this work and investigate the transfer of incompatible knowledge after adaptation, which significantly reduces the realism of generated images from adapted generators. We also provide a thorough evaluation of alternative approaches using various adaption scenarios. We believe our proposed method can also be applied to other generative models and adaptation setups, as it works by adaptively eliminating filters that are incompatible with the target domain while maintaining knowledge crucial to the target domain, which is robust to different generative model architectures and training strategies. In future research, we are also interested in exploring the knowledge transfer for other generative models, including Diffusion Models [16]. Furthermore, the effects of transferring incompatible knowledge on downstream tasks is also a interesting problem to study.

M. Broader impact

M.1. Potential social and ethical impact

Throughout the paper, we demonstrate effective target adaption results using extremely small target training sample(s). Although we have achieved new state-of-the-art performance across various FSIG setups, we caution that because the FSIG adaptation of our method is lightweight, it could be quickly and cheaply applied to a real person in practice, there may be potential social and ethical issues if it is used by malicious users. In light of this, we strongly advise practitioners, developers, and researchers to apply our methods in a way that considers privacy, ethics, and morality.

M.2. Amount of computation and CO_2 emission

Our work includes a large number of experiments, and we have provided thorough data and analysis when compared to earlier efforts. In this section, we include the amount of compute for different experiments along with CO_2 emission. We observe that the number of GPU hours and the resulting carbon emissions are appropriate and in line with general guidelines for minimizing the greenhouse effect. Compared to existing works in computer vision tasks that adopt large-scale pretraining [7, 16] and consume a massive amount of energy, our research is not heavy in computation. We summarize the estimated results in Table S9.

Table S9. Estimation of amount of compute and CO_2 emission in this work. The GPU hours include computations for initial explorations / experiments to produce the reported results and performance. CO_2 emission values are computed using Machine Learning Emissions Calculator: `https://mlco2.github.io/impact/` [12].

Experiments	**Hardware Platform**	**GPU Hours (h)**	**Carbon Emission (kg)**
Main paper : Table 1 (Repeated three times)	NVIDIA A100-PCIE (40 GB)	312	33.7
Main paper : Figure 1	NVIDIA A100-PCIE (40 GB)	64	6.91
Main paper : Figure 2	NVIDIA A100-PCIE (40 GB)	112	12.1
Main paper : Figure 4 & Figure 5	NVIDIA A100-PCIE (40 GB)	87	9.4
Supplementary : Additional Experiments & Analysis	NVIDIA A100-PCIE (40 GB)	58	6.26
Supplementary : Ablation Study	NVIDIA A100-PCIE (40 GB)	16	1.73
Additional Compute for Hyper-parameter tuning	NVIDIA A100-PCIE (40 GB)	24	2.59
Total	**–**	**673**	**72.69**

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6430–6439, 2019. 6

[2] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 9

[3] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *CVPR*, 2021. 12

[4] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in Neural Information Processing Systems*, 33:16481–16494, 2020. 3

[5] Dennis Crockett. The most famous painting of the "golden twenties"? otto dix and the trench affair. *Art Journal*, 51(1):72–80, 1992. 3

[6] 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, 2009. 3

[7] 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. 13

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[9] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 11, 12

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

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 11

[12] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 13

[13] Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15885–15896. Curran Associates, Inc., 2020. 3, 6, 7, 11, 12

[14] 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. 2, 3, 11

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 3

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 12, 13

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 6

[18] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1955–1967, 2008. 3, 8, 11

[19] 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. 7, 11, 12

[20] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2019. 7

[21] 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. 8

[22] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG)*, 38(4):1–15, 2019. 3

[23] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 11

[25] Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai man Cheung. Few-shot image generation via adaptation-aware kernel modulation. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 2, 3, 6, 7, 10, 11, 12

[26] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *CVPR*, 2022. 2, 3, 10, 12