Content-Aware GAN Compression

Yuchen Liu[1], Zhixin Shu[2], Yijun Li[2], Zhe Lin[2], Federico Perazzi[2], S.Y. Kung[1]
[1]Princeton University [2]Adobe Research
[1]{yl16, kung}@princeton.edu [2]{zshu, yijli, zlin, perazzi}@adobe.com

Abstract

Generative adversarial networks (GANs), e.g., StyleGAN2, play a vital role in various image generation and synthesis tasks, yet their notoriously high computational cost hinders their efficient deployment on edge devices. Directly applying generic compression approaches yields poor results on GANs, which motivates a number of recent GAN compression works. While prior works mainly accelerate conditional GANs, e.g., pix2pix and CycleGAN, compressing state-of-the-art unconditional GANs has rarely been explored and is more challenging. In this paper, we propose novel approaches for unconditional GAN compression. We first introduce effective channel pruning and knowledge distillation schemes specialized for unconditional GANs. We then propose a novel content-aware method to guide the processes of both pruning and distillation. With content-awareness, we can effectively prune channels that are unimportant to the contents of interest, e.g., human faces, and focus our distillation on these regions, which significantly enhances the distillation quality. On StyleGAN2 and SN-GAN, we achieve a substantial improvement over the state-of-the-art compression method. Notably, we reduce the FLOPs of StyleGAN2 by 11× with visually negligible image quality loss compared to the full-size model. More interestingly, when applied to various image manipulation tasks, our compressed model forms a smoother and better disentangled latent manifold, making it more effective for image editing.

1. Introduction

Generative adversarial networks (GANs) [16] are the leading model for several crucial computer vision tasks like image generation [7, 30] and image editing [3, 4, 18, 47]. Due to their growing popularity and convincing performance, there is an increasing interest in deploying them on edge devices like mobile phones. However, state-of-the-art GANs often require large storage space, high computational cost, and great memory utility, which disallows them for efficient deployment. For example, StyleGAN2 [30] requires 45.1B/74.3B FLOPs to generate a 256px/1024px image, around 150×/250× more than MobileNet [43].

A number of network compression techniques have been developed for classification models, including weight quantization [12, 26], network pruning [17, 24, 34, 55], and knowledge distillation [22, 41]. Nonetheless, these methods are not directly applicable for GANs. For example, although removing channels with low activations [24] is effective for classifier compression, we find it not better than training a smaller GAN from scratch (Tab. 1).

As such, several specialized GAN compression mechanisms are introduced to learn efficient GAN models with the techniques of channel pruning and knowledge distillation [45, 10, 48, 8, 35]. For example, Wang et al. [48] propose GAN-Slimming (GS), which unifies losses of channel pruning and knowledge distillation and achieves the state-of-the-art compression results. However, these methods mainly target on conditional GANs (pix2pix [25], CycleGAN [58], etc.) compression, and there is little study to compress unconditional GANs (StyleGAN2, etc.). While conditional GANs normally have paired training data and perform translation from images to images, unconditional GANs are trained under completely unpaired setting and have much different source domains (white noises), which adds extra challenges for the compression. Therefore, a redesign of channel pruning and knowledge distillation schemes is required for effective unconditional GAN compression. In addition, these works also miss a significant trait of GANs that the output of GANs are images with strong spatial correlation and meaningful semantic contents. While they prune channels by weights norm or scaling factors and distill images over all spatial locations, they pay no attention on the generated contents and just treat output images as normal 3D tensors.

To combat these issues, we propose novel approaches to effectively compress unconditional GANs. We first develop an effective pruning metric to remove redundant channels and explore several distillation losses for unconditional GANs compression. Different from prior works [10, 8, 35, 48] where either a norm-based loss or a perceptual loss is used for knowledge distillation, we find that a combination

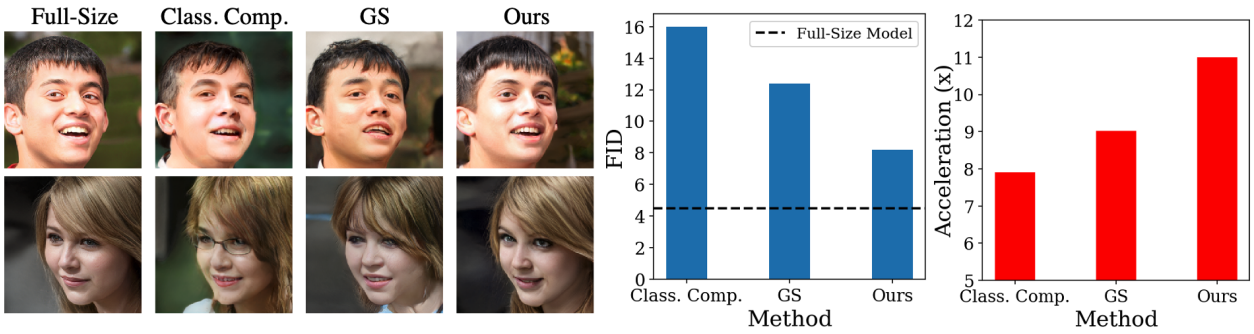

Figure 1: We demonstrate the advantage of our compression approach on StyleGAN2 over two baseline methods: (1) a conventional classification compression (Class. Comp.) approach with low activation based channel pruning [24] and norm-based knowledge distillation [41]. (2) the state-of-the-art GAN compression method, GAN-Slimming [48] (GS). **Left:** Images generated by full-size model and three compression approaches. Our results show the least artifacts and the best distillation quality. **Right:** Model statistics for three compression schemes. Our model achieves the best FID performance with the highest FLOPs acceleration ratio from the full-size model compared to two baseline methods.

of these losses improves GAN compression results. With our new pruning and distillation scheme, we achieve a major improvement in quantitative measurements over GS.

We then make the first attempt of leveraging the semantic contents in the generated images to guide the GAN compression process of both pruning and distillation. Specifically, we leverage a content-parsing network to identify contents of interest (COI), a set of spatial locations with salient semantic concepts, within the generated images. We design a content-aware pruning metric to remove channels that are least sensitive to COI in the generated images. For knowledge distillation, we focus our distillation region only to COI of the teacher's outputs which further enhances target contents' distillation. The advantage of the content-aware scheme over conventional method is not only demonstrated by a clear improvement in numerical statistics, but also visually explained in Fig. 7 and 9. Compared to a classification compression approach and GS, our compressed model enjoys better generation quality and higher computational acceleration, as shown in Fig. 1.

Our contributions are four-fold: (1) We develop a new framework of channel pruning and knowledge distillation for unconditional GANs compression, which achieves a clear improvement over prior methods quantitatively. (2) We propose a novel content-aware compression paradigm, which leverages generated contents to guide the process of pruning and distillation. Such a scheme further enhances both visual quality and numerical statistics of the compressed generators. (3) Compared to the state-of-the-art GAN compression method, GAN-Slimming [48], our method shows a major advancement in image generation, embedding, and editing on SN-GAN and StyleGAN2. (4) We find that our compressed generators not only have a better resource-performance tradeoff, but also own a smoother latent space manifold compared to the uncompressed model, which is beneficial to image editing tasks.

2. Related Work

Network Compression. To accelerate a classification network, researchers have developed techniques of weight quantization [12, 40], tensor factorization [27, 33], and network pruning [17, 34, 24, 38, 19, 55, 50, 20]. Among the network pruning approaches, a common method is to remove channels with lower activations [24] or smaller incoming weights [34, 19]. For instance, [24] removes channels with low activations by averaging their percentage of zeros, and [34, 19] use the $\ell 1$-norm of channels' incoming weights as saliency metric. However, simply applying [24] for GANs pruning would achieve merely the same performance as training from scratch. Therefore, a more specific approach to identify redundancy in GANs is needed.

Knowledge Distillation. The idea of knowledge distillation is pioneered by Hinton et al. [22] to allow a student classifier to mimic the output of its teacher. Romero et al. [41] later propose FitNets which additionally learns from teacher model's intermediate representation. While norm-based distillation scheme is widely used in distilling recognition models [51, 9, 36, 11] and has been tried on small GANs [5], applying it on the state-of-the-art unconditional GAN would result in an inferior distillation performance. Thus, a better design of knowledge distillation is required.

Content Awareness. In addition to model compression, our work is also related to image saliency/content detection [57, 44, 23, 49, 56] and semantic segmentation [6, 52, 15], which use deep networks to extract spatial information in images. Zhou et al. [57] propose a class activation mapping (CAM) mechanism, enabling a network to localize class-specific image regions. This method is generalized in [44] to more network structures and more visual tasks. Yu et al. [52] introduce BiSeNet for image segmentation and is adopted for human face parsing[1].

[1] https://github.com/zllrunning/face-parsing.PyTorch

While the notion of content awareness has been applied for image/video compression [59, 13], it was rarely used under the context of network compression. Zagoruyko et al. [53] propose an attention transfer scheme to distill the hidden layers of a student classification network. However, rather than doing a more focused distillation solely on the generated contents as in our proposal, they just treat the attention map as an additional feature map and ask the student to match this map in all spatial locations. While their method is only applied on classifiers' distillation, our content-aware scheme is designed for more challenging generative models and uses the image contents to guide both processes of pruning and distillation.

GAN Compression. A number of GAN compression works [45, 8, 10, 35, 48, 48] have been developed to address the issue of efficient GAN deployment, mainly by pruning and distillation. While [45] only uses channel pruning and [8, 10] use knowledge distillation singly, [35, 48] combine both techniques to enhance compression efficacy. In particular, Li et al. [35] select channels with large incoming weights and distill knowledge with a norm-based loss. Wang et al. [48] propose a novel GAN-Slimming (GS) approach, where they impose sparsity constraint on scaling factors for pruning and leverage a style transfer loss [28] for distillation together, with a quantization option.

Although GS achieves state-of-the-art performance on conditional GAN compression, directly applying it to unconditional GANs, such as StyleGAN2, shows sub-optimal results (Tab. 4). Hence, we propose a different pruning metric as in [35, 48] and a different knowledge distillation scheme, which leverages both norm-based loss and perceptual loss for distillation, rather than using a single loss [35, 48]. With them, we advance GS on several compression tasks. We then initiate a novel content-aware compression strategy for both pruning and distillation, which further improves model's quantitative measurement. Not surprisingly, this enhances the visual quality of both generated and edited images, especially for the contents of interest region. To the best of our knowledge, we are the first one to leverage content awareness in GAN compression.

3. Methodology

An unconditional noise-to-image GAN G maps random noises from domain $\mathcal{Z}$ to the real world images, $\mathcal{I}$. We aim to learn a compact and efficient generator G' such that: (1) their generated images $\{G(z), z \in \mathcal{Z}\}$ and $\{G'(z), z \in \mathcal{Z}\}$ have similar visual quality; (2) their embeddings of the real world images $\{Proj(G, I), I \in \mathcal{I}\}$, $\{Proj(G', I), I \in \mathcal{I}\}$ are similar. Specifically, we leverage the techniques of channel pruning and knowledge distillation, where we incorporate content awareness into both processes.

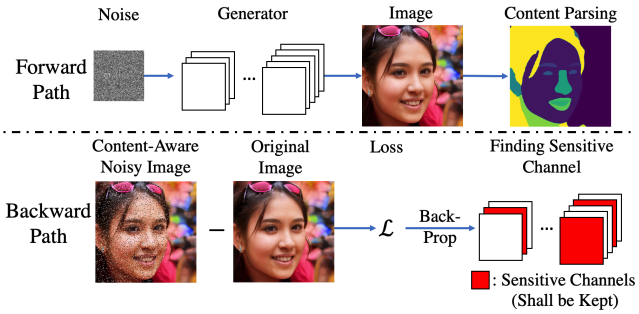

Figure 2: The content-aware pruning metric with a forward and backward path to identify informative channels.

3.1. Channel Pruning

3.1.1 Weight-Based Pruning

Let $\mathbf{W} \in \mathbb{R}^{n_{in} \times n_{out} \times h \times w}$ denote the convolutional kernel of a layer in G and we aim to quantitate the importance of the ith channel, C_i. Unlike Li et al. [35] which uses $\ell 1$-norm of C_i's incoming weights, we find that $\ell 1$-norm of C_i's outgoing weights is a better saliency indicator (shown in Fig. 7 and 8), and denote the quantity as $\ell 1$-out:

$$\ell 1\text{-out}(C_i) = ||\mathbf{W}_i||_1, \mathbf{W}_i \in \mathbb{R}^{n_{out} \times h \times w} \quad (1)$$

The channels with a high $\ell 1$-out are more informative, while the ones with lower values are redundant.

3.1.2 Content-Aware Pruning

To make the pruned model retain more information on the content of interest, we further develop a content-aware version of $\ell 1$-out (named CA-$\ell 1$-out), which is shown in Fig. 2 with a forward path and a backward path.

In the forward path, we first feedforward a latent variable $z \in \mathcal{Z}$ to G and obtain the generated image $G(z) \in \mathbb{R}^{H \times W \times 3}$. We then run a content-parsing neural network Net_p on $G(z)$, which returns a content-mask $m \in \mathbb{R}^{H \times W}, m_{h,w} \in \{0, 1\}$, where $COI = \{(h, w) | m_{h,w} = 1\}$ denotes the content of interest in the generated image $G(z)$.

For the backward path, we first add a random image noise $\mathcal{N}$ only on COI of $G(z)$ to obtain a COI-noisy images $G_{\mathcal{N}}(z)$. A differentiable loss $L_{CA}(G(z), G_{\mathcal{N}}(z))$ is then constructed between the original image $G(z)$ and the COI-noisy images $G_{\mathcal{N}}(z)$. We then back-propagate $L_{CA}(G(z), G_{\mathcal{N}}(z))$ to the network's convolution kernel $\mathbf{W}$ and get its gradient $\nabla \mathbf{g} \in \mathbb{R}^{n_{in} \times n_{out} \times h \times w}$.

Such a forward-backward procedure is iterated with multiple samples z to derive the expectation of the content-aware gradient $\mathbb{E}[\nabla \mathbf{g}]$. Finally, we measure the $\ell 1$-norm of each channel's outgoing filters' gradient as the saliency indicator and denote it as CA-$\ell 1$-out:

$$\text{CA-}\ell 1\text{-out}(C_i) = ||\mathbb{E}[\nabla \mathbf{g}]_i||_1, \ \mathbb{E}[\nabla \mathbf{g}]_i \in \mathbb{R}^{n_{out} \times h \times w} \quad (2)$$

Intuitively, channels with larger CA-$\ell 1$-out are more sensitive to COI of the generated images and shall be kept in the

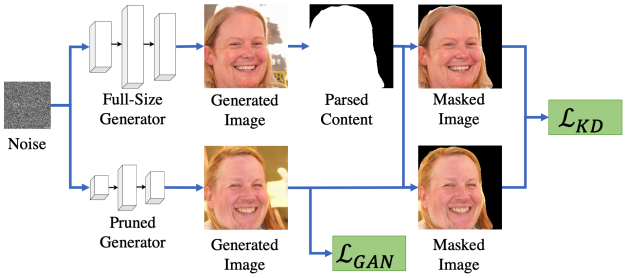

Figure 3: Our content-aware knowledge distillation (CA-KD) scheme, where the minmax GAN loss is applied on the pruned model's generated images while the knowledge distillation loss is imposed on the content-masked images from the full-size teacher and pruned student.

pruning process. Such a content-aware metric selects more informative channels, as shown in Fig. 7 and 8.

3.2. Knowledge Distillation

3.2.1 Pixel-Level Distillation

Intuitively, we can impose losses to reduce the norm-difference between the activations and outputs of G' and G. Based on where the distillation losses are inserted, we can categorize the norm-based distillation loss into two types: output only distillation and intermediate distillation. For output only distillation, we construct our loss as:

$$\mathcal{L}_{KD}^{norm} = \mathbb{E}_{z\in\mathcal{Z}}[||G(z), G'(z)||_1] \tag{3}$$

We can also do intermediate distillation as:

$$\mathcal{L}_{KD}^{norm} = \sum_{t=1}^{T} \mathbb{E}_{z\in\mathcal{Z}}[||G_t(z), f_t(G'_t(z))||_1], \tag{4}$$

where $G_t(z)$ /$G'_t(z)$ are the intermediate activations of layer t and $G_T(z)$/$G'_T(z)$ are the output images. f_t is a linear transform to match the depth dimension of the activations.

3.2.2 Image-Level Distillation

Apart from learning low-level details from the teacher, we also want the student to generate perceptually similar outputs. To achieve this, we adopt neural network based perceptual metrics. Unlike GAN-Slimming [48], which uses a style transfer loss [28] for distillation, we propose to use LPIPS [54] as our perceptual distillation loss, which measures the perceptual distance between output images from two generators:

$$\mathcal{L}_{KD}^{per} = \mathbb{E}_{z\in\mathcal{Z}}[LPIPS(G(z), G'(z))] \tag{5}$$

In our experiments, we find that LPIPS is a better than [28] for unconditional GAN compression.

3.2.3 Content-Aware Distillation

Similar to pruning, we introduce content-awareness to knowledge distillation as shown in Fig. 3, where we focus our distillation on specific contents. We first feedforward a latent variable z to obtain both networks' generated images, $G(z)$ and $G'(z)$. Then, we run the content-parsing network Net_p on $G(z)$ and get its content mask m. Based on COI of m, we compute two masked images, $G(z)_m = G(z) \odot m$ and $G'(z)_m = G'(z) \odot m$, where $\odot$ denotes the element-wise multiplication in images' spatial domain. Under this scheme, $\mathcal{L}_{GAN}$ will be imposed on the unmasked generated image of the pruned network, $G'(z)$, while the distillation loss for the output images are measured between the masked images $G(z)_m$ and $G'(z)_m$. Such a content-aware scheme allows a more attentive distillation for the generated contents, as evidenced in Fig. 9.

3.2.4 Training Objectives

In summary, our training loss for the pruned generator G' can finally be formulated as:

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda\mathcal{L}_{KD}^{norm} + \gamma\mathcal{L}_{KD}^{per} \tag{6}$$

where $\mathcal{L}_{GAN}$ is the minmax objective for GAN training, and λ and γ are the weights for the knowledge distillation losses. Unlike prior works [35, 48] singly uses either pixel-level distillation loss $\mathcal{L}_{KD}^{norm}$ or image-level distillation loss $\mathcal{L}_{KD}^{per}$, we find that it is necessary to combine both of them for an enhanced distillation performance. Moreover, applying $\mathcal{L}_{KD}^{norm}$ and $\mathcal{L}_{KD}^{per}$ under the content-aware scheme further improves distillation quality.

For $\mathcal{L}_{GAN}$, we derive the student generator G' by pruning G, while initializing the student discriminator D' with the same architecture and pre-trained weights as teacher discriminator D. We fine-tune both G' and D' by a standard minmax optimization scheme [16].

4. Experimental Results

We carry out compression experiments on models with different computation budgets and datasets with diverse image resolutions to show the general effectiveness of our approach. Specifically, we investigate into the following tasks: SN-GAN [39] on CIFAR-10 [31] at 32px, StyleGAN2 [30] on FFHQ dataset [29] at 256px and at 1024px.

4.1. Evaluation Metrics

We use the following five quantitative metrics to evaluate the image generation and image projection performance of a GAN: Inception Score (IS) [42], Fréchet Inception Distance (FID) [21], Perceptual Path Length (PPL) [29], and PSNR/LPIPS between real and projected images.

Inception Score.[2] IS is proposed to measure the classification quality of the generated images. Specifically, it awards high scores to a generator whose generated images

[2]We use https://github.com/tsc2017/Inception-Score

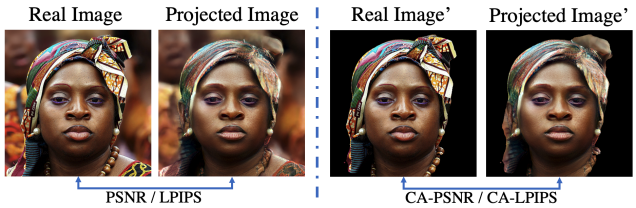

Figure 4: Image projection evaluation. We measure the PSNR/LPIPS for the pair of real and projected images, as well as their content-aware masked images.

could be classified by an inception classifier [46] with high confidence while having a diverse label distribution.

Fréchet Inception Distance.[3] FID quantitates the similarity between the synthetic images from a generator and the real-world images. It is computed by feed-forwarding two sets of images to an inception network followed by a Fréchet Distance [14] measurement between their corresponding activation features.

Perceptual Path Length. PPL is proposed to measure the smoothness of a StyleGAN's latent space [29]. It is derived by calculating the LPIPS distance between two images generated by a pair of little-perturbed latent codes. Our PPL implementation has two slight modifications from [29]: (1) instead of using a reimplemented LPIPS in TensorFlow [2], we use the original LPIPS implementation[4]; (2) we fix the perturbation factor $\epsilon = 10^{-4}$, and drop the scaling term $\frac{1}{\epsilon^2}$ to make the score more intuitive.

Image Projection. Our image projection evaluation method is shown in Fig. 4. We sample 55 real human face images from Helen [32] (not appeared in any training dataset) with various lighting conditions, genders, ethnicities, and ages, and name the dataset as Helen-Set55, shown in the Supplementary. We run an L-BFGS optimizer [37] with 200 iterations to find a real image's latent code in a StyleGAN2. We feed-forward this latent code and obtain a projected image. We average LPIPS and PSNR for all pairs of Helen-Set55 images and projected images.

Moreover, in many image editing applications [1], we mostly care about the projection quality of the content of interest region. Thus, we further propose a content-aware projection evaluation scheme, where we measure the PSNR and LPIPS between the content-aware masked real and projected images, denoted as CA-PSNR and CA-LPIPS.

4.2. Pruning Effectiveness

We first examine the effectiveness of our channel pruning metric on a 256px StyleGAN2. Given a pretrained generator, we uniformly remove 30% of channels from each layer by $\ell 1$-out or other metrics. Three baselines are included to show our effectiveness: (1) training from scratch by keeping the pruned network structure while re-initializing the weights; (2) a conventional classification pruning method, to remove channels with low activations[5] [24]; (3) random pruning. These pruned networks are then fine-tuned by the vanilla training loss, $\mathcal{L}_{GAN}$, where the discriminators see 2.9M real examples in total.

Model	Image Size	FLOPs	FID (↓)
Original Full-Size	256	45.1B	4.5
Compressed Models			
Low-Act Pruning [24]	256	22.3B	7.9
Training from Scratch	256	22.3B	8.1
Random Pruning	256	22.3B	6.2
$\ell 1$-out Pruning	256	22.3B	**5.4**

Table 1: Pruning metric effectiveness investigation. We use FID (the lower the better) to compare $\ell 1$-out pruning with three other baselines.

As shown in Tab. 1, the low-act pruned model has merely the same FID as training from scratch, even worse than random pruning. This indicates that directly applying classification pruning metric can fail on GAN compression. Moreover, we find that the $\ell 1$-out pruned generator has an only 0.9 FID loss from the full-size model with 50% less FLOPs, and it achieves the best FID among compared methods.

4.3. Knowledge Distillation Schemes

We then analyze the effectiveness of different knowledge distillation losses under a high acceleration ratio. We use $\ell 1$-out to uniformly remove 80% of channels from each layer of a full-size StyleGAN2, resulting in an 25× FLOPs-accelerated generator. We then retrain the pruned model with 4.3M real examples by six combinations of distillation losses, which are specified by λ, γ, and the type of norm-based distillation. Moreover, we include the VGG style transfer loss [28] used in GAN-Slimming distillation [48] for our comparison. For intermediate distillation, we choose the outputs of the to_rgb modules to construct the loss in Eqn. 3. As the depth dimension of to_rgb outputs are always 3, we can fix $f_t(x) = x$.

As shown in Tab. 2, the VGG style transfer loss yields inferior results to LPIPS which suggests the need for redesign distillation loss for unconditional GANs compression. Moreover, the output only $\mathcal{L}_{KD}^{norm}$ + LPIPS $\mathcal{L}_{KD}^{per}$ distillation scheme achieves the best quantitative results for both image generation projection: (1) it achieves the best FID score of 12.5 which has an FID improvement of 2.6 over the no KD scheme. (2) it achieves the best image projection results with a PSNR of 31.03 and LPIPS of 0.170. Such results are interesting that: (1) rather than prior work which only adopts a single loss distillation scheme [35, 48], we find that it is necessary to distill knowledge both at the pixel-level and image-level for enhanced results; (2) while

[3] We use https://github.com/mseitzer/pytorch-fid

[4] https://github.com/richzhang/PerceptualSimilarity

[5] Rather than counting average percentage of zeros, we choose to measure activations' $\ell 1$-norm to remove low activation maps, as the activation function in StyleGAN2 is leaky ReLU, not ReLU.

Model	Image Size	$\mathcal{L}_{KD}^{norm}$ Type	λ	$\mathcal{L}_{KD}^{per}$ Type	γ	FLOPs	FID (↓)	PSNR (↑)	LPIPS (↓)
Original Full-Size	256	-	-	-	-	45.1B	4.5	32.02	0.113
Compressed Models									
No KD	256	-	-	-	-	1.9B	15.1	30.88	0.182
$\mathcal{L}_{KD}^{norm}$ Only	256	Output Only	3	-	-	1.9B	14.2	30.79	0.186
$\mathcal{L}_{KD}^{norm}$ Only	256	Intermediate	3	-	-	1.9B	14.4	30.78	0.187
$\mathcal{L}_{KD}^{per}$ Only	256	-	-	VGG	3	1.9B	14.6	30.81	0.186
$\mathcal{L}_{KD}^{per}$ Only	256	-	-	LPIPS	3	1.9B	13.6	30.91	0.177
$\mathcal{L}_{KD}^{norm}$ + $\mathcal{L}_{KD}^{per}$	256	Intermediate	3	LPIPS	3	1.9B	13.6	30.76	0.180
$\mathcal{L}_{KD}^{norm}$ + $\mathcal{L}_{KD}^{per}$	256	Output Only	3	LPIPS	3	1.9B	**12.5**	**31.03**	**0.170**

Table 2: Results of different knowledge distillation schemes.

Model	Img Size	FLOPs	Param.	IS (↑)
Ori. Full-Size	32	1.60B	4.27M	8.37±0.11
Compressed Models				
GS [48]	32	1.11B	3.38M	8.01
Ours	32	0.83B	2.23M	8.36±0.12
Ours-CA	32	0.83B	2.23M	**8.36±0.08**
GS [48]	32	0.51B	2.19M	7.65
Ours	32	0.42B	1.20M	8.21±0.11
Ours-CA	32	0.42B	1.20M	**8.31±0.08**

Table 3: Comparison to the state-of-the-art method, GAN-Slimming, with SN-GAN on CIFAR-10.

distilling the knowledge in the intermediate features improves conditional GAN's performance [35], it does not help for the case of unconditional GAN like StyleGAN2.

We regard the output only $\mathcal{L}_{KD}^{norm}$ + LPIPS $\mathcal{L}_{KD}^{per}$ as our best KD loss which are used in the following experiments.

4.4. Comparison to the State of the Art

To further demonstrate the effectiveness of our approach, we compare our scheme with the state-of-the-art GAN compression method, GAN Slimming (GS) [48]. The results are summarized in Tab. 3 and 4, where we consistently outperform GS in all measurements.

4.4.1 Experimental Settings

We compare compressed models by three different approaches: **Ours**, **Ours-CA** and **GS**.

Ours. We use $\ell 1$-out for channel pruning with the best KD loss in Sec. 4.3 for fine-tuning. We set $\lambda = 3$ and $\gamma = 3$.

Ours-CA. We use CA-$\ell 1$-out as the pruning metric and apply the best KD loss with content-aware distillation (Fig. 3). On CIFAR-10, we use the class activation mapping (CAM) [57] to detect generated images' COI. For FFHQ, we use a BiSetNet [52] to parse human faces. The COI is the entire human face, i.e., excluding the clothes and the image background. For CA-$\ell 1$-out, we use the salt-and-pepper noise as $\mathcal{N}$, and use an $\ell 1$-loss for L_{CA}. For content-aware distillation, we use the same value for λ and γ.

GS. We use the numbers reported in the original paper for the SN-GAN comparison. As GS has not been applied to StyleGAN2 in the previous work, we make our own implementation following their three-step method: (1) fine-tune the full-size network with a combination of a minmax GAN loss, an $\ell 1$ sparsity loss on scaling factors, and a VGG distillation loss [28]; (2) remove channels with zero scaling factors in the tuned network; (3) fine-tune the pruned network with the GAN loss and the VGG style transfer loss.

On SN-GAN, our discriminator sees 1.6M images in the fine-tuning process. On StyleGAN2, the discriminators in both our scheme and **GS** see the same number of images, 7.5M, ensuring the fairness for the comparison. Extra experiment details are included in the Supplementary.

4.4.2 Results

On SN-GAN, both of our compressed models have no IS loss at 2× acceleration with around 0.35 IS gain over **GS**, as shown in Tab. 3. At 4× acceleration level, our method is even more promising: while **Ours** has a 0.56 IS increase compared to **GS**, **Ours-CA** can further improve **Ours** by 0.1 on IS. Such results clearly show the advantage of our content-aware GAN compression scheme.

On 256px StyleGAN2, **Ours** enjoys an 11× acceleration from the full-size model and achieves 3.5 FID gain over the 9×-accelerated **GS** model, shown in Tab. 4. **Ours-CA** further improves **Ours** by 1.0 on FID and advances the projection performance, especially for the COI projection measured by CA-PSNR/CA-LPIPS. We also note **Ours-CA** achieves a much lower PPL compared to **GS**, and even smaller than that of the full-size model. This indicates that our content-aware GAN compression scheme can not only improve model's efficiency, but also the smoothness of its latent space. This PPL improvement is further examplified by visual evidence in Fig. 5 and 6.

At 1024px resolution, we can only obtain a 3.1×-accelerated generator by **GS** where overpruning would not guarantee the generator to converge in the learning process. With 3.4× acceleration over **GS**, **Ours/Ours-CA** enjoy an FID improvement of 2.0/2.5 and better image projection performance. **Ours-CA** again achieve the best performance in image generation and image projection.

4.5. Image Editing

We further demonstrate the benefit of our content-aware compressed StyleGAN2 for editing tasks of style mixing, latent space image morphing, and a recent proposed tech-

Model	Image Size	FLOPs	FID(↓)	PPL(↓)	PSNR(↑)	LPIPS(↓)	CA-PSNR(↑)	CA-LPIPS(↓)
Original Full-Size	256	45.1B	4.5	0.162	32.02	0.113	33.03	0.076
Compressed Models								
GS [48] (Our Impl.)	256	5.0B	12.4	0.313	31.02	0.177	32.39	0.117
Ours	256	4.1B	8.9	0.145	31.37	0.149	32.67	0.099
Ours-CA	256	4.1B	**7.9**	**0.143**	**31.41**	**0.144**	**32.75**	**0.096**
Original Full-Size	1024	74.3B	2.7	0.162	31.38	0.149	32.67	0.096
Compressed Models								
GS [48] (Our Impl.)	1024	23.9B	10.1	0.211	30.74	0.189	32.17	0.121
Ours	1024	7.0B	8.1	0.157	30.94	0.174	32.31	0.113
Ours-CA	1024	7.0B	**7.6**	**0.157**	**30.96**	**0.170**	**32.33**	**0.111**

Table 4: Comparison to the state-of-the-art method, GAN-Slimming, with 256px/1024px StyleGAN2 on FFHQ.

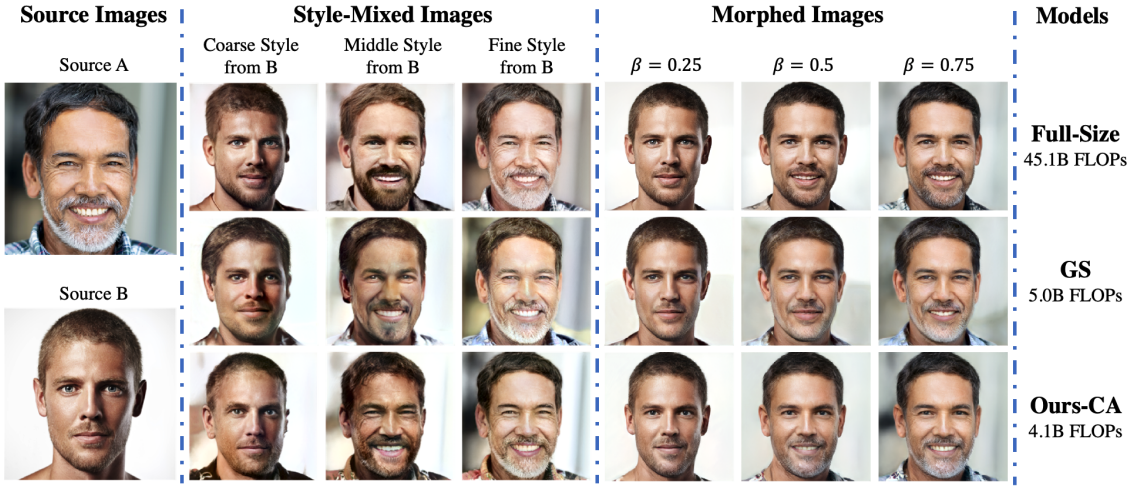

Figure 5: An example typifying the effectiveness of our compressed StyleGAN2 for image style-mixing and morphing. When we mix middle styles from B, GS produces a blurred face with uneven skin texture while the original full-size model has a significant identity loss. In contrast, our approach better preserves the person's identity with less artifacts. We observe that our morphed images have a smoother expression transition compared to GS's in mouth and teeth, and our beard transition is even better than the full-size model, substantiating our advantage in latent space smoothness.

nique, GANSpace [18]. More results are included in the Supplementary.

Style Mixing and Morphing. Given two real world images, we run our image projection algorithm in Sec. 4.1 to find their W^+ embedding latent codes, followed by two manipulations of these codes [29, 3]: (1) crossover the codes at layer $l \in [1{:}L]$ for image style-mixing; (2) interpolate the codes with parameter $\beta \in [0, 1]$ for image morphing.

We show an example in Fig. 5 with a 256px model where we set $l \in \{4, 8, 11\}$ and $\beta \in \{0.25, 0.5, 0.75\}$. While GS yields a number of artifacts (skin texture, hair, etc.) for style-mixing, our model performs comparatively to the full-size model in image quality and even preserves a better identity (in middle style mixing). We find visual evidence for our advantage of having a smoother latent space in the morphed images, where our expression transition is much smoother than GS in mouth and teeth and our beard transition is even better than the full-size model. This agrees well with the fact that our model have a lower PPL score.

GANSpace Editing. We further deploy our compressed 1024px model for GANSpace [18] editing. Specifically, we use PCA to find principle components of the W space, and traverse a latent code in the direction of a component to generate a sequence of images.

We show an example in in Fig. 6, where we use the same latent code as in the original paper [18] and traverse it in the direction of the first principal component, $\mathbf{u}_0$. While it is claimed that $\mathbf{u}_0$ is a direction for gender editing, we find that the full-size model also changes person's age along $\mathbf{u}_0$. The full-size model also produces artifacts at the chin of the generated images at large deviation. In contrast, our compressed model only changes the person's gender and generates more natural images at different variation scales. This again visually indicates that our compressed model has a smoother and more disentangled latent manifold.

5. Ablation Study

Pruning Effectiveness. We conduct a channel selection analysis on StyleGAN2 with 5 pruning metrics: low-

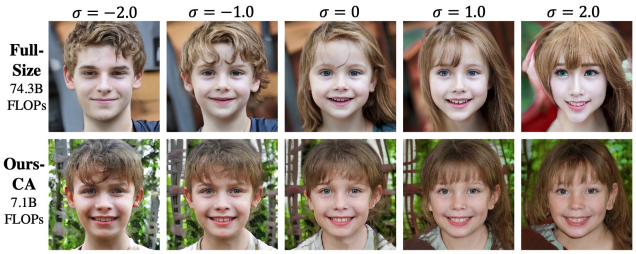

Figure 6: A demonstration of more effective GANSpace direction discovery with our compressed model. The direction is suggested for change of gender [18], yet the full-size model also changes the age significantly along the direction. In contrast, our compressed model retains the age and produces less artifact at large variation ($\sigma = 2.0$). These results suggest that our GANSpace direction is more disentangled, which indicates our latent space is more linearly separable.

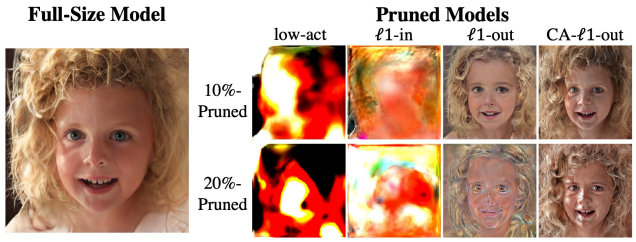

Figure 7: Effectiveness of our pruning metrics. **Left:** An image generated by full-size model. **Right:** Images generated by 8 pruned models (without retraining) varying layer pruning ratios (rows) and pruning metrics (columns). We find that both of our metrics $\ell 1$-out and CA-$\ell 1$-out achieves much better visual quality than low-act [24] and $\ell 1$-in [35], while CA-$\ell 1$-out enjoys the best perceptual performance.

act [24], $\ell 1$-in [35], random, $\ell 1$-out, and CA-$\ell 1$-out. Specifically, we create 20 pruned models by pruning the original full-size generators with these 5 metrics at 4 layer remove ratios (10%, 20%, 30%, 40%) without fine-tuning.

We randomly sample a latent variable and obtain the output images from the full-size model and the pruned models as shown in Fig. 7. We can clearly find that pruning low-activation (Col. 1) channels would distort the output images completely, and the prior conditional GAN compression metric, $\ell 1$-in (Col. 2), also fails to identify informative channels. We find that CA-$\ell 1$-out (Col. 4) preserves the most informative channels for image generation. We also plot the FID scores of these pruned models in Fig. 8, which correlates well with our visual judgement that CA-$\ell 1$-out achieves the best performance.

Distillation Effectiveness. We further demonstrate the advantage of our content-aware knowledge distillation (CA-KD) scheme over the all spatial locations distillation (AS-KD) method. As shown in Fig. 9, although the generated images from AS-KD might have similar backgrounds and clothes to the full-size model, our CA-KD scheme generates images with closer COI features (beard, eyes, glasses, etc.) as its teacher. Such characteristic not only improves model's generation quality (FID), but also explains the enhancement in model's image embedding quality (PSNR/LPIPS), especially for the COI region (CA-PSNR/LPIPS), where it owns a much better distillation.

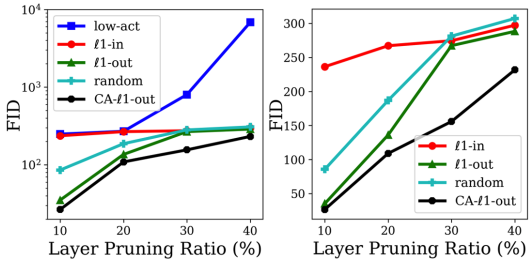

Figure 8: FID of pruned models obtained by different pruning metrics at different pruning ratios. **Left:** FID in log scale with low-act [24]. **Right:** FID in normal scale without low-act. Our CA-$\ell 1$-out metric best identifies informative channels quantitatively.

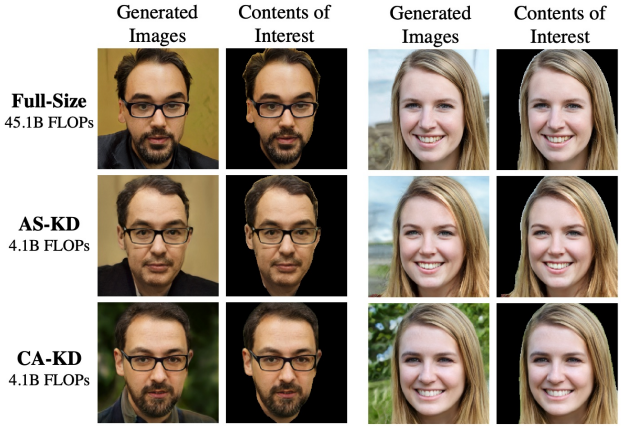

Figure 9: Effectiveness of content-aware distillation (CA-KD) scheme. Compared to all spatial locations distillation (AS-KD), the model learned by the CA-KD has better identity preservation, more similar glasses, beard and eyes in the content of interest region as its full-size teacher.

6. Conclusion

In this work, we propose a novel content-aware compression pipeline to learn efficient GANs. While prior works mainly focus on conditional GANs compression, we study a new approach of channel pruning and knowledge distillation under the context of unconditional GANs, and further introduce a content-aware version for both compression techniques. We carry out experiments on SN-GAN and StyleGAN2 to show the effectiveness of our scheme, where we outperform the state-of-the-art method on all tasks. Moreover, our compressed models not only enjoy a better resource-performance tradeoff compared to the full-size one, but also owns an extra advantage of smoother latent space manifold by numerical and visual evidences.

References

[1] *Adobe Photoshop*. https://www.adobe.com/products/photoshop.html.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019.

[4] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[5] Angeline Aguinaldo, Ping-Yeh Chiang, Alex Gain, Ameya Patil, Kolten Pearson, and Soheil Feizi. Compressing gans using knowledge distillation. *arXiv preprint arXiv:1902.00159*, 2019.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[8] Ting-Yun Chang and Chi-Jen Lu. Tinygan: Distilling biggan for conditional image generation. *arXiv preprint arXiv:2009.13829*, 2020.

[9] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017.

[10] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu. Distilling portable generative adversarial networks for image translation. *arXiv preprint arXiv:2003.03519*, 2020.

[11] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv preprint arXiv:1707.01220*, 2017.

[12] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[13] Yiping Duan, Yaqiang Zhang, Xiaoming Tao, Chaoyi Han, Mai Xu, Cheng Yang, and Jianhua Lu. Content-aware deep perceptual image compression. In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE, 2019.

[14] Maurice Fréchet. Sur la distance de deux lois de probabilité. *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES*, 244(6):689–692, 1957.

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[17] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[18] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

[19] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018.

[20] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019.

[21] 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. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[23] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[24] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[26] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.

[27] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014.

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

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

[30] 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.

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[32] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[33] V Lebedev, Y Ganin, M Rakhuba, I Oseledets, and V Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. In *3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings*, 2015.

[34] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[35] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. Gan compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5284–5294, 2020.

[36] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14639–14647, 2020.

[37] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[38] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2736–2744, 2017.

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

[40] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

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

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

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

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[45] Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, and Chang Xu. Co-evolutionary compression for unpaired image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3235–3244, 2019.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[47] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. *arXiv preprint arXiv:2003.03581*, 2020.

[48] Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang. Gan slimming: All-in-one gan compression by a unified optimization framework. *arXiv preprint arXiv:2008.11062*, 2020.

[49] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[50] Jianbo Ye, Xin Lu, Zhe Lin, and James Z Wang. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. *arXiv preprint arXiv:1802.00124*, 2018.

[51] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

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

[53] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[55] Tianyun Zhang, Shaokai Ye, Kaiqi Zhang, Jian Tang, Wujie Wen, Makan Fardad, and Yanzhi Wang. A systematic dnn weight pruning framework using alternating direction method of multipliers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–199, 2018.

[56] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019.

[57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[59] Fabio Zünd, Yael Pritch, Alexander Sorkine-Hornung, Stefan Mangold, and Thomas Gross. Content-aware compression using saliency-driven image retargeting. In *2013 IEEE International Conference on Image Processing*, pages 1845–1849. IEEE, 2013.