AutoGAN-Distiller: Searching to Compress Generative Adversarial Networks

Yonggan Fu [1] **Wuyang Chen** [2] **Haotao Wang** [2] **Haoran Li** [1] **Yingyan Lin** [1] **Zhangyang Wang** [2]

Abstract

The compression of Generative Adversarial Networks (GANs) has lately drawn attention, due to the increasing demand for deploying GANs into mobile devices for numerous applications such as image translation, enhancement and editing. However, compared to the substantial efforts to compressing other deep models, the research on compressing GANs (usually the generators) remains at its infancy stage. Existing GAN compression algorithms are limited to handling specific GAN architectures and losses. Inspired by the recent success of AutoML in deep compression, we introduce AutoML to GAN compression and develop an *AutoGAN-Distiller* (**AGD**) framework. Starting with a specifically designed efficient search space, AGD performs an end-to-end discovery for new efficient generators, given the target computational resource constraints. The search is guided by the original GAN model via knowledge distillation, therefore fulfilling the compression. AGD is **fully automatic**, **standalone** (i.e., needing no trained discriminators), and **generically applicable** to various GAN models. We evaluate AGD in two representative GAN tasks: image translation and super resolution. Without bells and whistles, AGD yields remarkably lightweight yet more competitive compressed models, that largely outperform existing alternatives. Our codes and pretrained models are available at: https://github.com/TAMU-VITA/AGD.

1. Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014; Zhu et al., 2017) nowadays have empowered many real applications such as image stylization, image editing and enhancement. Those applications have driven the growing demand to deploy GANs, usually their trained **generators**, on resource-constrained platforms, e.g., for real-time style transfer and super resolution (Shi et al., 2016) in mobile APPs. However, just like other deep models, GAN generators require heavy memory and computation resources to run, which challenge most mobile devices. For example, the renowned image-to-image translation network, CycleGAN (Zhu et al., 2017), would cost 54 GFLOPs to process one 256×256 image. For most mobile phones, that latency would notably degrade user experience, if ever feasible.

To reduce this gap, it is a natural motive to refer to model compression techniques (Han et al., 2015). The current mainstream of model compression focuses on deep image classification or segmentation, which, as demonstrated in (Shu et al., 2019), cannot be easily extended to compress GANs (by default, defined as compressing their trained generators), That is mainly due to the vulnerability of their learned mappings (i.e., between high-dimensional structured image spaces), and the notoriously training instability (that challenges the re-training practice in compression). To our best knowledge, *at the time of submission*, (Shu et al., 2019) is the only published algorithm in GAN compression. They specifically focus on GANs with the cycle-consistency loss, and simultaneously compress generators of both directions. Despite the promising performance, their algorithm cannot be straightforwardly extended to other popular GAN types, such as encoder-decoder GANs (Wang et al., 2018a; Chen et al., 2018b) which are widely used in image editing/enhancement. Moreover, the proposed co-evolution algorithm was built on pruning only, while many other compression techniques, such as quantization (Jacob et al., 2018), knowledge distillation (Polino et al., 2018), and even AutoML (He et al., 2018), were left unexplored. Besides, their algorithm's loss function relied on the original trained discriminator to be available. Yet in practice, this might not always be realistic: as most applications only use trained generators, the discriminators might often have been discarded or no longer available after training.

This paper aims to *significantly push forward the application frontier of GAN compression*. For this new problem, we extensively refer to the state-of-the-art compression techniques including AutoML and knowledge distillation, while striving for more general techniques that can compress any

[1]Rice University, Houston, Texas, USA [2]Texas A&M University, College Station, Texas, USA. Correspondence to: Zhangyang Wang <atlaswang@tamu.edu >, Yingyan Lin <yingyan.lin@rice.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

common GAN generator. The resulting framework, dubbed *AutoGAN-Distiller* (**AGD**), is the first AutoML framework dedicated to GAN compression (alongside the concurrent work (Li et al., 2020)), and is also among a few earliest works that explore AutoML for GANs (Gong et al., 2019). AGD is established on a specifically designed search space of efficient generator building blocks, leveraging knowledge from state-of-the-art GANs for different tasks. It then performs differentiable neural architecture search under the target compression ratio (computational resource constraint), which preserves the original GAN generation quality via the guidance of knowledge distillation (Polino et al., 2018). AGD makes no assumption towards GAN structures, loss forms, nor even the availability of trained discriminators. It is therefore broadly applicable to compressing both CycleGANs and other GANs without the cycle loss. We demonstrate AGD on two representative mobile-based GAN applications: unpaired image translation (using a CycleGAN), and super resolution (using an encoder-decoder GAN). In both tasks, AGD overwhelms the state-of-the-arts (Li et al., 2016; Shu et al., 2019; Ledig et al., 2017; Kim et al., 2016).

2. Related Work

2.1. AutoML: Neural Architecture Search

As one of the most significant sub-fields of AutoML (Hutter et al., 2019), Neural Architecture Search (NAS) (Zoph & Le, 2016) seeks an optimal neural network architecture from data, instead of using hand-crafting. Already by now, NAS methods have outperformed manually designed architectures on a range of tasks such as image classification (Tan & Le, 2019; Tan et al., 2019; Howard et al., 2019; Liu et al., 2018a), and segmentation (Chen et al., 2018a; Liu et al., 2019; Chen et al., 2019b). A comprehensive survey of NAS could be found in (Elsken et al., 2018)

Introducing NAS to GANs appears to more challenging due to the training instability and lack of direct performance measures. (Gong et al., 2019) developed the first NAS framework for GANs, on the task of unconditional image generation from random noise. It adopted a multi-level search strategy with an RNN controller to search for the generator architecture using reinforcement learning. However, the existing NAS for GAN framework is **not directly applicable** in our case due to the following reasons:

- The framework in (Gong et al., 2019) searched for a GAN from scratch, without referring to a trained model. There is also no computational resource constraint in their search. It is hence not a "compression" technique and lacks the ability to distill knowledge from those already-trained, high-performance GANs.
- (Gong et al., 2019) only demonstrated to find relatively small GAN models that synthesize low-resolution images from noise (e.g., a 6-layer generator on 32×32 CIFAR-10 images), and also did not consider image-to-image GANs. In comparison, the desired GAN compression shall apparently be able to handle the heavily-parameterized GANs on high-resolution images, with image-to-image applications (stylization, editing, enhancement, etc) as the primary focuses.

2.2. Model Compression

With the unprecedented demand for deploying deep models on resource-constrained platforms, many compression methods are proposed to narrow the gap between the model complexity and on-device resources (Han et al., 2020; Wang et al., 2018b; Chen et al., 2020a). Popular techniques include knowledge distillation (Polino et al., 2018), pruning (Han et al., 2015) and quantization (Jacob et al., 2018).

Knowledge distillation was proposed in (Hinton et al., 2015) to transfer the knowledge from one model to another by imitating the soft labels generated by the former. It was widely applied in learning a compact network by fitting the soft labels generated by a pretrained large model (Lopez-Paz et al., 2015; Bulò et al., 2016; Wang et al., 2018c; Chen et al., 2019a; 2020b), therefore compressing the latter.

Pruning (Han et al., 2015) refers to sparsifying the model weights by thresholding the unimportant weights. All of them follow a similar workflow: (iteratively) pruning the model to a smaller size and then retraining to recover accuracy. Structured pruning, e.g. channel pruning (Wen et al., 2016; He et al., 2017), was widely adopted as a hardware-friendly compression means (Gui et al., 2016). (Luo et al., 2017) also suggested to prune filters based on the statistics from the next layer to evaluate a filter's importance.

Quantization reduces the float-number representations of weights and activations to lower numerical precision (Wu et al., 2016; Han et al., 2015). The extreme case could even just use binary values (Courbariaux et al., 2015; Rastegari et al., 2016). Beyond scalar quantization, vector quantization was widely adopted too in model compression for parameter sharing (Gong et al., 2014; Wu et al., 2018).

Recently, (He et al., 2018; Liu et al., 2018b) pointed out that conventional model compression techniques rely on domain experts to explore the large design space trading off, which could be sub-optimal and tedious. The idea of AutoML for model compression has been extended by a number of works, e.g, (Wu et al., 2019; Cheng et al., 2018; Tan & Le, 2019; Tan et al., 2019; Cheng et al., 2018), although most of them focused on compressing deep classifiers.

2.3. GANs and GAN Compression

GANs have witnessed prevailing success in many practical applications built on image-to-image translation. *CycleGAN* (Zhu et al., 2017) and its many successors pioneered

in demonstrating GAN's power in image-to-image translation without paired training data. They explored cycle-consistency loss to avoid mode collapse, enabling both diverse and semantically consistent synthesis, therefore making them popular choices in style transfer-type applications. However, GANs with a cycle loss are often costly and difficult to train. Recent works have found that *encoder-decoder GANs*, i.e., fully feedforward generators and discriminators with no cycle structures, can also perform on par, with or without paired training data. These encoder-decoder GANs also gain popularity in image enhancement (Wang et al., 2018a; Kupyn et al., 2019; Jiang et al., 2019) and editing (Sanakoyeu et al., 2018; Yang et al., 2019).

Substantial evidences reveal that the generative quality of GANs consistently benefits from larger-scale training (Karras et al., 2017; Brock et al., 2018; Gui et al., 2020). However, the GAN training is notoriously unstable, and therefore numerous techniques were proposed to stabilize the training of increasingly larger GANs, including spectral normalization (Miyato et al., 2018), gradient penalty (Gulrajani et al., 2017) and progressive training (Karras et al., 2017).

Despite their performance boost, the growing complexity of those large GANs conflicts the demands of mobile deployments, calling for compression techniques. To our best knowledge, (Shu et al., 2019) is the only published method in GAN compression at our submission time. The authors experimentally showed that classical channel pruning would fail when applied to compress GANs, due to the sensitivity of GAN training. They thus proposed a co-evolution algorithm to regularize pruning, that relied on the cycle-consistency loss. As a result, their algorithm was restricted to compressing CycleGAN or its variants, and cannot be applied to encoder-decoder GANs. It also needed the trained discriminator of original GANs as part of the pruning loss.

A concurrent work (available after our submission time) that we shall credit is (Li et al., 2020), which also presents AutoML for GAN compression. Its uniqueness lies in searching the channel width for an existing generator. In a "once-for-all" manner, their searched models achieved impressive image translation performance without fine-tuning. In comparison, our proposed AGD framework customizes the search space for tasks with different properties, searches for operators types in addition to channel widths, and additionally evaluates on a super-resolution task.

3. Our AutoGAN-Distiller Framework

Given a pretrained generator G_0 over the data $\chi = \{x_i\}_{i=1}^N$, our goal is to obtain a more compact and hardware-friendly generator G, from G_0 while the generated quality does not sacrifice much, e.g., $G_0(x) \approx G(x), x \in \chi$.

We adopt the differentiable design (Liu et al., 2018a) for our NAS framework. A NAS framework consists of two key components: the search space and the proxy task. AGD customizes both these two components for the specific task of GAN compression, as described in the following sections.

3.1. The Proposed Customized Search Space

General Design Philosophy. Many NAS works adopt the directed acyclic graph (DAG)-type search spaces (Liu et al., 2018a) due to its large volume and flexibility. However, those can result in lots of irregular dense connections. As discussed in (Ma et al., 2018), such network fragmentation is hardware-unfriendly, as it will significantly hamper the parallelism and induce latency/energy overheads. Therefore, we choose a sequential search space, for the ease of implementing the desired parallelism on real mobile devices. In a sequential pipeline, each individual searchable module (a.k.a. a node) only contains a single operator, and they are sequentially connected without any branching.

We jointly search for the operator type and the width of each layer. Different from previous differentiable search (Wu et al., 2019) that selects from predefined building block options where the combinations of operator types and widths are manually picked, we enable each layer's width and operators to be independently searchable.

Application-specific Supernet. On top of the above general design philosophy, we note that different tasks have developed their own preferred and customized components, calling for instantiating different network architectures per application needs. For example, among common image-to-image GAN tasks, image enhancement prefers deeper networks to recover detailed textures and being precise at pixel-level predictions, while style transfer often refers to shallower architectures, and focuses on consistency of the global style more than local details.

We next derive two specific supernet structures below, for two common tasks: image translation and image super-resolution, respectively. Both follow the general sequential search space design with the efficiency constraint in mind.

- *Image Translation.* Our supernet architecture inherits original CycleGAN's structure (see Fig. 1). We keep the same operators from the stem (the first three convolutional layers) and the header (the last two transposed convolutional layers and one convolutional layer), except searching for their widths. This is motivated by the fact that their downsampling/upsampling structures contribute to lowering the computational complexity of the backbone between them, which have been proved to be effective (Zhu et al., 2017).

  For the backbone part between the stem and header, we adopt nine sequential layers with both searchable operators and widths to trade off between model per-

formance and efficiency. For the normalization layers, we adopt the instance normalization (Ulyanov et al., 2016) that is widely used in image translation (Zhu et al., 2017) and style transfer (Johnson et al., 2016).

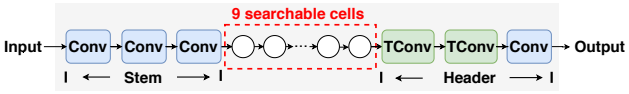

Figure 1. Our supernet for Image Translation, where "TConv" denotes a transposed convolutional layer. We search for both the operators and widths for the layers inside the red box, leaving the stem and header to only search for widths.

- *Super Resolution.* Our supernet architecture is inspired by SRResNet (Ledig et al., 2017) where most computation is performed in the low resolution feature space for efficiency (see Fig. 2). We fix the stem and header parts in the original design which amount to $< 1\%$ computation, and search the remaining residual layers.

  For the residual network structure, ESRGAN (Wang et al., 2018a) introduces residual-in-residual (RiR) blocks with dense connections which have higher capacity and is easier to train. Despite improved performance, such densely-connected blocks are hardware-unfriendly. Therefore, we replace the dense blocks in residual-in-residual modules with five sequential layers with both searchable operators and widths. We also remove all the batch normalization layers in the network to eliminate artifacts (Wang et al., 2018a).

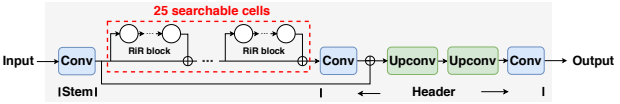

Figure 2. Our supernet for Super Resolution, where "Upconv" denotes bilinearly upsampling followed by a convolutional layer. We search for both the operators and widths for the layers inside the red box, leaving the stem and header fixed.

Operator Search. We search for the following operators:

- Conv 1×1; Conv 3×3;
- Residual Block ("ResBlock") (2 layers of Conv 3×3, with a skip connection);
- Depthwise Block ("DwsBlock") (Conv1 ×1 + DepthwiseConv 3×3 + Conv 1×1, with a skip connection).

These cover popular and efficient building blocks that have been used to construct GAN generators. We did not include dilated convolutions, motivated by their hardware unfriendliness as analyzed in (Ma et al., 2018).

We search for the operator for each layer in a differentiable manner (Liu et al., 2018a; Wu et al., 2019). For the i-th layer, we use an architecture parameter α_i to determine the operator for the current layer, and the softmax value of α_i denotes the probability of choosing this operator. All the candidate operators are activated for every iteration in searching process and the output is the weighted sum of all the operators determined by all the softmax values of all α.

Width Search. The balance between the capacity of an operator and its width (i.e., number of input/output channels) is a crucial trade-off for a compressed model. Therefore, we also allow for the width search to build a slimmer generator. By searching the widths, we merge the pruning step into the searching process for an end-to-end GAN compression.

However, it is non-trivial to naively build a set of independent convolutional kernels with different widths for each operator, due to the exponentially possible combinations. We thus set a single convolution kernel with a maximal width, named *superkernel*. Then we search for the expansion ratio ϕ to make use of only a subset of the input/output dimensions of the superkernel. We set $\phi \in [\frac{1}{3}, \frac{1}{2}, \frac{3}{4}, \frac{5}{6}, 1]$ and use the architecture parameter γ_i to control the probability of choosing each expansion ratio in the i-th layer. We apply Gumbel-Softmax (Gumbel, 1948) to approximate differentiable sampling for ϕ based on γ as (Cai et al., 2018). Therefore, during the searching process only one most likely expansion ratio will be activated each time, which saves both memory and computation cost.

3.2. The Proposed Customized Proxy Task

During the search process, the NAS framework is guided by the proxy task to optimize the architecture towards the desired patterns. As our objective is to obtain both superior performance and extreme efficiency, we need to customize our proxy task to supervise our search process in a proper way. We formulate the training objective of our AGD framework in Eq. 1:

$$\min_{G,\alpha,\gamma} \frac{1}{N} \sum_{i=1}^{N} d(G(x_i, \alpha, \gamma), G_0(x_i)) + \lambda F(\alpha, \gamma). \quad (1)$$

Here, $d(\cdot,\cdot)$ is a distance metric for the knowledge distillation (Polino et al., 2018) between the compact generator G and the pretrained one G_0, F is the computational budget determined by the network architecture, α and γ are the architecture parameters controlling the operator and width of each layer, respectively, and λ is the trade-off parameter. Note that both $d(\cdot,\cdot)$ and F are functions of α and γ.

AGD searches for an efficient generator G under the guidance of distillation from the original model G_0, through optimizing Eq. 1. Note that, the objective function in Eq. 1 is free of any trained discriminator. We choose so because: (1) in practice, the discriminator is often discarded after the generator is trained, therefore not necessarily available when compressing the generator in future; and (2) we also

tried to add a discriminator loss term into Eq. 1, similarly to (Shu et al., 2019) did, but found no improvement in this way (actually, the search becomes more unstable).

One extra hassle arises from our decoupled search of operators and widths. During the search process, if we jointly update α and γ, we observe that our model is prone to suffering from the same "architecture collapse" problem observed in (Chen et al., 2019b), i.e., NAS is quickly biased towards some low-latency yet low-performance models. Therefore, we decouple the computational budget calculation into two individual terms, i.e., the operator-aware and width-aware parts, and weight the two differently:

$$F(\alpha, \gamma) = \omega_1 F(\alpha|\gamma) + \omega_2 F(\gamma|\alpha) \tag{2}$$

In order for more flexible model space exploration while enforcing the budget, we set an upper bound and a lower bound for the target computational budget, and double (half) λ when the computational budget of the derived architecture determined by the current architecture parameters is larger (smaller) than the upper (lower) bound.

The last remaining pieces of our framework are the specific choices of $d(\cdot, \cdot)$ and F in Eq. 1. For $d(\cdot, \cdot)$, we use a combination of content loss L_c (avoid color shift) and perceptual loss L_p (preserve visual and semantic details) as defined in (Johnson et al., 2016), in addition to a total variation loss L_{tv} (enforce pixel-level similarity) (Aly & Dubois, 2005):

$$d(G, G_0) = \beta_1 L_c(G, G_0) + \beta_2 L_p(G, G_0) + \beta_3 L_{tv}(G, G_0), \tag{3}$$

where $\beta_1, \beta_2, \beta_3$ are hyperparameters. For F, we apply FLOPs as the computational budget in Eq. 1.

We summarize our AGD framework in Algorithm 1.

Algorithm 1 The Proposed AutoGAN-Distiller Framework

Input: dataset $\chi = \{x_i\}_{i=1}^N$, pretrained generator G_0, search space and supernet G, epochs to pretrain (T_1), search (T_2) and train-from-scratch (T_3)
Output: trained efficient generator G^*
Equally split χ into χ_1 and χ_2 and initialize supernet weight w and architecture parameters $\{\alpha, \gamma\}$ with uniform distribution
 # First Step: Pretrain
for $t \leftarrow 1$ to T_1 **do**
  Get a batch of data X_1 from χ_1
  for γ in $[\gamma_{\max}, \gamma_{\min}, \gamma_{\text{random1}}, \gamma_{\text{random2}}]$[1] **do**
    $g_w^{(t)} = \nabla_w d(G(X_1, \alpha, \gamma), G_0(X_1))$
    $w^{(t+1)}$ = update($w^{(t)}$, $g_w^{(t)}$)
  end for
end for
 # Second Step: Search
for $t \leftarrow 1$ to T_2 **do**
  Get a batch of data X_1 from χ_1
  $g_w^{(t)} = \nabla_w d(G(X_1, \alpha, \gamma), G_0(X_1))$
  $w^{(t+1)}$ = update($w^{(t)}$, $g_w^{(t)}$)
  Get a batch of data X_2 from χ_2
  $g_\alpha^{(t)} = \nabla_\alpha d(G(X_2, \alpha, \gamma), G_0(X_2)) + \lambda \cdot \omega_1 \nabla_\alpha F(\alpha|\gamma)$
  $g_\gamma^{(t)} = \nabla_\gamma d(G(X_2, \alpha, \gamma), G_0(X_2)) + \lambda \cdot \omega_2 \nabla_\gamma F(\gamma|\alpha)$
  $\alpha^{(t+1)}$ = update($\alpha^{(t)}$, $g_\alpha^{(t)}$)
  $\gamma^{(t+1)}$ = update($\gamma^{(t)}$, $g_\gamma^{(t)}$)
end for
 # Third Step: Train from scratch
Derive the searched architecture G^* with maximal $\{\alpha, \gamma\}$ for each layer and re-initialize weight w.
for $t \leftarrow 1$ to T_3 **do**
  Get a batch of data X from χ
  $g_w^{(t)} = \nabla_w d(G^*(X), G_0(X))$
  $w^{(t+1)}$ = update($w^{(t)}$, $g_w^{(t)}$)
end for

4. Experiment Results

4.1. Considered Tasks & Models.

Unpaired Image-to-image Translation. We apply AGD on compressing CycleGAN (Zhu et al., 2017) and consider two datasets, horse2zebra (Zhu et al., 2017) and summer2winter (Zhu et al., 2017). In particular, we individually conduct the architecture search for each task in one dataset. We also consider a special case for AGD that all the weights and activations are quantized to 8-bit integers for hardware-friendly implementation.

Super Resolution. We apply AGD on compressing ESRGAN (Wang et al., 2018a) on a combined dataset of DIV2K and Flickr2K (Timofte et al., 2017) with a upscale factor of 4×, following (Wang et al., 2018a). We evaluate the searched/compressed model on several popular SR benchmarks, including Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), BSD100 (Martin et al., 2001) and Urban100 (Huang et al., 2015).

[1]We train each operator with the largest width, the smallest width and two random widths during pretraining, following the "sandwich rule" in (Yu & Huang, 2019).

4.2. Evaluation Metrics.

Both tasks are mainly evaluated by visual quality. We also use FID (Frechet Inception Distance) (Heusel et al., 2017) for the unpaired image-to-image translation task, and PSNR for super resolution, as quantitative metrics.

For the efficiency aspect, we measure the model size and the inference FLOPs (floating-point operations). As both might not always be aligned with the hardware performance, we further measure the real-device inference latency using NVIDIA GEFORCE RTX 2080 Ti (NVIDIA Inc.).

4.3. Training details.

NAS Search. The AGD framework adopts the differential search algorithm (Liu et al., 2018a). We split the training dataset into two halves: one for updating supernet weight and the other for updating architecture parameters. The entire search procedure consists of three steps:

- Pretrain. We only train the supernet weights on the first half dataset. To adapt the supernet for different expansion ratios, we train each operator with the largest width, the smallest width, and two random widths, following the "sandwich rule" in (Yu & Huang, 2019).
- Search. We then search the architecture by alternatively updating supernet weights and the architecture parameters $\{\alpha, \gamma\}$. During the training, the width of each layer is sampled through Gumbel-Softmax based on γ and the output is the weighted sum of the results of all operators based on the softmax value of α.
- Derive. We derive the final architecture by choosing the operator and width with the maximal probability for each layer, determined by α and γ. We then train the derived architecture from scratch.

Unpaired Image-to-image Translation. For AGD on CycleGAN, λ in Eq. 1 is 1×10^{-17}, ω_1 and ω_2 in Eq. 2 are set to 1/4 and 3/4, and β_1, β_2 and β_3 in Eq. 3 are set to be 1×10^{-2}, 1, and 5×10^{-8}, respectively. We pretrain and search for 50 epochs, with batch size 2. We use an SGD optimizer with a momentum of 0.9 and the initial learning rate 1×10^{-1} for the weights, which linearly decays to 0 after 10 epochs, and an Adam optimizer with a constant 3×10^{-4} learning rate for architecture parameters. We train the searched architecture from scratch for 400 epochs, with a batch size of 16 and an initial learning rate of 1×10^{-1}, which linearly decays to 0 after 100 epochs.

Super Resolution. For AGD on ESRGAN, λ in Eq. 1 is 1×10^{-12}, ω_1 and ω_2 in Eq. 2 are 2/7 and 5/7, and β_1, β_2, and β_3 in Eq. 3 are set to be 1×10^{-2}, 1, and 5×10^{-8}, respectively. We pretrain and search for 100 epochs, with each batch of 6 cropped patches with 32×32. We set the initial learning rate for the weights to be 1×10^{-4} with an Adam optimizer, decayed by 0.5 at the 25-th, 50-th, and 75-th epoch, and that for the architecture parameters is a constant of 3×10^{-4}. To train the searched architecture from scratch, we train for 1800 epochs with a learning rate of 1×10^{-4}, decayed by 0.5 at 225-th, 450-th, 900-th, and 1300-th epoch, following (Wang et al., 2018a).

4.4. AGD for Efficient Unpaired Image Translation

For unpaired image translation, the original performance and efficiency statistics of CycleGAN are reported in Table 1. We compare our AGD with CEC (Shu et al., 2019), the existing GAN compression algorithm specifically for CycleGAN; we also include classical structural pruning (Li et al., 2016) developed for classification models[2].

Table 1. Statistics of the original CycleGAN (Zhu et al., 2017) on different unpaired image translation tasks, where the latency is measured on NVIDIA GEFORCE RTX 2080 Ti. Note that the FID of CycleGAN is from CEC (Shu et al., 2019) since FID (Heusel et al., 2017) is proposed after CycleGAN (Zhu et al., 2017).

GFLOPs	54.17	Latency (ms)	7.25
Memory (MB)	43.51		
FID	horse2zebra		74.04
	zebra2horse		148.81
	summer2winter		79.12
	winter2summer		73.31

Quantitative Results. Table 2 demonstrates that for all tasks, AGD consistently achieves significantly better FID than the state-of-the-art CEC, with even higher savings of model sizes and FLOPs. Besides, the performance of structural pruning, which is effective for compressing classification models, lags far behind with CEC, not to say AGD. That re-confirms (Shu et al., 2019)'s finding that existing classification-oriented compression algorithms cannot be directly plugged into compressing GANs.

Furthermore, the gain of AGD also manifests in real-device latency measurements. For example, on the summer2winter dataset, the per-image inference latency of the AGD-compressed model is reduced by 59.86% with an even slightly better FID than the original CycleGAN.

Table 2. Quantitative comparison with the structural pruning (Li et al., 2016) and CEC (Shu et al., 2019) on CycleGAN compression in different unpaired image translation tasks.

Dataset	Method	FID	GFlops	Memory (MB)	Latency (ms)
horse2zebra	Prune	220.91	4.95	3.91	6.19
	CEC	96.15	13.45	10.16	-
	AGD	**83.6**	**6.39**	**4.48**	**4.07**
zebra2horse	Prune	206.56	4.95	3.91	6.19
	CEC	157.9	13.06	10	-
	AGD	**137.2**	**4.84**	**3.2**	**3.99**
summer2winter	Prune	120.94	4.95	3.91	6.19
	CEC	78.58	12.98	7.98	-
	AGD	**78.33**	**4.34**	**2.72**	**2.91**
winter2summer	Prune	106.2	4.95	3.91	6.19
	CEC	79.26	16.45	7.61	-
	AGD	**77.73**	**4.26**	**2.64**	**4.26**

Visualization Results. We compare the visual quality from the original CycleGAN, CEC, and AGD in Fig. 3, using examples from the *horse2zebra* and *zebra2horse* tasks. The

[2]We prune 70% filters with smallest ℓ_1-norm values in each layer, and then finetune the remaining weights.

Figure 3. Visualization examples of CycleGAN compression on the horse2zebra task (rows 1 and 2), zebra2horse task (row 3), and summer2winter task (row 4). Columns from left to right: source images, translation results by original CycleGAN, CEC, and AGD, respectively, and the FLOPs and memory (model size) of each method on each task are annotated above the images.

AGD-compressed CycleGAN yields comparable visual quality to the uncompressed model in all cases. With zooming-in, AGD also appears to visually outperform CEC, in producing sharper edges and textural details, suppressing checkboard artifacts, as well as eliminating color distortion.

AGD Searched Architecture. The searched architecture by AGD on the horse2zebra dataset is outlined in Table 3. AGD selects the "heavier" DwsBlocks and ResBlocks with narrower channel widths for the early layers; while leaving the last three layers to be Conv 1x1, with larger widths. The searched design remains to be overall light-weighted, while not "collapsing" to overly under-parameterized, poor

Table 3. The searched architecture (operator, width) by our AGD framework on CycleGAN.

Block ID	Stem0	Stem1	Stem2	B1	B2
(OP, Width)	(-, 88)	(-, 88)	(-, 88)	(DwsBlock, 88)	(ResBlock, 88)
Block ID	B3	B4	B5	B6	B7
(OP, Width)	(ResBlock, 88)	(ResBlock, 128)	(ResBlock, 88)	(DwsBlock, 128)	(Conv1x1, 216)
Block ID	B8	B9	Header1	Hearder2	Header3
(OP, Width)	(Conv1x1, 88)	(Conv1x1, 128)	(-, 216)	(-, 88)	(-, 3)

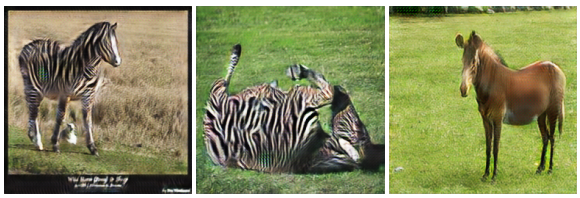

Figure 4. Visualization results of CycleGAN compressed by AGD with quantization, where the first two figures are from the *horse2zebra* task,and the last figure is from the *zebra2horse* task.

performance models. Similar architectures searched on the other three tasks are shown in our supplementary materials.

Quantization. We also conduct an extended study on how our AGD compression work with quantization. On the AGD-compressed CycleGAN, we further quantize all the weights and activations to 8-bit integers, following (Banner et al., 2018). The quantitative and visualization results are shown in Table 4 and Fig. 4, respectively: after being further aggressively compressed, the models maintain both competitive FID and comparable visualization quality.

Table 4. Performance of AGD framework on CycleGAN compression with 8-bit quantization.

Datasets	FID	Memory (MB)
horse2zebra	85.74	1.00
zebra2horse	140.08	1.24

4.5. AGD for Efficient Super Resolution

We then apply AGD on compressing ESRGAN (Wang et al., 2018a), a start-of-the-art GAN for super resolution. ESRGAN does not adopt a CycleGAN-type structure (its generator is a feedforward encoder-decoder), therefore CEC (Shu et al., 2019) is not applicable. We compare AGD with a structural pruning (Li et al., 2016) baseline, and also include two other high-performance and lighter SR models: SRGAN (Ledig et al., 2017) and VDSR (Kim et al., 2016), to better evaluate the utility-efficiency trade-off of our compressed models. Besides, since visual quality is the ultimate demand in SR, we first adopt the visualization-oriented ESRGAN[3] as the teacher model in Eq. 3.

[3]ESRGAN (Wang et al., 2018a) provides two pretrained models, including the PSNR-oriented one trained only with the content loss, targeted for higher PSNR performance; and the visualization-oriented one trained with both content loss and perceptual loss, targeted for higher perceptual quality.

Table 5. Quantitative comparison with the state-of-the-art GAN-based visualization-oriented SR models, where the FLOPs are calculated as processing a 256×256 low resolution image with a scale factor of four.

Model		ESRGAN	ESRGAN-Prune	SRGAN	AGD
GFLOPs (256x256)		1176.61	113.07	166.66	**108.6**
Memory (MB)		66.8	6.40	6.08	**1.64**
Set5	PSNR	**30.47**	28.07	29.40	**30.44**
	Latency (ms)	84.62	64.75	**4.92**	**5.08**
Set14	PSNR	26.29	25.21	26.02	**27.28**
	Latency (ms)	80.03	78.47	**5.37**	**6.91**
BSD100	PSNR	25.32	24.74	25.16	**26.23**
	Latency (ms)	81.34	69.43	**4.15**	**5.11**
Urban100	PSNR	24.36	22.67	24.39	**24.74**
	Latency (ms)	184.79	113.82	25.28	**21.60**

Quantitative Results. Table 5 shows the quantitative comparison in PSNR of a series of GAN-based super resolution methods, re-confirming the little-to-no performance loss of AGD on ESRGAN (interestingly, sometimes even PSNR improvement is seen). Compared to the off-the-shelf alternatives (e.g., SRGAN), AGD compressed ESRGAN obtains consistently superior PSNRs, with a $4\times$ smaller model size and comparable real-device latency.

Visualization Results. As shown in Fig. 5, AGD compression dramatically reduces the model size by over 40 times, without sacrificing the visual quality. The real device latency is also shrunk by 10-18 times on the four sets.

Table 6. Searched architecture (operator, width) by AGD on visualization-oriented ESRGAN, where we search for five operators and their widths in the five Residual-in-Residual blocks.

RiR Block ID	OP1	OP2	OP3	OP4	OP5
1	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 64)
2	(DwsBlock, 24)	(DwsBlock, 24)	(ResBlock, 64)	(Conv3x3, 24)	(ResBlock, 64)
3	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 64)
4	(Conv3x3, 32)	(DwsBlock, 32)	(DwsBlock, 24)	(DwsBlock, 24)	(ResBlock, 64)
5	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 64)

AGD Searched Architecture. As shown in Table 6, we find that among the original five RiR blocks, AGD assigns more computationally powerful operators to the 2nd and 4th one, while keeping the other three simple by only using Conv1 $\times$ 1 operations, achieving a proper balance between heavy and efficient operations.

Table 7. Searched architecture (operator, width) by our proposed AGD on the PSNR-oriented ESRGAN (Wang et al., 2018a), where RiR denotes a Residual-in-Residual block using the operators in our search space. We search for five operators and their widths in each RiR block (totally five RiR blocks).

RiR Block ID	**OP1**	**OP2**	**OP3**	**OP4**	**OP5**
1	(DwsBlock, 32)	(DwsBlock, 32)	(DwsBlock, 48)	(DwsBlock, 24)	(ResBlock, 64)
2	(DwsBlock, 24)	(ResBlock, 32)	(Conv1x1, 32)	(Conv1x1, 24)	(Conv3x3, 64)
3	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 64)
4	(DwsBlock, 32)	(DwsBlock, 48)	(DwsBlock, 24)	(DwsBlock, 24)	(ResBlock, 64)
5	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 24)	(Conv1x1, 64)

Comparison with PSNR-oriented SR Methods. We also search for a new architecture as shown in Table 7 under the

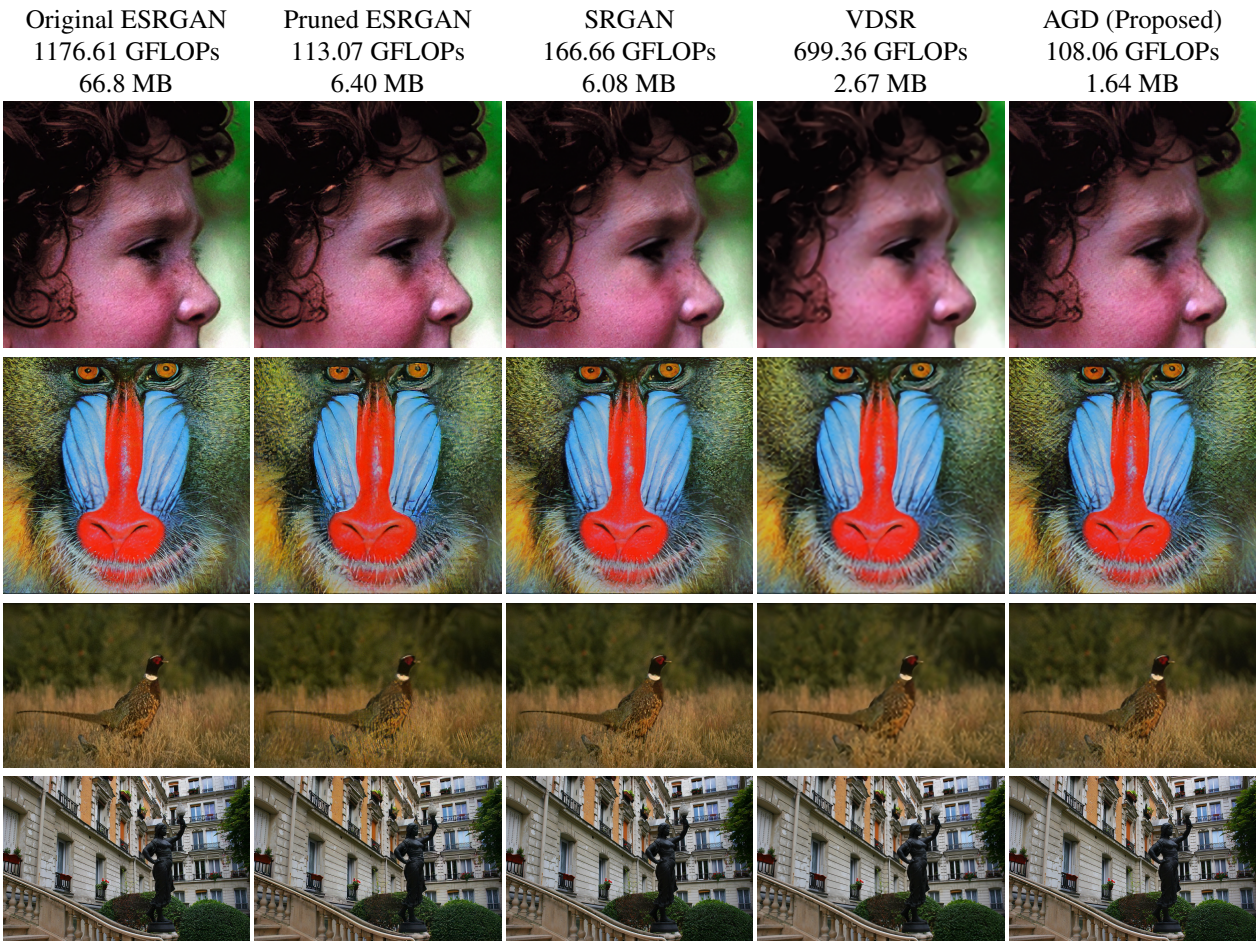

Figure 5. Visualization results of different super resolution methods (Better viewed after zooming in). From top to bottom rows are examples from Set5, Set14, BSD100, and Urban100. Columns from left to right: original ESRGAN (Wang et al., 2018a), ESRGAN after structural pruning (Li et al., 2016), SRGAN (Ledig et al., 2017), VDSR (Kim et al., 2016), and ESRGAN compressed by AGD. FLOPs (calculated as processing a 256×256 low resolution image with a scale factor of four) and memory (model size) of each method on each task are annotated above the images.

distillation of a PSNR-oriented ESRGAN with the content loss L_c only in Eq. 3 (following (Wang et al., 2018a)) within the same search space. The quantitative results of the original ESRGAN (Wang et al., 2018a), the searched architecture of AGD, and the PSNR-oriented model VSDR (Kim et al., 2016) are shown in Table 8. Compared with VDSR, AGD achieves better PSNR (up to 0.44) on all the four datasets with 84.1% fewer FLOPs and 32.6% smaller model size.

5. Conclusion

We propose the AGD framework to significantly push forward the frontier of GAN compression by introducing AutoML here. Starting with a specially designed search space of efficient building blocks, AGD performs differential neural architecture search under both the guidance of knowledge distillation and the constraint of computational resources. AGD is automatic with no assumptions about GAN structures, loss forms, or the availability of discriminators, and can be applicable to various GAN tasks. Experiments show that AGD outperforms existing options with aggressively reduced FLOPs, model size, and real-device latency, without degrading model performance. Our future work include evaluating AGD in more GAN-based applications such as data augmentation (Zhang et al., 2019).

Table 8. Quantitative comparison with the state-of-the-art PSNR-oriented SR models. FLOPs are calculated as processing a 256×256 low resolution image with a scale factor of four.

Model	GFLOPs (256x256)	Memory (MB)	PSNR			
			Set5	Set14	BSD100	Urban100
ESRGAN	1176.61	66.8	32.73	28.99	27.85	27.03
VDSR	699.36	2.67	31.35	28.01	27.29	25.18
AGD	110.9	1.8	31.79	28.36	27.41	25.55

Acknowledgements

The work is supported by the National Science Foundation (NSF) through the Energy, Power, Control, and Networks (EPCN) program (Award number: 1934755, 1934767).

References

Aly, H. A. and Dubois, E. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659, 2005.

Banner, R., Hubara, I., Hoffer, E., and Soudry, D. Scalable methods for 8-bit training of neural networks. In *Advances in neural information processing systems*, pp. 5145–5153, 2018.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Bulò, S. R., Porzi, L., and Kontschieder, P. Dropout distillation. In *International Conference on Machine Learning*, pp. 99–107, 2016.

Cai, H., Zhu, L., and Han, S. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., and Tian, Q. Data-free learning of student networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3514–3522, 2019a.

Chen, H., Wang, Y., Shu, H., Tang, Y., Xu, C., Shi, B., Xu, C., Tian, Q., and Xu, C. Frequency domain compact 3d convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1641–1650, 2020a.

Chen, H., Wang, Y., Shu, H., Wen, C., Xu, C., Shi, B., Xu, C., and Xu, C. Distilling portable generative adversarial networks for image translation. *arXiv preprint arXiv:2003.03519*, 2020b.

Chen, L.-C., Collins, M., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., and Shlens, J. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in neural information processing systems*, pp. 8699–8710, 2018a.

Chen, W., Gong, X., Liu, X., Zhang, Q., Li, Y., and Wang, Z. Fasterseg: Searching for faster real-time semantic segmentation. *arXiv preprint arXiv:1912.10917*, 2019b.

Chen, Y., Lai, Y.-K., and Liu, Y.-J. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9465–9474, 2018b.

Cheng, A.-C., Lin, C. H., Juan, D.-C., Wei, W., and Sun, M. Instanas: Instance-aware neural architecture search. *arXiv preprint arXiv:1811.10201*, 2018.

Courbariaux, M., Bengio, Y., and David, J.-P. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, pp. 3123–3131, 2015.

Elsken, T., Metzen, J. H., and Hutter, F. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

Gong, X., Chang, S., Jiang, Y., and Wang, Z. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3224–3234, 2019.

Gong, Y., Liu, L., Yang, M., and Bourdev, L. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gui, J., Sun, Z., Ji, S., Tao, D., and Tan, T. Feature selection based on structured sparsity: A comprehensive study. *IEEE transactions on neural networks and learning systems*, 28(7):1490–1507, 2016.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Gumbel, E. J. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1948.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

He, Y., Zhang, X., and Sun, J. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017.

He, Y., Lin, J., Liu, Z., Wang, H., Li, L.-J., and Han, S. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019.

Huang, J.-B., Singh, A., and Ahuja, N. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Hutter, F., Kotthoff, L., and Vanschoren, J. *Automated Machine Learning*. Springer, 2019.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., and Wang, Z. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Kim, J., Kwon Lee, J., and Mu Lee, K. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

Kupyn, O., Martyniuk, T., Wu, J., and Wang, Z. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8878–8887, 2019.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J.-Y., and Han, S. Gan compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5284–5294, 2020.

Liu, C., Chen, L.-C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., and Fei-Fei, L. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 82–92, 2019.

Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018a.

Liu, S., Lin, Y., Zhou, Z., Nan, K., Liu, H., and Du, J. On-demand deep model compression for mobile devices: A usage-driven model selection framework. In *Proceedings of International Conference on Mobile Systems, Applications, and Services*, pp. 389–400. ACM, 2018b.

Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Luo, J.-H., Wu, J., and Lin, W. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pp. 5058–5066, 2017.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

Martin, D., Fowlkes, C., Tal, D., Malik, J., et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Iccv Vancouver:, 2001.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

NVIDIA Inc. NVIDIA GEFORCE RTX 2080 Ti. https://www.nvidia.com/en-us/geforce/

graphics-cards/rtx-2080-ti/, accessed 2020-06-14.

Polino, A., Pascanu, R., and Alistarh, D. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018.

Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pp. 525–542, 2016.

Sanakoyeu, A., Kotovenko, D., Lang, S., and Ommer, B. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 698–714, 2018.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Shu, H., Wang, Y., Jia, X., Han, K., Chen, H., Xu, C., Tian, Q., and Xu, C. Co-evolutionary compression for unpaired image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3235–3244, 2019.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, 2019.

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114–125, 2017.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018a.

Wang, Y., Nguyen, T., Zhao, Y., Wang, Z., Lin, Y., and Baraniuk, R. Energynet: Energy-efficient dynamic inference. *NeurIPS workshop*, 2018b.

Wang, Y., Xu, C., Xu, C., and Tao, D. Adversarial learning of portable student networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018c.

Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. In *Advances in neural information processing systems*, pp. 2074–2082, 2016.

Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., and Keutzer, K. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10734–10742, 2019.

Wu, J., Leng, C., Wang, Y., Hu, Q., and Cheng, J. Quantized convolutional neural networks for mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4820–4828, 2016.

Wu, J., Wang, Y., Wu, Z., Wang, Z., Veeraraghavan, A., and Lin, Y. Deep k-means: Re-training and parameter sharing with harder cluster assignments for compressing deep convolutions. In *International Conference on Machine Learning*, pp. 5359–5368, 2018.

Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., and Guo, Z. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4442–4451, 2019.

Yu, J. and Huang, T. S. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1803–1811, 2019.

Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pp. 711–730. Springer, 2010.

Zhang, X., Wang, Z., Liu, D., and Ling, Q. Dada: Deep adversarial data augmentation for extremely low data regime classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2807–2811. IEEE, 2019.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.