Image Difficulty Curriculum for Generative Adversarial Networks (CuGAN)

Petru Soviany[1], Claudiu Ardei[1], Radu Tudor Ionescu[1], Marius Leordeanu[2,3]
[1]University of Bucharest
[2]Institute of Mathematics of the Romanian Academy, [3]University "Politehnica" of Bucharest

Abstract

Despite the significant advances in recent years, Generative Adversarial Networks (GANs) are still notoriously hard to train. In this paper, we propose three novel curriculum learning strategies for training GANs. All strategies are first based on ranking the training images by their difficulty scores, which are estimated by a state-of-the-art image difficulty predictor. Our first strategy is to divide images into gradually more difficult batches. Our second strategy introduces a novel curriculum loss function for the discriminator that takes into account the difficulty scores of the real images. Our third strategy is based on sampling from an evolving distribution, which favors the easier images during the initial training stages and gradually converges to a uniform distribution, in which samples are equally likely, regardless of difficulty. We compare our curriculum learning strategies with the classic training procedure on two tasks: image generation and image translation. Our experiments indicate that all strategies provide faster convergence and superior results. For example, our best curriculum learning strategy applied on spectrally normalized GANs (SNGANs) fooled human annotators in thinking that generated CIFAR-like images are real in 25.0% *of the presented cases, while the SNGANs trained using the classic procedure fooled the annotators in only* 18.4% *cases. Similarly, in image translation, the human annotators preferred the images produced by the Cycle-consistent GAN (CycleGAN) trained using curriculum learning in* 40.5% *cases and those produced by CycleGAN based on classic training in only* 19.8% *cases,* 39.7% *cases being labeled as ties.*

1. Introduction

Generative Adversarial Networks (GANs) [11] represent a hot topic in computer vision, drawing the attention of many researchers who developed several improvements of the standard architecture [1, 6, 14, 18, 24, 25, 28, 31, 32, 33, 40, 43]. Yet, this kind of neural models are still very hard to train [27]. In this paper, we study the hypothesis of improving the training process of GANs in terms of both accuracy and time, by employing curriculum learning [3]. *Curriculum learning* is the process of training machine learning models by presenting the training examples in a meaningful order which gradually illustrates more complex concepts. Although many curriculum learning approaches [10, 12, 13, 17, 19, 23, 30, 42, 41, 44] have been proposed for training deep neural networks, to our knowledge, there are only a few studies that apply curriculum learning to GANs [7, 9].

In this paper, we propose three novel curriculum learning strategies that provide faster convergence during GANs training, as well as improved results. Our curriculum learning strategies are general enough to be applied to any GAN architecture, as shown in Figure 1. They rely on a state-of-the-art image difficulty predictor [17], which scores the (real) training images with respect to the difficulty of solving a visual search task. After receiving the image difficulty scores as input, we employ one of our curriculum learning strategies listed below:

- Divide the training images into m easy-to-hard batches and start training the GAN with the easy batch. The other batches are added into the training process, in increasing order of difficulty, after a certain number of iterations.
- Add another component to the discriminator loss function which makes the loss value proportional to the easiness (inverse difficulty) score of the images. The impact of this new component is gradually attenuated, until the easiness score has no more influence in the last training iterations.
- Change the discriminator loss function by including probabilities of sampling real images from a biased distribution that strongly favors easier images during the first training iterations. The probability distribution is continuously updated with each iteration, until it becomes uniform in the last training iterations.

Our three curriculum learning strategies follow two important principles. First, we keep the easier images until the end of the training process, to prevent catastrophic forgetting [21, 26]. Second, we want all training examples to receive equal importance *in the end* (when training is finished), as we have no reason to favor the easy or the difficult images. However, during the initial stages of training, we emphasize easier images in order to achieve faster convergence and possibly a better local minimum.

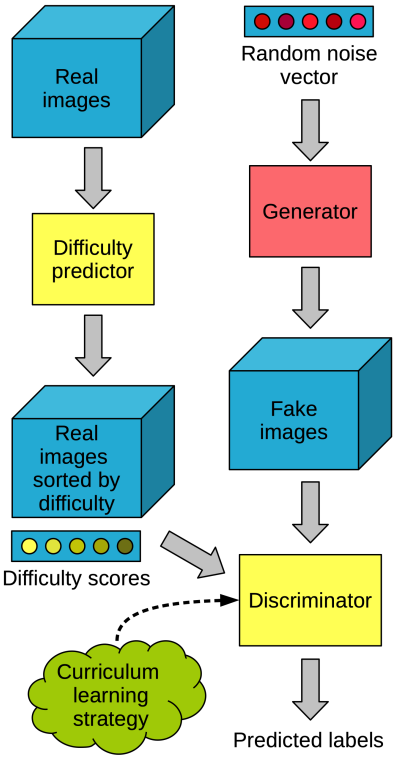

Figure 1. Our GAN training pipeline based on curriculum learning. The real training images are passed to an image difficulty predictor which provides a difficulty score for each image. A curriculum learning strategy that takes into account the difficulty scores is employed to train the discriminator. Best viewed in color.

We perform image generation experiments using the spectrally normalized GAN (SNGAN) model [29], and image translation experiments using the Cycle-consistent GAN (CycleGAN) model [45]. The goal of our experiments is to compare the standard training process, in which examples are presented in a random order, with the training process based on curriculum. The image generation results on CIFAR-10 [22] indicate that all the proposed curriculum learning strategies improve the Inception Score (IS) [35] and the Fréchet Inception Distance (FID) [15] over the state-of-the-art SNGAN model. Furthermore, we conducted several human annotations studies, to determine whether our generated or translated images are better than those produced by the baselines SNGAN and CycleGAN, respectively. Our best curriculum learning strategy fooled human annotators in thinking that generated CIFAR-like images are real in 25.0% of the presented cases (on average), while the SNGAN fooled the annotators in only 18.4% cases. This represents an absolute gain of 6.6% over SNGAN. We obtain significant improvements in image translation as well. For example, in the horse2zebra [45] experiment, the human annotators opted for our method in 52.5% of the presented cases and for the baseline CycleGAN in only 11.9% cases, 35.6% cases being labeled as draws. This represents an absolute gain of 40.6% over CycleGAN. We thus conclude that employing curriculum learning for training GANs is useful.

We organize the rest of this paper as follows. In Section 2, we present related works and how our approach is different. In Section 3, we describe our curriculum learning strategies for training GANs. We present the image generation and image translation experiments in Section 4. We draw our conclusion and discuss future work in Section 5.

2. Related work

Generative Adversarial Networks. Generative Adversarial Networks [11] are composed of two neural networks, a generator and a discriminator, which are trained for generating new images, similar to those provided in a training set. Since 2014, many variations of GANs have been proposed in order to improve the quality of the generated samples [1, 6, 14, 18, 24, 25, 28, 31, 32, 33, 40, 43]. Mirza and Osindero [28] introduced a conditional version of GANs, termed CGAN, which is based on feeding label information to both the generator and the discriminator. As CGAN, the Auxiliary Classifier GAN (AC-GAN) [31] is a class conditional model, which in addition, leverages side information through an auxiliary decoder that is responsible for reconstructing class labels. Deep convolutional GANs (DCGANs) [32] include a set of constraints to the architectural topology of the classic GAN, to improve training stability. Wasserstein GANs (WGANs) [1] use the Earth Mover distance instead of other popular metrics to provide easier training, while lowering the chances of entering mode collapse. Still, WGAN employs a weight clipping technique which can result in failure to converge and bad outputs. This problem is addressed in WGAN-GP (Wasserstein GAN with Gradient Penalty) [14], where weight clipping is replaced by gradient penalty, providing better performance on different architectures. SNGAN [29] introduces spectral normalization, another normalization technique used to stabilize the training of the discriminator. Compared to the other regularization methods, spectral normalization provides better results and lower computational costs.

CycleGAN [45] performs image translation without requiring paired images to learn the mapping. Instead, it learns the relevant features of two domains and how to translate between these domains. It uses cycle consistency, which encodes the idea that translating from one domain to another and back again should take you back to the same place. Choi et al. [6] introduced StarGAN, a conditional solution that has the advantage of providing good results when translating between more than two domains, using a single discriminator and generator network.

Some studies [18, 33, 43] showed that providing additional information to a GAN model can result in performance improvements in a wide range of common generative

tasks. Similar to these approaches, we use an external difficulty score, trying to constrain the order and the importance of the training samples, in order to imitate the easy-to-hard (curriculum) learning paradigm from humans.

Curriculum learning. Bengio et al. [3] studied easy-to-hard strategies to train machine learning models, showing that machines can also benefit from learning by gradually adding more difficult examples. They introduced a general formulation of the easy-to-hard training strategies known as *curriculum learning*. In the past few years curriculum learning has been applied to semi-supervised image classification [10], language modelling [12], question answering [12], object detection [38, 39, 42, 44], person re-identification [41], weakly supervised object detection [17, 23]. Other works proposed refined techniques for improving neural network training components, e.g. dropout [30], or training frameworks, e.g. teacher-student [19], using curriculum learning. Ionescu et al. [17] considered an image difficulty predictor trained on image difficulty scores produced by human annotators. Similar to Ionescu et al. [17], we use an image difficulty predictor, but with a completely different purpose, that of training GANs. In addition, we explore several curriculum learning strategies that enable end-to-end training by defining new curriculum loss functions.

Curriculum GANs. To our knowledge, there are a just few works that propose curriculum learning approaches for training GANs [7, 9]. Doan et al. [7] introduced an adaptive curriculum learning strategy for training GANs, called acGAN, which uses multiple discriminators with different architectures of various depths. The authors proposed a reward function that uses an online multi-armed bandit algorithm. The reward function measures the progress made by the generator and uses it to update the weights of each discriminator, ensuring that the generator and the discriminators learn at the same pace. Different from the approach of Doan et al. [7], we consider the difficulty of the training samples and propose strategies to train GANs gradually, from the easy images to the hard ones. While the approach of Doan et al. [7] uses multiple discriminators, increasing the training time, our approach does not require any additional training time.

Ghasedi et al. [9] proposed ClusterGAN, an easy-to-difficult approach for image clustering. ClusterGAN is an unsupervised model composed of three elements: a generator, a discriminator and a clustering network. The samples are introduced gradually in the training, from the easy ones to the hard ones. The values of the loss function are used as difficulty scores for the corresponding image samples. Their curriculum learning strategy leads to good results when training clustering networks with large depth. While Ghasedi et al. [9] study the problem of clustering images, we apply curriculum learning in order to generate or translate images. Furthermore, we propose and study three curriculum learning strategies instead of a single one.

3. Method

3.1. Preliminaries and notations

Generative Adversarial Networks [11] are composed of two neural networks, the generator (G) and the discriminator (D), which are trained to compete against each other in an adversarial game. The generator learns to generate image samples from a Gaussian noise density p_z, such that the generated (fake) images (from the learned density p_g) are difficult to distinguish from real images for the discriminator. Meanwhile, the discriminator is trained to differentiate between real images from a density p_r and fake images from the density p_g learned by G. The two networks, G and D, compete in a minimax game with the objective function $V(G, D)$ defined as follows:

$$V(G, D) = \mathbb{E}_{x \sim p_r}[l(D(x))] + \mathbb{E}_{z \sim p_z}[l(-D(G(z)))], \quad (1)$$

where x is a real image sampled from the true data density p_r, z is the random noise vector sampled from the density p_z, and l is a loss function, e.g. cross-entropy [11] or Hinge loss [29]. The goal of the generator G is to minimize this error, while the goal of the discriminator D is to maximize it. Hence, during training, we aim to optimize the objective function as follows:

$$\min_G \max_D V(G, D). \quad (2)$$

The two networks are alternatively trained until the generator learns the probability density function of the training data p_r, i.e. until $p_g \approx p_r$.

3.2. Curriculum GANs based on image difficulty

While machines are commonly trained by presenting examples in a random order, humans learn new concepts by organizing them in a meaningful order which gradually illustrates higher complexity. To this end, Bengio et al. [3] proposed curriculum learning for training machine learning models, specifically neural networks, which are influenced by the order in which the examples are presented during training. Since deep neural networks are models that essentially try to mimic the brain, it seems natural to also adopt curriculum learning from humans [37]. We hypothesize that a curriculum learning strategy for training GANs can bring several benefits, e.g. faster convergence, improved stability and superior results. To demonstrate our hypothesis we explore three curriculum learning strategies that are generic enough to be applied to any GAN architecture. In order to learn in the increasing order of difficulty (from easy to hard), we first need to apply an image difficulty predictor on the training set of real images. This allows us to change the distribution of the real images p_r in order to introduce curriculum when training the discriminator D. Since the generator G tries to learn a distribution p_g that closely follows

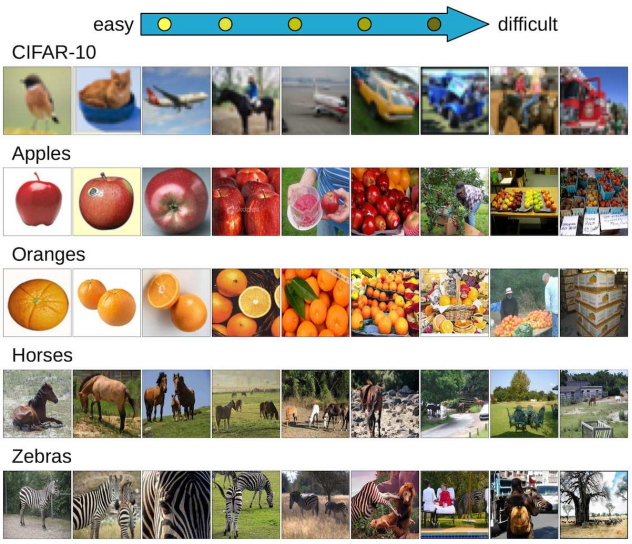

Figure 2. From left to right, images in increasing order of difficulty selected from CIFAR-10 [22], apple2orange [45] and horse2zebra [45] data sets, respectively. Best viewed in color.

p_r, G is implicitly influenced by the curriculum learning strategy. Therefore, it is not necessary to apply the image difficulty predictor on the generated images, saving the additional training time. Moreover, the difficulty predictor needs to be applied only once on the real images, before starting to train the GANs. We next present the image difficulty predictor and our three curriculum learning strategies.

Image difficulty prediction. Ionescu et al. [17] defined image difficulty as the human response time for solving a visual search task, collecting corresponding difficulty scores for the PASCAL VOC 2012 data set [8]. We follow the approach proposed in [17] to build a state-of-the-art image difficulty predictor. The model is based on concatenating deep features extracted from two Convolutional Neural Networks (CNN), VGG-f [5] and VGG-verydeep-16 [36], which are pre-trained on ImageNet [34]. We remove the softmax layer of each CNN model and use the output of the penultimate fully-connected layer, resulting in a feature vector of 4096 components. We divide each image into 1×1, 2×2 and 3×3 bins in order to obtain a pyramid representation, which leads to performance improvements [17]. We concatenate the feature vectors corresponding to each bin into a single vector corresponding to the entire image. We L_2-normalize the concatenated feature vectors before training a ν-Support Vector Regression (ν-SVR) [4] model to regress to the ground-truth difficulty scores provided for PASCAL VOC 2012 [8]. We use the learned predictor P as an image difficulty scoring function that provides difficulty scores on a continuous scale:

$$s_i = \frac{P(x_i) - \min_{x_j \in X}\{P(x_j)\}}{\max_{x_j \in X}\{P(x_j)\}} \cdot 2 - 1, \quad (3)$$

where s_i is the difficulty score for the image x_i in a set of images $X = \{x_1, x_2, ..., x_n\}$, where $n = |X|$. Eq. (3) maps the predicted difficulty scores for the set X to the interval $[-1, 1]$. Our predictor attains a Kendall's τ correlation coefficient of 0.471 on the same test set of [17]. In Figure 2, we present images in increasing order of difficulty from the data sets considered in our experiments from Section 4.

Learning using image difficulty batches. Our first curriculum learning strategy is based on dividing the real images into m equally-sized batches indexed from 1 to m, of increasing difficulty, such that images in each batch $i+1$ have higher difficulty scores than the images in the batch i, $\forall i \in \{1, 2, ..., m-1\}$. Thus, the first batch contains the easiest images and the last batch contains the hardest ones. After dividing the images into batches of increasing difficulty, we start training the GANs using only images from the first batch. After a fixed number of iterations, we include images from the second batch into the training. This process continues until all m batches are included into the training. In this way, the generator learns a progressively complex density p_g. Since images in the former batches can be learned faster (due to their easiness), we consider a smaller number of iterations during the early training stages. The number of iterations increases as more difficult batches are added.

Learning by weighting according to difficulty. Our second curriculum learning strategy is based on integrating the difficulty scores into the discriminator loss function, by weighting the real images according to their difficulty scores. In the first training iterations, we aim to provide higher weights to the easy images and lower weights to the difficult images. With each training iteration, the weights of both easy and difficult images gradually converge to a single value. The weights are computed using the following scoring function w_P:

$$w_P(x_i, t) = 1 - k \cdot s_i \cdot e^{-\gamma \cdot t}, \quad (4)$$

where x_i is an image from the set of real images X, s_i is the image difficulty score as in Eq. (3), t is the current training iteration index, γ is a parameter that controls how fast the scores converge to the value 1 and k is a parameter that controls the impact of the difficulty weights to the overall loss value. Figure 3 illustrates the behavior of w_P for various easiness scores in the interval $[0, 2]$. In the first iteration (when $t = 0$), the easiness scores are equal to $1 - s$, for $k = 1$. Note that in the last iterations all images have basically the same weight, regardless of their difficulty. By including the scoring function w_P from Eq. (4) into the objective function V defined in Eq. (1), we obtain a novel objective (loss) function $V^{(1)}$ based on curriculum learning, defined as follows:

$$V^{(1)}(G, D, P) = \mathbb{E}_{x \sim p_r}[l(D(x)) + w_P(x, t)] + \mathbb{E}_{z \sim p_z}[l(-D(G(z)))]. \quad (5)$$

We note that when t approaches infinity, the objective func-

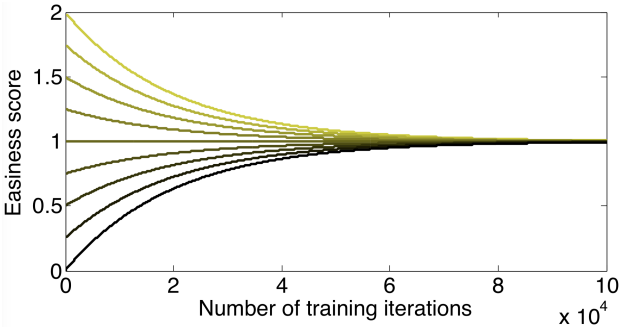

Figure 3. Easiness scores between 0 and 2 converge to 1 as the number of training iterations increases, by applying the scoring function defined in Eq. (4) with $k = 1$ and $\gamma = 5 \cdot 10^{-5}$. Each curve represents the evolution of the weight for a given image, which starts with the weight equal to its easiness score $(1 - s)$ at the first iteration and ends with a weight equal to 1, regardless of its initial easiness score. Best viewed in color.

tion $V^{(1)}$ converges asymptotically to $V + 1$, i.e.:

$$\lim_{t \to \infty} V^{(1)}(G, D, P) = V(G, D) + 1. \quad (6)$$

This can be immediately demonstrated by considering that:

$$\lim_{t \to \infty} w_P(x, t) = \lim_{t \to \infty} 1 - k \cdot s \cdot e^{-\gamma \cdot t} = 1. \quad (7)$$

Learning by sampling according to difficulty. Our third curriculum learning strategy is based on changing the probability density function of the real images p_r, by multiplying it with another probability density function that is proportional to w_P defined in Eq. (4):

$$p_{r,w_P} = p_r \cdot p_{w_P} \propto w_P(x, t). \quad (8)$$

By including the novel density p_{r,w_P} into the objective function V defined in Eq. (1), we effectively obtain a novel loss function $V^{(2)}$ based on curriculum learning, defined as follows:

$$V^{(2)}(G, D, P) = \mathbb{E}_{x \sim p_{r,w_P}}[l(D(x))] + \mathbb{E}_{z \sim p_z}[l(-D(G(z)))]. \quad (9)$$

We use the weights $w_P(x_i, t)$ to define a distribution over the training images. We then sample training images from this distribution during training. We define a discrete random variable R with possible values associated to indexes of images in the training set X, such that the probability $Prob(R = i)$ of sampling an index for real image x_i from X is equal to the weight $w_P(x_i, t)$ divided by the sum of all weights, making all probabilities sum up to 1:

$$Prob(R = i) = \frac{w_P(x_i, t)}{\sum_{x_j \in X} w_P(x_j, t)}, \forall i \in \{1, ..., n\}, \quad (10)$$

where $n = |X|$. Consequently, easier images have a higher chance of being sampled in the first learning iterations. When $k > 1$ in Eq. (4), we need to add the constant $k - 1$ to each term in Eq. (10), i.e. we replace $w_P(x, t)$ with $w_P(x, t) + k - 1$, to obtain positive values. Towards the end of the training process, as w_P converges asymptotically to 1 (see Eq. (7)), it becomes equally likely to sample an easy or a difficult image, i.e.:

$$\lim_{w_P \to 1} Prob(R = i) = \frac{1}{n}, \forall i \in \{1, \dots, n\}, \quad (11)$$

where $n = |X|$. At the limit, p_{w_P} converges to a uniform density and p_{r,w_P} becomes equal to p_r.

Observation. $V^{(2)}$ can be seen as a continuous version of our first curriculum learning approach, in which the probability of sampling a real image x from the set of training images X is given by a step function, where the number of steps is equal to the number of batches m.

4. Experiments

4.1. Data sets

We perform image generation experiments on the CIFAR-10 data set [22]. It consists of 50000 color train images of 32×32 pixels, equally distributed into 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Our translation experiments include two of the data sets used in [45]. Horse2zebra contains 939 horse images and 1177 zebra images downloaded from ImageNet [34] using the keywords *wild horse* and *zebra*. Apple2orange has 996 apple images and 1020 orange images from the same source, labeled with *apple* and *navel orange*. All images are 256×256 pixels in size.

4.2. Baselines, evaluation and parameter choices

Baselines. For the image generation experiments on CIFAR-10, we employ a state-of-the-art baseline, SNGAN [29], which is based on the Hinge loss. We consider SNGAN as the most relevant baseline, since we use it as starting point for our curriculum learning approaches. However, we include additional models from the recent literature, namely DCGAN [32], WGAN-GP [14], Parallel Optimal Transport GAN (POT-GAN) [2] and Generative Latent Nearest Neighbors (GLANN) [16]. We also include the results of the acGAN proposed by Doan et al. [7], which uses adaptive curriculum. Since our curriculum SNGAN-based models are unsupervised, we do not compare with class conditional (supervised) baselines. For the image translation experiments, we employ CycleGAN [45] as baseline.

Evaluation metrics. Evaluating the quality and realism of generated content as perceived by humans is not an easy task. At this moment, there is no universally agreed metric able to measure the outputs of GANs, each having its own shortcomings. To automatically quantify the performance, we use the Inception Score (IS) [35] and the Fréchet Inception Distance (FID) [15], which are computed over 10000 generated images (not used in the training process). The reported scores are averaged over 5 runs. A higher IS or a lower FID indicates higher performance. Along with the

automatic metrics, we evaluate the results by asking humans to annotate images, in order to determine if they prefer the baseline GANs or our Curriculum-GANs (CuGANs).

Implementation details. In the image generation experiments, we used the SNGAN implementation available at https://github.com/watsonyanghx/GAN_Lib_Tensorflow, which can reproduce the results reported in [29]. The model is based on ResNet. We trained the model for 80000 iterations using mini-batches of 64 samples. We observed that the Inception Score stabilizes much sooner (before 50000 iterations) using the Adam optimizer [20]. The learning rate is $2 \cdot 10^{-4}$. For the first curriculum learning approach, we split the training set in $m = 3$ batches, as Ionescu et al. [17]: an easy batch, a medium batch, and a difficult batch. Each batch contains the same number of samples. For the second and the third curriculum learning approaches, we set $\gamma = 5 \cdot 10^{-5}$, which is chosen with respect to the total number of iterations (80000). We conducted preliminary experiments to tune the other parameters. For the first curriculum learning approach, we experimented with three different numbers of iterations to train on the easy batch (5000, 10000 and 15000), and another three numbers of iterations to train on the easy and medium batches together (15000, 20000 and 25000). We obtained slightly better results for training on the easy batch for 15000 iterations, and on both easy and medium batches for 25000 iterations. The rest of the iterations (40000) include all three batches in the training. For the second and the third curriculum learning approaches, we conducted experiments with $k \in \{1, 2, 4\}$. When we weight the training images with the corresponding difficulty scores (as in Eq. (5)), we obtain optimal results with $k = 2$. When we sample the training images according to the difficulty scores (as in Eq. (9)), we obtain optimal results with $k = 4$.

In the image translation experiments, we used the CycleGAN implementation available at https://github.com/leehomyc/cyclegan-1. The model is trained for 25000 iterations, using a mini-batch size of 8 samples. As for SNGAN, we employ the Adam optimizer [20] with a learning rate of $2 \cdot 10^{-4}$. The weight of the cycle consistency loss term in the full objective function is set to $\lambda = 10$, as in the original paper [45]. We apply linear weight decay after the first 12500 iterations. We compare the baseline CycleGAN with the Curriculum-CycleGAN based on weighting the training images with the corresponding difficulty scores, since the weighting strategy provides the best FID score in the image generation experiments on CIFAR-10. We did not evaluate the other two curriculum learning approaches to avoiding tripling the human annotation time and costs.

4.3. Image generation results

Faster convergence. In Figure 4, we present the evolution of the Inception Scores for the standard SNGAN and three other SNGAN versions, that are enhanced through one of our curriculum learning strategies, on CIFAR-10. We note that the performance of each model stabilizes after 50000 iterations. However, the curriculum-based models reach higher IS values, right from the first iterations. This indicates that the Curriculum-SNGANs converge faster than the standard SNGAN. For instance, the Curriculum-SNGAN based on sampling (corresponding to Eq. (9)) achieves about the same IS value as the baseline SNGAN, in only 20000 iterations instead of 65000 iterations.

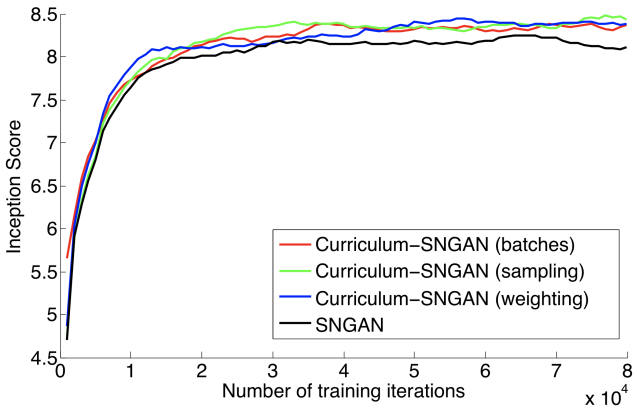

Figure 4. Inception Scores (IS) of SNGAN (baseline) versus three Curriculum-SNGAN models based on various curriculum learning strategies (batches, sampling, weighting), on CIFAR-10. The scores are computed on generated images, not used in the training process. Best viewed in color.

Method	IS	FID
DCGAN [32]	6.16 ± 0.07	71.07 ± 1.06
DCGAN (with ResNet)* [32]	6.64 ± 0.14	-
WGAN-GP* [29]	6.68 ± 0.06	40.20
WGAN-GP (with ResNet) [14]	7.86 ± 0.07	-
GLANN [16]	-	46.50 ± 0.20
POT-GAN [2]	6.87 ± 0.04	32.50
acGAN [7]	6.22 ± 0.04	49.81 ± 0.23
SNGAN* [29]	8.22 ± 0.05	21.70 ± 0.21
Curriculum-SNGAN (batches)	8.46 ± 0.13	14.64 ± 0.31
Curriculum-SNGAN (weighting)	8.44 ± 0.11	$\mathbf{14.41} \pm 0.24$
Curriculum-SNGAN (sampling)	$\mathbf{8.51} \pm 0.09$	14.48 ± 0.26

Table 1. Inception Scores (IS) and Fréchet Inception Distances (FID) on CIFAR-10. Several unsupervised GAN models [2, 7, 14, 16, 29, 32] are compared with our SNGAN-based approaches, each employing a different curriculum learning strategy proposed in this paper. The results marked with an asterisk are taken from the SNGAN paper [29]. The results of acGAN are based on the source code provided by Doan et al. [7]. The best IS (higher is better) and the best FID (lower is better) scores are marked in bold.

Superior results. In Table 1, we compare our Curriculum-SNGANs with the standard SNGAN, as well as DCGAN [32], WGAN-GP [14], GLANN [16], POT-GAN [2] and acGAN [7], on CIFAR-10. First, we note that SNGAN

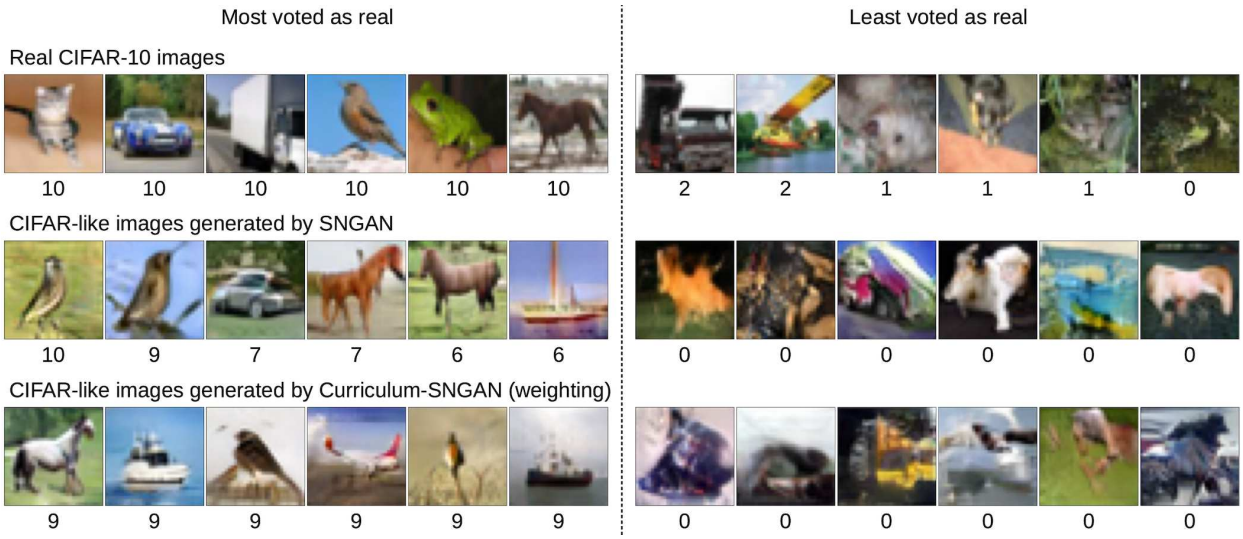

Figure 5. Most voted and least voted images from the set of 600 images labeled by human annotators. Images on each row are selected from different subsets: real images, generated by SNGAN and generated by Curriculum-SNGAN with weighting. Best viewed in color.

	A	B	C	D	E	F	G	H	I	J	Avg.
CIFAR-10	156	195	124	168	151	142	167	160	95	127	74.3%
SNGAN [29]	36	111	26	18	30	34	40	38	24	11	18.4%
Curriculum	48	123	27	27	39	60	67	58	28	23	25.0%

Table 2. Number of images voted as real by 10 human annotators (identified by letters from A to J). The annotators were asked to label 600 images (200 real CIFAR-10 images, 200 images generated by SNGAN and another 200 images generated by Curriculum-SNGAN with weighting) as real or fake.

Option	H→Z	Z→H	A→O	O→A	Avg.
CycleGAN [45]	11.9%	13.9%	20.6%	32.7%	19.8%
Curriculum-CycleGAN	52.5%	37.4%	37.1%	35.1%	40.5%
Ties	35.6%	48.7%	42.3%	32.2%	39.7%

Table 3. Average percentage of cases in which 6 human annotators consider images generated by CycleGAN as better, images generated by Curriculum-CycleGAN (weighting) as better, or both equally good. Evaluations are provided for 4 test sets of images: horse2zebra (H→Z), zebra2horse (Z→H), apple2orange (A→O), orange2apple (O→A). The overall average is also included.

achieves the best results among all baselines, confirming that SNGAN is indeed representative for the state-of-the-art. We observe that all our curriculum learning strategies can further boost the performance of SNGAN. When we divide the images into easy-to-hard batches, we achieve an IS of 8.46. The best IS score (8.51) is obtained when we sample the train images according to difficulty. For an unsupervised model, we believe that an IS of 8.51 is noteworthy. Furthermore, our improvements in terms of FID are much higher than all baselines, even compared to the adaptive curriculum approach of Doan et al. [7]. While easy-to-hard batches and sampling provide better IS values, we observe that the Curriculum-SNGAN based on weighting according to difficulty (corresponding to Eq. (5)) achieves the best FID value (14.41). For this reason we choose this curriculum learning strategy for the human evaluation experiments.

We asked 10 human annotators to label images either as real or fake. We provided the same set of 600 images (presented in a random order) to each annotator. We randomly selected 200 real CIFAR images, 200 images generated by SNGAN and 200 images generated by Curriculum-SNGAN (weighting). The goal of the annotation study is to determine the percentage of generated images that fool the annotators, using the real images as a control set (preventing evaluators from labeling every image as fake). In Table 2, we report the number of images labeled as real by each annotator. We note that in 25.7% cases, the annotators labeled real CIFAR images as being fake. Nevertheless, the humans largely figured out what images are generated. The standard SNGAN fooled annotators in 18.4% cases, while the Curriculum-SNGAN fooled annotators in 25.0% cases. Interestingly, each and every human labeled more images generated by Curriculum-SNGAN as real than images generated by the baseline SNGAN. As illustrated in Figure 5, there are several images generated by Curriculum-SNGAN, which are labeled as real by 9 out of 10 annotators. The number of votes drops faster for the standard SNGAN approach. All results indicate that the images generated by Curriculum-SNGAN are superior to those generated by SNGAN.

4.4. Image translation results

The image translation results are evaluated only by human annotators. There are four test sets of images [45], corresponding to the following translations: horse2zebra (120 images), zebra2horse (140 images), apple2orange (266 images) and orange2apple (248 images). We asked 6 hu-

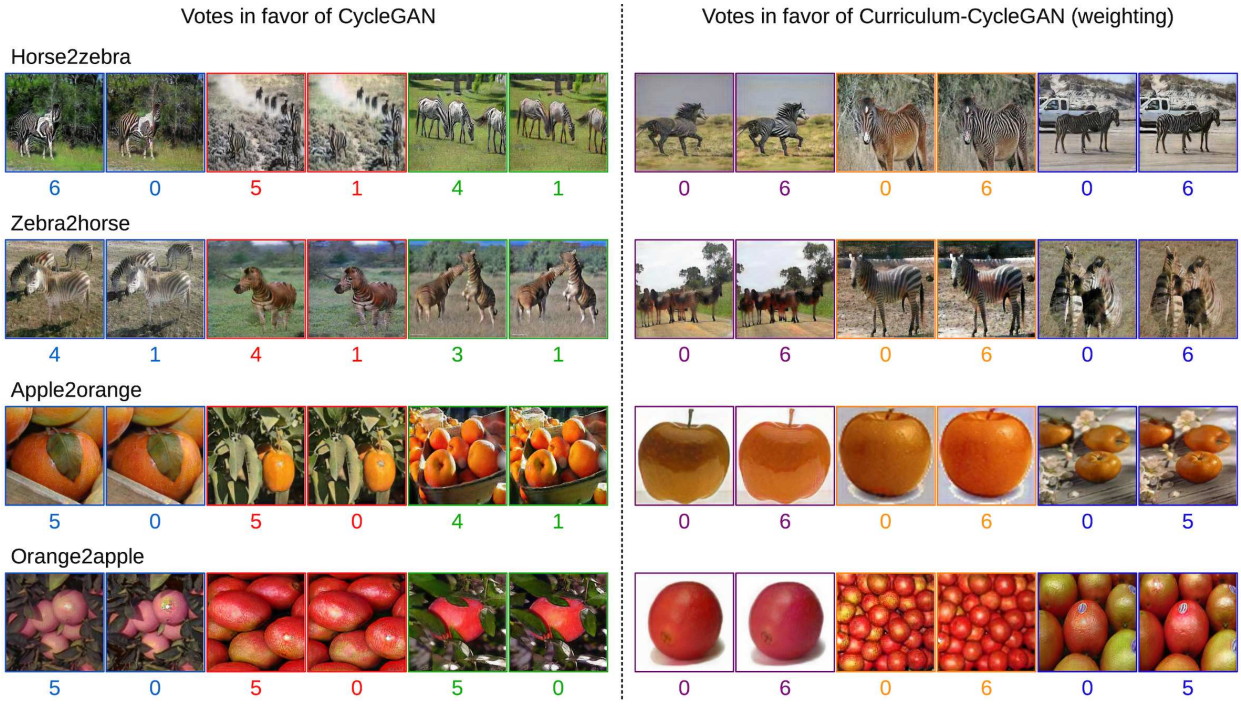

Figure 6. Side by side image pairs generated by CycleGAN (left image in each pair) and Curriculum-CycleGAN (right image in each pair) with the corresponding number of votes provided by 6 human annotators. When the sum of the number of votes in a pair is lower than 6, it means that the missing votes correspond to ties. Image pairs that received most votes in favor of CycleGAN are presented in the left-hand side of the figure, while image pairs that received most votes in favor of Curriculum-CycleGAN are presented in the right-hand side. Best viewed in color.

man annotators to choose between the images translated by CycleGAN and those translated by Curriculum-CycleGAN (weighting), without disclosing any information about the models. In each case, we also provided the original (source) image. Since the test images are fixed for both models, the random chance factor is eliminated. In Figure 6, we show several images translated by both models side by side. We notice that there are several image pairs in which all 6 annotators opted for Curriculum-CycleGAN. For horse2zebra, the baseline CycleGAN wins when our model produces brownish zebras. For apple2orange, annotators prefer the baseline when our model produces artifacts, but they prefer our model when it produces the right tone of orange.

In Table 3, we present the average percentage of cases (computed on the 6 annotators) in which the annotators prefer either the CycleGAN output images or the Curriculum-CycleGAN output images, as well as the percentage of tied cases (images are labeled as equally good). We note that on three sets of images (horse2zebra, zebra2horse and apple2orange), the annotators show significant preference for our Curriculum-CycleGAN based on weighting. Furthermore, in these three test sets, all humans prefer our model over the baseline CycleGAN (the individual percentages are not shown in Table 3 due to lack of space). For orange2apple, only 2 out of 6 annotators prefer our model, although our model has a higher average preference (35.1%) compared to the baseline (32.7%), as seen in Table 3. All in all, the human annotators seem to prefer the curriculum-based approach in 20.7% more cases than the baseline CycleGAN, confirming once more that the curriculum strategy is indeed useful.

5. Conclusion

In this paper, we presented three curriculum learning strategies for training GANs. The empirical results indicate that our curriculum learning strategies achieve faster convergence during training, i.e. the number of training iterations can be reduced by a factor of three without affecting the quality of the generative results. Furthermore, using a similar number of training iterations, our curriculum learning strategies can boost the quality of the generative and translation results, surpassing all considered baselines [2, 7, 14, 16, 29, 32, 45] on CIFAR-10. Both automatic measures and human evaluators confirm our findings. While we conducted experiments on images of 32×32 and 256×256 of pixels in size, in future work, we aim to apply our curriculum learning strategies in order to generate larger images, containing a natural variety of object classes.

Acknowledgements. Work funded from a grant of Ministry of Research and Innovation, CNCS-UEFISCDI, project no. PN-III-P1-1.1-PD-2016-0787, and from the EEA Grants 2014-2021, project no. EEA-RO-NO-2018-0496.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017. 1, 2

[2] G. Avraham, Y. Zuo, and T. Drummond. Parallel Optimal Transport GAN. In *Proceedings of CVPR*, pages 4411–4420, 2019. 5, 6, 8

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of ICML*, pages 41–48, 2009. 1, 3

[4] C.-C. Chang and C.-J. Lin. Training ν-Support Vector Regression: Theory and Algorithms. *Neural Computation*, 14:1959–1977, 2002. 4

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of BMVC*, 2014. 4

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of CVPR*, pages 8789–8797, 2018. 1, 2

[7] T. Doan, J. Monteiro, I. Albuquerque, B. Mazoure, A. Durand, J. Pineau, and R. D. Hjelm. On-line Adaptative Curriculum Learning for GANs. In *Proceedings of AAAI*, 2019. 1, 3, 5, 6, 7, 8

[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 Results, 2012. 4

[9] K. Ghasedi, X. Wang, C. Deng, and H. Huang. Balanced Self-Paced Learning for Generative Adversarial Clustering Network. In *Proceedings of CVPR*, pages 4391–4400, 2019. 1, 3

[10] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. 1, 3

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of NIPS*, pages 2672–2680, 2014. 1, 2, 3

[12] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of ICML*, pages 1311–1320, 2017. 1, 3

[13] L. Gui, T. Baltrušaitis, and L.-P. Morency. Curriculum learning for facial expression recognition. In *Proceedings of FG*, pages 505–511, 2017. 1

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Proceedings of NIPS*, pages 5767–5777, 2017. 1, 2, 5, 6, 8

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of NIPS*, pages 6626–6637, 2017. 2, 5

[16] Y. Hoshen, K. Li, and J. Malik. Non-Adversarial Image Synthesis with Generative Latent Nearest Neighbors. In *Proceedings of CVPR*, pages 5811–5819, 2019. 5, 6, 8

[17] R. Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari. How hard can it be? estimating the difficulty of visual search in an image. In *Proceedings of CVPR*, pages 2157–2166, 2016. 1, 3, 4, 6

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of CVPR*, pages 1125–1134, 2017. 1, 2

[19] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *Proceedings of ICML*, pages 2309–2318, 2018. 1, 3

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015. 6

[21] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 1

[22] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2, 4, 5

[23] S. Li, X. Zhu, Q. Huang, H. Xu, and C.-C. J. Kuo. Multiple Instance Curriculum Learning for Weakly Supervised Object Detection. In *Proceedings of BMVC*, 2017. 1, 3

[24] Z. Lin, A. Khetan, G. Fanti, and S. Oh. PacGAN: The power of two samples in Generative Adversarial Networks. In *Proceedings of NeurIPS*, pages 1505–1514, 2018. 1, 2

[25] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of ICCV*, pages 2794–2802, 2017. 1, 2

[26] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. 1989. 1

[27] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of GANs. In *Proceedings of NIPS*, pages 1825–1835, 2017. 1

[28] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of ICLR*, 2018. 2, 3, 5, 6, 7, 8

[30] P. Morerio, J. Cavazza, R. Volpi, R. Vidal, and V. Murino. Curriculum dropout. In *Proceedings of ICCV*, pages 3544–3552, 2017. 1, 3

[31] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceeding of ICML*, pages 2642–2651, 2017. 1, 2

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of ICLR*, 2016. 1, 2, 5, 6, 8

[33] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. In *Proceedings of ICML*, pages 1060–1069, 2016. 1, 2

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, K. A., A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 4, 5

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Proceedings of NIPS*, pages 2234–2242, 2016. 2, 5

[36] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of ICLR*, 2014. 4

[37] L. B. Smith, S. Jayaraman, E. Clerkin, and C. Yu. The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, 22(4):325–336, 2018. 3

[38] P. Soviany and R. T. Ionescu. Frustratingly Easy Trade-off Optimization between Single-Stage and Two-Stage Deep Object Detectors. In *Proceedings of CEFRL Workshop of ECCV*, pages 366–378, 2018. 3

[39] P. Soviany and R. T. Ionescu. Optimizing the Trade-off between Single-Stage and Two-Stage Deep Object Detectors using Image Difficulty Prediction. In *Proceedings of SYNASC*, pages 209–214, 2018. 3

[40] I. O. Tolstikhin, S. Gelly, O. Bousquet, C.-J. Simon-Gabriel, and B. Schölkopf. AdaGAN: Boosting Generative Models. In *Proceedings of NIPS*, pages 5424–5433, 2017. 1, 2

[41] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. MANCS: A Multi-task Attentional Network with Curriculum Sampling for Person Re-identification. In *Proceedings of ECCV*, pages 365–381, 2018. 1, 3

[42] J. Wang, X. Wang, and W. Liu. Weakly-and Semi-supervised Faster R-CNN with Curriculum Learning. In *Proceedings of ICPR*, pages 2416–2421, 2018. 1, 3

[43] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *Proceedings of ECCV*, pages 318–335, 2016. 1, 2

[44] D. Zhang, J. Han, L. Zhao, and D. Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, 127(4):363–380, 2019. 1, 3

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of ICCV*, pages 2223–2232, 2017. 2, 4, 5, 6, 7, 8