RareGAN: Generating Samples for Rare Classes

Zinan Lin, Hao Liang, Giulia Fanti, Vyas Sekar
Carnegie Mellon University
zinanl@andrew.cmu.edu, hl106@rice.edu, gfanti@andrew.cmu.edu, vsekar@andrew.cmu.edu

Abstract

We study the problem of learning generative adversarial networks (GANs) for a rare class of an unlabeled dataset subject to a labeling budget. This problem is motivated from practical applications in domains including security (e.g., synthesizing packets for DNS amplification attacks), systems and networking (e.g., synthesizing workloads that trigger high resource usage), and machine learning (e.g., generating images from a rare class). Existing approaches are unsuitable, either requiring fully-labeled datasets or sacrificing the fidelity of the rare class for that of the common classes. We propose RareGAN, a novel synthesis of three key ideas: (1) extending conditional GANs to use labelled and unlabelled data for better generalization; (2) an active learning approach that requests the most useful labels; and (3) a weighted loss function to favor learning the rare class. We show that RareGAN achieves a better fidelity-diversity tradeoff on the rare class than prior work across different applications, budgets, rare class fractions, GAN losses, and architectures.

1 Introduction

Many practitioners in diverse domains such as security, networking, and systems require samples from rare classes. For example, operators often want to generate queries that force servers to send undesirable responses (Moon et al. 2021), or generate packets that trigger high CPU/memory usage or processing delays for performance evaluation (Petsios et al. 2017).

Prior domain-specific solutions to these problems rely heavily on prior knowledge (e.g., source code) of the systems, which is often unavailable (Lin et al. 2019). Indeed, in response to a recent executive order Improving the Nation’s Cybersecurity1, the U.S. National Institute of Standards and Technology published guidance highlighting the importance of creating ‘black box’ tests for device and software security that do not rely on the implementation or source code of systems (Black, Guttmann, and Okun 2021).

Given the success of generative adversarial networks (GANs) (Goodfellow et al. 2014) on data generation, we ask if we can use GANs to generate samples from a rare class (e.g., attack packets, packets that trigger high CPU usage) without requiring prior knowledge about the systems. Note that there are two unique characteristics in our problem:

C1. High labeling cost. Labels (whether a sample belongs to the rare class) are often not available a priori, and getting labels is often resource intensive. For example, for a new system, we often do not know a priori which packets will trigger high CPU usage, and evaluating the CPU usage of a packet (for labeling it) can be time consuming.

C2. Rare class only. We only need samples from the rare class (e.g., attack packets); system operators are often less concerned about common class samples (e.g., benign packets).

To the best of our knowledge, no prior GAN paper considers both constraints. Prior related work (see § 3.2) often assumes that the labels are available (failing C1), or tries to generate both rare and common samples, which sacrifices the fidelity on the rare class (failing C2). We will see in § 4 that these new characteristics bring unique challenges.

Contributions. We propose RareGAN, a generative model for rare data classes, given an unlabeled dataset and a labeling budget. It combines three ideas: (1) It modifies existing conditional GANs (Odena, Olah, and Shlens 2017) to use both labelled and unlabelled data for better generalization. (2) It uses active learning to label samples; we show theoretically that unlike prior work (Xie and Huang 2019), our implementation does not bias the learned rare class distribution. (3) It uses a weighted loss function that favors learning the rare class over the common class; we propose efficient optimization techniques for realizing this reweighting.

We show that RareGAN achieves a better fidelity-diversity tradeoff on the rare class than baselines across different use cases, budgets, rare class fractions, GAN losses, and architectures. Table 1 shows that RareGAN achieves better fidelity and diversity (with a smaller labeling budget) when gen-

<table>
<thead>
<tr>
<th></th>
<th>Budget↓</th>
<th>Fidelity↓</th>
<th>Diversity↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmpMAP</td>
<td>14,788,089</td>
<td>16.60</td>
<td>1.68%</td>
</tr>
<tr>
<td>RareGAN (ours)</td>
<td>200,000</td>
<td>4.16</td>
<td>98.07%</td>
</tr>
</tbody>
</table>

Table 1: RareGAN achieves better fidelity and diversity with lower budget on DNS amplification attacks than domain-specific techniques. See § 5.1 for the definition of metrics.

1 https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity
Motivating scenario 1: amplification attacks (security). Many widely deployed public servers and protocols like DNS, NTP, and Memcached are vulnerable to amplification attacks (Moon et al. 2021; Rossow 2014), where the attacker send requests (network packets) to public servers with spoofed source IP addresses, so the response goes to the specified victims. These requests are designed to maximize response size, thus exhausting victims’ bandwidth. Server operators want to know which requests trigger these amplification attacks to e.g., drop attack requests. Prior solutions require detailed information about the server, such as source code (Rossow 2014), which may be unavailable (Lin et al. 2019).

Our problem formulation. As in (Moon et al. 2021), we treat the rare class as all requests with an amplification factor $\text{size of response packet} / \text{size of request packet} \geq T$, a pre-defined threshold. All other requests belong to the common class. To label a request (i.e., check its amplification factor), we send the request through the server, which can be costly. Hence, we want to limit the number of label queries. We want to learn a uniform distribution over high-amplification requests to maximize coverage of the input space.

Motivating scenario 2: performance stress testing (systems & networking). Many deployed systems and networks today rely on black-box components (e.g., lacking source code, detailed specifications). System operators may therefore want to understand worst-case system performance (e.g., CPU/memory usage or delay in the presence of congestion) and optimize for such scenarios (Pedrosa et al. 2018). However, current tools for generating such workloads often rely on a system’s source code (Caballero et al. 2007; Petsios et al. 2017; Pedrosa et al. 2018), which may be unavailable (Lin et al. 2019; Black, Guttmann, and Okun 2021).

Our problem formulation. We treat the rare class as packets with resource usage (e.g., CPU/memory/processing delay) $\geq T$, a pre-defined threshold. The operator can use the trained $\hat{p}_r$ to synthesize such workloads. For the same reasons as the previous case, we want to limit the number of label queries and learn the rare class faithfully.

Motivating scenario 3: inspecting rare class images (ML). Prior GANS on unbalanced image datasets focus on generating samples from both rare and common classes for improving downstream classification accuracy (§ 3.2). However, in some other applications, we may only need rare class samples. For example, in federated learning, we may want to inspect the samples from specific client/class slices that have bad accuracy for debugging (Augenstein et al. 2019).

We will see in § 5.2 that RareGAN outperforms baselines across all these very different use cases and data types.

3 Background and Related Work

3.1 Background

Generative Adversarial Networks (GANs). GANs (Goodfellow et al. 2014) are a class of deep generative models that have spurred significant interest in recent years. GANs involve two neural networks: a generator $G$ for mapping a random vector $z$ to a random sample $G(z)$, and a discriminator $D$ for guessing whether the input image is generated or from the real distribution $p$. The vanilla GAN loss (Goodfellow et al. 2014) is given by

$$L_{GAN}(G, D) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log (1 - D(G(z))].$$

Real
GAN
ACGAN
ALCG
Elastic-InfoGAN
BAGAN
RareGAN (ours)

Figure 1: Random generated samples (no cherry-picking) on MNIST class ‘0’ with $B = 1,000$ and $\alpha = 1\%$. The red channel plots a generated image, and the green channel the nearest real image from the training set. Yellow pixels show where the two overlap. RareGAN achieves the high sample quality and diversity without memorizing training data.
et al. 2014) is \( \min_G \max_D L_{\text{GAN}}^W (D; \hat{p}; p) \), where

\[
L_{\text{GAN}}^W (D; \hat{p}; p) = \mathbb{E}_{x \sim p} \left[ \log D(x) \right] + \mathbb{E}_{z \sim \mathcal{Z}} \left[ \log (1 - D(x)) \right],
\]

and \( \hat{p} \) denotes the generated distribution induced by \( G(z) \) where \( z \) is sampled from a fixed prior distribution \( p_z \) (e.g., Gaussian or uniform). It has been shown that under some assumptions, Eq. (1) is equivalent to \( \min_G d_{\text{JS}} (p, \hat{p}) \), where \( d_{\text{JS}} (\cdot, \cdot) \) denotes Jensen-Shannon divergence between the two distributions. Several other distance metrics have later been proposed to improve the stability of training (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Nowozin, Cseke, and Tomioka 2016; Mao et al. 2017).\footnote{The Wasserstein distance \( d_W (\cdot, \cdot) \) is one of the most widely used metrics (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017). The loss of Wasserstein GAN is \( \min_G \max_{\|D\| \leq 1} L_{\text{GAN}}^W (D; \hat{p}; p) \), where

\[
L_{\text{GAN}}^W (D; \hat{p}; p) = \mathbb{E}_{x \sim p} [D(x)] - \mathbb{E}_{z \sim \mathcal{Z}} [D(x)]
\]

and \( \|D\| \) denotes the Lipschitz constant of \( D \). RareGan works well with both of these losses.}

**Auxiliary Classifier GANs (ACGAN).** Conditional GANs (CGANs) (Mirza and Osindero 2014; Odena, Olah, and Shlens 2017) are a variant of GANs that support conditional sampling. Besides \( z \), the generators in CGANs have an additional input \( c \) which controls the properties (e.g., category) of the generated sample. For example, in a face image dataset (with male/female labeled), instead of only sampling from the entire face distribution, generators in CGANs could allow us to control whether to generate a male or female by specifying \( c \). Several different techniques have been proposed to train such a conditional generator (Mirza and Osindero 2014; Odena, Olah, and Shlens 2017; Salimans et al. 2016; Mariani et al. 2018). ACGAN (Odena, Olah, and Shlens 2017) is one such widely-used variant (Kong et al. 2019; Xie and Huang 2019; Choi et al. 2018). ACGAN adds a classifier \( C \) which discriminates the labels for both generated data and real data. The ACGAN loss function is defined by

\[
\min_G \min_C \max_D L_{\text{GAN}} (D, \hat{p}; p) + L_{\text{classification}} (C; \tilde{p}_z; p_z),
\]

where \( L_{\text{GAN}} (D, \hat{p}; p) \) is the regular GAN loss (e.g., Eq. (1) (Kong et al. 2019; Xie and Huang 2019; Choi et al. 2018) or Eq. (2) (Lee 2018)), except that \( \hat{p} \) is induced by \( G(z; c) \), where \( z \sim p_z \) and \( c \sim p_l \), where \( p_l \) is the ground truth label. \( L_{\text{classification}} \) is defined by

\[
L_{\text{classification}} (C; \tilde{p}_z; p_z) = -\mathbb{E}_{(x, c) \sim p_z} \log C(x, c) - \mathbb{E}_{(x, c) \sim \hat{p}_z} \log C(x, c),
\]

where \( C(x, c) \) denotes classifier \( C \)'s probability prediction for class \( c \) on input sample \( x \), \( p_z \) denotes the real joint distribution of samples and labels, and \( \hat{p}_z \) denotes the joint distribution over labels and generated samples in \( \hat{p} \). In practice, \( D \) and \( C \) usually share some layers. Note that in Eq. (4) the classifier is trained to match not only the real data, but also the generated data, as in (Kong et al. 2019; Lee 2018; Odena, Olah, and Shlens 2017). However, in some other implementations, the second part of loss only applies on \( G \), so that the classifier will not be misled by errors in the generator (Xie and Huang 2019; Choi et al. 2018).

### 3.2 Related Work

Depending on the availability of the labels, prior related works can be classified into fully-labeled, unsupervised, semi-supervised, and self-supervised GANs.

**Fully supervised GANs.** Prior work has studied how to use GANs (particularly ACGAN) to augment imbalanced, labeled datasets, e.g., for downstream classification tasks (Mullick, Datta, and Das 2019; Douzas and Bacao 2018; Ren, Liu, and Liu 2019; Ali-Gombe and Elyan 2019; Mariani et al. 2018; Rangwani, Mopuri, and Babu 217; Asokan and Seelamantula 2020; Yang and Zhou 2021). For example, EW-GAN (Ren, Liu, and Liu 2019), MFC-GAN (Ali-Gombe and Elyan 2019), Douzas and Bacao (2018), Wei et al. (2019), and BAGAN (Mariani et al. 2018) all augment the original dataset by generating samples from the minority class with a conditional GAN. Wei et al. (2019) utilizes known mappings between images in different classes (e.g., mapping an image of normal colon tissue to precancerous colon tissue). BAGAN (Mariani et al. 2018) instead trains an autoencoder on the entire dataset, learns a Gaussian latent distribution for each class, and uses that as the input noise for each class to the GAN generator. We cannot utilize these approaches because we lack labels.

**Unsupervised GANs.** Unsupervised GANs (Chen et al. 2016; Ojha et al. 2019; Lin et al. 2020) do not control which factors to learn. Hence, there is no guarantee that they will learn to separate samples along the desired factor and threshold (e.g., classification time of the generated packets).

**Semi-supervised GANs.** Our proposed approach in § 4.1 is one instance of semi-supervised GANs (Odena 2016; Dai et al. 2017; Kumar, Sattigeri, and Fletcher 2017; Haque 2020; Zhou et al. 2018). Other semi-supervised GANs could also be used, like the seminal one (Salimans et al. 2016), which uses a single modified discriminator both to separate fake from real samples (as in classical GANs) and to classify the labels of real data. We choose to use separated classifier (as in ACGAN) as the classifier is not influenced by real/fake objective and therefore provides cleaner signal for our active learning technique in § 4.2. The closest prior work, ALCG (Xie and Huang 2019), is also an instance of semi-supervised GANs. Like us, they are training conditional GANs in an active learning setting. However, their goal is to synthesize high-quality samples from all classes, whereas we want to faithfully reproduce only the rare class. We show how this distinction requires different algorithmic designs (§ 4), and leads to poor performance by ALCG on our problems (§ 5).

**Self-supervised GANs.** Self-supervision has been used in both unsupervised and semi-supervised GANs (Sun, Bhat-tarai, and Kim 2020; Chen et al. 2019; Ojha et al. 2019). It is also unclear how to apply self-supervised GANs in our problem, as they are most useful when we have some prior understanding of the physical or semantic of a system, but in our problem we are given an arbitrary system whose internal structure is unknown. For example, for disentangling digit types (e.g., 0 v.s. 1), Elastic-InfoGAN (Ojha et al. 2019) applies operations like rotation on images to construct positive pairs for self-supervised loss, as we know these operations do not change digit types. However, it is unclear what cor-
responding operations should be in our problem due to the ‘black box’ nature of the systems. For example, in motivating scenario 2 (§ 2), it is unclear what operations on packets would keep the CPU/memory usage of the system.

Our work is also related to prior work on weighted loss and data augmentation for GANs, which we discuss below.

**Weighted loss.** Our proposed approach involves a weighted loss for GANs (§ 4.3). Although weighted loss for GANs and classification has been proposed in prior work (Zadorozhnyy, Cheng, and Ye 2021; Cui et al. 2019), the weighting schemes and the goals are completely different to ours. For example, awGAN (Zadorozhnyy, Cheng, and Ye 2021) adjusts the weights on the real/fake losses, with the goal of balancing the gradient directions of these two losses and making training more stable. Instead, as we see in § 4.3, the proposed RareGAN uses different weights for samples in the rare and the common classes (regardless of whether they are fake or real), with the goal of balancing the learning of the rare and the common classes.

**Data augmentation.** Data augmentation techniques have been proposed for improving GANs’ performance in limited data regimes (Karras et al. 2020; Zhao et al. 2020). However, these techniques usually focus on image datasets and the augmentation operations (e.g., random cropping) do not extend to other domains (e.g., our networking datasets).

## 4 Approach

As mentioned in § 2 and § 3.2, our problem is distinguished by two factors: (1) We want to learn only the rare class distribution; (2) The rare/common labels are not available in advance, and we have a fixed labeling budget (can be used online during training).

The most obvious straw-man solution to our problem is to randomly and uniformly draw \( B \) samples from the dataset and request their labels. Then, we train a vanilla GAN on the packets with label “rare”. However, since the rare class could have a very low fraction, the number of training samples will be small and the GAN is likely to overfit to the training dataset and generalize poorly.

As in prior work (Xie and Huang 2019; Ren, Liu, and Liu 2019; Ali-Gombe and Elyan 2019; Mariani et al. 2018), we can use conditional GANs like ACGAN (§ 3.1) to incorporate common class samples into training, because they could actually be useful for learning the rare class. For example, in face image datasets, the rare class (e.g., men with long hair) and common classes share some characteristics (i.e., faces). However, due to the small number of rare samples, ACGAN still has bad fidelity and diversity (§ 5.2).

In the following, we progressively discuss the design choices we make in RareGAN to address the challenges, and highlight the differences to ALCG (§ 3.2), the most closely-related work.

### 4.1 Better Distribution Learning with Unlabeled Samples

In the above process, the majority of samples from \( D \) are unlabeled because of the labeling budget. Those samples are not used in training (as in ALCG). However, they contain information about the mixture distribution of rare and common classes, and could therefore help learn the rare class.

Our proposed approach relies on carefully altering the ACGAN loss. Recall that the loss has two parts: a classification loss separating rare and common samples, and a GAN loss evaluating their mixture distribution (§ 3.1). Note that the GAN loss does not require labels. We propose a modified ACGAN training that uses labeled samples for the classification loss, and all samples for the GAN loss. However, when training the GAN loss, we need to know the fraction of rare/common classes in order to feed the condition input to the generator. This can be estimated from the labeled samples. The maximum likelihood estimate \( \hat{\alpha} = \frac{1}{n} \) has variance \( \alpha(1-\alpha) \), which is small for reasonable \( n \).

### 4.2 Improving Classifier Performance with Active Learning

Because the rare and common classes are highly imbalanced, the classifier in ACGAN could have a bad accuracy. In classification literature, confidence-based active learning has been widely used for solving this challenge (Li and Sethi 2006; Joshi, Porikli, and Papanikolopoulos 2009; Sivaraman and Trivedi 2010), which expends the labeling budget on samples about which the classifier is least confident.

Inspired by these works, our approach is to divide the training into \( S \) stages. At the beginning of each stage, we pass all unlabeled samples through the classifier, and request the labels for the \( B/s \) samples that have the lowest \( \max \{ C(x, \text{rare}), C(x, \text{common}) \} \), where \( C(x, c) \) denotes the classifier’s (normalized) output. This sample selection criterion is called “least confidence sampling” in prior literature (Lewis and Catlett 1994). There are other criteria like margin of confidence sampling and entropy-based sampling (Vlachos 2008; Kong et al. 2019; Xie and Huang 2019). Since we have only two classes, they are in fact equivalent.

While ALCG also uses confidence-based active learning, it does so in a diametrically opposite way: they request the labels for the most certain samples. These two completely different designs result from having different goals: ALCG aims to generate high-quality images, whereas we aim to faithfully reproduce the rare class distribution. The most certain samples usually have better image quality, and therefore ALCG wants to include them in the training. The least certain samples are more informative for learning distribution boundaries, and therefore we label them.

**Relation to § 4.1.** Naively using active learning is actually counterproductive; if we only use labeled samples in the training (as is done in ALCG), the learned rare distribution will be biased, which partially explains ALCG’s poor performance in our setting (§ 5). As discussed in § 4.1, we instead use all unlabeled samples for training the GAN loss. The following proposition explains how this circumvents the problem (proof in App. A (Lin et al. 2022)).

---

\( \text{In the first stage, the samples for labeling are randomly chosen from the dataset.} \)
Proposition 1 The optimization

\[ p^* \in \arg\min_{\hat{p}} \min_{C \in C} d(\hat{p}, p) + \mathcal{L}_{\text{classification}}(C; \hat{p}_x; p'_x) \]

satisfies \( d(p^*, p_x) = 0 \), where: (a) \( p^* \) is \( p \) under condition "rare"; (b) \( p'_x \) is any joint (sample, label) distribution where the support of \( p'_x \) covers the entire sample space; (c) \( p_x \) is the generated joint distribution of samples and labels; and (d) \( C \) is the set of measurable functions.

The above optimization is a generalization of the ACGAN loss function, where \( d(\cdot, \cdot) \) denotes an appropriately-chosen distance function for the GAN loss (e.g., Wasserstein-1). Note that the classification loss is computed over the (possibly biased) distribution induced by active learning, \( \hat{p} \), whereas the GAN loss is computed with respect to the true distribution \( p \), which uses all samples, labeled or not. The proposition is stating that even though we use the biased \( \hat{p} \) for the classification loss, we can still learn \( p_x \). On the other hand, if we had used \( \hat{p}' \) for the GAN loss (as is done in ALCG), we would recover a biased version of \( p_x \).

4.3 Better Rare Class Learning with Weighted Loss

Because the rare class has low mass, errors in the rare distribution have only bounded effect on the GAN loss in Eq. (3). Next, we propose a re-weighting technique for reducing this effect at the expense of learning the common class. Let \( \hat{p} \) be the learned sample mixture distribution (without labels). Let \( \hat{p} = \int_{\text{Support}(p)} \hat{p}(x) dx \) be the fraction of rare samples under \( \hat{p} \), and let \( p'_x \) be \( \hat{p} \) restricted to and normalized over Support(\( p_x \)). (Recall that \( p_x \) is the generated distribution under condition \( y = "\text{rare}"; \) it need not be the case that Support(\( p_x \)) = Support(\( \hat{p}_x \)).) Similarly, let \( p'_x \) be \( \hat{p} \) over \( \lambda' = \text{Support}(\hat{p}_x) \) where \( \lambda' \) is the entire sample space. Recall that the original GAN loss in Eq. (3) tries to minimize

\[
d(p, \hat{p}) = d(\alpha \cdot p_x + (1 - \alpha) \cdot \hat{p}_x, \hat{p} + (1 - \alpha) \cdot \hat{p}_x)
\]

where \( d \) is \( d_{\lambda} \) or \( d_{\hat{p}} \). We propose to modify this objective function to instead minimize

\[
d(\omega \cdot p_x + (1 - \omega) \cdot \hat{p}_x, w' \cdot \hat{p}_x + \frac{1 - w}{s} \cdot \hat{p}_x) \cdot \hat{p}_x
\]

where \( \omega \in (1, 1/\alpha) \) is the additional multiplicative weight to put on the rare class; and \( s = \frac{1 - \omega}{1 - \alpha} \) is the normalization constant, which is 1 when \( \alpha = \hat{\alpha} \). It is straightforward to see that Eq. (5) \( = 0 \Leftrightarrow d(p, \hat{p}) = 0 \). However, these two objective functions have different effects in training: this modified loss will more heavily penalize errors in the rare distribution. Consider two extremes: (1) When \( w = 1 \), Eq. (5) is reduced to the original \( d(p, \hat{p}) \), placing no additional emphasis on the rare class; (2) When \( w = 1/\alpha \), Eq. (5) is reduced to \( d(p_x, \hat{p}_x) \), focusing only on the rare class and placing no constraint on the common class. This completely loses the benefit of learning the classes jointly (§ 4). For a \( \omega \in (1, 1/\alpha) \), we can achieve a better trade-off between the information from the common class and the penalty on the error of rare class.

To implement the above idea, we propose to add a multiplicative weight to the loss of both real and generated samples according to their label, i.e. changing Eqs. (1) and (2) to

\[
\mathcal{L}_{\text{GAN}}(D, \hat{p}; p) = \mathbb{E}_{x \sim p}[W(x) \cdot \log D(x)] + \frac{1}{2} \cdot \mathbb{E}_{z \sim \hat{p}}[W(x) \cdot \log (1 - D(x))]
\]

for Jensen-Shannon divergence and

\[
\mathcal{L}_{\text{GAN}}(D, \hat{p}; p) = \mathbb{E}_{x \sim p}[W(x) \cdot D(x)] - \frac{1}{2} \cdot \mathbb{E}_{z \sim \hat{p}}[W(x) \cdot D(x)]
\]

for Wasserstein distance, where

\[
W(x) = \begin{cases} 
\frac{w}{1 - w} & (x \in \text{Support}(p_x)) \\
\frac{1 - w}{1 - \alpha} & (x \notin \text{Support}(p_x)) 
\end{cases}
\]

Proposition 2 For any \( D, p, \) and \( \hat{p} \), we have

\[
\mathcal{L}_{\text{GAN}}^w(D, \hat{p}; p) = \mathcal{L}_{\text{GAN}}^w(D, \hat{q}; q)
\]

\[
\mathcal{L}_{\text{GAN}}^w(D, \hat{p}; p) = \mathcal{L}_{\text{GAN}}^w(D, \hat{q}; q)
\]

where \( \hat{q} = \frac{w}{s} \cdot \hat{p}_x + \frac{1 - w}{s} \cdot \hat{p}_c \), and \( q = w \cdot p_x + (1 - w) \cdot p_c \).

Implementing this weighting is nontrivial, however: (1) The above implementation requires the ground truth labels of all real and generated samples for evaluating Eq. (8), and we do not want to waste labeling budget on weight estimation. For this, we use the ACGAN rare/common classifier as a surrogate labeler. Although this classifier is imperfect, weighting real and generated samples according to the same labeler is sufficient to ensure that the optimum is still \( d(p, \hat{p}) = 0 \). (2) Evaluating the normalization constant \( s \) in Eqs. (6) and (7) requires estimating \( \hat{\alpha} \), which is inefficient as \( \hat{\alpha} \) changes during training. Empirically, we found that setting \( s = 1 \) gave good and stable results.

5 Experiments

We conduct experiments on all three applications in § 1. The code can be found at https://github.com/fjxmlzn/RareGAN.

Use case 1: DNS amplification attacks. DNS is one of the most widely-used protocols in amplification attacks (Rossow 2014). DNS requests that trigger high amplification have been extensively analyzed in the security community (Kambourakis et al. 2007; Anagnostopoulos et al. 2013; Rossow 2014), though most of the those analyses are manual or use tools specifically designed for (DNS) amplification attacks. We show that RareGAN, though designed for a more general set of problems, can also be effectively used for finding amplification attack requests. In this setting, we define the rare class as DNS requests that have \( \frac{\text{size of response}}{\text{size of request}} \geq T \), where \( T \) is a threshold specified by users. For the request space, we follow the configuration of (Lin et al. 2019; Moon et al. 2021): we let GANs generate 17 fields in the DNS request; for 5 fields among them (qr, opcode, rdata type, rdata class, and url), we provide candidate values; for all other 12 fields, we let GANs explore all possible bits. The entire search space is \( 3.6 \times 10^{17} \). Unlike image datasets where samples from the
mixture distribution $p$ are given, here we need to define $p$. Since our goal is to find all DNS requests with amplification $\geq T$, we define $p$ as a uniform distribution over the search space. More details are in App. C (Lin et al. 2022).

Note on ethics: For this experiment, we needed to make many DNS queries. To avoid harming the public DNS resolvers, we set up our own DNS resolvers on Cloudlab (Duplyakin et al. 2019) for the experiments.

Use case 2: packet classification. Network packet classification is a fundamental building block of modern networks. Switches or routers classify incoming packets to determine what action to take (e.g., forward, drop) (Liang et al. 2019). An active research area in networking is to propose classifiers with low inference latency (Chiu et al. 2018; Liang et al. 2019; Soylu, Erdem, and Carus 2020; Rashelbach, Rottenstreich, and Silberstein 2020). We take a recently-proposed packet classifier (Liang et al. 2019) for example, which was designed to optimize classification time and memory footprint. We define the rare class as network packets that have classification time $\geq T$, a threshold specified by users. GANs generate the bits of 5 fields: source/destination IP, source/destination port, and protocol. The search space is $1.0 \times 10^{31}$. As before, $p$ is a uniform distribution over the entire search space. To avoid harming network users, we ran all measurements on our own infrastructure rather than active switches. More details are in App. C (Lin et al. 2022).

Use case 3: inspecting rare images. Although RareGAN is primarily designed for the above use cases, we also use images for visualizing the improvements. Following the settings of related work (Mariani et al. 2018), we simulate the imbalanced dataset with widely-used datasets: MNIST (LeCun et al. 1998) and CIFAR10 (Krizhevsky 2009). For MNIST, we treat digit 0 as the rare class, and all other digits as the common class. For CIFAR10, we treat airplane as the rare class, and all other images as the common class. In both cases, the default class fraction is 10%. To simulate a smaller rare class, we randomly drop images from the rare class.

5.1 Evaluation Setup

Baselines. To demonstrate the effect of each design choice, we compare all intermediate versions of RareGAN: vanilla GAN (§ 4), ACGAN (§ 4), ACGAN trained with all unlabeled samples (§ 4.1), plus active learning (§ 4.2), plus weighted loss (§ 4.3). In all figures and tables, they are called: “GAN”, “ACGAN”, “RareGAN (no AL)”, “RareGAN” annotated with 1.0, and “RareGAN” annotated with weight (i.e., $1.0 > p$). All the above baselines and RareGAN use the same network architectures. For the first two applications, the generators and discriminators are MLPs. The GAN loss is Wasserstein distance (Eq. (2)), as it is known to be more stable than Jensen-Shannon divergence on categorical variables (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017). For the image datasets, we follow the popular public ACGAN implementation (Lee 2018), where the generator and discriminator are CNNs, and the GAN loss is Jensen-Shannon divergence (Eq. (1)).

We also evaluate representative prior work on three directions: (1) GANs with active learning: ALCG (using only labeled samples in training and using the most certain samples for labeling) (Xie and Huang 2019); (2) GANs for imbalanced datasets: BAGAN (Mariani et al. 2018); (3) Unsupervised/Self-supervised GANs: Elastic-InfoGAN (Ojha et al. 2019). We only evaluate the last two on MNIST, as they only released codes for that. As we cannot directly apply the last two in our problem (§ 3.2); we make minimal modifications to make them suitable (App. C (Lin et al. 2022)).

Metrics. We aim to minimize the distance between real and generated rare class $d \left( \hat{p}_r, p_r \right)$. In practice, generative models are often evaluated along two axes: fidelity and diversity (Naeem et al. 2020). Because the data types differ across applications, we use different ways to quantify them:

Network packets (use cases 1, 2). (1) Fidelity. Network packets are a high-dimensional list of categorical variables. We lack the ground truth $p_r$ to estimate fidelity. Instead, we estimate the true distribution of “scores” in $p_r$ (i.e., $\frac{\text{size of response}}{\text{size of request}}$ in DNS amplification attacks, and classification time in packet classifiers). This surrogate distribution is operationally meaningful, e.g., for quantifying mean or maximum security risk. We define $h_r$ as the ground truth distribution of this number over the rare class (estimated by drawing random samples from the entire search space, computing their scores, and then filtering out the scores that belong to the rare class), and $\hat{h}_r$ as its corresponding generated distribution. We use $d_{W1}(h_r, \hat{h}_r)$ as the fidelity metric, where $d_{W1}(\cdot, \cdot)$ denotes Wasserstein-1 distance, as it has a simple, interpretable geometric meaning (integrated absolute error between the 2 CDFs). (2) Diversity. When GANs overfit, many generated packets are duplicates. Therefore, we count the fraction of unique rare packets (i.e., those with a threshold score $\geq T$) in a set of 500,000 generated samples as the diversity metric.

Images (use case 3). (1) Fidelity. We use widely-used Fréchet Inception Distance (FID) (Heusel et al. 2017) between generated data and real rare data to measure fidelity. (2) Diversity. The previous diversity metric is not applicable here, as duplicate images are very rare. Instead, we take a widely used heuristic (Wang, Zhang, and Van De Weijer 2016) to check if GAN overfits to the training data: for each generated image, we find its nearest neighbor (in L2 pixel distance) in the training dataset. We then compute the average of nearest distances among a set of generated samples as the diversity metric.

1These experiments did not involve collecting any sensitive data. Such “penetration testing” of services is common practice in the security literature and we followed best practices (Matwyshyn et al. 2010). Two leading guidelines are responsible disclosure and avoid unintentional harm. We avoided harming the public Internet by running our experiments in sandboxed environments. Since we only reproduced synthetic (known) attack modes (Moon et al. 2021), we did not need to disclose new vulnerabilities.
Figure 2: Generated samples (no cherry-picking) on CIFAR10 ‘airplanes’ with $B = 10,000$ and $\alpha = 10\%$. Each baseline’s upper row is generated samples; lower row is the closest real sample.

5.2 Results

Unless otherwise specified, the default configurations are: the number of stages $S = 2$ (for RareGAN and ALCG), weight $w = 3$ (for RareGAN); in DNS, labeling budget $B = 200,000$, rare class fraction $\alpha = 0.776\%$ (corresponding to $T = 10$); in packet classification, $B = 200,000$, $\alpha = 1.150\%$ (corresponding to $T = 0.055$); in MNIST, $B = 5,000$, $\alpha = 1\%$; in CIFAR10, $B = 10,000$, $\alpha = 10\%$. Note that the choice of these default configurations do not influence the ranking of different algorithms too much, as we will show in the studies later. All experiments are run over 5 random seeds. More details are in App. C (Lin et al. 2022).

Robustness across applications. We start with a qualitative comparison between baselines. We show randomly-generated samples on MNIST and CIFAR10. GAN produces high-quality MNIST images in Fig. 1 by memorizing the labeled rare data (9 samples). Other baselines do not memorize, but either produce mode-collapsed samples (e.g., BAGAN (Mariani et al. 2018)), low-quality and mode-collapsed samples (e.g., ACGAN (Odena, Olah, and Shlens 2017), ALCG (Xie and Huang 2019)), or samples from wrong classes (e.g., Elastic-InfoGAN (Ojha et al. 2019)). RareGAN produces samples that are of the same quality as GANs, but with better diversity. On CIFAR10, Fig. 2 shows for each baseline randomly-generated samples (top row) and the closest real samples (bottom row). Again, GAN memorizes the training data, ACGAN and ALCG have poor image quality, and RareGAN trades off between the two (its sample quality is slightly worse than GAN, but its diversity is much better).

Quantitatively, Fig. 3 plots the fidelity-diversity tradeoff of each baseline on our datasets. Lower fidelity scores (left) and higher diversity scores (upwards) are better. The main takeaway of Fig. 3 is that RareGAN has the best tradeoff in our experiments. We discuss each method. (a) GAN. In all cases, GANs have poor diversity due to memorization. In network applications, GAN fidelity is good due to overfitting. In the image datasets, FID is bad, as FID scores capture overfitting. (b) ALCG and BAGAN. Generally, ALCG and BAGAN have much worse fidelity than other methods, consistent to the qualitative results. (c) Elastic-InfoGAN. Elastic-InfoGAN has higher diversity but much worse fidelity (Fig. 3c). Note that the higher diversity metric here is an artifact of Elastic-InfoGAN incorrectly generating digits that are mostly not 0 (Fig. 1), as it is not able to learn the boundary between rare and common classes well. (d) ACGAN. ACGAN has better diversity metrics and less overfitting than GAN, at the cost of sample fidelity. (e) Using unlabeled data. Comparing “RareGAN (no AL)” with “ACGAN”, we see that unlabeled data significantly helps the image datasets, but not the network datasets. This may be because of problem dimensionality: the dimension of the images are much larger than the other two cases, so additional data gives a more prominent benefit. (f) Active learning and weighted loss. Comparing RareGAN (3.0 and 1.0) with RareGAN (no AL), we see that weighted loss benefits the network packet datasets, but
Distance to the nearest training point (Diversity) 2 4 8 14 (App. E (Lin et al. 2022)). As we increase the weight from $w$ to 0.2%, 0.5%, 1.0%, and 2.0%, the performance of RareGAN improves upon ACGAN and ALCG for almost all S and w.

**Ablation study on RareGAN components.** RareGAN has three parts: (1) using unlabeled samples ($\S$ 4.1), (2) active learning ($\S$ 4.2), and (3) weighted loss ($\S$ 4.3). Active learning only makes sense with unlabeled samples, so there are 6 possible combinations. Fig. 15 (App. F (Lin et al. 2022)) shows each variant on DNS. Including all components, RareGAN yields the best diversity-fidelity tradeoff and low variance.

**Comparison to domain-specific techniques.** We compare to AmpMAP, the state-of-the-art work on (DNS) amplification attacks in the security community (Moon et al. 2021), in Table 1. AmpMAP finds high amplification packets by drawing random packets and requesting their amplification factors, and then doing random field perturbation on high amplification packets. AmpMAP uses amplification threshold 10, and the same packet space as ours. Note that AmpMAP is specifically designed for amplification attacks, not applicable for other applications we did. Even in that case, our proposed RareGAN still achieves much better fidelity and diversity with a fraction of the budget.

### 6 Conclusions

We propose RareGAN for generating samples from a rare class subject to a limited labeling budget. We show that RareGAN has good, stable diversity and fidelity in experiments covering different loss functions (e.g., Jensen-Shannon divergence (Goodfellow et al. 2014), Wasserstein distance (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017)), architectures (e.g., CNN, MLP), data types (e.g., network packets, images), budgets, and rare class fractions.

**Acknowledgements**

We thank Yucheng Yin for the help on baseline comparison, and Sekar Kulandaivel, Wenyu Wang, Bryan Phee, and Shruti Datta Gupta for their help with earlier versions of RareGAN. This work was supported in part by faculty research awards from Google, JP Morgan Chase, and the Sloan Foundation, as well as gift grants from Cisco and Siemens AG. This research was sponsored in part by National Science Foundation Convergence Accelerator award 2040675 and the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Zinan Lin acknowledges the support of the Siemens FutureMakers Fellowship, the CMU Presidential Fellowship, and the Cylab Presidential Fellowship. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al. 2014), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system (Nystrom et al. 2015), which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).
References


Lin, Z.; Liang, H.; Fanti, G.; and Sekar, V. 2022. RareGAN: Generating Samples for Rare Classes. arXiv:2203.10674.

Proceedings of the 18th ACM Workshop on Hot Topics in Networks, 43–51.


Naeem, M. F.; Oh, S. J.; Uhl, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In ICML, 7176–7185. PMLR.


Yang, H.; and Zhou, Y. 2021. IDA-GAN: A Novel Imbalanced Data Augmentation GAN. In ICPR. IEEE.

