Paper

限られたデータで尤度ベースの生成AIを学習させるための段階的なデータ拡張

見村 優太（正会員）

あらまし　Generativeモデルは現実的な画像を作成することに優れているが、学習のために広範なデータセットに依存するため、特にデータ収集にコストがかかる、あるいは困難な領域では大きな課題がある。現在のデータ効率の良い手法は、主にGenerative Adversarial Network (GAN)アーキテクチャに焦点を当てており、他のタイプの生成モデルの学習にはギャップが残っている。本研究では、本質的なデータ分布を変更することなく、限られたデータシナリオで学習を最適化することで、このギャップを解決する新しい手法として、「　段階的データ増強」を導入する。学習フェーズを通して増強強度を制限することで、我々の手法は限られたデータから学習するモデルの能力を向上させ、忠実度を維持する。PixelCNNとVector Quantized Variational AutoEncoder 2 (VQ-VAE-2)を統合したモデルに適用した結果、我々のアプローチは、多様なデータセットにおける定量的・定性的評価の両方において優れた性能を示した。これは尤度ベースモデルの効率的な学習における重要な前進であり、データ補強技術の有用性をGANだけにとどまらない。

キーワード: 限られたデータでの学習、PixelCNNs、 VQ-VAE-2、データ拡張、生成AI

1. まえがき

生成モデルは、説得力のある画像を生成する能力で有名であり、従来は最適な学習のために大規模なデータセットに依存していた。しかし、特に「医療画像など多くの領域でデータ収集が非常に高価なプロセスであることは広く知られている」[1]ことを考えると、領域固有のデータを大量に蓄積するという課題は依然として大きな障壁となっている。この限界に対処することは、限られたデータセットで生成モデルを効率的に学習するための簡単なアプローチを紹介する、我々の研究の基礎を形成する。

ディープラーニングのより広い分野では、特に大規模なデータセットが利用できない場合に、転移学習が強力なツールとして登場している。転移学習は、関連するソースタスクのデータを用いてターゲットドメインのパフォーマンスを向上させることができるが、「関連性の低いソースから知識を転移させると、ターゲットのパフォーマンスを逆に低下させる可能性があり、これは負の転移として知られる現象である」[2] 。同時に、データ増強はデータセットを人為的に拡張する戦略として受け入れられているが、時折、元のデータ分布を歪めてしまう。

Received March 17, 2024; Revised June 15, 2024; Accepted July 17, 2024
† Department of Cosmosciences, Graduate School of Science, Hokkaido University
(Sapporo, Japan)

さらに、既存のデータ効率の良い学習手法[1)3)4)]の多くは、主にGenerative Adversarial Networks (GANs)ドメインに対応しており、他の生成モデルは比較的未開拓のままである。

提案手法は「フェーズド・データ補強」と呼ばれ、この方向への最初のステップを提供する。モデルの学習フェーズに合わせて、データ増強の強度を低減することができる。初期状態では、データセットの有効サイズを増幅し、モデルが一般的なパターンを把握するのを助ける。学習が進むにつれて、モデルが元の学習データに内在する顕著な特徴に焦点を当てるように、補強パラメータが強化される。標準的なデータ補強に基づくこの方法論は、GAN以外の生成モデルにも適用可能である。

本論文では、PixelCNNとVector Quantized Variational AutoEncoder 2 (VQ-VAE-2)を統合したモデル(PCVQ2[5])と呼ぶ)に本手法を適用する。PC-VQ2はGANのアーキテクチャを持たない尤度ベースの生成モデルである。PixelCNNは、多くの対応する画像よりもシャープで多様な画像を生成する有望な可能性を持っている。VQ-VAE-2との統合により、画像の精度をさらに向上させることができ、その階層構造から得られる利点がある。

我々の実験から得られた経験的証拠は、定量的評価と定性的評価の両方において、標準的なデータ増強技術に対する我々のアプローチの優位性を一貫して示しており、

Paper

Phased Data Augmentation for Training a Likelihood-Based Generative Model with Limited Data

Yuta Mimura (member)†

Abstract Generative models excel in creating realistic images, yet their dependency on extensive datasets for training presents significant challenges, especially in domains where data collection is costly or challenging. Current data-efficient methods largely focus on Generative Adversarial Network (GAN) architectures, leaving a gap in training other types of generative models. Our study introduces "phased data augmentation" as a novel technique that addresses this gap by optimizing training in limited data scenarios without altering the inherent data distribution. By limiting the augmentation intensity throughout the learning phases, our method enhances the model's ability to learn from limited data, thus maintaining fidelity. Applied to a model integrating PixelCNNs with Vector Quantized Variational AutoEncoder 2 (VQ-VAE-2), our approach demonstrates superior performance in both quantitative and qualitative evaluations across diverse datasets. This represents an important step forward in the efficient training of likelihood-based models, extending the usefulness of data augmentation techniques beyond just GANs.

Keywords: training with limited data, PixelCNNs, VQ-VAE-2, data augmentation, generative models

1. Introduction

Generative models, renowned for their ability to generate compelling images, traditionally rely on large datasets for optimal training. However, the challenge of amassing substantial, domain-specific data remains a significant barrier, especially given that "it is widely known that data collection is an extremely expensive process in many domains, e.g. medical images"[1]. Addressing this limitation forms the cornerstone of our study, wherein we introduce a straightforward approach for training generative models efficiently on limited datasets.

In the broader landscape of deep learning, transfer learning has emerged as a potent tool, particularly when large datasets are unavailable. While transfer learning can enhance performance in a target domain using data from a related source task, "when transferring knowledge from a less related source, it may inversely hurt the target performance, a phenomenon known as negative transfer"[2]. Concurrently, while data augmentation has been embraced as a strategy to expand datasets artificially, it occasionally distorts the original data distribution. Furthermore, most existing data-efficient training methods[1)3)4)] primarily cater to the Generative Adversarial Networks (GANs) domain, leaving other generative models relatively unexplored.

Our proposed method, termed "phased data augmentation", provides the first step in this direction. It reduces the intensity of data augmentation in line with the model's learning phases. Initially, it amplifies the dataset's effective size, aiding the model in grasping the general patterns. As training advances, the augmentation parameters are tightened to ensure the model focuses on salient features intrinsic to the original training data. This methodology, based on standard data augmentation, is applicable to generative models other than GANs.

This paper applies our method to a model that integrates PixelCNNs with Vector Quantized Variational AutoEncoder 2 (VQ-VAE-2), which we refer to as PC-VQ2[5]. PC-VQ2 is a likelihood-based generative model without GANs' architecture. PixelCNNs have promising potential to generate images that are both sharper and more varied than many of their counterparts. Their integration with VQ-VAE-2 can further enhance image precision, a benefit derived from its hierarchical structure.

Empirical evidence from our experiments consistently demonstrates the superiority of our approach over standard data augmentation technique in both quantitative and qualitative assessments, underscoring its potential

Received March 17, 2024; Revised June 15, 2024; Accepted July 17, 2024

† Department of Cosmosciences, Graduate School of Science, Hokkaido University
(Sapporo, Japan)

限られたデータセットで尤度ベースのモデルを訓練するための効果的な戦略としての可能性を強調している。様々なデータドメインとサンプリングされたデータセットで検証されたこの有効性の頑健性は、限られたデータリソースのコンテキストであっても、従来のデータ増強よりもこの手法の一貫した性能向上を強調している。

本論文の残りの部分は以下のように構成されている：セクション2では、PC-VQ2を生成モデルのより広い範囲に位置づけ、実験に利用したPC-VQ2の利点を明らかにする。セクション3では、本研究で検討するモデルの概要を説明する。セクション4では、我々の提案する戦略と前述のモデルへの適用について掘り下げる。実験結果はセクション5で詳述する。セクション6では、データ補強に関連する研究について述べる。このセクションでは、フェーズド・データ補強の導入につながる包括的な背景を読者に提供することを目的とする。最後に、セクション7では、我々の発見と貢献の簡潔な要約と結論を述べる。

2. 生成モデルにおけるPC-VQ2の位置づけ

生成モデルの広大な領域を考慮し、本稿では、いくつかの説得力のある属性に支えられた、PixelCNNとVQ-VAE-2[5]の統合に焦点を当てる。

PixelCNNは、GAN[6][7]のような著名な生成モデルと比較して、より多様な画像セットを生成する能力を持っている。この多様性の能力はPixelCNNに限ったことではなく、他の尤度ベースのモデルでも見られる。ここで注目すべきは拡散モデル(Diffusion Models: DMs)[8] であり、最近採用が増加し、この有利な特徴を共有している。
しかし、PixelCNNを際立たせているのは、DMを含む他の尤度ベースモデルと比較して、多様であるだけでなく、より鮮明な画像を生成するユニークな能力である。このエッジはPixelCNNのピクセル単位の学習アプローチによるもので、他の尤度ベースモデルで見られるより一般的な画像単位の学習とは対照的である。高忠実度画像を生成する場合、この区別は特に重要である。
VQ-VAE-2の統合は、その階層構造と離散的な潜在空間により、シャープネスをさらに高める。これらの空間内での符号化は、PixelCNNモデルが自己回帰的に各画像の画素を生成する場合、計算効率を大幅に向上させる。

3. PC-VQ2で利用されるPixelCNNとVQVAE-2の概要

本節では、PC-VQ2フレームワークで利用されるPixelCNNとVQ-VAE-2の包括的な概要を説明する。

VQ-VAE-2isanAutoエンコーダとベクトル量子化潜在空間、エンコーダE()とデコーダD()から構成される。連続ベクトルを符号化するとき、これらは最も近い量子化されたベクトルeにマッピングされる：

量子化(E(x)) = e k ここで、k = arg min ¦E(x)-e j ¦(l) j |(1)

[5]．VQ-VAEの原論文[9]では、事後カテゴリ分布は決定論的であると考えられ、z = kに対する単純な一様事前分布によって一貫したKLダイバージェンス正則化項を導く。モデルは尤度を最大化し、量子化のためにマッチングを調整するように学習される。損失関数は以下の通りである：

$$\mathcal{L}(\mathbf{x}, D(\mathbf{e})) = ||\mathbf{x}-D(\mathbf{e})||_2^2+||\mathrm{sg}[E(\mathbf{x})]-\mathbf{e}||_2^2 +\beta||\mathrm{sg}[\mathbf{e}]-E(\mathbf{x})||_2^2 \quad (2)$$

ここで、sgは「停止勾配」演算を表す。最後の項は、エンコーダの出力が選択された量子化ベクトルに向かって引き寄せられることを促す。さらに、VQ-VAE-2は、2つの異なるサイズの量子化潜在空間を持つ階層構造を利用する:大きい「ボトムレベル」潜在マップは、符号化された連続ベクトルと小さい「トップレベル」潜在マップの離散化された出力を受け取る。
PixelCNNは画像xのピクセルの結合分布を、条件付き分布の後続の積としてモデル化する(x iは個々のピクセルを表す)。

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \quad (3)$$

[10]．画素列はラスタースキャン順序に忠実である。依存関係はマスクされた畳み込みフィルタを用いてモデル化され、RNN構造よりも学習効率が優れている。学習目的は尤度を最大化することであり、損失関数はクロスエントロピーに代表される再構成損失である。
図1に示すように、PC-VQ2アーキテクチャは、トップレベルにグローバルな画像情報を割り当てるが、ボトムレベルはローカルな詳細情報に焦点を当てる。サンプリング画像には、付録Aの図A.1に示すように、潜在マップに対してPixelCNNを採用し、PC-VQ2サンプリング処理の概要を説明する。

各潜在マップについて、VQ-VAE2論文で言及された特徴を持つ特定のPixelCNNモデルを選択した。

as an effective strategy for training a likelihood-based model with limited datasets. The robustness of this efficacy, validated across various data domains and sampled datasets, underscores the method's consistent performance improvements over the traditional data augmentation, even in the context of limited data resources.

The remainder of this paper is organized as follows: Section 2 situates PC-VQ2 within the broader landscape of generative models, elucidating the merits of PC-VQ2, utilized in our experiments. In Section 3, we provide an overview of the model under consideration in this study. Section 4 delves into our proposed strategy and its application to the aforementioned model. The experimental results are detailed in Section 5. Section 6 discusses related works pertaining to data augmentation. This section aims to furnish the reader with a comprehensive background leading to the introduction of phased data augmentation. Finally, Section 7 offers a concise summary and conclusion of our findings and contributions.

2. The Position of PC-VQ2 in Generative Models

Given the expansive realm of generative models, our focus in this paper is on the integration of PixelCNNs with VQ-VAE-2[5], underpinned by several compelling attributes.

PixelCNNs have the capability to generate a more diverse set of images compared to prominent generative models, such as GANs[6][7]. This capacity for diversity is not exclusive to PixelCNNs but is also seen in other likelihood-based models. A notable mention here is Diffusion Models (DMs)[8], which have recently witnessed a rise in adoption and share this advantageous trait.

What sets PixelCNNs apart, however, is their unique capacity to produce images that are not only diverse but also sharper compared to other likelihood-based models, including DMs. This edge is ascribed to PixelCNN's pixel-wise learning approach, in stark contrast to the more common image-wise learning seen in other likelihood-based models. When producing high-fidelity images, this distinction is particularly significant.

The integration of VQ-VAE-2 further enhances the sharpness, due to its hierarchical structure and discrete latent spaces. The encoding within these spaces significantly improves the computational efficiency, when PixelCNN models generate pixels for each image in an autoregressive manner.

3. Overview of the PixelCNNs and VQ-VAE-2 Utilized in PC-VQ2

This section provides a comprehensive overview of PixelCNNs and VQ-VAE-2 as utilized in the PC-VQ2 framework.

VQ-VAE-2 is an AutoEncoder with vector-quantized latent spaces, comprising an encoder $E()$ and decoder $D()$. When encoding continuous vectors, these are mapped to their nearest quantized ones $\mathbf{e}$:

$$\text{Quantize}(E(\mathbf{x})) = \mathbf{e}_k \text{ where } k = \arg\min_j ||E(\mathbf{x}) - \mathbf{e}_j|| \quad (1)$$

[5]. In the original VQ-VAE paper[9], the posterior categorical distribution was considered deterministic, leading to a consistent KL divergence regularization term through a simple uniform prior over z = k. The model is trained to both maximize the likelihood and tune the matching for quantization. The loss function is as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, D(\mathbf{e})) \;=\; & ||\mathbf{x} - D(\mathbf{e})||_2^2 + ||\text{sg}[E(\mathbf{x})] - \mathbf{e}||_2^2 \\ & + \beta ||\text{sg}[\mathbf{e}] - E(\mathbf{x})||_2^2 \end{aligned} \quad (2)$$

[5]. Here, sg denotes the "stop gradient" operation. The last term encourages the encoder's output to gravitate toward the selected quantized vector. Additionally, VQ-VAE-2 utilizes a hierarchical structure with two different sizes of quantized latent spaces: the larger "bottom-level" latent map receives the encoded continuous vectors and the discretized output of the smaller "top-level" latent map.

PixelCNNs model the joint distribution of pixels in an image $\mathbf{x}$ as the subsequent product of conditional distributions, where x_i denotes an individual pixel:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}) \quad (3)$$

[10]. The pixel sequence adheres to raster scan ordering. The dependencies are modeled using masked convolution filters, offering a training efficiency advantage over RNN structures. The training objective is to maximize the likelihood, with the loss function being a reconstruction loss typified by cross-entropy.

As depicted in Fig. 1, the PC-VQ2 architecture assigns global image information to the top-level, whereas the bottom-level focuses on local details. For sampling images, PixelCNNs are employed over the latent maps, as illustrated in Fig. A.1 of Appendix A, which provides an overview of the PC-VQ2 sampling process.

For each latent map, we chose specific PixelCNN

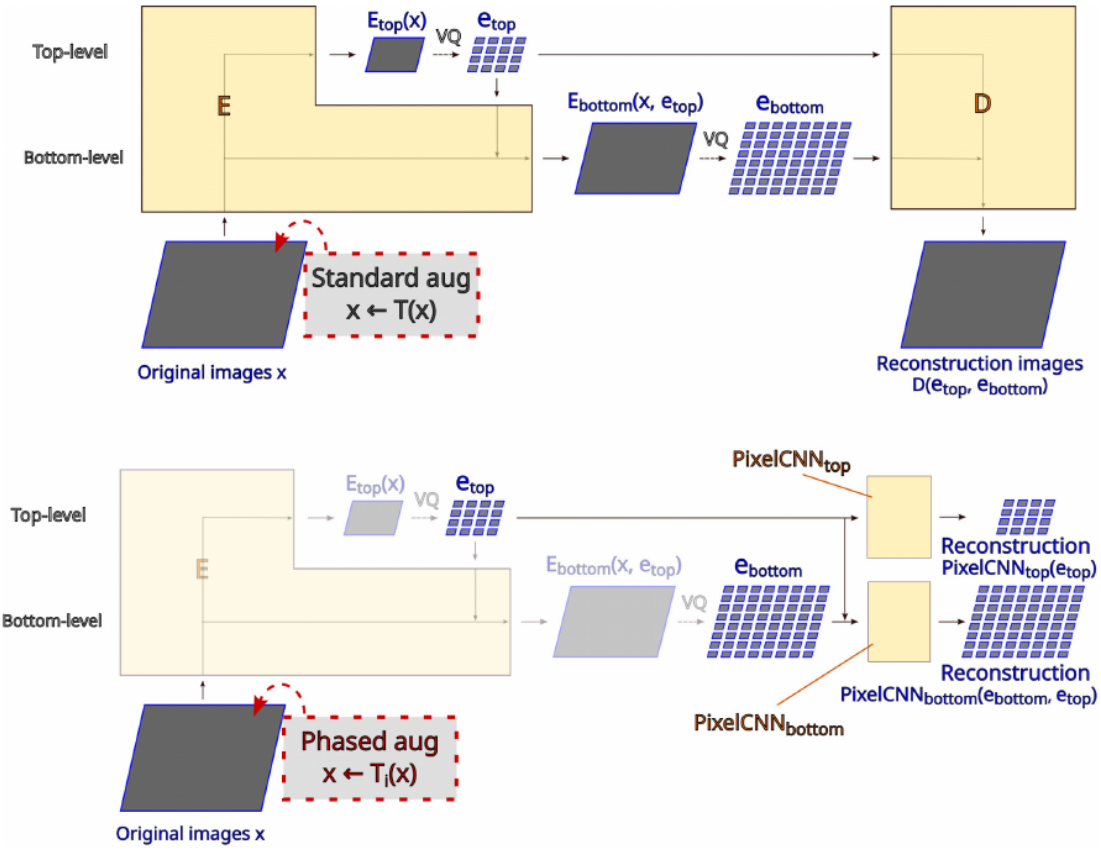


Fig. 1 PC-VQ2の学習構造の概要と、それに段階的補強を適用した。上図はVQ-VAE-2の学習、下図はPixelCNNの個別学習を示している。

ボトムレベルは、ゲート活性化ユニットとトップレベルとの条件付けを特徴とする条件付きGated PixelCNN[10]を利用する。一方、トップレベルはPixelSNAIL[11]を採用しており、これはゲート活性化ユニットを継承しながら注意層を統合する。VQ-VAE-2エンコーダデコーダモデル、条件付きGated PixelCNNモデル、PixelSNAILモデルを原著論文[5)10)11)]とオープンソース実装に基づいて実装した。

VQ-VAE-2で学習を進め、トップレベルのPixelSNAILとボトムレベルのGated PixelCNNで個別に学習を行う。

4. フェーズデータ補強

本節では、「フェーズドデータ補強」と呼ばれる我々の提案手法を紹介し、さらにPC-VQ2におけるその応用を説明する。

4.1 生成モデルにおけるオーバーフィッティング

生成モデルは、最適な性能を確保するために、通常数万枚の画像にまたがる大規模な学習データセットを必要とすることが多い。これらのモデルが限られたデータセットにしかアクセスできない場合、オーバーフィッティングが重大な懸念となる。

例えば、GANを標準的な、あるいは拡張を行わない制約付きデータセットで学習させる場合、一般的に高品質な画像を生成するのに苦労する[1)3)4)]。

PC-VQ2モデルでも同様の問題が生じる。PC-VQ2が限られたデータとデータ補強なしで学習された場合、ノイズに似た画像が主に生成される。標準的なデータ補強技術を取り入れた場合でも、PC-VQ2はしばしば不自然に歪んだように見える画像を生成する。標準的なデータ補強は、オーバーフィッティングをある程度緩和し、生成画像の全体的な品質を向上させることができるが、それでも出力はしばしば不自然に見える。これらの現象の詳細な比較は、「結果と考察」のセクション、特に図4と図5に記載されている。

4.2 フェーズド・データ補強の方法論

この問題に対処するために、我々は「フェーズドデータ増強」と呼ばれる、新規かつ素直な学習戦略を導入する。このアプローチの基本原理は以下のように記述される：

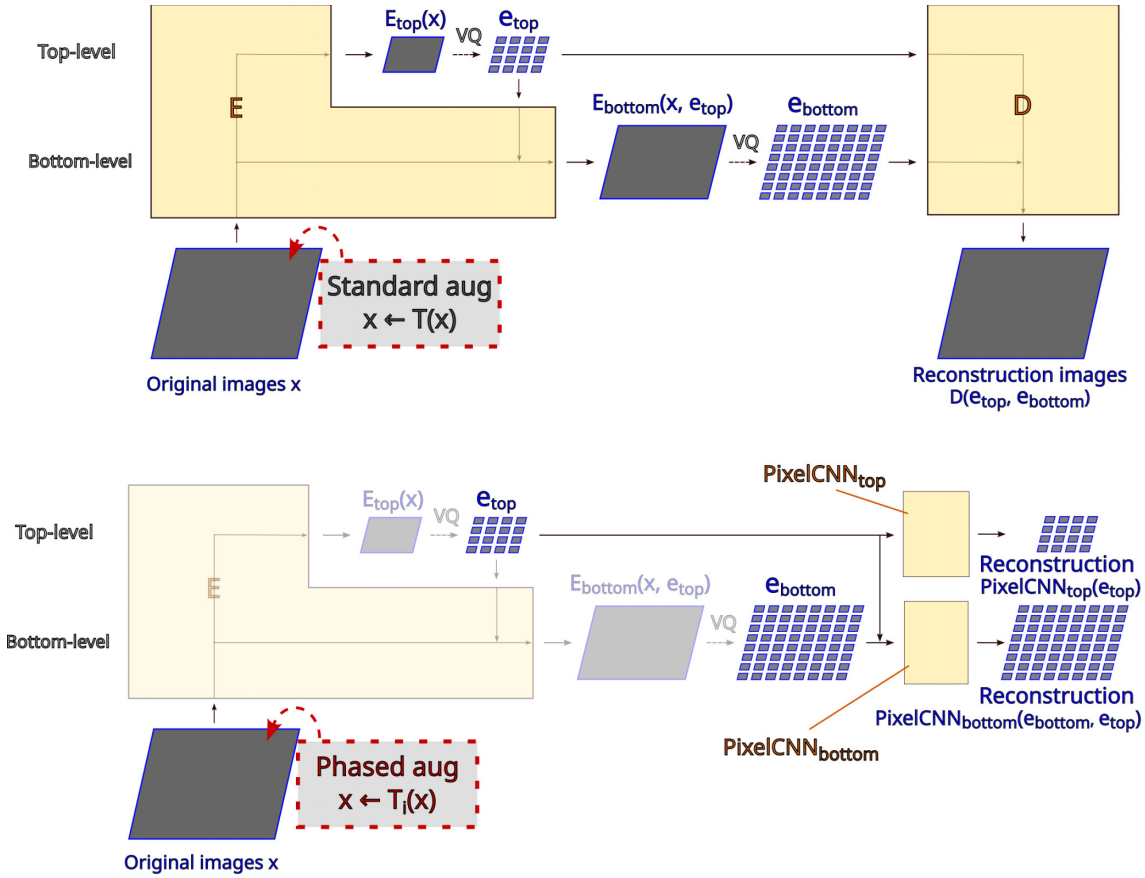


Fig. 1 Overview of a PC-VQ2 training structure and application of phased augmentation to it. The upper figure illustrates the training of VQ-VAE-2, and the lower figure illustrates the individual training of PixelCNNs.

models that have features mentioned in the VQ-VAE-2 paper. The bottom-level utilizes a conditional Gated PixelCNN[10] featuring gated activation units and conditioning with the top-level. In contrast, the top-level employs PixelSNAIL[11], which integrates attention layers while inheriting the gated activation units. We implemented the VQ-VAE-2 encoder-decoder model, the conditional Gated PixelCNN model, and the PixelSNAIL model on the basis of the original papers[5)10)11)] and open-source implementations.

Training proceeds initially with VQ-VAE-2, followed by individual training for the top-level PixelSNAIL and the bottom-level Gated PixelCNN.

4. Phased Data Augmentation

In this section, we introduce our proposed method, termed "phased data augmentation", and further explain its applications within PC-VQ2.

4.1 Overfitting in Generative Models

Generative models often necessitate extensive training datasets, commonly spanning tens of thousands of images, to ensure optimal performance. When these models have access only to a limited dataset, overfitting becomes a significant concern.

For instance, when GANs are trained on a constrained dataset using standard or no augmentation, they typically struggle to produce high-quality images [1)3)4)].

A similar issue arises with the PC-VQ2 model. When PC-VQ2 is trained with limited data and no data augmentation, it predominantly generates images that resemble noise. Even with the incorporation of standard data augmentation techniques, the PC-VQ2 often produces images that appear unnaturally distorted. Although standard data augmentation can mitigate overfitting to some extent and enhance the overall quality of the generated images, the output still often appears unnatural. A detailed comparison of these phenomena is provided in the "Results and Discussions" Section, specifically in Fig. 4 and 5.

4.2 Methodology of Phased Data Augmentation

To address the issue, we introduce a novel yet straightforward training strategy, termed "phased data augmentation". The fundamental principle of this approach is described as follows:

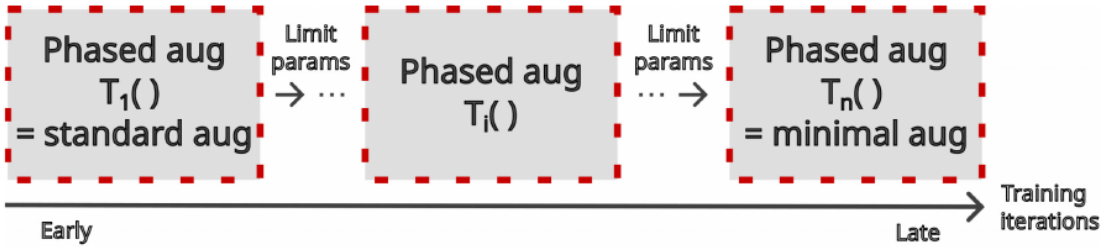


Fig. 2　段階的データ補強のグラフィカルな表現。

（）限られたデータセットでのラベル保存標準データ補強の全範囲から始める。

（）トレーニングが進むにつれて、これらの範囲を段階的に制限する。

3 ）標準的なオーグメンテーションから、学習セットの目標分布を維持する最小オーグメンテーションへのシフト。

図 2 は、この戦略をグラフ化したものである。

我々の段階的データ補強は、標準的なデータ補強技術を基礎としているため、GANの構造を必要とせず、標準的なデータ補強技術の代替として、我々の段階的データ補強を導入することで、GANだけでなく、幅広いモデルアーキテクチャに適用可能なアプローチを提供する。

本論文では、反転、回転、異方性ランダム整数アップスケーリングと一定範囲の切り出しからなるズームなどの基本変換と、明るさ、彩度、コントラスト操作などの色空間変換を用いた。フリッピングは、目標とするトレーニングセットの分布を維持する傾向があるため、すべてのフェーズで一貫して使用された。残りの3つの変換、回転変換、ズーム変換、色空間変換は、データセットの分布を元の状態に段階的に近づけるために、前述の順序で段階的に制限された。元の学習データの分布を変更することへの影響を考慮し、操作に制限を課す順序を決定した。具体的には、回転、ズーム、色空間変換は、データの分布に徐々に強い影響を与えるため、この特定の順序に制限を課した。

この転移学習に対する戦略の明確な点は、元の分布を反映した分布を持つデータセットで学習できることである。これは利点と見なすことができる。

先に強調したように、最初の学習段階では、データ増強により、モデルは増強された学習データに共通する包括的なパターンをゼロから把握することができる。学習が進み、オーグメンテーションが減少すると、モデルは元のデータセットの明確な属性に細かく適応するようになる。この観察は、PC-VQ2トレーニングの各フェーズで生成された画像を示す図3によって裏付けられている。

4.3 PC-VQ2への段階的補強の適用

このセクションでは、前述の方法論に基づいて、PC-VQ2モデルに適用される段階的データ増強の具体的なセットアップを説明する。経験的な観察に基づいて決定されたことはあるものの、可能な限り、理性的な判断に慎重に基づいたものであった。重要なことは、我々の実験では、同一のセットアップを複数のデータドメインとサンプルデータセットに適用し、標準的なデータ増強アプローチに対する明確で一貫した優位性を示したことである。これらの実験結果の詳細な分析については、「結果と考察」のセクションでさらに詳しく説明する。

初期段階では、以下のような補強が施された。

- フリッピング
- 回転させ、最大範囲を±180とした。　±180 degrees
- ズーム

– パラメータ値0.30で色空間変換を行う。

その後の段階を経て、トレーニングが進むにつれて

– 第2フェーズでは、回転の最大次数を±18に制限した。
- 第3段階では、回転は許されなかった。
- 第4フェーズでは、ズーム操作を除外した。

– 第5段階では、色変換パラメータを0.15に設定した。

– 最終段階では、色変換は適用されなかった。

初期設定について、開始の決定は以下の通りである。

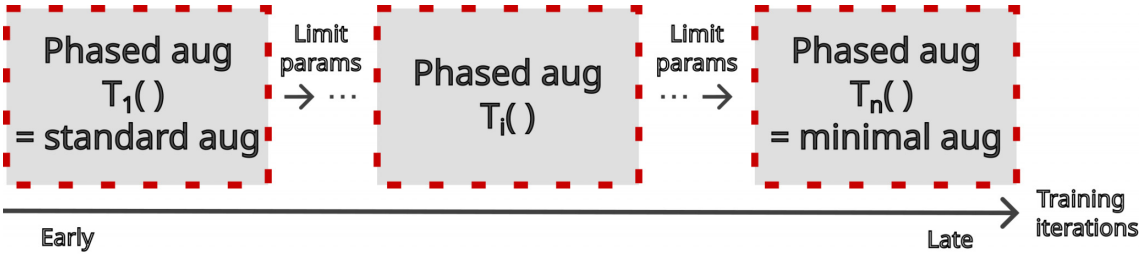


Fig. 2 Graphical representation of phased data augmentation.

(1) Begin with the full range of label-preserving standard data augmentations on a limited dataset.

(2) As training progresses, restrict these ranges in phases.

(3) Shift from standard augmentation to minimal augmentation that maintains a target distribution of the training set.

Figure 2 provides a graphical representation of this strategy.

Because our phased data augmentation builds upon the standard data augmentation technique, it does not require GANs' structures, offering an approach that is applicable across a wide range of model architectures beyond just GANs, by introducing our phased data augmentation as a replacement for standard data augmentation technique.

In this paper, we used basic transformations such as flipping, rotation, zooming that consisted of anisotropic random integer upscaling and constant range cropping, and color-space transformations such as brightness, saturation, and contrast manipulation. Flipping was consistently used throughout all phases because it tends to maintain a target training set distribution. The remaining three transformations, rotation, zooming, and color-space transformations, were gradually limited in phases in the order mentioned, to bring the dataset distribution incrementally close to its original state. We determined the sequence of imposing limitations on the operations by considering their impact on altering the distribution of the original training data. Specifically, because rotation, zooming, and color-space transformations progressively exert a stronger influence on the data's distribution, we imposed limitations in this specific order.

A distinct point of this strategy over transfer learning is its capacity to train on a dataset with a distribution that mirrors the original. This can be seen as an advantage.

As highlighted earlier, during the initial training stages, data augmentation enables a model to grasp the overarching patterns shared by augmented training data from scratch. As training advances and augmentation is reduced, the model becomes finely attuned to the distinct attributes of the original dataset. This observation is supported by Fig. 3, showcasing images generated at each phase in a PC-VQ2 training.

4.3 Application of Phased Augmentation to PC-VQ2

In this section, we delineate the specific setup of phased data augmentation applied to the PC-VQ2 model, based on the methodology described earlier. Although certain decisions were made based on empirical observations, they were, as much as possible, carefully grounded in reasoned judgment. Importantly, in our experiments, the identical setup was applied across multiple data domains and sample datasets, demonstrating clear and consistent superiority over standard data augmentation approach. Detailed analyses of these experimental outcomes are further elaborated in the "Results and Discussion" section.

During the initial phase, the following augmentations were applied:

- Flipping
- Rotation, allowing for a maximum range of ±180 degrees
- Zooming
- Color-space transformations with a parameter value of 0.30.

As the training progressed through subsequent phases:

- In the second phase, the rotation's maximum degree was restricted to ±18.
- In the third phase, no rotation was allowed.
- The fourth phase excluded the zooming operation.
- During the fifth phase, the color transformation parameter was set to 0.15.
- In the final phase, no color transformation was applied.

Regarding the initial settings, the decision to start

Fig. 3 ボトムレベルのPixelCNN学習により各フェーズで再構成された画像。Change'行の各要素は、前のフェーズに対して課された具体的な制限を示し、より詳細な説明はセクション4.3で行う。

データ補強の利点を最大限に発揮するために、360度までの回転数でズーム変換と色変換については、被写体の一部しか見えない過度にズームインされた画像や、被写体が識別できなくなる過度に明るくされた画像など、元の分布からデータを大きく逸脱させるような極端な変換を避けるよう注意した。

回転に関しては、拡張された訓練データの分布への影響を抑制するために、2つの段階で実装し、最初は2番目の段階での回転を全範囲の10分の1に制限した。1回の回転を行うと、訓練データの分布が大きく変化する可能性があるからである。色変換に関しては、最後のフェーズで限られたデータに対する増強の効果が減少することを考慮し、過剰な色変換情報を削除しながら学習内容を保持することで学習を促す2フェーズアプローチを選択した。オーグメンテーション効果を10分の1に減らすことは過度に制限的であることを考慮し、代わりに2分の1に減らすことにした。回転やズーム操作によるマージン導入を防ぐため、一定の整数アップスケーリングを利用した。回転処理では、アップスケーリング後、元の解像度を維持するために一貫したセンタークロッピングを使用した。具体的には、ズーム処理では異方性ランダム整数アップスケーリングを採用し、高さと幅を独立に1.05～1.30のファクターでスケーリングした。

アップスケーリング後、画像を最小アップスケーリング画素領域内に保つためにクロッピングを行った。

位相進行に関しては、各位相は10,000回の反復の後に遷移したが、第5位相から最終位相への遷移と最終位相の終了は例外で、最終位相の限られたデータに対する増強の効果が減少していることを考慮すると、どちらも5,000回の反復の後に起こる。

すべてのランダム値は一様分布からサンプリングされた。

コードはTensorflow 2.10.1とOpenCV 4.8.0を用いて実装した。
図1に示すように、最上位と最下位のPixelCNNsモデルを学習する際に、限られた学習データセットに対して段階的補強アプローチを利用する。VQ-VAE-2の学習では、PixelCNNsモデルの学習時に拡張されたデータセットが活用されることを考慮し、標準的な拡張手法が利用される。VQ-VAE-2は、各データドメインとサンプリングされたデータセットに対して個別に学習される。比較実験では、公平な比較を保証するために、手法のみが変化し、PixelCNNs 部分のみが再学習される場合に同じ VQ-VAE-2 が使用される。しかし、データ領域が変わると、PixelCNNとVQ-VAE-2の両方が再学習される。

5. 結果および考察

本節では、限られたデータでPC-VQ2を学習させた場合の、フェーズドデータ補強と他のデータ効率化手法との比較実験結果を示す。

PC-VQ2 を用いた実験では、

Fig. 3 Images reconstructed in each phase through a bottom-level PixelCNN training. Each element in the 'Change' row indicates the specific limitation imposed relative to the previous phase, with more detailed explanations provided in Section 4.3.

with rotations up to 360 degrees aimed to maximize the benefits of data augmentation. For zooming and color transformations, we exercised caution to avoid extreme transformations that could significantly deviate the data from its original distribution, such as overly zoomed-in images where only a part of the subject is visible or overly brightened images where the subject becomes indiscernible.

Concerning rotation, we implemented it in two stages to suppress the impact on the augmented training data's distribution, initially limiting the rotation in the second phase to one-tenth of the full range, because performing the rotation in one go could significantly alter the training data distribution. As for color transformations, considering the diminished effect of augmentation on limited data in the last phases, we opted for a two-phase approach to encourage learning by retaining learned content while removing excess color transformed information. Considering that reducing the augmentation effect to one-tenth would be excessively restrictive, we chose to reduce it to one-half instead. To prevent the introduction of margins due to the rotation and zooming operations, constant integer upscaling was utilized. In the rotation process, after upscaling, a consistent center cropping was used to maintain the original resolution. Specifically, the zooming process employed anisotropic random integer upscaling, wherein the height and width are independently scaled by a factor ranging between 1.05 and 1.30. After upscaling, cropping was performed to keep the image within the minimum upscaled pixel region.

In terms of phase progression, each phase transitioned after 10,000 iterations, with the exception of the transition from the fifth to the last phase and the conclusion of the final phase, both taking place after 5,000 iterations, considering the diminished effect of augmentation on limited data in the last phases.

All the random values were sampled from uniform distributions.

The codes were implemented using Tensorflow 2.10.1 and OpenCV 4.8.0.

The phased augmentation approach is utilized on the limited training dataset when training the top-level and bottom-level PixelCNNs models, depicted in Fig. 1. For VQ-VAE-2 training, the standard augmentation technique is utilized, given that the augmented dataset gets leveraged during the training of the PixelCNNs models. VQ-VAE-2 is trained separately for each data domain and sampled dataset. In the comparative experiments, to ensure a fair comparison, the same VQ-VAE-2 is used when only the technique changes, and only the PixelCNNs part is retrained. However, when the data domain changes, both PixelCNNs and VQ-VAE-2 are retrained.

5. Results and Discussions

This section represents the experimental results comparing phased data augmentation to the other data-efficient technique when training PC-VQ2 with limited data.

For the experiments involving PC-VQ2, standard

Table 1 学習済みPC-VQ2モデルに対するFIDの結果。

Method	FFHQ 0-99	10,000-10,099	20,000-20,099	AFHQ v2 Cat
標準的なデータ補強	263.21	240.66	252.87	259.24
段階的データ補強	**169.62**	**140.46**	**149.63**	**177.33**

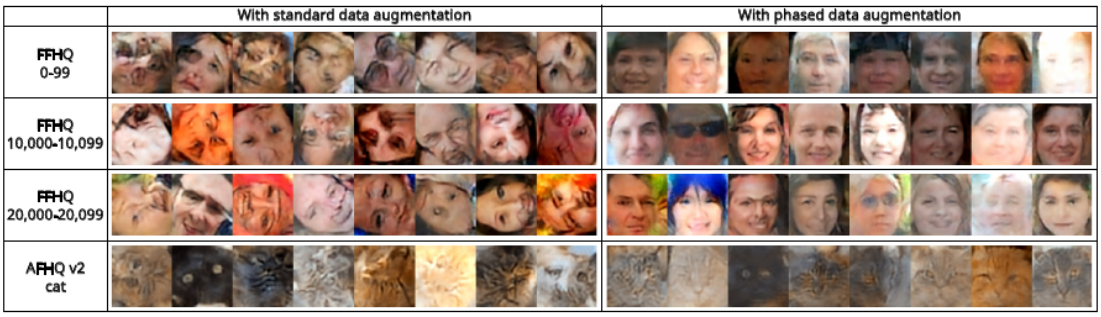


Fig. 4 学習したPC-VQ2モデルにより、人間の顔画像と猫の顔画像を生成した。

図5 PC-VQ2モデルによるデータ補強を行わない場合の生成画像。

標準的なデータ補強手法を採用した。この判断は、データ補強を行わない場合、PC-VQ2はノイズに似た無視できる品質の画像を生成するという観察に基づいている。

FIDスコアは、データ効率の良いGANに関する先行論文[3]で用いられているように、5,000枚の生成画像と100枚の実画像を用いて性能を測定するために採用された。生成された画像の数値には、最初の8枚のサンプリング画像を利用した。実験設定の詳細は、付録Bに記載されている。

FFHQデータセットは高品質な人間の顔データセット[12]であり、AFHQ v2 catデータセットは高品質な猫の顔データセット[13]–[15]である。FFHQデータセットから3つの異なるサブセット、具体的にはインデックス0-99、10,000-10,099、20,000-20,099を抽出した。AFHQ v2 catデータセットでは、100枚の画像をランダムに選択した。

表1にまとめたように、すべてのトレーニングデータセットにおいて、フェーズドアグメンテーションは、標準的なオーグメンテーション技術と比較して、有意に優れたFIDスコアを示した。各手法のスコアは、類似の値を中心にクラスタリングされている。この観察は、様々なデータドメインとサンプリングされたデータセットにおいて、標準的なデータ補強に対する提案手法の一貫した明確な優位性と相まって、限られたデータリソースのコンテキストにおいても、その有効性の頑健性を強調するものである。

段階的データ補強手法は、すべての評価データセットにおいて、標準的なデータ補強手法を一貫して上回り、81.91点から103.24点の範囲でFIDスコアの有意な減少を示した。これは、学習済みPC-VQ2モデルの品質向上におけるフェーズドデータ補強の優れた有効性を示している。

図4は、生成された画像を示し、定量的な知見を裏付けている。例えば、左から2列目を見ると、標準的な方法で生成された画像は、一方向に統一されていない複数の顔パーツを示し、歪んだ不自然な画像になる。一方、フェーズド法では、顔の特徴がよく一致した画像を生成するため、より自然で認識しやすい顔となる。図5は、オーグメンテーションを行わずに学習中に生成された画像がノイズに似ていることを示している。

生成された画像は若干のぼやけが見られる。その要因の一つは、低解像度の潜在空間16-8を用いることである。実際、図A.2に示すように、より高解像度の潜在空間32-16を利用すると、多少の歪みはあるものの、より鮮明な画像が得られる。しかし、この解像度でさえ、元の研究の構成には及ばないままである。その他の要因としては、オリジナルの512と比較してコードブックのサイズが256と小さいこと、3つの潜在空間(128-64-32)で1024×1024という高い解像度を利用したオリジナルのFFHQ実験と比較して、より小さな潜在空間構成を使用したことが挙げられる。これらの修正は、費用対効果を高めるために行われたが、観察されたぼやけにつながったと思われる。の大きな潜在空間構成32-16のもとで生成された画像は、以下のようになる。

Table 1 FID results over trained PC-VQ2 models.

Method	FFHQ 0-99	10,000-10,099	20,000-20,099	AFHQ v2 Cat
Standard data augmentation	263.21	240.66	252.87	259.24
Phased data augmentation	**169.62**	**140.46**	**149.63**	**177.33**

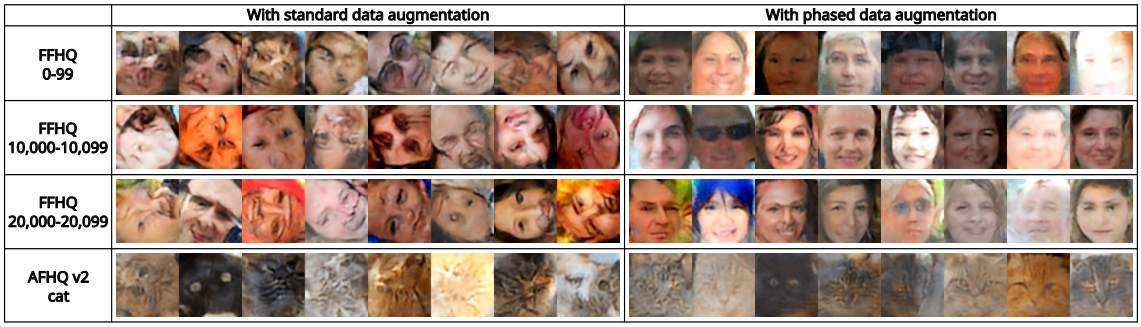


Fig. 4 Generated human-face and cat-face images by the trained PC-VQ2 models.

Fig. 5 Generated images by the trained PC-VQ2 model with no data augmentation.

data augmentation technique was adopted for the comparison. This decision was based on the observation that using no data augmentation, the PC-VQ2 generated images of negligible quality, resembling noise.

The FID score was employed to measure performance, utilizing 5,000 generated images and 100 real images, as used in a prior paper on data-efficient GANs [3]. We utilized the first eight sampled images for figures of the generated images. Details on our experimental settings, can be found in Appendices B.

We utilized the FFHQ dataset, which is the high-quality human-face dataset[12], and the AFHQ v2 cat dataset, which is the high-quality cat-face dataset [13]-[15] during the training. Three distinct subsets were drawn from the FFHQ dataset, specifically indices 0-99, 10,000-10,099, and 20,000-20,099. For the AFHQ v2 cat dataset, a set of 100 images was selected at random.

As summarized in Table 1, across all training datasets, phased augmentation demonstrated significantly better FID scores compared to standard augmentation technique. The scores for each method were clustered around similar values. This observation, coupled with the consistent and clear superiority of the proposed method over standard data augmentation across various data domains and sampled datasets, underscores the robustness of its efficacy, even in contexts of limited data resources. The phased data augmentation technique consistently outperforms the standard data augmentation method across all evaluated datasets, showing a significant reduction in FID scores ranging from 81.91 to 103.24 points. This demonstrates the superior efficacy of phased data augmentation in improving the quality of the trained PC-VQ2 models.

Figure 4 showcases the generated images, corroborating the quantitative findings. For instance, when examining the second column from the left, the images generated using the standard method exhibit multiple facial parts that are not unified in one direction, resulting in distorted and unnatural images. In contrast, the phased method produces images with well-aligned facial features, leading to a more natural and recognizable face. Figure 5 illustrates that the generated images during training with no augmentation resemble noise.

The generated images exhibit some blur. One of the factors is the employment of low-resolution latent spaces, 16-8. In fact, leveraging higher-resolution latent spaces, 32-16, results in sharper images, albeit with some distortion, as illustrated in Fig. A.2. However, even this resolution remains below the original study's configuration. Other factors contributing to the blurriness include a smaller codebook size of 256 compared to the original 512, and the use of a smaller latent space configuration compared to the original FFHQ experiments, which utilized higher resolutions of 1024 x 1024 with three latent spaces (128-64-32). These modifications were made to enhance cost-effectiveness but likely led to the observed blurriness. The images generated under the larger latent space configuration, 32-16, in

図A.2は、モデルが鮮明な画像を生成する可能性を示している。

6. 関連作品

このセクションでは、データ補強の複雑な領域を探求し、その様々な側面と、他のデータ効率の良い戦略との絡み合いについて議論する。さらに、生成モデルにおけるオーグメンテーションに基づくデータ効率の良い手法の関連性を明らかにし、我々の研究で提案された手法と明確に対比させる。

6.1 データ効率の良いアルゴリズムにおけるデータ補強

この小節では、データ効率の良いアルゴリズムの領域におけるデータ補強の役割を明らかにする。さらに、データ補強に関連する様々な概念と方法について議論し、段階的データ補強とそれに対応するものとの区別を強調する。

データ効率の良いアルゴリズムは、特に限られたデータに制約されたシナリオにおいて、ディープラーニングモデルにとって大きな利点として浮上している。による包括的な調査では、これらのアルゴリズムを、教師なしパラダイム、データ増強、知識共有、ハイブリッドシステムという明確なカテゴリーに分類している。知識共有にはいくつかの戦略がある。これらは、転移学習、マルチタスク学習、生涯学習、メタ学習を包含する。我々の研究はデータ補強を優先している。このアプローチは、ラベルを保持したまま画像を補強することに長けており、特定のドメイン内でデータが乏しい場合でも、しばしば有益である。

画像データ補強技術の広範なレビューを[17]に示す。カーネルフィルタ、幾何変換、色空間変換など、多様な手法がある。我々の研究は幾何学的変換と色空間変換に集中しており、その主な理由はこれらの手法がデータの元のラベルを保持するためである。目的がターゲット領域内の画像のボリュームを増加させることである場合、カーネルフィルタは最適な選択ではないかもしれない。従来は画像強調に用いられており、強調画像のみを利用する前処理段階に限定して用いるべきであることが示唆されている。

様々なデータ増強戦略を、本稿で紹介する段階的データ増強と対比することで進める。[18]は、増強されたデータで事前に訓練され、その後元のデータで微調整されたモデルが有益な情報を学習できることに注目する。

しかし、我々の提案する方法は、学習データ分布を1つのインスタンスではなく、段階的に変更することで差別化を図っている。特定の変換(例えば、回転)を徐々に制限または排除することで、拡張されたデータ分布は、元のデータ分布とより密接に整合する。したがって、段階的なデータ補強を行うことで、モデルは元の分布をより効果的に学習することができる。さらに、このアプローチは、変換関連情報の抽出を強化することができる。

データ補強パラメータの管理という文脈では、自動制御はAutoAugment[19]に類似していると考えることができる。しかし、我々の方法は、自動制御が一般的に追加の計算資源とデータを必要とすること、特に、検証データに基づく損失関数計算を必要とすることを認識し、パラメータ範囲の段階的な削減を採用している。

6.2 生成モデルにおけるデータ補強

このサブセクションでは、生成モデルの範囲内で、データ補強に根ざした従来のデータ効率の良い方法について議論し、本研究で提案した方法との比較分析を提供する。
データ効率の良い学習、特に制約のあるデータセットでの学習の分野では、最近、生成モデルに焦点を当てた研究が行われている。GANに関するいくつかの研究[1)3)4)]は、生成器の最適化中だけでなく、識別器の最適化プロセス全体を通して、識別器の入力にデータ補強を組み込むことを独立に提唱している。この戦略は、合成分布の望ましくないシフトを先取りすることを目的としており、GANに有効であることが証明されている。しかし、識別器を持たない生成モデルには適用できない。例えば、PixelCNNはこの手法を使うことができない。

また、[4)]で詳述した革新的なアプローチは、前述の戦略と組み合わせて、特定の確率の下でのデータ増強の実装を提案するものである。この技法は、GANのユニークなアーキテクチャに基づいて計算された関数を通してこの確率を適応的に調節することによってさらに改良され、GANに対する特異性を強化する。これらの方法とは逆に、我々の研究は、データ増強パラメータを特定のフェーズに制限し、徐々に減少する確率の使用を禁止することを選択する。この選択は、最小でありながらゼロでない確率であっても、変換パラメータの全スペクトルにさらされるデータを包含する可能性があるという理解からきている。

Fig. A.2, demonstrate the model's potential to generate sharpened images.

6. Related Works

This section explores the intricate domain of data augmentation, discussing its various aspects and how it intertwines with other data-efficient strategies. Furthermore, we clarify the connection between augmentation-based data-efficient methods in generative models, distinctly contrasting them with the method proposed in our research.

6.1 Data Augmentation in Data-Efficient Algorithms

This subsection delineates the role of data augmentation within the sphere of data-efficient algorithms. It further discusses various concepts and methods pertinent to data augmentation, highlighting the distinctions between phased data augmentation and its counterparts.

Data-efficient algorithms have emerged as a significant boon for Deep Learning models, particularly in scenarios constrained by limited data. A comprehensive survey by[16] classifies these algorithms into distinct categories: non-supervised paradigms, data augmentation, knowledge sharing, and hybrid systems. Knowledge sharing includes several strategies. These encompass transfer learning, multi-task learning, lifelong learning, and meta-learning. Our research prioritizes data augmentation. This approach is proficient in augmenting images while preserving their labels, often beneficial even when data is scarce within a specific domain.

An extensive review of image data augmentation techniques is presented in[17]. The methods are diverse, including kernel filters, geometric transformations, color space transformations, and more. Our study concentrates on geometric transformations and color space transformations, primarily because these methods maintain the original labels of the data. When the objective is to augment the volume of images within a targeted domain, kernel filters might not be the optimal choice. They are conventionally employed for image enhancement, implying their use should be confined to preprocessing stages where only enhanced images are utilized.

We proceed by contrasting various data augmentation strategies with the phased data augmentation introduced in this paper.[18] notes that models pre-trained on augmented data and subsequently fine-tuned with original data can learn beneficial information. However, our proposed method distinguishes itself by altering training data distributions in stages, rather than in a single instance. Through the gradual limitation or elimination of specific transformations (e.g., rotation), the augmented data distribution incrementally aligns more closely with the original. Therefore, phased data augmentation can enable models to learn the original distribution more effectively. Additionally, this approach can enhance the extraction of transformation-related information.

In the context of managing data augmentation parameters, one might consider automated control analogous to AutoAugment[19]. Yet, our method adopts a phased reduction in parameter ranges, acknowledging that automated control typically necessitates additional computational resources and data — specifically, it demands loss function computation based on validation data.

6.2 Data Augmentation in Generative Models

This subsection discusses traditional data-efficient methods rooted in data augmentation within the scope of generative models, providing a comparative analysis with the method proposed in our study.

In the sphere of data-efficient learning, particularly learning with constrained datasets, research focusing on generative models has recently been performed. Several studies on GANs[1)3)4)] have independently advocated for the incorporation of data augmentation into the inputs of the discriminator, not only during the optimization of the generator but also throughout the discriminator's optimization process. This strategy aims to preempt any undesirable shifts in the synthesis distribution and has proven effective for GANs. However, it is not applicable to generative models without a discriminator. For instance, PixelCNNs cannot use this technique.

Another innovative approach detailed in[4] suggests the implementation of data augmentation under a specific probability, coupled with the aforementioned strategy. This technique is further refined by adaptively modulating this probability through a function calculated based on the unique architecture of GANs, reinforcing its specificity to GANs. Contrary to these methods, our research opts to constrict the data augmentation parameters within certain phases, shunning the use of a gradually diminishing probability. This choice stems from the understanding that even a minimal yet non-zero probability could encompass data subjected to the full spectrum of transformation parameters.

生成モデルに関する論文[20]では、データ増強とデータ増強パラメータを持つモデルの条件付けを用いることが提案されている。しかし、この研究は、例えば90度回転のような、限られたデータ補強タイプの配列でしか実験しておらず、我々の論文で議論されている限られたデータと比較して、補強なしでより実質的なデータセットに依存していた。わずか1000枚の画像で生成モデルを学習させるというタスクには、より多様なデータ補強技術が必要である。

7. むすび

本研究は、GANアーキテクチャを主な対象とするデータ効率の良い学習手法の既存の状況の中で、代替生成モデルのためのデータ効率の良い学習を探求するための先駆的な一歩を示すものである。我々は、限られたデータセットで動作する尤度ベースの生成モデルに特化した、「段階的データ増強」と名付けられた新しい学習戦略を導入した。我々の実験結果は、フェーズド・データ補強の有効性を一貫して実証しており、従来の補強アプローチに対する明らかな優位性を示している。これは、様々なデータドメインとサンプリングされたデータセットで検証され、データリソースが限られたコンテキストでも、PC-VQ2モデルの一貫した性能向上と頑健性を示している。公平性に関しては、フェーズド補強と標準補強の最適なパラメータは、データセットによって異なる可能性があることに注意することが重要である。複数のデータ条件を選択し、各手法に固定パラメータセットを適用することで、どちらの手法も不当に有利にも不利にもならないことを保証する。このアプローチにより、両方の補強方法をバランスよく公平に評価することができる。

提案手法は、様々なデータ条件下で既存のロバスト手法と比較して十分な有効性を示すが、データセットごとに最適なパラメータが異なることは一般に事実である。したがって、提案手法に基づく更なるチューニングを行うことで、さらに良い結果が得られる可能性がある。

我々の実験では、フェーズド・データ補強プロセスのどの側面が性能向上に最も大きく寄与しているかを明確に特定することはできないが、これらの構成要素をさらに調査することで、貴重な洞察が得られる可能性がある。このような調査は、大規模なデータセットを扱う場合に特に重要である。小規模なデータセットで最も効果的なパラメータを特定することは、計算コストの削減に役立つからである。しかし、この焦点は、フェーズド・オーグメンテーション・プロセスの全体的な有効性を実証するという、我々の論文の主目的から若干逸脱している。

従来のデータ補強の原理から、我々の方法論は、従来の補強技術の広範な有用性を反映し、多様な生成モデルと転移学習コンテキストへの幅広い適用性を約束する。理論的には正しいが、他の尤度ベースモデルに対するこのアプローチの実用的な有効性は、さらなる実証的な調査が必要である。本研究がPCVQ2の能力を向上させるだけでなく、尤度ベースモデルのさらなる研究と応用を活性化し、限られたデータからの効果的な学習を促進することを期待している。

Acknowledgments

This study was supported by JST SPRING, Grant Number JPMJSP2119. We would like to thank my supervisor, Associate Professor Kazuhiko SUEHIRO at Hokkaido University for his advice on this paper and review of it throughout this study. We would like to thank Associate Professor Sho TAKAHASHI at Hokkaido University for his advice. For enhancing the clarity and accuracy of our paper, as well as for re-viewing the code utilized in our evaluation metrics, we employed the assistance of ChatGPT21).

付録

A. PC-VQ2生成構造の概要

本節では、PC-VQ2 の生成について概説する。図A.1にPC-VQ2生成プロセスの構造レイアウトを示す。

トップレベルとボトムレベルのそれぞれの離散潜在マップにおいて、左上のピクセルの値は、一様分布からランダムにサンプリングされた離散ベクトルインデックスを表す。各マップの後続ピクセルは、ラスター走査パターンに従って、それぞれの学習済みPC-VQモデルに基づいてサンプリングされる。ボトムレベルでの生成過程では、左上レベルで生成された離散的な潜在マップが、左上ピクセルのランダムにサンプリングされた値とともに、条件付けに利用される。

B. 実験設定の詳細について

本節では、PC-VQ2モデルのハイパーパラメータとその学習に関する実験セットアップを詳述する。PC-VQ2では、StyleGAN3の公式PyTorchコードに基づき、

A paper on generative models[20] has proposed using data augmentation and conditioning a model with the data augmentation parameters. However, the study only experimented with a limited array of data augmentation types, for instance, 90-degree rotations, and relied on a more substantial dataset without augmentation compared to the limited data discussed in our paper. The task of training a generative model with a mere 100 images necessitates a more diverse set of data augmentation techniques.

7. Conclusions

Within the existing landscape of data-efficient training methods, which predominantly cater to GAN architectures, this study represents a pioneering step towards exploring data-efficient training for alternative generative models. We have introduced a novel training strategy named "phased data augmentation," tailored specifically for a likelihood-based generative model operating with limited datasets. Our experimental findings consistently demonstrate the efficacy of phased data augmentation, showcasing its evident superiority over the traditional augmentation approach. This was validated across various data domains and sampled datasets, indicating consistent performance improvements and robustness for the PC-VQ2 model, even in contexts with limited data resources. Regarding fairness, it is important to note that the optimal parameters for phased and standard augmentations can vary depending on the dataset. By selecting multiple data conditions and applying fixed parameter sets for each method, we ensure that neither method is unfairly advantaged or disadvantaged. This approach allows for a balanced and fair evaluation of both augmentation methods.

While our proposed method demonstrates sufficient effectiveness compared to existing robust techniques under varying data conditions, it is generally true that the optimal parameters can differ for each dataset. Therefore, further tuning based on our proposed method could potentially yield even better results.

While our experiments do not explicitly identify which aspects of the phased data augmentation process contribute most significantly to performance improvements, investigating these components further could provide valuable insights. Such investigations are particularly important when working with large datasets, as identifying the most effective parameters on smaller datasets can help reduce computational costs. However, this focus deviates slightly from the main objective of our paper, which is to demonstrate the overall efficacy of the phased augmentation process.

Drawing from the principles of traditional data augmentation, our methodology promises broad applicability to a diverse array of generative models and transfer learning contexts, mirroring the wide-ranging utility of traditional augmentation techniques. While theoretically sound, the practical effectiveness of this approach for other likelihood-based models warrants further empirical investigation. It is our hope that this research not only advances the capabilities of the PC-VQ2 but also invigorates further study and application of likelihood-based models, facilitating their effective learning from limited data.

Acknowledgments

This study was supported by JST SPRING, Grant Number JPMJSP2119. We would like to thank my supervisor, Associate Professor Kazuhiko SUEHIRO at Hokkaido University for his advice on this paper and review of it throughout this study. We would like to thank Associate Professor Sho TAKAHASHI at Hokkaido University for his advice. For enhancing the clarity and accuracy of our paper, as well as for reviewing the code utilized in our evaluation metrics, we employed the assistance of ChatGPT[21].

Appendix

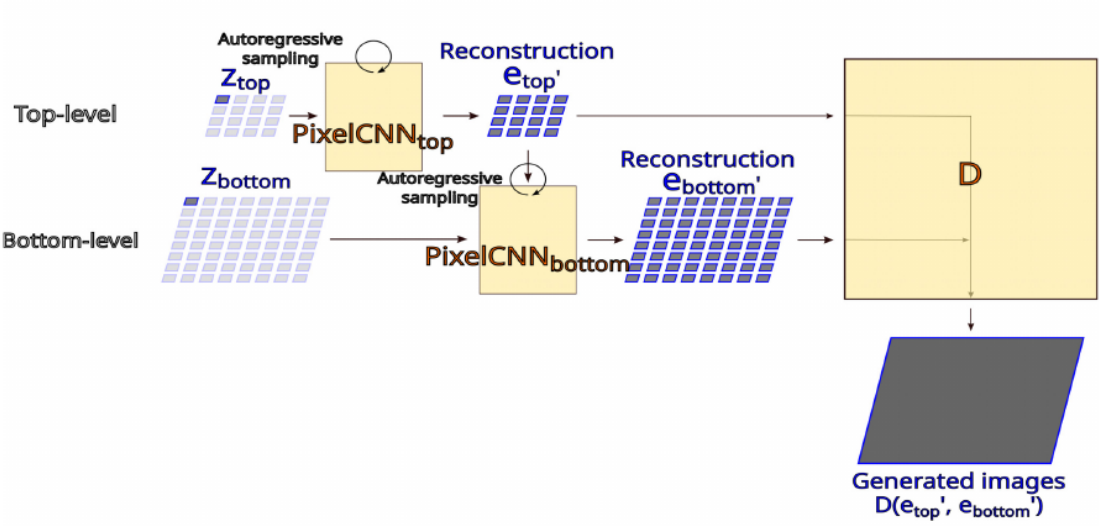
A. Overview of a PC-VQ2 Generation Structure

This section overviews a PC-VQ2 generation. Figure A.1 illustrates the structural layout of the PC-VQ2 generation process.

In each of the discrete latent maps for both the top-level and bottom-level, the value of the top-left pixel represents a discrete vector index, which is randomly sampled from a uniform distribution. Subsequent pixels in each map are sampled based on the respective trained PC-VQ models, following a raster scanning pattern. During the generation process at the bottom-level, the discrete latent maps generated at the top-level are utilized for conditioning, along with the randomly sampled value of the top-left pixel.

B. Details of the Experimental Settings

This section details the experiment setups regarding the hyperparameters of PC-VQ2 model and its training. We implemented FID metrics to evaluate in PC-VQ2,



app.Fig. 1　Overview of a PC-VQ2 generation structure.

FIDメトリクスを実装して評価した。
VQ-VAE-2エンコーダ・デコーダ、再構成モジュール、生成モジュールをDeepMindのソネットライブラリに基づいて実装した。さらに、VQ-VAE-2、PixelSNAIL、GatedPixelCNNのVectorQuantizerEMAを実装し、tf2-published-modelsのSarusの実装をベースにした。

VQ-VAE-2ハイパーパラメータでは、コードブックサイズを離散ベクトルの数を表す256に設定し、コードブックの次元を各ベクトルの次元を表す64に設定した。PixelCNNsモデルでは、ドロップアウト率を0.2とした。
次に、トレーニングの詳細を説明する。学習データセットは解像度256×256の画像から構成される。

VQ-VAE-2の学習に関しては、学習率0.0003、バッチサイズ32、合計4,000回の学習反復でAdam optimizerを使用した。
PixelCNNsの学習に関しては、基本的に学習率0.0003、バッチサイズ32、合計50,000回の学習反復でAdam optimizerを使用した。段階的なデータ増強は転移学習に似ていることを考慮し、オプティマイザは各段階遷移でリセットされ、学習率はその後の段階において調整され、それぞれ10、40、100、500、1,000のファクターで減少した。これらの因子は試行錯誤によって経験的に決定された。

公正な比較を確実にするため、標準的なデータ補強実験と補強なし実験の両方で同じ設定を採用した。標準的なデータ補強のために、色演算のパラメータを0.15に変更した。

データ補強なしのトレーニングと、FID計算に実画像を使用する場合の、さらに公平な比較のために、段階的データ補強で使用されたアプローチを反映した、一定のアップスケーリングを採用した。

B.1　VQ-VAE-2エンコーダデコーダ実装モデルの主な変更点

我々は、オリジナルの実装に以下の変更を加えた：

- エンコーダ、デコーダ、残差スタックのオリジナルのSonnet実装をTensorFlow Kerasレイヤーに変換。

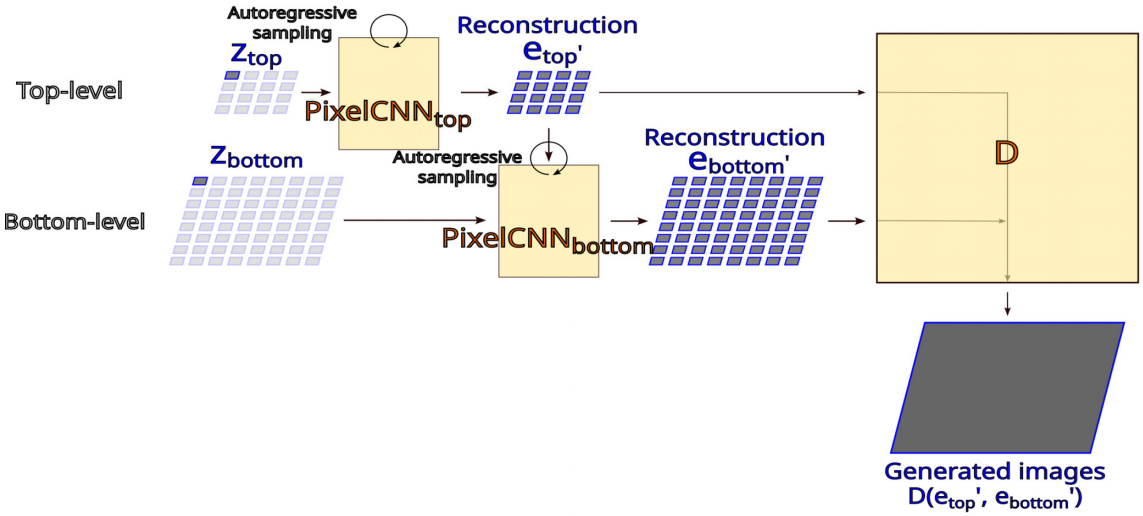
- 256x256から16x16の解像度の画像をエンコードするために、より深いアーキテクチャを作成するために、畳み込み層を追加してエンコーダクラスを拡張した。

- 階層的な符号化・復号化処理を行うために、エンコーダートップクラスとデコーダートップクラスを追加。

- 階層型VQ-VAEアーキテクチャをモジュール化するためのクラスを導入。
- VQVAEクラスを強化し、トップレベル、ボトムレベルのレイヤーとともに、階層的なエンコードとデコードのプロセスを統合した。

B.2　PixelCNN実装モデルの主な変更点

VQ-VAE-2の原著論文[5]に記載されている内容や機能拡張を反映し、性能を向上させ、特定のユースケースにモデルを適応させるために、以下の重要な修正を行った：
- 残差の強化：モデル内の残差接続を改善するために、新しいクラスを追加した。この変更は、VQ-VAE-2の原著論文に基づくもので、学習時間やメモリ使用量に大きな影響を与えることなく、



app.Fig. 1 Overview of a PC-VQ2 generation structure.

based on the official PyTorch code of StyleGAN3.

We implemented VQ-VAE-2 encoder-decoder, reconstruction, and generation modules on the basis of DeepMind's sonnet library. Additionally, we implemented VectorQuantizerEMA of VQ-VAE-2, PixelSNAIL, and GatedPixelCNN, based on Sarus's implementation from tf2-published-models.

For the VQ-VAE-2 hyperparameters, we set the codebook size to 256, representing the number of discrete vectors, and the codebook dimension to 64, indicating the dimension of each vector. For the PixelCNNs models, we used a dropout rate of 0.2.

Next, we detail the pieces of training. Our training datasets consisted of images with a resolution of 256 x 256.

Regarding VQ-VAE-2 training, we used Adam optimizer with a learning rate of 0.0003, a batch size of 32, and a total of 4,000 training iterations.

Regarding the PixelCNNs training, basically, we used Adam optimizer with a learning rate of 0.0003, a batch size of 32, and a total of 50,000 training iterations. Considering that phased data augmentation is akin to transfer learning, the optimizer was reset at each phase transition, and the learning rate was adjusted in subsequent phases, decreased by factors of 10, 40, 100, 500, and 1,000 respectively. These factors were empirically determined through trial and error.

To ensure a fair comparison, both the standard data augmentation and the no augmentation experiments employed the same settings. For standard data augmentation, we changed the parameters of color operations to 0.15.

For a further equitable comparison in the no data augmentation training and when using real images for FID calculations, we employed constant upscaling, mirroring the approach used in phased data augmentation.

B. 1 Key Modifications to the VQ-VAE-2 Encoder Decoder Implemented Model

We made the following modifications to the original implementations:

• Converted the original Sonnet implementation of the encoder, decoder, and residual stack to TensorFlow Keras layers.

• To encode images from a resolution of 256x256 to 16x16, we extended the Encoder class by adding additional convolutional layers to create a deeper architecture.

• Added Encoder_Top and Decoder_Top classes to handle the hierarchical encoding and decoding process.

• Introduced classes for modularizing the hierarchical VQ-VAE architecture.

• Enhanced the VQVAE class to integrate the hierarchical encoding and decoding process, along with the top-level and bottom-level layers.

B. 2 Key Modifications to the PixelCNN Implemented Models

The following key modifications were made to enhance performance and adapt the models for our specific use case, reflecting contents and enhancements described in the original VQ-VAE-2 paper[5]:

• Residual Enhancements: Added a new class to improve the residual connections in the model. This change is based on the original VQ-VAE-2 paper, which discusses the use of deep residual networks consisting of

app.Fig. 2 より高解像度の潜在空間、32-16、および位相補強を持つPC-VQ2モデルによって生成された人間の顔画像。

モデルの尤度を向上させるために、1w1畳み込みからなる深い残差ネットワークの使用について論じている。

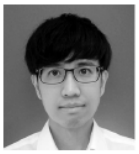
- CausalAttentionBlock: CausalAttentionBlock内にドロップアウト層を統合し、学習中にランダムにユニットをドロップすることでオーバーフィッティングを防ぐ。

- PixelSNAILとGatedPixelCNNモデル: 両モデルを修正し、残差ブロックを追加し、メインネットワークに通す前に入力を処理する初期畳み込み層を追加した。両モデルともVQ-VAE-2から離散ベクトルを入力として受け取り、それをワンホットエンコードし、最初の畳み込み層を通して処理し、後続の層の表現を調整する。モデルの深さと容量をさらに向上させるために、追加の残差ブロックを含めた。

- GatedPixelCNNモデル: 具体的には、条件付きGatedPixelCNN論文[10]に基づき、文脈処理をDense層からConv2D層に変換し、空間情報をより良く扱うために、階層的文脈条件付けのためのアップサンプリング層と畳み込み層を追加した。汎化を改善するために、モデル内の統合されたドロップアウト層。

References

1) N.T. Tran, V.H. Tran, N.B. Nguyen, T.K. Nguyen, and N.M. Cheung, "On data augmentation for gan training," IEEE Transactions on Image Processing, vol.30, pp.1882–1897, 2021.
2) Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.11293–11302, 2019.
3) S. Zhao, Z. Liu, J. Lin, J.Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient gan training," Advances in Neural Information Processing Systems, vol.33, pp.7559–7570, 2020.
4) T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," Advances in Neural Information Processing Systems, vol.33, pp.12104– 12114, 2020.
5) A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," Advances in neural information processing systems, vol.32, 2019.
6) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol.27, 2014.
7) A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," International Journal of Information Management Data Insights, vol.1, no.1, p.100004, 2021.
8) J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," International conference on machine learning, pp.2256–2265, PMLR, 2015.
9) A. Van Den Oord, O. Vinyals, et al., "Neural discrete representation learning," Advances in neural information processing systems, vol.30, 2017.
10) A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., "Conditional image generation with pixelcnn decoders," Advances in neural information processing systems, vol.29, 2016.
11) X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," International Conference on Machine Learning, pp.864–872, PMLR, 2018.
12) T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.4401– 4410, 2019.
13) Y. Choi, Y. Uh, J. Yoo, and J.W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
14) G. Parmar, R. Zhang, and J.Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11410– 11420, 2022.
15) T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," Proc. NeurIPS, 2021.
16) A. Adadi, "A survey on data-efficient algorithms in big data era," Journal of Big Data, vol.8, no.1, pp.1–54, 2021.
17) C. Shorten and T.M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of big data, vol.6, no.1, pp.1–48, 2019.
18) A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018 international interdisciplinary PhD workshop (IIPhDW), pp.117–122, IEEE, 2018.
19) E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q.V. Le, "Autoaugment: Learning augmentation policies from data," arXiv preprint arXiv:1805.09501, 2018.
20) H. Jun, R. Child, M. Chen, J. Schulman, A. Ramesh, A. Radford, and I. Sutskever, "Distribution augmentation for generative modeling," International Conference on Machine Learning, pp.5006–5019, PMLR, 2020.
21) OpenAI, "ChatGPT (September 25 Version)." https://chat.openai.com/chat, 2023. [Large language model].



Yuta Mimura received the degree of Master in the field of Cosmosciences at the Graduate School of Science, Hokkaido University. He completed his coursework in the doctoral program in the field of Cosmosciences at the Graduate School of Science, Hokkaido University, and subsequently withdrew from the program. His research interests include generative models and applications of machine learning to particle physics.

app.Fig. 2 Human-face images generated by the PC-VQ2 model with higher-resolution latent spaces, 32-16, and phased augmentation.

1 × 1 convolutions to improve model likelihood without significantly impacting training time or memory usage.

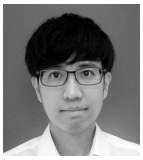
• CausalAttentionBlock: Integrated a dropout layer within the CausalAttentionBlock to prevent overfitting by randomly dropping units during training.

• PixelSNAIL and GatedPixelCNN Models: Modified both models to include additional residual blocks and an initial convolution layer to process the input before passing it through the main network. Both models receive discrete vectors from VQ-VAE-2 as input, one-hot encode them, and process them through the initial convolutional layer to adjust the representation for subsequent layers. Included the additional residual blocks to further improve the models' depth and capacity.

• GatedPixelCNN Model: Specifically, converted context processing from Dense layers to Conv2D layers and added upsampling and convolutional layers for hierarchical context conditioning to better handle spatial information, based on the conditional GatedPixelCNN paper[10]. Integrated dropout layers within the model to improve generalization.

References

1) N.T. Tran, V.H. Tran, N.B. Nguyen, T.K. Nguyen, and N.M. Cheung, "On data augmentation for gan training," IEEE Transactions on Image Processing, vol.30, pp.1882–1897, 2021.
2) Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.11293–11302, 2019.
3) S. Zhao, Z. Liu, J. Lin, J.Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient gan training," Advances in Neural Information Processing Systems, vol.33, pp.7559–7570, 2020.
4) T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," Advances in Neural Information Processing Systems, vol.33, pp.12104– 12114, 2020.
5) A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," Advances in neural information processing systems, vol.32, 2019.
6) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol.27, 2014.
7) A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," International Journal of Information Management Data Insights, vol.1, no.1, p.100004, 2021.
8) J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," International conference on machine learning, pp.2256–2265, PMLR, 2015.
9) A. Van Den Oord, O. Vinyals, et al., "Neural discrete representation learning," Advances in neural information processing systems, vol.30, 2017.
10) A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., "Conditional image generation with pixelcnn decoders," Advances in neural information processing systems, vol.29, 2016.
11) X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," International Conference on Machine Learning, pp.864–872, PMLR, 2018.
12) T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.4401– 4410, 2019.
13) Y. Choi, Y. Uh, J. Yoo, and J.W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
14) G. Parmar, R. Zhang, and J.Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11410– 11420, 2022.
15) T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," Proc. NeurIPS, 2021.
16) A. Adadi, "A survey on data-efficient algorithms in big data era," Journal of Big Data, vol.8, no.1, pp.1–54, 2021.
17) C. Shorten and T.M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of big data, vol.6, no.1, pp.1–48, 2019.
18) A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018 international interdisciplinary PhD workshop (IIPhDW), pp.117–122, IEEE, 2018.
19) E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q.V. Le, "Autoaugment: Learning augmentation policies from data," arXiv preprint arXiv:1805.09501, 2018.
20) H. Jun, R. Child, M. Chen, J. Schulman, A. Ramesh, A. Radford, and I. Sutskever, "Distribution augmentation for generative modeling," International Conference on Machine Learning, pp.5006–5019, PMLR, 2020.
21) OpenAI, "ChatGPT (September 25 Version)." https://chat. openai.com/chat, 2023. [Large language model].



Yuta Mimura received the degree of Master in the field of Cosmosciences at the Graduate School of Science, Hokkaido University. He completed his coursework in the doctoral program in the field of Cosmosciences at the Graduate School of Science, Hokkaido University, and subsequently withdrew from the program. His research interests include generative models and applications of machine learning to particle physics.