NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Dmitry Baranchuk

APPLICATIONS OF DEEP GENERATIVE MODELS FOR MACHINE LEARNING PREDICTION PROBLEMS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2024

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Artem Babenko, Candidate of Sciences, National Research University Higher School of Economics.

1 Introduction

Topic of the thesis

For the past decade, deep neural networks have continuously grown in capability and capacity and excelled in various machine learning tasks, such as language processing, image recognition, speech synthesis, video generation and others. The deep learning methods can be grouped into two main classes: *discriminative* and *generative* approaches.

Discriminative models aim to answer specific questions about the data objects. For example, determine what is depicted in a picture, count the number of people on CCTV snapshots, and suggest an effective treatment for a patient given their measurements. More formally, the discriminative methods model the conditional distribution $p(y|x)$ given the observed pairs (x, y), where x is an input object and y is a target label. Neural networks have rapidly demonstrated remarkable performance in a wide range of predictive tasks due to the emergence of large labeled datasets and the development of specialized hardware, e.g., graphics processing units (GPU). However, there are still many practical challenges in discriminative problems. For example, the data objects can have missing observations that could be informative for more accurate model predictions. Sometimes, collecting a large labeled dataset can be challenging and costly and hence, one requires the top-performing methods that have access only to few labeled samples during training. Also, in some areas and applications, data may be subject to the General Data Protection Regulation (GDPR) and contain private or sensitive user data. This problem might limit the use and collection of such data for developing machine learning methods.

ontrary to discriminative modeling, the fundamental goal of generative models is to approximate the data distribution p_{data} given the finite set of observed objects $\mathcal{D} = \{x_0, ..., x_N\}$ from this distribution. Deep generative methods approximate p_{data} using a deep neural network with parameters θ. The parameters are learned to minimize the distance between the model distribution p_θ and p_{data}: $\theta^* = \min_\theta d(p_{data}, p_\theta)$. The distance $d(\cdot, \cdot)$ may be an arbitrary similarity measure between distributions, e.g., KL divergence. An illustrative example of the generative problem: given Vincent van Gogh's paintings, learn the model θ to draw the new paintings in the same style. Compared to the similar predictive problem "Who is the author of the painting?", one can correctly conclude that generative tasks are usually significantly more sophisticated than discriminative ones.

There exist many classes of deep generative models, and they can be grouped into two major categories: *likelihood-based models* and *implicit generative models*. Likelihood-based models explicitly learn p_θ via maximizing the likelihood directly or its lower bound. The examples of likelihood-based methods include autoregressive models [1], diffusion models [2, 3], normalizing flows [4, 5, 6], variational autoencoders [7]. On the other hand, implicit models do not have direct access to the density function but can still produce plausible samples from the target distribution. The prominent representative is generative adversarial networks (GANs) [8]. Each class of generative models has its strengths and weaknesses. For this reason, different kinds of generative models can be preferable in different practical applications and domains. We refer the reader to the comprehensive overviews of the existing generative models [9, 10, 11] for details.

Deep generative modeling has been thoroughly investigated in recent years and achieved impressive results in various areas. Today, people can generate highly realistic images based on text descriptions [12, 13], produce advertising videos [14] and chat with "intelligent" systems like GPT4 [15]. These successes raise the question of whether so powerful deep generative models can complement established discriminative solutions. Many research works have already provided an affirmative answer to this question in different areas [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. This thesis extends this line of works and considers deep generative models for the following practical applications: i) missing data imputation in time series to improve the performance of classification and regression methods; ii) image semantic segmentation when the amount of labeled data is scarce; iii) tabular data generation to design high-quality and private synthetic datasets for downstream tabular tasks.

Relevance

The thesis addresses the applications of deep generative models for three different fundamental machine learning problems. Below, we briefly discuss each of them in more detail.

The first work focuses on the time series imputation problem. The task aims to fill in missing observations in real-world time series data. Multivariate time series with missing values are prevalent in areas such as healthcare and finance and have increased in number and complexity over the past years. Missing values often occur due to faulty measurement

devices, costly procedures, and human mistakes. As a result, the data can lack some informative features, causing machine learning methods to make incorrect predictions. Recent research has shown that accurate time series imputation significantly enhances the performance on downstream tasks [27, 28, 29].

Popular deep learning imputation approaches usually apply recurrent neural networks (RNNs) for sequence modeling [27, 30, 31, 29]. Other works combine RNNs with an adversarial objective [32, 28, 33] to improve the imputation performance. In this thesis, we make the first attempt to use deep probabilistic generative models for time series imputation. Specifically, we propose a variational autoencoder (VAE) with Gaussian process (GP) prior and demonstrate its effectiveness on the image and healthcare datasets with a temporal component. The follow-up work [20] proposes the probabilistic imputation method based on diffusion models and further enhances the imputation performance. Moreover, the appealing property of probabilistic imputation methods is that they can provide uncertainty estimations for the predicted values. This property is crucial for interpretive estimates and the trustworthiness of the method, especially if one aims to integrate it into medical applications.

The second work investigates generative models in the context of image semantic segmentation. Semantic segmentation is a fundamental computer vision problem that aims to recognize elements in an image at the pixel level. Opposed to image classification, where the model typically predicts a single label for an image, semantic segmentation seeks to assign *each pixel* to the class label. This makes semantic segmentation a highly challenging problem that would benefit from large labeled datasets. However, the accurate and consistent annotation of many images requires tremendous human effort and cost. For this reason, the methods that can provide strong segmentation performance given only few labeled images are in high demand [18, 19, 34].

Deep generative models have already been applied for semantic segmentation. Most methods leverage state-of-the-art GANs [35] as infinite generators of synthetic labeled data. This data is then used to train the semantic segmentation models. Some methods [36, 37, 38] exploit the evidence that the latent space of the GANs contains a direction that allows producing synthetic images along with foreground/background segmentation masks. Other works [18, 19] exploit intermediate pixel-level representations of GANs to predict segmentation masks for generated images. These methods demonstrate promising results in the setting when there is a limited number of human-annotated images.

Diffusion probabilistic models (DPMs) demonstrate state-of-the-art image generation in terms of both image quality and diversity [39, 12, 13]. The advantages of DPM are successfully exploited in generative tasks such as image colorization [40], inpainting [40], super-resolution [41, 42], and semantic editing [43], where DPMs often achieve more impressive results than GANs. However, it has not been explored whether DPMs can be effectively applied to discriminative vision problems. We have investigated intermediate representations of DPMs and revealed that they contain the pixel-level semantic information of the input image. Following [18], we propose a novel semantic segmentation method that exploits these image representations. We demonstrate its superiority over GAN-based and self-supervised approaches in the label-efficient setting.

Finally, we extend the framework of diffusion probabilistic models to the tabular domain. Tabular datasets are usually isolated and limited in size, as opposed to textual or image data that is massively available on the Internet. Often, tabular data contains personal, private or sensitive information and hence cannot be publicly shared without violating GDPR-like regulations. Deep generative models in the tabular domain are mainly used to mitigate this problem by replacing real user data with synthetic data. At the same time, the synthetic dataset has to inherit the properties of the real distribution to be useful for downstream applications. The recent works have developed many generative modeling methods, including tabular VAEs [44] and GAN-based approaches [44, 22, 23, 24, 25, 26, 45, 46, 47, 48] Motivated by the success of diffusion models in other domains, we introduce TabDDPM — a diffusion model that can be applied to arbitrary tabular datasets and handles various feature distributions. We extensively evaluate TabDDPM on a wide set of benchmarks and demonstrate its superiority over existing GAN/VAE alternatives.

2 Main Results and Conclusions

Contribution. The main results of the work are formulated below.

1. We propose a novel probabilistic model: a variational autoencoder with a Gaussian process prior in the latent space for effective time series data modeling. The designed model is applied for the time series imputation task. We demonstrate that our approach outperforms several classical and deep learning-based data imputation methods on multivariate time series from the computer vision and healthcare do-

mains. In addition, the method improves the smoothness of the imputations and provides interpretable uncertainty estimates.

2. We reveal that the state-of-the-art diffusion models have meaningful pixel-level image representations. Based on this knowledge, we propose a novel semantic segmentation approach that outperforms previous state-of-the-art generative and self-supervised methods when few annotated images are available.

3. We propose TabDDPM — a diffusion model for tabular data generation. This model outperforms other generative models for this task and can be useful for practitioners to replace private and sensitive data with generated data. This potentially takes a step toward the safe sharing of a company's internal data to develop high-quality prediction methods.

Theoretical and practical significance. The proposed methods and empirical findings contribute to the increasing prevalence of generative models for predictive tasks in machine learning. In scenarios characterized by a scarcity of labeled data, we demonstrate that a pretrained diffusion model can serve either as an effective data engine or as a strong discriminative model out of the box. For missing data imputation, we provide evidence that deep probabilistic modeling is a promising paradigm in healthcare applications, where it can recover missing patient measurements in an interpretable manner. Moreover, the thesis introduces a novel state-of-the-art approach for tabular data synthesis, enabling the training of highly effective machine learning methods in privacy-concerned scenarios.

Key aspects/ideas to be defended:

1. A deep probabilistic time series imputation method based on a variational autoencoder that uses a Gaussian process prior for better time series modeling;

2. Investigation of the internal representations of diffusion models, revealing the presence of useful fine-grained semantic information about input images. A semantic segmentation method that effectively utilizes the image representations extracted from pretrained diffusion models when labeled data is limited;

3. A diffusion-based generative approach for tabular data modeling.

Personal contribution. In the first work, the author was responsible for the technical contribution of the paper: developing the method and conducting most experiments and

analysis. In the second work, the author proposed the core scientific ideas, collected the datasets, implemented the method, conducted most experiments and analysis and wrote the text. In the third work, the author formulated the key ideas, organized the research project, designed the experiment pipelines, and contributed to writing the paper.

Publications and probation of the work

Top-tier publications

1. *Vincent Fortuin**, **Dmitry Baranchuk***, *Gunnar Rätsch, Stephan Mandt* GP-VAE: Deep probabilistic time series imputation. International Conference on Artificial Intelligence and Statistics, 2020 (AISTATS 2020). CORE *A* conference;
2. **Dmitry Baranchuk**, *Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, Artem Babenko* Label-Efficient Semantic Segmentation with Diffusion Models. International Conference on Learning Representations, 2022 (ICLR 2022). CORE A^* conference;
3. *Akim Kotelnikov,* ***Dmitry Baranchuk****, Ivan Rubachev, Artem Babenko* TabDDPM: Modelling Tabular Data with Diffusion Models. International Conference on Machine Learning, 2023 (ICML 2023). CORE A^* conference.

Reports at seminars

1. Seminar of the research group in Biomedical Informatics at ETH Zurich, Zurich, August 20, 2019. Topic: "Variational Autoencoders with Gaussian Process Priors for Time Series Modeling";
2. Christmas Colloquium on Computer Vision. Moscow, December 27, 2021. Topic: "Label-Efficient Semantic Segmentation with Diffusion Models";
3. Yandex Research Seminar, Moscow, July 24, 2022. Topic: "Applications of Diffusion Probabilistic Models in Practical Machine Learning Problems".

Volume and structure of the work. The thesis contains an introduction, the content of publications, a conclusion and includes the text of publications. The total volume of the thesis is 61 pages.

3 Content of the work

3.1 GP-VAE: Deep Probabilistic Time Series Imputation

This work addresses the problem of multivariate time series imputation, i.e., filling in missing values in time series data. Multivariate time series consist of multiple correlated univariate time series ("channels") and require imputation models that consider both temporal correlations within each channel and correlations across channels.

We denote a multivariate time series of length τ_T as $\mathbf{X} \in \mathbb{R}^{T \times d}$. A data point $\mathbf{x}_t = [x_{t1}, \ldots, x_{tj}, \ldots, x_{td}]^\top \in \mathbb{R}^d$ is measured at T consecutive time points $\tau = [\tau_1, \ldots, \tau_T]^\top$ with $\tau_t < \tau_{t+1} \, \forall t$ and $\tau_1 = 0$.

Then, we assume that any number of these data features x_{tj} can be missing. Thus, each data point can be partitioned into observed and unobserved features: $\mathbf{x}_t^o := [x_{tj} \,|\, x_{tj} \text{ is observed}]$ and $\mathbf{x}_t^m := [x_{tj} \,|\, x_{tj} \text{ is missing}]$ with $\mathbf{x}_t^o \cup \mathbf{x}_t^m \equiv \mathbf{x}_t$, respectively.

Missing value imputation describes the problem of estimating the true values of the missing features $\mathbf{X}^m := [\mathbf{x}_t^m]_{1:T}$ given the observed features $\mathbf{X}^o := [\mathbf{x}_t^o]_{1:T}$. Many methods assume the different data points to be independent, in which case the inference problem reduces to T separate problems of estimating $p(\mathbf{x}_t^m \,|\, \mathbf{x}_t^o)$. In the time series setting, this independence assumption is not satisfied, which leads to the more complex estimation problem of $p(\mathbf{x}_t^m \,|\, \mathbf{x}_{1:T}^o)$.

Method overview. We design a method based on variational autoencoders (VAEs) that map multivariate time series with missing observations into a latent space in which every dimension is fully determined. In the latent space, the temporal dynamics are modeled with a Gaussian process (GP). Since many features in the data might be correlated, the latent representation captures these correlations and uses them to reconstruct the missing values. Moreover, the GP prior in the latent space encourages the model to embed the data into a representation in which the temporal dynamics are smoother and more easily explainable than in the original data space. Finally, the decoder transforms the learned latent representation to estimate the missing values in the original feature space. The scheme of the proposed model is presented in Figure 1. The model comprises *generative* and *infenrence* models and we describe them in more detail below.

V. Fortuin*, D. Baranchuk*, G. Rätsch, S. Mandt. GP-VAE: Deep probabilistic time series imputation. AISTATS2020

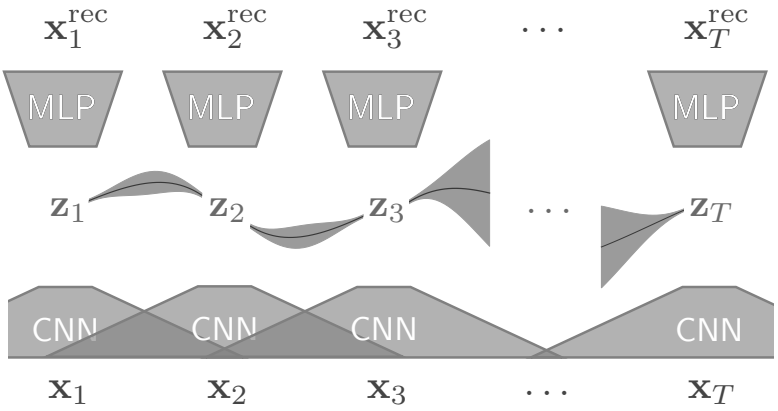


Figure 1: Overview of the GP-VAE model consisting of a convolutional inference network, a deep feed-forward generative network and a Gaussian process prior with mean function $m(\cdot)$ and kernel function $k(\cdot, \cdot)$ in latent space. The several CNN blocks as well as the MLP blocks are sharing their parameters.

Generative model. First, we apply GP in the latent space of a variational autoencoder. Specifically, we assign a latent variable $\mathbf{z}_t \in \mathbb{R}^k$ for every $\mathbf{x}_t$, and model temporal correlations in this reduced representation using a GP, $\mathbf{z}(\tau) \sim \mathcal{GP}(m_z(\cdot), k_z(\cdot, \cdot))$. This way, we decouple the step of filling in missing values and capturing instantaneous correlations between the different feature dimensions from modeling dynamical aspects.

A remaining practical difficulty we encountered is that many multivariate time series display dynamics at multiple time scales. One of our main motivations is to model time series that arise in medical setups where doctors measure different patient variables and vital signs, such as heart rate, blood pressure, etc. To model data that varies at multiple time scales, we consider the Cauchy kernel for our Gaussian process prior: This kernel has previously been successfully used in robust dynamic topic modeling where similar multi-scale time dynamics occur [49]. Given the latent time series $\mathbf{z}_{1:T}$, the observations $\mathbf{x}_t$ are generated time-point-wise by

$$p_\theta(\mathbf{x}_t \,|\, \mathbf{z}_t) = \mathcal{N}\left(g_\theta(\mathbf{z}_t), \sigma^2 \boldsymbol{I}\right) \ , \tag{1}$$

where $g_\theta(\cdot)$ is a potentially nonlinear function parameterized by the parameter vector θ. In our experiments, the function g_θ is implemented by a multilayer perceptron (MLP).

Inference model. To learn the parameters of the deep generative model described above and to efficiently infer its latent state, we are interested in the posterior distribution $p(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T})$. Since the exact posterior is intractable, we use variational inference [50, 51, 52] and amortize it using deep neural network [7]. To make our variational distribution

more expressive and capture the temporal correlations of the data, we employ a structured variational distribution [53] with efficient inference that leads to an approximate posterior, which is also a GP. We approximate the true posterior $p(\mathbf{z}_{1:T,j} \mid \mathbf{x}_{1:T})$ with a multivariate Gaussian variational distribution:

$$q(\mathbf{z}_{1:T,j} \mid \mathbf{x}^o_{1:T}) = \mathcal{N}\left(\mathbf{m}_j, \mathbf{\Lambda}_j^{-1}\right) \ , \tag{2}$$

where j indexes the dimensions in the latent space. Our approximation implies that our variational posterior can reflect correlations in time, but breaks dependencies across the different dimensions in $\mathbf{z}$-space (which is typical in VAE training [7, 54]).

We choose the variational family to be the family of multivariate Gaussian distributions in the time domain, where the precision matrix $\mathbf{\Lambda}_j$ is parameterized as a tridiagonal matrix. Samples from q can thus be generated in linear time in T [55, 56, 57] as opposed to the cubic time complexity for a full-rank matrix. Moreover, compared to a fully factorized variational approximation, the number of variational parameters are merely doubled. Note that while the precision matrix is sparse, the covariance matrix can still be dense, allowing to reflect long-range dependencies in time.

We amortize the inference over $\mathbf{m}_j$ and $\mathbf{\Lambda}_j$ using an inference network $q_\psi(\cdot)$. Following VAE training, the parameters of the generative model θ and inference network ψ can be jointly trained by optimizing the evidence lower bound (ELBO),

$$\log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \mathbb{E}_{q_\psi(\mathbf{z}_t \mid \mathbf{x}_{1:T})}\left[\log\, p_\theta(\mathbf{x}^o_t \mid \mathbf{z}_t)\right] - \beta\, D_{KL}\left[q_\psi(\mathbf{z}_{1:T} \mid \mathbf{x}^o_{1:T}) \,\|\, p(\mathbf{z}_{1:T})\right] \tag{3}$$

We evaluate the ELBO only on the observed features of the data since the remaining features are unknown, and set these missing features to a fixed value (zero) during inference.

Results. We performed experiments on the benchmark data set *Healing MNIST* [60], which combines the classical MNIST data set [61] with properties common to medical time series, the SPRITES data set [62], and on a real-world medical data set from the 2012 Physionet Challenge [63]. We compared our model against conventional single imputation methods [58], GP-based imputation [64], VAE-based methods that are not specifically designed to handle temporal data [7, 59], and modern state-of-the-art deep learning methods for temporal data imputation [65, 27].

We observe strong quantitative (Tab. 1, 2) and qualitative (Fig. 2) evidence that our proposed model outperforms most baseline methods in terms of imputation quality on all three tasks and performs comparable to the state of the art (BRITS) on the medical data.

Table 1: Performance of the different models on the Healing MNIST test set and the SPRITES test set in terms of negative log likelihood [NLL] and mean squared error [MSE] (lower is better), as well as downstream classification performance [AUROC] (higher is better).

	Healing MNIST			SPRITES
Model	NLL	MSE	AUROC	MSE
Mean imputation [58]	-	0.168 ± 0.000	0.938 ± 0.000	0.013 ± 0.000
Forward imputation [58]	-	0.177 ± 0.000	0.935 ± 0.000	0.028 ± 0.000
VAE [7]	0.599 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	0.034 ± 0.000
HI-VAE [59]	0.372 ± 0.008	0.134 ± 0.003	**0.962 ± 0.001**	0.035 ± 0.000
GP-VAE (proposed)	**0.341 ± 0.007**	**0.117 ± 0.002**	**0.960 ± 0.002**	**0.002 ± 0.000**

Table 2: Performance of the different models on the Physionet data set in terms of AUROC of a logistic regression trained on the imputed time series.

Model	AUROC
Mean imputation [58]	0.703 ± 0.000
Forward imputation [58]	0.710 ± 0.000
GP [64]	0.704 ± 0.007
VAE [7]	0.677 ± 0.002
HI-VAE [59]	0.686 ± 0.010
GRUI-GAN [65]	0.702 ± 0.009
BRITS [27]	**0.742 ± 0.008**
GP-VAE (proposed)	**0.730 ± 0.006**

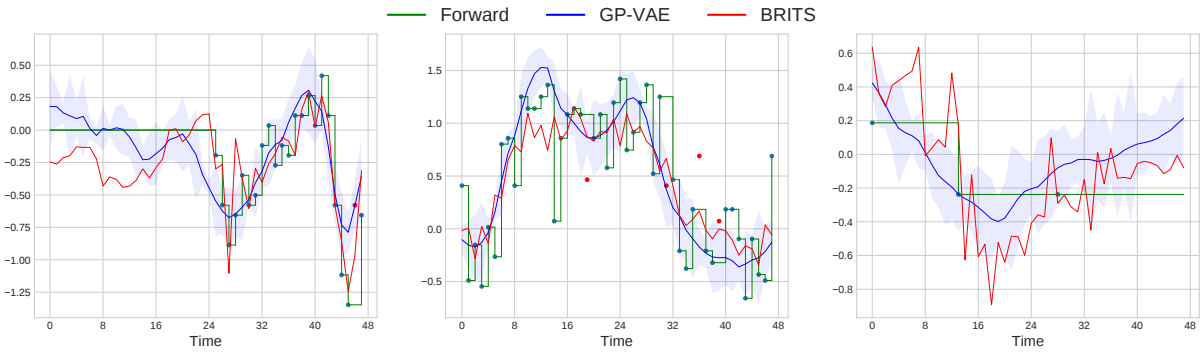


Figure 2: Imputations of several clinical variables with different amounts of missingness. BRITS (red) and forward imputation (green) yield single imputations, while the GP-VAE (blue) allows to draw samples from the posterior. The GP-VAE produces smoother curves, reducing noise from the original input, and exhibits an interpretable posterior uncertainty.

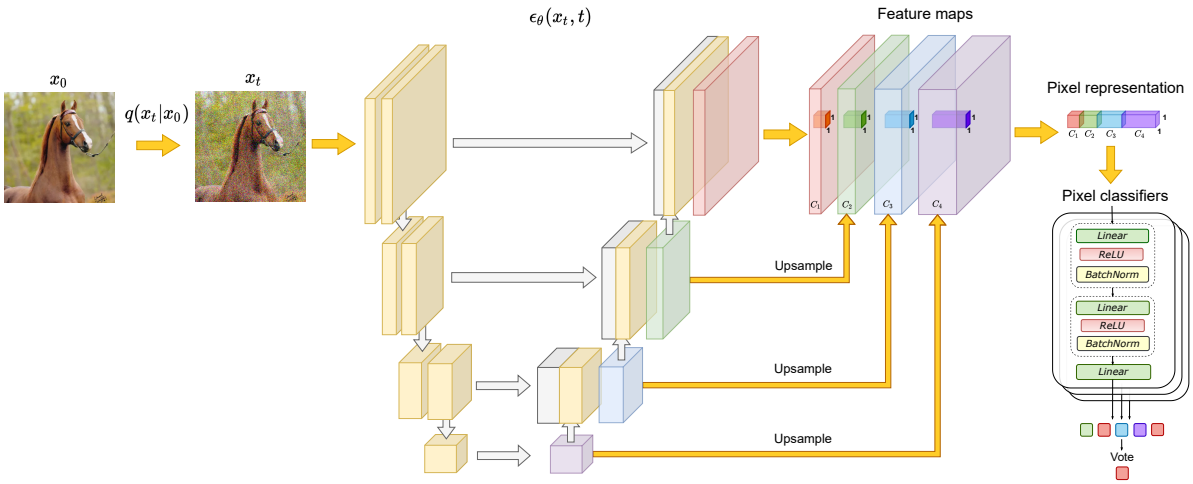


Figure 3: **Overview of the proposed method.** (1) $x_0 \rightarrow x_t$ by adding noise according to $q(x_t|x_0)$. (2) Extracting feature maps from a noise predictor $\epsilon_\theta(x_t, t)$. (3) Collecting pixel-level representations by upsampling the feature maps to the image resolution and concatenating them. (4) Using the pixel-wise feature vectors to train an ensemble of MLPs to predict a class label for each pixel.

3.2 Label-efficient Semantic Segmentation with Diffusion Models

In the following work, we investigate the representations learned by the state-of-the-art diffusion probabilistic models (DPMs) and show that they capture high-level semantic information valuable for semantic segmentation and outperforms the alternatives in the few-shot operating point.

Background. Typically, diffusion models transform noise $x_T \sim N(0, I)$ to the sample x_0 by gradually denoising x_T to less noisy samples x_t. Following a forward diffusion process, a noisy sample x_t can be obtained directly from a data point x_0:

$$\begin{aligned} q(x_t|x_0) &:= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \\ x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}, \ \sim \mathcal{N}(0,1), \end{aligned} \tag{4}$$

where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and define the schedule of the diffusion process.

A pretrained DDPM approximates a reverse diffusion process:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{5}$$

In practice, the neural network $\epsilon_\theta(x_t, t)$ predicts the noise component at a time step t; the mean is a linear combination of this noise component and x_t. The covariance predictor $\Sigma_\theta(x_t, t)$ is usually a constant scalar value for the particular step t.

D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, A. Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. ICLR2022

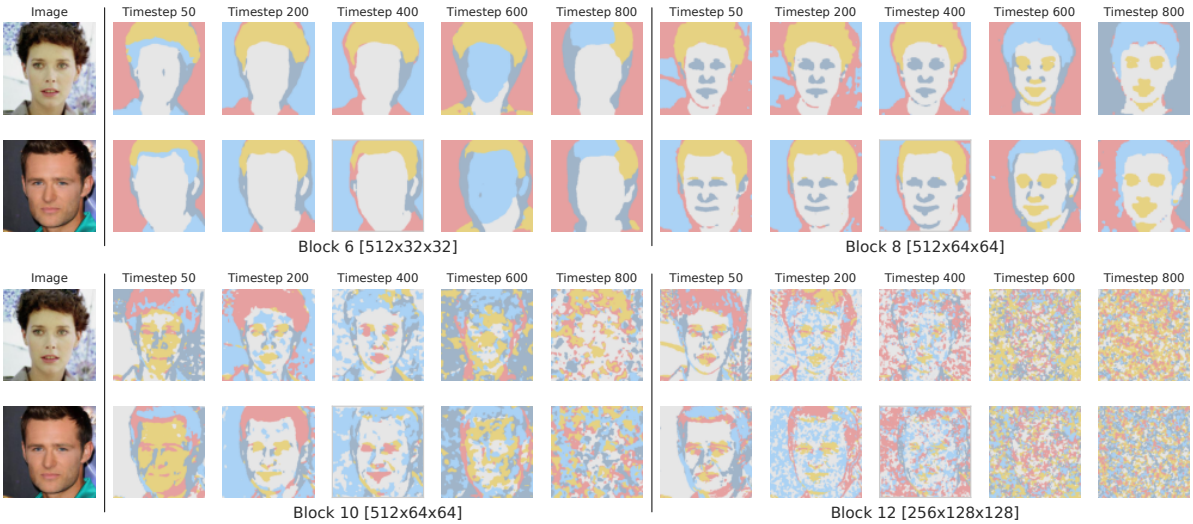


Figure 4: Examples of k-means clusters (k=5) formed by the features extracted from the UNet decoder blocks $\{6, 8, 10, 12\}$ on the diffusion steps $\{50, 200, 400, 600, 800\}$. The clusters from the middle blocks spatially span coherent semantic objects and parts.

The denoising model $\epsilon_\theta(x_t, t)$ is typically parameterized by different variants of the UNet architecture [66], and we investigate the state-of-the-art one proposed in [39].

Extracting representations. For a given real image $x_0 \in \mathbb{R}^{H \times W \times 3}$, one can compute T sets of activation tensors from the noise predictor network $\epsilon_\theta(x_t, t)$. The overall scheme for a timestep t is presented in Figure 3. First, we corrupt x_0 by adding Gaussian noise according to Equation (4). The noisy x_t is used as an input of $\epsilon_\theta(x_t, t)$ parameterized by the UNet model. The UNet's intermediate activations are then upsampled to $H \times W$ with bilinear interpolation. This allows treating them as pixel-level representations of x_0.

DPM representation analysis. Figure 4 shows the k-means clusters (k=5) formed by the features extracted by the FFHQ checkpoint from the different blocks and diffusion steps can span coherent semantic objects and object-parts. In the deeper blocks, the features correspond to coarse semantic masks, while the shallow ones can discriminate between fine-grained face parts but exhibit less semantic meaningness for coarse fragmentation. Across different diffusion steps, the most meaningful features correspond to the later ones. We attribute this behavior to the fact that on the earlier steps of the reverse process, the global structure of a DDPM sample has not yet emerged, therefore, it is hardly possible to predict segmentation masks at this stage.

Few-shot semantic segmentation method. The potential effectiveness of the intermediate DDPM activations observed above implies their usage as image representations for dense prediction tasks. Figure 3 schematically presents our overall approach for im-

Method	Bedroom-28	FFHQ-34	Cat-15	Horse-21	CelebA-19*	ADE Bedroom-30*
ALAE	20.0 ± 1.0	48.1 ± 1.3	—	—	49.7 ± 0.7	15.0 ± 0.5
VDVAE	—	57.3 ± 1.1	—	—	54.1 ± 1.0	—
GAN Inversion	13.9 ± 0.6	51.7 ± 0.8	21.4 ± 1.7	17.7 ± 0.4	51.5 ± 2.3	11.1 ± 0.2
GAN Encoder	22.4 ± 1.6	53.9 ± 1.3	32.0 ± 1.8	26.7 ± 0.7	53.9 ± 0.8	15.7 ± 0.3
SwAV	42.4 ± 1.7	56.9 ± 1.3	45.1 ± 2.1	54.0 ± 0.9	52.4 ± 1.3	30.6 ± 1.6
MAE	45.0 ± 2.0	**58.8 ± 1.1**	**52.4 ± 2.3**	63.4 ± 1.4	57.8 ± 0.4	31.7 ± 1.8
DDPM (Ours)	**49.4 ± 1.9**	**59.1 ± 1.4**	**53.7 ± 3.3**	**65.0 ± 0.8**	**59.9 ± 1.0**	**34.6 ± 1.7**

Table 3: The comparison of the segmentation methods in terms of mean IoU. **(*)** On CelebA-19 and ADE Bedroom-30, we evaluate models pretrained on FFHQ-256 and LSUN Bedroom, respectively.

age segmentation, which exploits the discriminability of these representations. In more detail, we consider a few-shot semi-supervised setup, when a large number of unlabeled images $\{X_1, \dots, X_N\} \subset \mathbb{R}^{H\times W\times 3}$ from the particular domain are available, and only for n training images $\{X_1, \dots, X_n\} \subset \mathbb{R}^{H\times W\times 3}$ the groundtruth K-class semantic masks $\{Y_1, \dots, Y_n\} \subset \mathbb{R}^{H\times W\times\{1,\dots,K\}}$ are provided.

The pretrained diffusion model is used to extract the pixel-level representations of the labeled images using the subset of the UNet blocks and diffusion steps t. In this work, we use the representations from the middle blocks B=$\{5, 6, 7, 8, 12\}$ of the UNet decoder and later steps t=$\{50, 150, 250\}$ of the reverse diffusion process. The extracted representations from all blocks B and steps t are upsampled to the image size and concatenated, forming the feature vectors for all pixels of the training images. Then, following [18], we train an ensemble of independent multi-layer perceptrons (MLPs) on these feature vectors, which aim to predict a semantic label of each pixel available for training images.

To segment a test image, we extract its DDPM-based pixel-wise representations and use them to predict the pixel labels by the ensemble. The final prediction is obtained by majority voting.

Datasets. In our evaluation, we mainly work with the "bedroom", "cat" and "horse" categories from LSUN [67] and FFHQ-256 [35]. As a training set for each dataset, we consider several images for which the fine-grained semantic masks are collected following the protocol from [18]. For each dataset, a professional assessor was hired to annotate

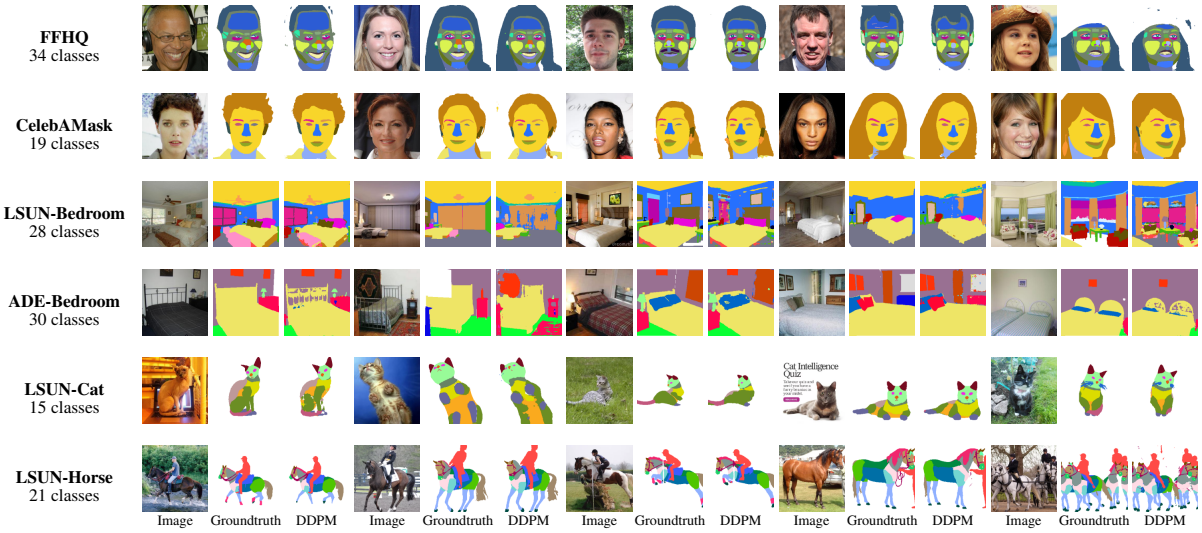


Figure 5: The examples of segmentation masks predicted by our method on the test images along with the groundtruth annotated masks.

train and test samples. We denote the collected datasets as Bedroom-28, FFHQ-34, Cat-15, Horse-21, ADE-Bedroom-30, CelebA-19 where the number corresponds to the number of semantic classes.

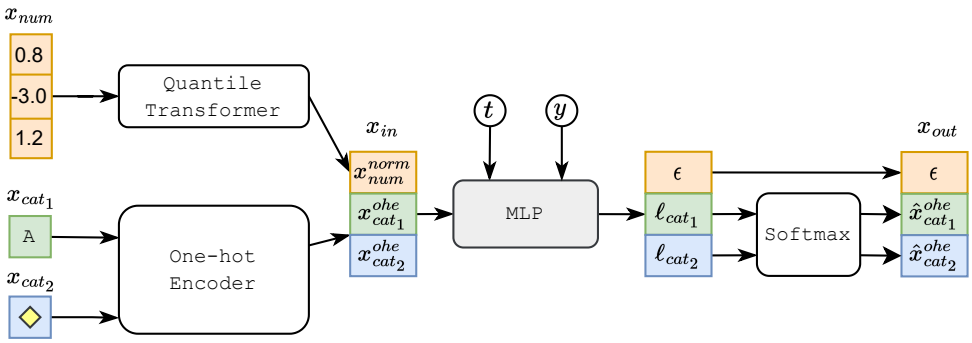
Methods. In the evaluation, we compare our DDPM-based method to the similar one but extract the features from various state-of-the-art self-supervised and generative models: MAE [68], SwAV [69], GAN Inversion + StyleGAN, GAN Encoder, VDVAE [70]

Main results. The comparison of the methods in terms of the mean IoU measure is presented in Table 3. The results are averaged over 5 independent runs for different data splits. Additionally, we provide several qualitative examples of segmentation with our method in Figure 5. Below, we highlight several key observations:

- The proposed method based on the DDPM representations significantly outperforms the alternatives on most datasets.
- The MAE baseline is the strongest competitor to the DDPM-based segmentation and demonstrates comparable results on the FFHQ-34 and Cat-15 datasets.

Overall, the proposed DDPM-based segmentation outperforms the baselines that exploit alternative generative models and also the baselines trained in the self-supervised fashion. This result highlights the potential of using state-of-the-art DDPMs as strong unsupervised representation learners.

Figure 6: TabDDPM scheme for classification tabular problems; t, y and ℓ correspond to a diffusion timestep, a class label, and logits, respectively.



3.3 TabDDPM: Modelling Tabular Data with Diffusion Models

Finally, we present TabDDPM, a novel DPM designed specifically for generating tabular data comprising both numerical and categorical features. TabDDPM surpasses existing tabular data generation methods based on GANs and VAEs. Additionally, we demonstrate that simple interpolation-based techniques, such as SMOTE [71], can produce remarkably effective synthetic data with high ML efficiency. However, in privacy-sensitive contexts where synthetic data is needed to replace real user data that cannot be shared, TabDDPM offers a preferable solution compared to SMOTE.

TabDDPM employs multinomial diffusion to model categorical and binary features and Gaussian diffusion for numerical features. Specifically, a tabular data sample $x = [x_{\text{num}}, x_{\text{cat}_1}, \ldots, x_{\text{cat}_C}]$ consists of N_{num} numerical features $x_{\text{num}} \in \mathbb{R}^{N_{\text{num}}}$ and C categorical features x_{cat_i} with K_i categories each. As for preprocessing, categorical features are one-hot encoded, i.e., $x_{\text{cat}_i}^{\text{ohe}} \in \{0,1\}^{K_i}$, and numerical features are normalized using the Gaussian quantile transformation from the scikit-learn library [72]. Consequently, the input x_0 has a dimensionality of $\left(N_{\text{num}} + \sum_{i=1}^{C} K_i\right)$. The reverse diffusion step in TabDDPM is modeled by an MLP architecture adapted from [73]:

$$\begin{aligned} \texttt{MLP}(x) &= \texttt{Linear}\left(\texttt{MLPBlock}\left(\ldots\left(\texttt{MLPBlock}(x)\right)\right)\right) \\ \texttt{MLPBlock}(x) &= \texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(x))) \end{aligned} \tag{6}$$

The model is trained by minimizing a sum of Gaussian and multinomial diffusion loss terms. The former corresponds to L_t^{simple}[2] objective for numerical features. The latter represents a sum of KL divergences between multinomial distributions, L_t^i, for each

A. Kotelnikov, D. Baranchuk, I. Rubachev, A. Babenko. TabDDPM: Modelling Tabular Data with Diffusion Models. ICML2023

categorical feature. The multinomial diffusion loss is additionally divided by the number of categorical features. The overall objective for a time step t can be described as follows:

$$L_t^{\text{TabDDPM}} = L_t^{\text{simple}} + \frac{\sum_{i \leq C} L_t^i}{C} \tag{7}$$

The model is parametarized to predict $\epsilon \sim N(0, 1)$ for numerical features and category probabilities $\hat{x}^{\text{ohe}}_{\text{cat}_i}$ for multinomial ones. For classification datasets, the model is conditioned on a class label, i.e., $p_\theta(x_{t-1}|x_t, y)$ is learned. For regression datasets, we consider a target value as an additional numerical feature and learn the joint distribution $p_\theta(x_{t-1}, y_{t-1}|x_t, y_t)$. TabDDPM for classification datasets is illustrated in Figure 6.

Datasets. For performance evaluation of tabular generative models, we consider a diverse set of 15 real-world public datasets, previously used for evaluating tabular models in [26, 73]. These datasets vary in size, nature, number of features, and their distributions.

Baselines. Given the large number of generative models proposed for tabular data, we evaluate TabDDPM against the leading approaches from each generative modeling paradigm: **TVAE** [44], **CTABGAN** [26], **CTABGAN+** [74]. Additionally, we include a "shallow" interpolation-based method **SMOTE** [71] and "generate" a synthetic point as a convex combination of a real data point and its k-th nearest neighbor from the dataset.

Evaluation measure. Our primary evaluation measure is machine learning (ML) efficiency [44]. In more detail, ML efficiency quantifies the performance of classification or regression models trained on synthetic data and evaluated on the real test set. In our experiments, we evaluate ML efficiency w.r.t. CatBoost [75], GBDT implementation providing state-of-the-art performance on tabular tasks [73].

Qualitative comparison. First, we investigate the ability of TabDDPM to model the individual and joint feature distributions. We visualize the typical individual feature distributions for real and synthetic data in Figure 7.

In most cases, TabDDPM produces more realistic feature distributions compared with TVAE and CTABGAN+. The advantage is more pronounced (1) for numerical features, which are uniformly distributed, (2) for categorical features with high cardinality, and (3) for mixed-type features that combine continuous and discrete distributions. Then, we also visualize the differences between the correlation matrices computed on real and synthetic data for different datasets, see Figure 8. In comparison with CTABGAN+ and TVAE, TabDDPM generates synthetic datasets with more realistic pairwise correlations.

Figure 7: The individual feature distributions for the real data and the data generated by TabDDPM, CTABGAN+, and TVAE. TabDDPM produces more realistic feature distributions in most cases.

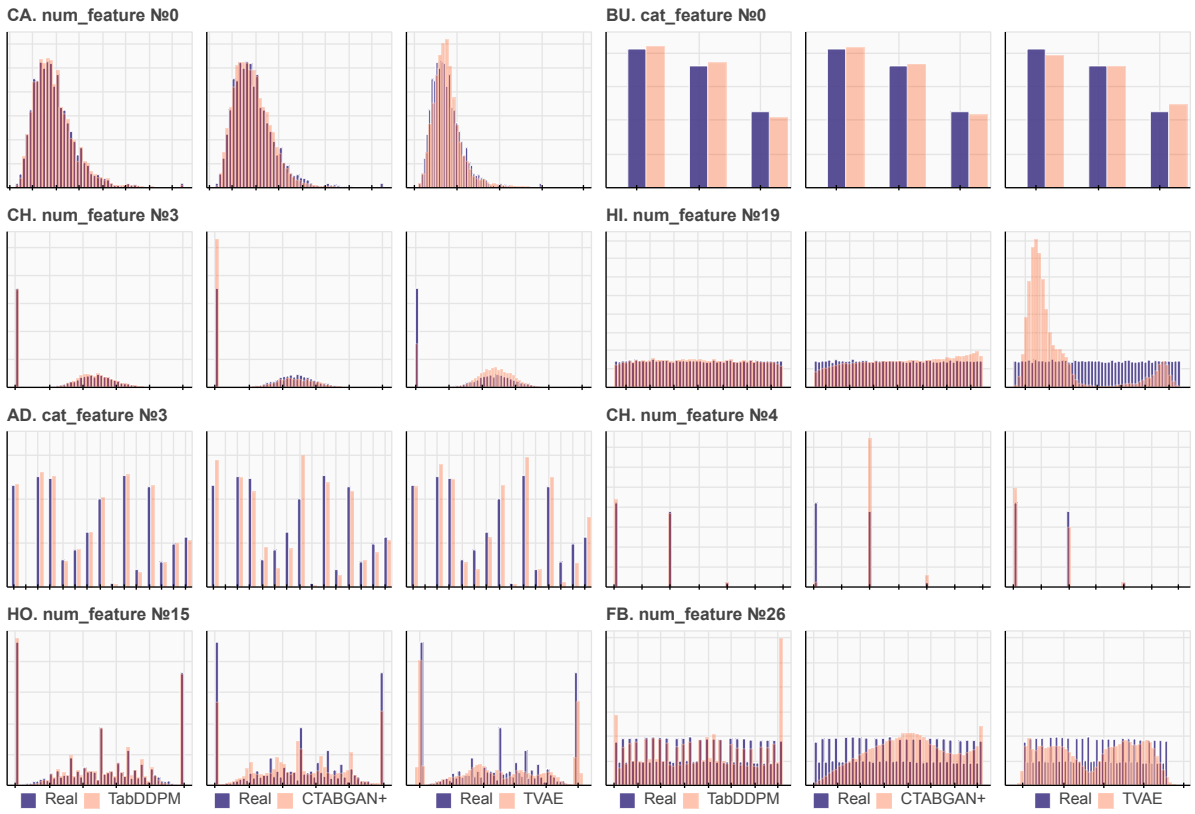


Figure 8: Absolute difference between correlation matrices computed on real and synthetic datasets. A more intensive red color indicates a higher difference between the real and synthetic correlation values. In most cases, TabDDPM captures feature correlations better than the alternatives.

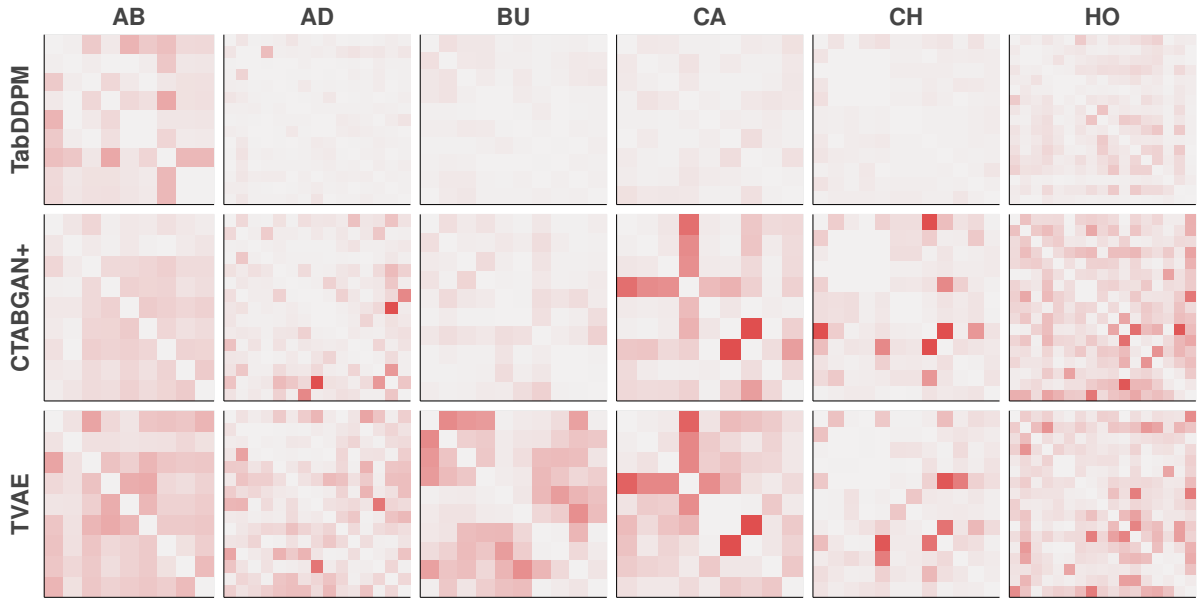


Table 4: The values of machine learning efficiency computed w.r.t. the state-of-the-art tuned CatBoost model.

	AB $_{(R2)}$	AD $_{(F1)}$	BU $_{(F1)}$	CA $_{(R2)}$	CAR $_{(F1)}$	CH $_{(F1)}$	DE $_{(F1)}$	DI $_{(F1)}$
CTGAN	$0.420_{\pm.004}$	$0.789_{\pm.001}$	$0.867_{\pm.003}$	$0.686_{\pm.003}$	$0.730_{\pm.001}$	$0.723_{\pm.006}$	$\mathbf{0.699_{\pm.002}}$	$0.459_{\pm.096}$
TVAE	$0.433_{\pm.008}$	$0.781_{\pm.002}$	$0.864_{\pm.005}$	$0.752_{\pm.001}$	$0.717_{\pm.001}$	$0.732_{\pm.006}$	$0.656_{\pm.007}$	$\mathbf{0.714_{\pm.039}}$
CTABGAN	–	$0.783_{\pm.002}$	$0.855_{\pm.005}$	–	$0.717_{\pm.001}$	$0.688_{\pm.006}$	$0.644_{\pm.011}$	$\mathbf{0.731_{\pm.022}}$
CTABGAN+	$0.467_{\pm.004}$	$0.772_{\pm.003}$	$0.884_{\pm.005}$	$0.525_{\pm.004}$	$0.733_{\pm.001}$	$0.702_{\pm.012}$	$0.686_{\pm.004}$	$\mathbf{0.734_{\pm.020}}$
SMOTE	$\mathbf{0.549_{\pm.005}}$	$0.791_{\pm.002}$	$0.891_{\pm.003}$	$\mathbf{0.840_{\pm.001}}$	$0.732_{\pm.001}$	$0.743_{\pm.005}$	$0.693_{\pm.003}$	$0.683_{\pm.037}$
TabDDPM	$\mathbf{0.550_{\pm.010}}$	$\mathbf{0.795_{\pm.001}}$	$\mathbf{0.906_{\pm.003}}$	$0.836_{\pm.002}$	$\mathbf{0.737_{\pm.001}}$	$\mathbf{0.755_{\pm.006}}$	$0.691_{\pm.004}$	$\mathbf{0.740_{\pm.020}}$
Real	$0.556_{\pm.004}$	$0.815_{\pm.002}$	$0.906_{\pm.002}$	$0.857_{\pm.001}$	$0.738_{\pm.001}$	$0.740_{\pm.009}$	$0.688_{\pm.003}$	$0.785_{\pm.013}$

	FB $_{(R2)}$	GE $_{(F1)}$	HI $_{(F1)}$	HO $_{(R2)}$	IN $_{(R2)}$	KI $_{(R2)}$	MI $_{(F1)}$	WI $_{(F1)}$
CTGAN	$0.443_{\pm.005}$	$0.333_{\pm.013}$	$0.575_{\pm.006}$	$0.433_{\pm.005}$	$0.745_{\pm.009}$	$0.772_{\pm.005}$	$0.783_{\pm.005}$	$0.749_{\pm.015}$
TVAE	$0.685_{\pm.003}$	$0.434_{\pm.006}$	$0.638_{\pm.003}$	$0.493_{\pm.006}$	$0.784_{\pm.010}$	$0.824_{\pm.003}$	$0.912_{\pm.001}$	$0.501_{\pm.012}$
CTABGAN	–	$0.392_{\pm.006}$	$0.575_{\pm.004}$	–	–	–	$0.889_{\pm.002}$	$\mathbf{0.906_{\pm.019}}$
CTABGAN+	$0.509_{\pm.011}$	$0.406_{\pm.009}$	$0.664_{\pm.002}$	$0.504_{\pm.005}$	$0.797_{\pm.005}$	$0.444_{\pm.014}$	$0.892_{\pm.002}$	$0.798_{\pm.021}$
SMOTE	$\mathbf{0.803_{\pm.002}}$	$\mathbf{0.658_{\pm.007}}$	$\mathbf{0.722_{\pm.001}}$	$0.662_{\pm.004}$	$\mathbf{0.812_{\pm.002}}$	$\mathbf{0.842_{\pm.004}}$	$0.932_{\pm.001}$	$\mathbf{0.913_{\pm.007}}$
TabDDPM	$0.713_{\pm.002}$	$0.597_{\pm.006}$	$\mathbf{0.722_{\pm.001}}$	$\mathbf{0.677_{\pm.010}}$	$0.809_{\pm.002}$	$\mathbf{0.833_{\pm.014}}$	$\mathbf{0.936_{\pm.001}}$	$\mathbf{0.904_{\pm.009}}$
Real	$0.837_{\pm.001}$	$0.636_{\pm.007}$	$0.724_{\pm.001}$	$0.662_{\pm.003}$	$0.814_{\pm.001}$	$0.907_{\pm.002}$	$0.934_{\pm.000}$	$0.898_{\pm.006}$

These illustrations indicate that TabDDPM is more flexible than alternatives and produces superior synthetic data.

Machine Learning efficiency. Then, we compare TabDDPM to alternative generative models in terms of ML efficiency. From each generative model, we sample a synthetic dataset with the size of a real train set. This synthetic data is then used to train a classification/regression model. In our experiments, classification performance is evaluated by the F1 score, and regression performance is evaluated by the R2 score.

We compute ML efficiency w.r.t. the current state-of-the-art model for tabular data. Specifically, we consider CatBoost [75] and the MLP architecture from [73] for evaluation.

Main results. The ML efficiency values are presented in Table 4. TabDDPM significantly outperforms TVAE and CTABGAN+ on most datasets, which highlights the advantage of diffusion models for tabular data as well as demonstrated for other domains in prior works. The interpolation-based SMOTE method demonstrates the performance competitive to TabDDPM and often significantly outperforms the GAN/VAE approaches.

Overall, TabDDPM provides state-of-the-art generative performance and can be used as a source of high-quality synthetic data. Interestingly, in terms of ML efficiency, a

simple "shallow" SMOTE method is competitive to TabDDPM, which raises the question if sophisticated deep generative models are needed.

Privacy. Here, we explore TabDDPM in privacy-concerned settings, e.g., sharing the data without disclosure of personal or sensitive information. In these setups, we aim to produce high-quality synthetic data that does not reveal the records from original data.

We measure the privacy of the generated data as a mean Distance to Closest Record (DCR) [26]. Low DCR values indicate that synthetic samples essentially mimic some real datapoints and can violate privacy requirements. Higher DCR values indicate that the generative model can produce "new" records rather than just near duplicates of the real data. Note that out-of-distribution data, e.g., random noise, will also provide high DCR. Therefore, DCR needs to be considered along with ML efficiency together.

Table 5 presents the DCR values for TabDDPM, SMOTE, CTABGAN+ and TVAE. TabDDPM is more private than SMOTE and less private than GAN/VAE alternatives. We attribute this to significantly lower ML efficiency of GAN/VAE-based baselines.

	AB	AD	BU	CA	CAR	CH	DE	DI
TVAE	0.088	0.220	0.226	0.056	0.010	0.241	0.096	0.146
CTABGAN+	0.081	0.400	0.242	0.070	0.020	0.235	0.131	0.204
SMOTE	0.018	0.082	0.080	0.016	0.007	0.099	0.054	0.074
TabDDPM	0.061	0.295	0.168	0.045	0.016	0.166	0.061	0.308
	FB	GE	HI	HO	IN	KI	MI	WI
TVAE	1.418	0.171	0.497	0.127	0.102	0.200	0.025	0.020
CTABGAN+	0.666	0.169	0.533	0.129	0.124	0.390	10.761	0.027
SMOTE	0.264	0.041	0.209	0.066	0.050	0.090	0.012	0.009
TabDDPM	0.785	0.076	0.473	0.096	0.050	0.252	0.574	0.023

Table 5: Comparison in terms of mean Distance to Closest Record (DCR) (higher is better). TabDDPM provides better DCR values compared with SMOTE but underperforms compared with TVAE and CTABGAN+. We attribute this to significantly lower ML efficiency of GAN/VAE-based alternatives.

4 Conclusion

The final section summarizes the main contributions of the thesis:

1. We have developed a novel multivariate time series imputation approach using a deep probabilistic model that combines variational autoencoders and Gaussian processes. The model maps the missing data from the input space into a latent space where each dimension is determined. In the latent space, a GP prior is used to better capture the temporal correlations of the data, resulting in more accurate imputations. The extensive experiments show that the proposed model produces state-of-the-art imputations, significantly improving the predictive methods on datasets with high missing rates.

2. Our research shows that pretrained diffusion models (DPMs) serve as effective representation learners for predictive computer vision tasks. Compared to GANs, DPMs offer simpler feature extraction without the need for an additional encoder and deliver superior generative quality. These benefits enable DPMs to achieve state-of-the-art performance in few-shot semantic segmentation, outperforming leading self-supervised approaches.

3. We have explored the diffusion framework for tabular data modeling and introduced TabDDPM, a method capable of generating data with various feature types, including numerical, ordinal, and categorical. Our method generates significantly more realistic tabular data compared to previous GAN- and VAE-based approaches. Consequently, our synthetic data can be used to train classification and regression models, particularly in scenarios where preserving user data privacy is crucial.

References

[1] Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17(1), 2016.

[2] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

[4] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf`.

[5] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *CoRR*, abs/1410.8516, 2015.

[6] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, 07–09 Jul 2015.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[9] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers, 2023.

[10] Harshvardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.

[11] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=08Yk-n5l2Al`.

[14] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

[15] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[16] Anonymous. Synthetic data from diffusion models improves imagenet classification. *Submitted to Transactions on Machine Learning Research*, 2023. URL `https://openreview.net/forum?id=DlRsoxjyPm`. Under review.

[17] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL `https://openreview.net/forum?id=laWYA-LXlNb`.

[18] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021.

[19] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021.

[20] YUSUKE TASHIRO, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[21] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *CVPR*, 2021.

[22] Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021.

[23] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[24] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. Relational data synthesis using generative adversarial networks: A design space exploration. *arXiv preprint arXiv:2008.12763*, 2020.

[25] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586: 485–500, 2022.

[26] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[27] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, pages 6775–6785, 2018.

[28] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: End-to-end generative adversarial network for multivariate time series imputation. IJCAI'19, page 3094–3100. AAAI Press, 2019. ISBN 9780999241141.

[29] Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=4c0J6lwQ4_`.

[30] Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 2019.

[31] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values, 2017.

[32] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. *International Conference on Learning Representations*, 2019.

[33] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *International Conference on Machine Learning*, 2018.

[34] Nico Catalano and Matteo Matteucci. Few shot semantic segmentation: a review of methodologies and open challenges, 2023.

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[36] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. *ICML*, 2021.

[37] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020.

[38] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021.

[39] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. 2021.

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. 2021.

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. 2021.

[42] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. 2021.

[43] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. 2021.

[44] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[45] Jayoung Kim, Jinsung Jeon, Jaehoon Lee, Jihyeon Hyeong, and Noseong Park. Oct-gan: Neural ode-based conditional tabular gans. In *Proceedings of the Web Conference 2021*, pages 1506–1515, 2021.

[46] Yishuo Zhang, Nayyar A Zaidi, Jiahui Zhou, and Gang Li. Ganblr: a tabular data generation model. In *2021 IEEE International Conference on Data Mining (ICDM)*, 2021.

[47] Richard Nock and Mathieu Guillame-Bert. Generative trees: Adversarial and copycat. *ICML*, 2022.

[48] Bingyang Wen, Yupeng Cao, Fan Yang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Modeling tabular data using causally-aware gan. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[49] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. Scalable generalized dynamic topic models. *Conference on Artificial Intelligence and Statistics*, 2018.

[50] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37 (2):183–233, 1999.

[51] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[52] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[53] Martin J Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008.

[54] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning*, 2014.

[55] Y Huang and WF McColl. Analytical inversion of general tridiagonal matrices. *Journal of Physics A: Mathematical and General*, 30(22):7919, 1997.

[56] Ranjan K Mallik. The inverse of a tridiagonal matrix. *Linear Algebra and its Applications*, 325(1-3):109–139, 2001.

[57] Robert Bamler and Stephan Mandt. Structured black box variational inference for latent time series models. *arXiv preprint arXiv:1707.01069*, 2017.

[58] Roderick JA Little and Donald B Rubin. Single imputation methods. *Statistical analysis with missing data*, pages 59–74, 2002.

[59] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

[60] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[61] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[62] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *International Conference on Machine Learning*, 2018.

[63] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[64] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[65] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1596–1607, 2018.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[67] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[68] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[69] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[70] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

[71] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[73] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[74] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan+: Enhancing tabular data synthesis. *arXiv preprint arXiv:2204.00401*, 2022.

[75] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.