Fast Particle-based Anomaly Detection Algorithm with Variational Autoencoder

Ryan Liu
University of California, Berkeley

Abhijith Gandrakota
Fermi National Accelerator Laboratory

Jennifer Ngadiuba
Fermi National Accelerator Laboratory

Maria Spiropulu
California Institute of Technology

Jean-Roch Vlimant
California Institute of Technology

Abstract

Model-agnostic anomaly detection is one of the promising approaches in the search for new beyond the standard model physics. In this paper, we present Set-VAE, a particle-based variational autoencoder (VAE) anomaly detection algorithm. We demonstrate a 2x signal efficiency gain compared with traditional subjettiness-based jet selection. Furthermore, with an eye to the future deployment to trigger systems, we propose the CLIP-VAE, which reduces the inference-time cost of anomaly detection by using the KL-divergence loss as the anomaly score, resulting in a 2x acceleration in latency and reducing the caching requirement.

1 Introduction

At the Large Hadron Collider (LHC), proton beams collide with each other at a frequency of $40\mathrm{MHz}$. The tremendous amount of data produced cannot be stored directly due to the limited capacity of downstream processing and storage systems. Therefore, an online processing system progressively reduces input data rates of three orders of magnitude [1, 2]. The first stage of this system consists of field-programmable gate arrays (FPGAs) where filters are executed with sub-microsecond latencies to retain the event data only if a specific set of criteria has been reached [3]. While this approach is very effective in discovering a new particle [4], it may be suboptimal when searching for new physics beyond the standard model that lacks a strong theoretical prior. Therefore, a model-agnostic approach to trigger in the detectors is of high interest to the high-energy physics community, and deep learning methods are among the most promising approaches [5].

Essentially, our goal is to find the out-of-distribution (OOD) events given the background distribution that is well-understood. This particular problem falls into the realm of anomaly detection (AD). One of the well-known architectures for anomaly detection is the autoencoder (AE) [6–8]. However, there exist a few challenges that we must tackle before deploying an autoencoder-based algorithm to the trigger in particle detectors. Firstly, we must develop a framework that can encode and decode a point cloud in an efficient way since a collision event is essentially a collection of particles [9, 10]. Secondly, we must control the number of operations and make the algorithm more parallelizable for future deployment to FPGAs [8, 11].

In this work, we propose an anomaly detection framework based on conditional variational autoencoders and Chamfer loss to address the first issue and propose a novel architecture called CLIP-VAE that is tailored to future deployment to online data processing systems. We present an evaluation of this framework on jet-level anomaly detection, and we envision that by demonstrating its capability on

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

jet-level particle-based anomaly detection, Set-VAE can be scaled to serve as an event-level anomaly detection algorithm at the CMS phase-2 level-1 trigger system. The code of this work is published in this Github repository.

2 Related Work

2.1 Anomaly Detection in High Energy Physics

Recently, anomaly detection has been widely studied in the high energy physics (HEP) community [5, 7, 8, 12, 13]. In particular, for jet-level anomaly detection, there are many works based on autoencoders [7, 9, 10, 13–15]. Traditionally, image-based autoencoders have been used for jet anomaly detection [7, 13, 15]. Recently, particle-based anomaly detection has gained more interest since it can exploit the sparse nature of jet data and gives better performance. Some of the examples include graph neural networks [10, 14] and deep sets [9]. However, the lack of a scalable decoding framework for particle-based autoencoders makes these algorithms infeasible for more realistic real-time trigger applications.

2.2 Permutation Invariant and Equivariant Models

As the input to the trigger is a set of particles with no particular ordering, it is important to guarantee that our model is permutation invariant or equivariant [16, 17]. To build a permutation invariant model, the common choices include a deep set [16] and a sequence of cross-attention layers with a destination length of one (the "class attention layer") [18]. As for permutation equivariant models, the self-attention block [19] and the deep set equivariant model [16, 17] can be used. However, attention-based algorithms are computationally intensive and thus infeasible for trigger applications.

3 Model Design

3.1 Set-VAE

Designing an autoencoder for point clouds poses a significant challenge, particularly in the decoding phase since generating a variable-size set from a fixed-dimensional latent space is non-trivial. Therefore, inspired by the neural translation model [20], we propose the Set-VAE to efficiently encode and decode a set. Given a set of pairs of continuous inputs (e.g. momentum, energy, etc.) and discrete labels (e.g. particle type) $\{(x_i, c_i)\}$, the encoder outputs a sample from the latent distribution $z \sim q_\phi(z|\{(x_i, c_i)\})$ for the whole set by using a permutation invariant model and the reparametrization trick [21]. To decode from a set-level embedding z to elements, we broadcast z to the same number of elements as the input. However, as the decoder is permutation equivariant, the output set will have identical elements. To break the degeneracy, we embed the particle type as well as an identification number (i.e. e_1^-, e_2^-, μ_1^+, etc.) and feed them to the decoder. This is done by using a superposition of categorical embeddings (particle type) and sinusoidal positional encoding (identification number). Finally, the reconstruction loss is defined as the Chamfer loss with the extra care that only elements with the same discrete label can be matched:

$$\mathcal{L}(\{(x_i, c_i)\}, \{(x'_j, c'_j)\}) = \frac{1}{2}\left(\sum_i \min_{j:c_i=c'_j} \|x_i - x'_j\|^2 + \sum_j \min_{i:c_i=c'_j} \|x_i - x'_j\|^2\right) \tag{1}$$

3.2 CLIP-VAE

Ref. [8, 9] used the KL-divergence loss of VAE as an anomaly score for OOD sample detection. However, in the Set-VAE paradigm, this approach proves ineffective due to the inconsistency between Chamfer loss and the assumptions used when deriving VAEs. In VAEs, to compute the evidence lower bound (ELBO), the log-likelihood is replaced with mean-squared error (MSE) loss by parametrizing $p_\theta(x|z)$ as a Gaussian distribution [21, 22]. However, in Set-VAE, due to the permutation invariance

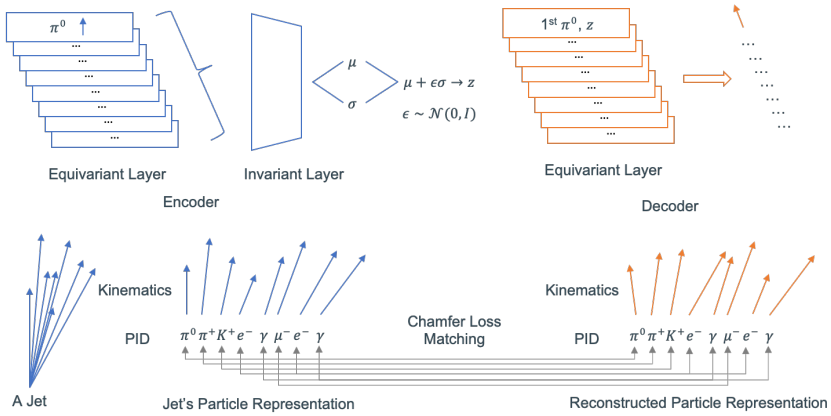


Figure 1: A sketch of the Set-VAE framework. The input to the encoder includes both kinematics and particle identification information, whereas the decoder receives particle identification and latent representation reconstructs kinematics.

of the set, the likelihood should be aggregated for all possible matches between the inputs and outputs.

$$\log p_\theta(\{x_i\}|z) \propto \log \left[\sum_{\sigma \in P(N)} \exp\left(-\sum_i^N \|x_i - \mu_{\sigma(i)}\|^2 \right) \right] + C \tag{2}$$

where the μ_i is the mean specified by the decoder (or the "reconstructed" sample) and C is the normalization constant. However, this expression is intractable and we approximate it with a lower bound which is the Chamfer loss. To ensure the Chamfer loss is a good approximation, two conditions must be met: (1) agreement with the leading term (Earth Mover's Distance) [23] and (2) exponential suppression of other terms. This holds true only when reconstructed particles closely match input particles. For poorly reconstructed samples, the reconstruction loss will be underestimated. Consequently, as the regularization of the reconstruction task becomes looser (underestimated), the KL-divergence reduces for these samples as it is a regularization. To address this issue, we propose the CLIP-VAE. In CLIP-VAE, to avoid over-regularization for the poorly reconstructed samples, we do not back-propagate the KL-divergence term for a fraction of the samples that have higher reconstruction loss. Our hypothesis is illustrated in Fig.2 and validated in Fig. 3.

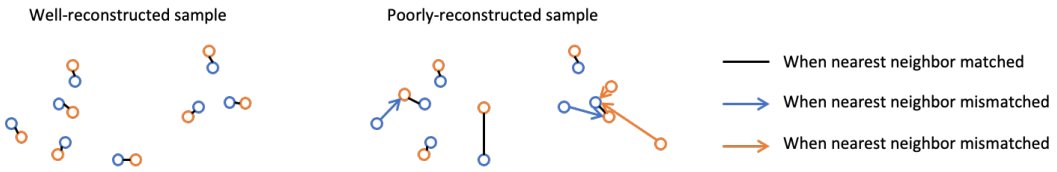
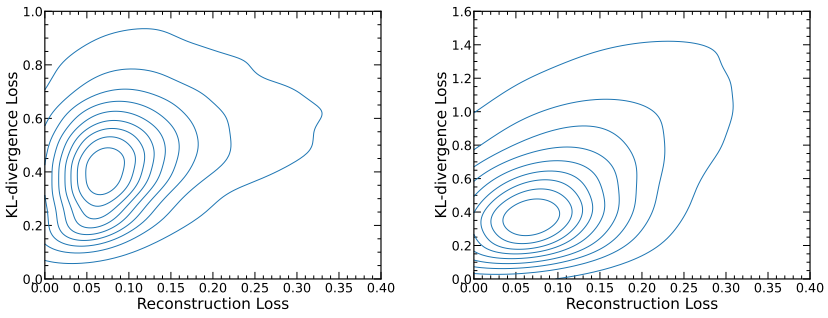


Figure 2: For well-reconstructed samples, the Chamfer loss agrees with the leading term in (2) and the contribution from all other matchings is exponentially suppressed. For poorly reconstructed samples, the Chamfer loss fails to produce a valid matching and underestimates the loss.

4 Experiments

4.1 Anomaly Detection with the `JetClass` Dataset

A *jet* is a collimated shower of particles that result from the decay and hadronization of quarks q and gluons g. The `JetClass` dataset is a large-scale jet dataset first introduced in Ref. [24]. The dataset contains 125M jets from ten different types of particles ranging from light quarks, gluons, various decay modes of top quarks (t) and Higgs bosons (H), and W and Z bosons decay to quarks (W/Z). We use it to train an anomaly detection model solely on q/g (QCD) jets and evaluate the anomaly detection performance on the other processes. Refer to [24] for additional information of



(a) Set-VAE (b) CLIP-VAE

Figure 3: The joint distribution of reconstruction loss and KL divergence loss. We can see that in the Set-VAE case, the highest reconstruction loss does not correspond to the highest KL-divergence loss, while in the CLIP-VAE case, a stronger correlation is observed.

the dataset. We train and evaluate our framework on two different architectures, the Deep Set and the Transformer. Each of the models takes as input the jet constituents' momenta and energies relative to the jet axis in cylindrical coordinate $(E_{rel}, p_{T,rel}, \Delta\eta, \Delta\phi)$ as continuous variables with appropriate shifting and scaling to ensure the magnitudes are comparable. As for discrete variables, we use the particle type information (PID). Here we consider eight types of particles: charged/neutral hadron (x3), photon, electron (x2), and muon (x2). As for the baseline, we train a logistic regression model with the n-subjettiness [25] (τ) observables to emulate a standard supervised search. The logistic regression is trained to optimize the nine types of signal simultaneously with an equal representation. The score is given by $s = \sigma(9.9 - 7.2\tau_{21} - 3.3\tau_{32} - 3.4\tau_{42})$. To make a fair comparison, since τ variables have no access to PID information, we train the transformer and deep set without PID information as an ablation study. All the numbers reported are averaged over five distinct runs. More details about the training procedure and model architecture can be found in the Github repository.

4.2 Results

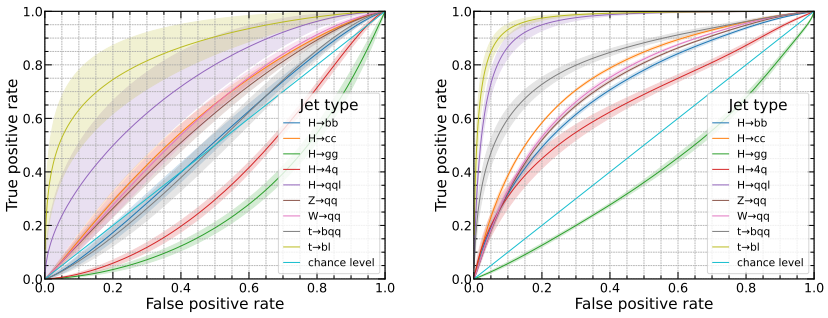
Set-VAE: Firstly, we train models with the Set-VAE paradigm and compare their performance. Since we focus on trigger applications, we report the signal efficiency $\mathrm{TPR}/(\mathrm{TPR}+\mathrm{FNR})$ of each type of jet at a background rejection $(\mathrm{TNR}+\mathrm{FPR})/\mathrm{FPR}$ of 100. As reported in Table. 1, the models

Table 1: Evaluation of models trained with Set-VAE paradigm.

	Model Profile		Signal efficiency (%) at Rej = 100								
	#params	FLOPs	$H \to 4q$	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to qql$	$W \to qq$	$Z \to qq$	$t \to bl$	$t \to bqq$
DeepSet w/ PID	**205K**	**13.8M**	5.9 ± 0.3	$\mathbf{7.1 \pm 0.8}$	$\mathbf{6.4 \pm 0.3}$	$\mathbf{0.6 \pm 0.1}$	$\mathbf{57 \pm 6}$	$\mathbf{6.7 \pm 0.3}$	$\mathbf{5.7 \pm 0.2}$	$\mathbf{77 \pm 9}$	$\mathbf{18.1 \pm 0.9}$
DeepSet w/o PID			4.2 ± 0.2	1.1 ± 0.1	2.6 ± 0.2	0.4 ± 0.2	28 ± 3	4.8 ± 0.6	3.4 ± 0.4	35 ± 7	9.1 ± 3.4
Transformer w/ PID	1.81M	171M	$\mathbf{6.3 \pm 0.7}$	6.1 ± 0.4	5.6 ± 0.4	$\mathbf{0.6 \pm 0.1}$	42 ± 3	5.2 ± 0.8	4.5 ± 0.5	54 ± 3	13.5 ± 0.7
Transformer w/o PID			3.0 ± 1.1	0.8 ± 0.3	1.9 ± 0.4	0.2 ± 0.1	15 ± 2	3.4 ± 0.3	2.4 ± 0.3	20 ± 8	4.4 ± 0.1
N-subjettiness	N/A	N/A	0.6	1.9	5.0	0.2	19	4.1	3.5	31	8.8

trained with the Set-VAE paradigm outperform the baseline of n-subjettiness logistic regression. Interestingly, we see no advantage in using transformers over deep sets. This can be explained by the fact that in an anomaly detection setting, we are not looking for a cumbersome model; instead, we want the model to be just enough expressive to encode the majority of training samples but not the OODs. More importantly, in terms of the number of operations (FLOPs), the deep set model is about thirteen times more efficient than the transformer.

CLIP-VAE Results: To illustrate the effectiveness of the approach, we first compare two models, one trained with Set-VAE and another trained with CLIP-VAE. We can see a significant difference in performance from Figure. 4 : We can see that with the KL-divergence clipping, it is possible to detect anomalies with the KL-divergence loss. Similarly, we train deep set and transformer CLIP-VAE with and without PID information. The performance is reported in Table. 2. Firstly, by comparing the number of operations to the ones reported in Table. 1, we can see that the CLIP-VAE paradigm can reduce the computational complexity by half. Furthermore, by comparing the signal efficiencies



(a) Set-VAE (b) CLIP-VAE

Figure 4: A comparison between the receiver operating curve of deep set models trained with Set-VAE and CLIP-VAE when using KL-divergence as anomaly scores. The same behavior is observed for the transformer architecture.

Table 2: Evaluation of models trained with CLIP-VAE paradigm.

	Model Profile		Signal efficiency (%) at Rej = 100								
	#params	FLOPs	$H \to 4q$	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to qql$	$W \to qq$	$Z \to qq$	$t \to bl$	$t \to bqq$
DeepSet w/ PID	**103K**	**6.95M**	5.8 ± 2.1	$\mathbf{5.1 \pm 1.2}$	5.2 ± 1.1	0.4 ± 0.1	35 ± 3	3.5 ± 0.6	3.3 ± 0.6	53 ± 8	$\mathbf{22 \pm 5}$
DeepSet w/o PID			1.0 ± 0.2	2.2 ± 0.2	$\mathbf{6.3 \pm 0.5}$	0.2 ± 0.1	19 ± 1	$\mathbf{6.0 \pm 0.6}$	$\mathbf{5.2 \pm 0.5}$	49 ± 2	4 ± 1
Transformer w/ PID	952K	78.9M	$\mathbf{6.5 \pm 0.8}$	4.0 ± 0.9	4.9 ± 0.7	$\mathbf{0.5 \pm 0.1}$	$\mathbf{43 \pm 4}$	3.8 ± 0.3	3.3 ± 0.3	$\mathbf{58 \pm 5}$	19 ± 1
Transformer w/o PID			3.1 ± 0.8	2.2 ± 0.3	5.7 ± 0.6	0.3 ± 0.1	23 ± 3	5.6 ± 0.9	5.0 ± 0.6	41 ± 3	11 ± 1
N-subjettiness	N/A	N/A	0.6	1.9	5.0	0.2	19	4.1	3.5	31	8.8

reported, we are able to conclude that CLIP-VAE has a similar anomaly detection performance. This makes CLIP-VAE very advantageous over other methods when it comes to FPGA deployment since CLIP-VAE does not require caching the inputs which can be very expensive on FPGAs.

5 Conclusion

In this paper, we present two novel architectures for jet-level anomaly detection. Firstly, the Set-VAE paradigm provides a general method to train an autoencoder for sets. We proposed a novel decoding framework for sets that can naturally produce a set of objects from a single latent representation. Furthermore, we utilize the idea of conditional autoencoder to incorporate PID information into our autoencoders. With this framework, we realized a significant improvement in terms of signal efficiency compared with the n-subjettiness methods. Secondly, we proposed the CLIP-VAE paradigm to resolve the problem that KL-divergence is not a good anomaly detector. By clipping some of the KL divergence, we are able to make the KL-divergence score a good indicator of anomalies and reduce the computational complexity by half while still retaining the high signal efficiencies seen in the Set-VAE case. We envision that the CLIP-VAE can be a very promising paradigm for deep learning algorithms for LHC triggers.

6 Broader Impact Statement

We expect that this work will stimulate further research and discussions in deep learning for anomaly detection. In particular, the Set-VAE paradigm provides a basic scalable framework for implementing particle-based autoencoders, which can serve as a basis for experimenting with different architectures. Furthermore, the CLIP-VAE paradigm enables fast anomaly detection without running the decoder, which can be very useful for anomaly detection at triggers.

7 Acknowledgement

This work is listed in Fermilab Technical Publications as FERMILAB-PUB-23-749-CMS. AG and JN are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the Department of Energy (DOE), Office of Science, Office of High Energy Physics. JN and RL are

also supported by the U.S. Department of Energy (DOE), Office of Science, Office of High Energy Physics "Designing efficient edge AI with physics phenomena" Project (DE-FOA-0002705). JN is also supported by the DOE Office of Science, Office of Advanced Scientific Computing Research under the "Real-time Data Reduction Codesign at the Extreme Edge for Science" Project (DE-FOA-0002501). This work was supported in part by the AI2050 program at Schmidt Futures (Grant G-23-64934).

References

[1] A Tapper and Darin Acosta. CMS Technical Design Report for the Level-1 Trigger Upgrade. Technical report, 2013. Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch.

[2] Tejinder Virdee, Achille Petrilli, and Austin Ball. CMS High Level Trigger. Technical report, CERN, Geneva, 2007. revised version submitted on 2007-10-19 16:57:09.

[3] Manfred Jeitler, A Taurok, H Bergauer, C Deldicque, J Erö, M Ghete, P Glaser, K Kastner, I Mikulec, T Nöbauer, et al. The level-1 global trigger for the cms experiment at lhc. *Journal of Instrumentation*, 2(01):P01006, 2007.

[4] Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

[5] Benjamin Nachman and David Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7):075042, 2020.

[6] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[7] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer Thompson. Qcd or what? *SciPost Physics*, 6(3):030, 2019.

[8] Ekaterina Govorkova, Ema Puljak, Thea Aarrestad, Thomas James, Vladimir Loncar, Maurizio Pierini, Adrian Alan Pol, Nicolò Ghielmetti, Maksymilian Graczyk, Sioni Summers, et al. Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 mhz at the large hadron collider. *Nature Machine Intelligence*, 4(2):154–161, 2022.

[9] Bryan Ostdiek. Deep set auto encoders for anomaly detection in particle physics. *SciPost Physics*, 12(1):045, 2022.

[10] Zichun Hao, Raghav Kansal, Javier Duarte, and Nadezda Chernyavskaya. Lorentz group equivariant autoencoders. *The European Physical Journal C*, 83(6):485, 2023.

[11] Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018.

[12] Gregor Kasieczka, Benjamin Nachman, David Shih, Oz Amram, Anders Andreassen, Kees Benkendorfer, Blaz Bortolato, Gustaaf Brooijmans, Florencia Canelli, Jack H Collins, et al. The lhc olympics 2020 a community challenge for anomaly detection in high energy physics. *Reports on progress in physics*, 84(12):124201, 2021.

[13] M Farina, Y Nakai, and D Shih. Searching for new physics with deep autoencoders (2018). *arXiv preprint arXiv:1808.08992*.

[14] Oliver Atkinson, Akanksha Bhardwaj, Christoph Englert, Vishal S Ngairangbam, and Michael Spannowsky. Anomaly detection with convolutional graph neural networks. *Journal of High Energy Physics*, 2021(8):1–19, 2021.

[15] Thorben Finke, Michael Krämer, Alessandro Morandini, Alexander Mück, and Ivan Oleksiyuk. Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High Energy Physics*, 2021(6):1–32, 2021.

[16] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[17] Horace Pan and Risi Kondor. Permutation equivariant layers for higher order interactions. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5987–6001. PMLR, 28–30 Mar 2022.

[18] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[22] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[23] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image, 2016.

[24] Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging, 2022.

[25] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with n-subjettiness. *Journal of High Energy Physics*, 2011(3), mar 2011.