Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

Volume 42 (2023), Number 7

CoC-GAN: Generate More and More Sets of Points for Image Synthesis

Abstract

Image generation tasks, a traditional domain within the field of computer vision, frequently incorporate either Convolutional Neural Networks (CNN) or Transformer architectures for feature extraction in models such as Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), and Diffusion Models. While convolution and attention mechanisms are often employed, they are not fundamentally indispensable. In this paper, we introduce a novel model premised on the Context Clustering (COC) method. This approach abandons both convolution and attention, relying instead entirely on clustering techniques paired with a multi-layer perception (MLP) generative countermeasure network. We introduce this architecture as the Context Clustering Generative Adversarial Network (COC-GAN). This clustering method, recently proposed for target detection tasks, interprets an image as a collection of unstructured points and extracts features via a simplified clustering algorithm. This algorithm offers a unique perspective in feature extraction. Building upon this foundation, we propose the incorporation of a Point Increaser module. This module is responsible for generating supplementary points for clustering, which we subsequently integrate within the Generative Adversarial Network (GAN) paradigm. Empirical evaluations reveal that our convolution-free and attention-agnostic GAN demonstrates remarkable performance on MNIST dataset. Given the feasibility demonstrated herein, future investigations can further explore the model's capacity to generate images of superior resolution and enhanced interpretability with increased efficiency.

1. Introduction

In recent developments, a pioneering approach known as Context Cluster (CoC) [MZW*23] has emerged, challenging the prevalent use of convolutional networks and attention mechanisms, which have traditionally dominated the field of Computer Vision (CV). This innovative methodology, introduced by Ma et al., underscores the critical role of image interpretation in the feature extraction process. In their work, they operationalize their philosophy by employing similarity measures to cluster sets of points. Despite their notable success in various dense-to-sparse prediction tasks, such as image classification, an intriguing question arises: can this intuitive method tackle the inverse problem, specifically in the domain of generative tasks? Current challenges in the realm of Generative Adversarial Networks (GAN) [GPM*20], including training instability and weak interpretability, persist unresolved. To address these concerns, we introduce a novel GAN architecture, underpinned by CoC, to evaluate its feasibility and potential impact on the field.

In the pursuit of refining the paradigm for implementing GANs, the goal has consistently been the evolution of superior generalization capabilities for image data. Goodfellow et al. introduced the intriguing concept of GANs that intuitively learn the prior for sampling from a statistical perspective. Models such as DCGAN [RMC16] and TransGAN [JCW21] were among the pioneers in applying convolution and transformer paradigms within GANs, leveraging the inherent image-friendly nature of convolutional networks and the scale-friendly attributes of attention-based methods, respectively.

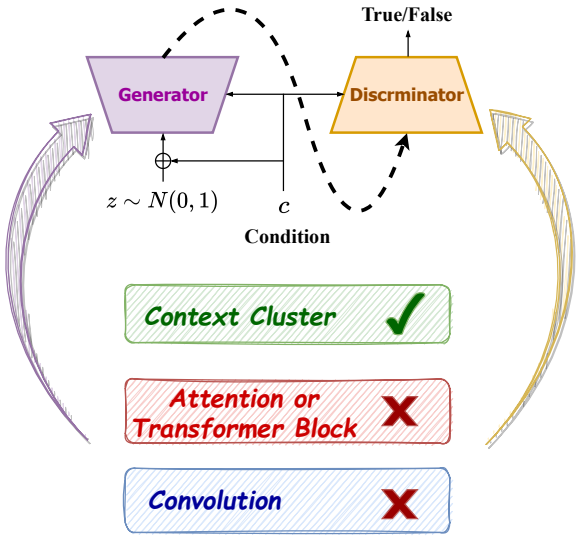

Figure 1: *Our motivation: Make an convolution & attention-free GAN by Context Cluster.*

Despite the advancements, these methods do not fundamentally address the inherent instability of training GANs, and issues surrounding weak interpretability persist, particularly when a high level of reliability is required. This paper introduces a unique GAN architecture, named CoC-GAN, which eschews the conventional use of convolution and attention mechanisms found in earlier GANs. Our motivation is illustrated in Fig. 1. For the discriminator in our proposed GAN, we straightforwardly apply the CoC. The central concept involves viewing the generative process as an augmentative operation on sets of points at different stages, facilitated by the Point Increaser module we have designed. This can be viewed as the functional opposite of the point reduction operation in the original CoC. We undertake empirical experiments to validate our method, encompassing both unconditional and conditional image synthesis.

The contributions of this paper are three-folded:

- Our study introduces CoC-GAN, a Context Cluster-based architecture for GAN image synthesis, eliminating the need for convolution and attention mechanisms.
- We propose a 'Point Increaser' block that utilizes a simplified clustering method and Multi-Layer Perception (MLP) for image generation from unordered point sets.
- We validate that the context clustering method significantly enhances the interpretability of CoC-GAN's image generation process.

2. Related Works

2.1. Paradigms in Computer Vision

Deep learning has profoundly changed the algorithm and benchmark in the Computer Vision area. To the best of our knowledge, convolution nets (ConvNets) and attention mechanisms make the main contribution to this progress. ConvNets have shown their ability to capture translation invariance of image data from LeNet [LBBH98]. Classic works in the deep learning era like AlexNet [KSH12], ResNet [HZRS16] etc. shows superiority of fewer parameters by its local parameter sharing property compared to the simple multi-layer perception. The drawback of ConvNets is the ability to generalize knowledge in the large-scale dataset. The emergence of ViT [DBK*21] marked attention-based Transformers gradually solve the problem of inductive bias in ConvNets, due to its bigger capacity and adaptive global interaction. Therefore more methods like SwinTransformer [LLC*21] explore utilizing both advantages of ConvNets and attention. However, researchers are also curious about methods other than these two mainstream ways. Before occurence of CoC, ViG [HWG*22] and MLP-mixer [THK*21] respectively use graph structure and pure multi-layer perception in a well-organized structure to realize this goal. And CoC uses clustering ideas to aggregate similar points in the image for the interaction of pixels and super-pixels , which fuse the information for simple downstream tasks. Our work is further exploring its traits in harder generative tasks.

2.2. Architecture in GAN

Adversarial method GAN is the most wide-used method for generative tasks like image synthesis tasks, which simply let adversaration between Generator and Discriminator learn the prior distribution. The first practice by Goodfellow et al. adopts the form of MLP, which shows its ability in small datasets like MNIST [Den12]. To Enhance the specific image generation ability, DC-GAN [RMC16] symbolizes the placement by better ConvNets, the significant effect was shown in different benchmarks. TransGAN [JCW21] is the first work to expand the strong fitting ability of the attention-based transformer for GANs. Same to the application combining ConvNets and attention in a wide CV area, GigaGAN [KZZ*23] also realizes this. In comparison, our novelty is that we are the first to attempt to apply Context clusting (CoC) method in GANs for generative tasks.

3. Methodology

3.1. Overview

Compared to the original CoC model, the most two prominent characteristics are that we propose the Point Increaser module to generate dense points and we employ the CoC module in both generator and discriminator of GAN. This convolution & attention-free architecture is demonstrated in Fig. 2.

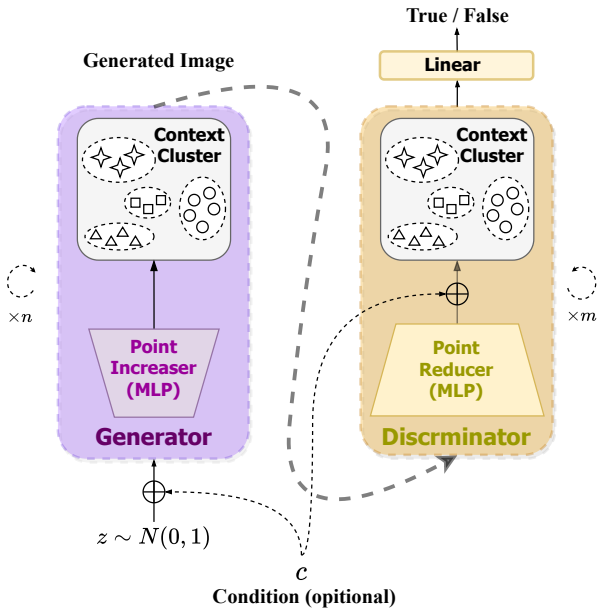

Figure 2: *Architecture of our CoC-GAN*

3.2. Point Increaser

In contrast to the downsampling effects produced by the Point Reducer Module in the original CoC model, our task necessitates the creation of a module with an antithetical function for generation tasks. It is evident that we can merely invert the Point Reducer by applying an analogous linear layer for upsampling. For the sake of simplicity and stable implementation, distinct input channels are transformed in a piecemeal manner via a linear layer specifically designed for upsampling.

3.3. CoC Module

We utilize original Context Clustering modules within our architecture, where the input is a set of feature points denoted by $\mathbf{P} \in \mathbf{R^{n \times d}}$. Initially, we project $\mathbf{P}$ into $\mathbf{P_s}$. Subsequently, we propose c centers in the space by computing the average of the nearest k points. Following this, based on the calculated pairwise cosine similarity matrix $\mathbf{S} \in \mathbf{R_{c \times n}}$, we allocate each point to the cluster represented by the center most similar to it. The aforementioned procedures constitute the clustering stage, setting the groundwork for the subsequent feature dispatching and aggregation stage.

In the feature aggregation stage, all points are adaptively aggregated based on their similarity. Given $\mathbf{m}$ points in a cluster and their similarity to the center of this cluster, we transform them into a value space to obtain $\mathbf{P_v} \in \mathbf{R^{m \times d'}}$ where $\mathbf{d'}$ denotes the dimension. The aggregated feature is generated through the following procedure:

$$g = \frac{1}{C\left(v_c + \sum_{i=1}^{m} \operatorname{sig}\left(\alpha s_i + \beta\right) * v_i\right)} \tag{1}$$

$$\text{s.t., } C = 1 + \sum_{i=1}^{m} \operatorname{sig}\left(\alpha s_i + \beta\right). \tag{2}$$

$\mathbf{v_c}$ is the center in value space, α, β are the learnable parameters for controling similarity and $\operatorname{sig}(\cdot)$ is the sigmoid function which scale the similarity in $(\mathbf{0}, \mathbf{1})$, and v_i refers to i-th point in $\mathbf{P_v}$.

Regarding the feature dispatching stage, the aggregated feature $\mathbf{g}$ is dispatched to every point in the cluster by the similarity. To be specific, each $\mathbf{p_i}$ is updated by following equation:

$$p_i' = p_i + \mathrm{FC}\left(\operatorname{sig}\left(\alpha s_i + \beta\right) * g\right) \tag{3}$$

The fully-connected(FC) layer takes the transform from $\mathbf{d'}$ to $\mathbf{d}$ into account.

The original CoC architecture also incorporates multi-head computation, which is used to transform $\mathbf{P_v}$ into $\mathbf{P_s}$ across $\mathbf{h}$ heads. A Fully-Connected (FC) layer is deployed to fuse the concatenated outcomes derived from each head.

Drawing from the work of Xu et al., as well as our own implementation, the CoC module enhances the global feature through clustering. In the case of downsampling, learnable parameters are trained to proficiently capture useful global features during aggregation and dimension-increasing transformations. However, our generator is designed to validate this concept, or its variation, in an upsampling scenario.

3.4. Whole Model in Unconditional/Conditional Generation Task

As illustrated in Fig.2, our novel approach employs CoC to create a convolution and attention-free architecture for GANs. Hence, we mimic the first GAN [GPM*20] for unconditional generation tasks, which involves the use of a Point Increaser in the generator to incrementally increase the number of points from a random seed. The discriminator, in this setup, is a straightforward implementation of the CoC model. The conditional case presents a more intricate scenario. We adhere to the design of GAN-INT-CLS [RAY*16] to ensure simplicity and conciseness in our approach. In this setup, half of the random seeds in the unconditional variant are replaced by features extracted from the condition. These extracted features also serve as the conditional input for the discriminator. Fully-Connected (FC) layers are employed to adjust the dimensionality of the condition to match the input.

3.5. Training objective

Loss is also a significant factor in validating. We regard two general kinds of loss for learning: hard ones and easy ones. To be representative, we select the most simple loss in [GPM*20] called min-max game as the former:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \tag{4}$$

$$\mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{5}$$

and its conditional version is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x \mid y)] + \tag{6}$$

$$\mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z \mid y)))] \tag{7}$$

Here, G and D are the discriminator and generator respectively, and x, y, z is the true sample, condition, and random seed separately. Given the latter should be helpful to stabilize training, we choose the loss in WGAN [**?**]. Although it is just the non-log version Eq. 4 (conditional case is Eq. 6) it mathmatically conforms to optimizing the best Wasserstein distance which effectively promotes the stableness of training.

4. Experiments

4.1. Task & Dataset

Our objective is to synthesize images in either unconditional or conditional scenarios. As a preliminary experiment, we seek to validate the feasibility of CoC for generative tasks, for which we apply our approach to the MNIST dataset [Den12].

The MNIST dataset is a well-recognized collection, encompassing 60,000 training images and 10,000 testing images. These images, sized 28 x 28 and single-channeled, represent handwritten digits. In the conditional scenario, the digit value serves as the condition guiding the generation process.

4.2. Settings of Training

The specifics of our model structure are detailed in Tab. 3.3. Notably, $sample_r$ denotes the rate of point increase or reduction in the upsampling or downsampling module, *regions* indicates the number of regions partitioned for our CoC operation, $local_centers$ signifies the quantity of proposed centers for clustering in each region, *heads* and $head_dim$ represent the number of heads in the CoC module and their corresponding dimensions, and mlp_r is the rate of dimension increase and reduction (to the same dimension before entering the modules) of the MLP, pertaining to transformation to and reduction from value space. For the WGAN version, the sigmoid activation in the discriminator is omitted.

To train the model, we employ the Adam optimizer [**?**] with a

Table 1: *Hyperparameter we select for concrete CoC-GAN in our tasks.*

	Points	Block	CoC-GAN:Generator		CoC-GAN:Discrminator	
Stage S1	1 * 1/28 * 28	Point Increaser/Reducer	$upsample\_r = 1$ $dim = 128$	$\times 1$	$downsample\_r = 2$ $dim = 3$	$\times 1$
	1 * 1/14 * 14	Context Cluster Blocks	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 16$ $mlp\_r. = 4$ $dim = 128$	$\times 1$	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 3$	$\times 1$
Stage S2	1 * 1/14 * 14	Point Increaser/Reducer	$upsample\_r = 2$ $dim = 64$	$\times 1$	$downsample\_r = 2$ $dim = 16$	$\times 1$
	2 * 2/7 * 7	Context Cluster Blocks	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 64$	$\times 2$	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 16$	$\times 2$
Stage S3	2 * 2/7 * 7	Point Increaser/Reducer	$upsample\_r = 7$ $dim = 32$	$\times 1$	$downsample\_r = 1$ $dim = 32$	$\times 1$
	4 * 4/ 1 * 1	Context Cluster Blocks	$regions = 1*1$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 32$	$\times 2$	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 64$	$\times 2$
Stage S4	4 * 4/ 1 * 1	Point Increaser/Reducer	$upwnsample\_r = 7$ $dim = 16$	$\times 1$	$downsample\_r = 1$ $dim = 128$	$\times 1$
	784	Context Cluster Blocks	$regions = 1*1$ $local\_centers = 1$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 3$	$\times 3$	$regions = 2*2$ $local\_centers = 4$ $heads = 4$ $head\_dim = 32$ $mlp\_r. = 8$ $dim = 128$	$\times 2$

Table 2: *The comparsion between our CoC-GAN and basic GAN model.*

model	DCGAN [RMC16]	WGAN [**?**]	WGAN-GP	CoC-GAN(**Uncondtional, Ours**)	CoC-GAN(**Conditional, Ours**)
IS	2.15+-0.039	2.39+-0.025	2.38+-0.041	**2.45+-0.032**	2.34+-0.034
FID	46.36	27.12	30.10	40.03	40.23

learning rate of $\mathbf{2e-4}$ and a batch size of 256. For the WGAN version, we utilize RMStrop [**?**] with identical learning rate and batch size. The cosine annealing decay policy is implemented to adjust the learning rate, and gradient cut-offs are used in the WGAN version. In all trials, the loss converges within 50 training epochs. Consequently, in addition to reviewing generated samples, we compute the best FID and IS to evaluate the model's performance.

4.3. Unconditional Generative Results on MNIST

We present the performance results for the unconditional scenario in Tab.3.4, with corresponding generated samples displayed in Fig.3. It is evident that the implementation of the WGAN version significantly enhances the performance of the CoC-GAN. Intuitively, the CoC-GAN seems well-suited to generative tasks, albeit requiring further refinement in model design and training settings, such as the incorporation of WGAN. Furthermore, these results also validate the practicality of the Point Increaser module in incrementally generating points.

4.4. Conditional Generative Results on MNIST

The performance results for the conditional scenario are displayed in Tab.3.4, with corresponding generated samples presented in Fig.4. These results reinforce the conclusion reached in Sec. 4.3 and further emphasize the complexity and instability inherent to training the CoC structure. A comparison with the unconditional

Figure 3: *Results on MNIST of unconditional case.*

Figure 4: *Results on MNIST of conditional case.*

scenario reveals that incorporating a condition into the generation process presents unexpected challenges.

4.5. The Interpretability of CoC-GAN

Considering the gifted interpretability of CoC-GAN, we take COC-GAN in unconditional case to embody this property. Left part of Fig. 5 visualizes generating process of CoC-GAN, the right part. Note that different color refer to different cluster in upsampling/downsampling stage.

5. Conclusion

In this paper, we introduce CoC-GAN, a novel approach that leverages the pure context clustering technique in both the generator and discriminator. In contrast to the point reducer used in downsampling processes, we propose the point increaser module to cater to the demand for generating an increasing number of points. Through empirical experiments conducted in both conditional and unconditional scenarios, we demonstrate that the combination of context clustering with the point increaser module can successfully tackle the generation task, albeit with careful architecture design and training settings. It is important to note that training the context clustering structure presents additional challenges compared to conventional MLP and convolution-based methods. While our model's performance may not surpass mainstream approaches, this

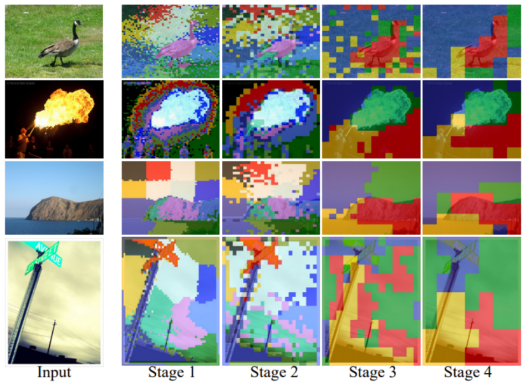

Figure 5: *Illustration of our visualization about different inference stage(both in generator and discriminator), in which the left is the generated image on MNIST domain.*

study represents the first attempt to apply context clustering in the context of image generation. Building upon the feasibility established in this work, future research can explore the interpretability advantages of the context clustering technique and optimize its efficiency further.

References

[DBK*21] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (2021), OpenReview.net. 2

[Den12] DENG L.: The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine 29*, 6 (2012), 141–142. 2, 3

[GPM*20] GOODFELLOW I. J., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A. C., BENGIO Y.: Generative adversarial networks. *Commun. ACM 63*, 11 (2020), 139–144. 1, 3

[HWG*22] HAN K., WANG Y., GUO J., TANG Y., WU E.: Vision GNN: an image is worth graph of nodes. *CoRR abs/2206.00272* (2022). 2

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016* (2016), IEEE Computer Society, pp. 770–778. 2

[JCW21] JIANG Y., CHANG S., WANG Z.: Transgan: Two pure transformers can make one strong gan, and that can scale up. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (2021), Ranzato M., Beygelzimer A., Dauphin Y. N., Liang P., Vaughan J. W., (Eds.), pp. 14745–14758. 1, 2

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States* (2012), Bartlett P. L., Pereira F. C. N., Burges C. J. C., Bottou L., Weinberger K. Q., (Eds.), pp. 1106–1114. 2

[KZZ*23] KANG M., ZHU J., ZHANG R., PARK J., SHECHTMAN E., PARIS S., PARK T.: Scaling up gans for text-to-image synthesis. *CoRR abs/2303.05511* (2023). 2

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. 2

[LLC*21] LIU Z., LIN Y., CAO Y., HU H., WEI Y., ZHANG Z., LIN S., GUO B.: Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021* (2021), IEEE, pp. 9992–10002. 2

[MZW*23] MA X., ZHOU Y., WANG H., QIN C., SUN B., LIU C., FU Y.: Image as set of points. *CoRR abs/2303.01494* (2023). 1

[RAY*16] REED S., AKATA Z., YAN X., LOGESWARAN L., SCHIELE B., LEE H.: Generative adversarial text to image synthesis. In *International conference on machine learning* (2016), PMLR, pp. 1060–1069. 3

[RMC16] RADFORD A., METZ L., CHINTALA S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (2016), Bengio Y., LeCun Y., (Eds.). 1, 2, 4

[THK*21] TOLSTIKHIN I. O., HOULSBY N., KOLESNIKOV A., BEYER L., ZHAI X., UNTERTHINER T., YUNG J., STEINER A., KEYSERS D., USZKOREIT J., LUCIC M., DOSOVITSKIY A.: Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (2021), Ranzato M., Beygelzimer A., Dauphin Y. N., Liang P., Vaughan J. W., (Eds.), pp. 24261–24272. 2