

EXPLORATION INTO TRANSLATION-EQUIVARIANT IMAGE QUANTIZATION

Woncheol Shin[†] Gyubok Lee[†] Jiyoung Lee[†] Eunyi Lyou[‡] Joonseok Lee^{‡*} Edward Choi[†]

[†] Graduate School of AI, KAIST, Daejeon, South Korea

[‡] Graduate School of Data Science, Seoul National University, Seoul, South Korea

* Google Research, Mountain View, California, USA

ABSTRACT

This is an exploratory study that discovers the current image quantization (vector quantization) do not satisfy translation equivariance in the quantized space due to aliasing. Instead of focusing on anti-aliasing, we propose a simple yet effective way to achieve translation-equivariant image quantization by enforcing orthogonality among the codebook embeddings. To explore the advantages of translation-equivariant image quantization, we conduct three proof-of-concept experiments with a carefully controlled dataset: (1) text-to-image generation, where the quantized image indices are the target to predict, (2) image-to-text generation, where the quantized image indices are given as a condition, (3) using a smaller training set to analyze sample efficiency. From the strictly controlled experiments, we empirically verify that the translation-equivariant image quantizer improves not only sample efficiency but also the accuracy over VQGAN up to +11.9% in text-to-image generation and +3.9% in image-to-text generation.

Index Terms— Vector Quantization, Translation Equivariance, Aliasing, Text-Image Generation, Sample Efficiency

1. INTRODUCTION

Vector quantization [1] has gained popularity in multimodal learning problems such as text-to-image generation. In particular, several works [2, 3, 4] demonstrated impressive text-conditioned image generation, only using image quantization and Transformer [5]. Methods that utilize vector quantization in image downstream tasks usually take a two-stage approach: first learning to quantize images, then solving a downstream task with the quantized representation. In the first step, an image quantizer (*e.g.*, VQVAE [1] or VQGAN [6]) encodes an image as a sequence of codebook indices using codebook embeddings (Stage 1). Then, the resulting indices are given as input to downstream image modeling tasks (Stage 2). By modeling images in this fashion, the input indices of the down-

This work was supported by the KAIST-NAVER Hyper-Creative AI Center and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, No.2022-0-00984) and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945, NRF-2021H1D3A2A03038607) funded by the Korea government (MSIT).

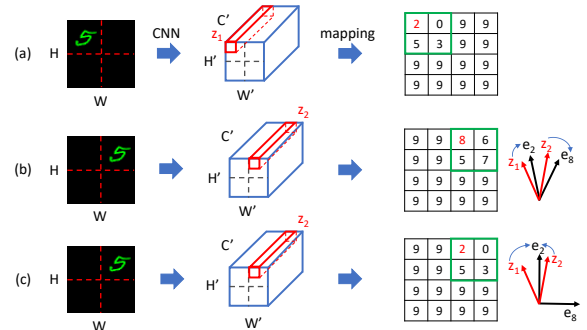


Fig. 1: (a) Image quantization before translation of the top-left 5; (b) Broken translation equivariance in the quantized space after the translation of 5; (c) Perfect translation equivariance in the quantized space with our orthogonal codebook embeddings. z_1, z_2 are the corresponding features in terms of shifted location. By enforcing orthogonality in the quantized space as in (c), a slight deviation due to aliasing can be ignored, and the same codebook index (*e.g.* 2) is given to z_1, z_2 .

stream task can be treated in the same manner as token indices in text. This discretization of images allows us to model a relatively shorter sequence of image representations, and to even jointly handle textual information more naturally.

Although the idea of representing an image with quantized indices has been appreciated, the representations learned by existing methods turn out to be far from ideal yet. Specifically, we focus on *translation equivariance*, a property for an image quantizer to represent semantically same objects with the same indices regardless of its location within the image. A toy example in Figure 1 (b) indicates, however, that the state-of-the-art image quantization method, VQGAN [6], breaks this property even for the most trivial setting (locating the exactly same image patch at the pixel level).

Referring to the literature, the cause of the broken translation equivariance is due to aliasing, caused by downsampling operations (*e.g.*, strided convolution, maxpooling) [7, 8]. In other words, the following can hold:

$$z_2 = z_1 + \epsilon, \quad (1)$$

where ϵ is a noise caused by aliasing, and z_1 and z_2 are the

corresponding features in terms of the shifted location (Figure 1). In an effort to reduce ϵ , one of the anti-aliasing methods in both signal processing and CNNs is to apply a low-pass filter during downsampling [9, 10, 8, 11]. In deep neural networks, it is challenging to enforce ϵ to be completely zero in the feature map space, as complete anti-aliasing still requires careful manipulation of filters. However, when it comes to image quantization, where images are represented as codebook indices, ϵ no longer needs to be zero as long as the features are mapped to the same codebook index. This implies that translation equivariance in the ‘quantized space’ can be achieved by a different approach than enforcing translation equivariance in the ‘feature map space.’

In this paper, we propose a simple yet effective way to achieve translation-equivariant image quantization by enforcing orthogonality among the codebook embeddings. Then, with carefully controlled datasets, we study how it affects the downstream image modeling tasks. For the Stage 2, we use the following three settings: (1) text-to-image generation, where the quantized image indices are the target to predict, (2) image-to-text generation, where the quantized image indices are given as a condition, (3) using a smaller training set to analyze sample efficiency. We report advantages and limitations of translation-equivariant image quantization in all three settings from thorough analysis of the experimental results.

Our contributions include: (1) To the best of our knowledge, this is the first work to explore the problem of *translation equivariance in the quantized image space*; (2) Instead of focusing on anti-aliasing, we take a direct approach to achieve translation equivariance in the quantized space by regularizing orthogonality in the codebook embedding vectors; and (3) We show that a translation-equivariant image quantizer improves not only sample efficiency but also the accuracy of text-to-image and image-to-text generation on text-augmented MNIST [12] by up to +11.9% and +3.9%, respectively. We discuss insights discovered from our analysis on the behavior of the quantized representations.

2. RELATED WORK

2.1. Image Quantization

Image quantization or vector quantization of images is an efficient encoding method that represents images in the discrete latent space via an encoder and decoder. Oord et al. [1] first applied this idea to the generative tasks, which can fully leverage the power of both encoder and decoder. In this manner, downstream models (generators) only need to handle a much shorter sequence of image tokens, compared to handling each pixel. Thanks to the other advantage of jointly handling image and text tokens more naturally, this idea quickly spread to multimodal tasks such as text-to-image generation. For example, DALL-E [2] and CogView [3] receive both the text and image tokens as a single stream, similar to GPT [13], and calculate self-attention. Parti [4] adopts an encoder-decoder

Transformer [5] structure, where the encoder takes text tokens as inputs and the decoder autoregressively predicts discrete image tokens. These models demonstrated the effectiveness of the quantization in the text-image multimodal problem.

Esser et al. [6] proposed VQGAN, an effective extension of VQVAE [1] by introducing an adversarial loss and perceptual loss. The addition of two objectives provides image representations that result in a sharper and detailed reconstruction of images. More recently, Yu et al. [14] further boosted the efficiency and reconstruction quality of image quantizers, replacing CNNs with the ViT architecture [15]. Since Convolutional Neural Network (CNN) is one of the most widely used and well-established networks, we utilize a CNN-based quantizer, VQGAN, as our base image quantizer in this work.

2.2. Translation Invariance and Equivariance

Translation invariance requires the output unchanged by the shifts in the input, while translation equivariance is a mapping which, when the input is shifted, leads to a shifted output. In other words, translation invariance is about the final representation after Global Average Pooling (GAP) in CNNs, and translation equivariance is about the feature map before GAP. A fundamental approach to handle translation invariance and equivariance is anti-aliasing. Simoncelli et al. [7] first formalized ‘shiftability’ and related it to aliasing. Since then, careful calibration of sampling rate according to the Nyquist sampling theorem [16] or applying a low-pass filter has been a natural choice to avoid aliasing in downsampling.

It is only recently that deep learning has started to explore translation invariance and equivariance [17, 8, 18, 19]. Zhang [8] applied a low pass filter between a stride-one operator and naive subsampling to improve translation equivariance in the latent feature map space, but Azulay and Weiss [17] pointed out that nonlinearity such as ReLU still hinders anti-aliasing even with low-pass filtering. Karras et al. [11] proposed an ideal sampling method and nonlinearity to avoid aliasing in the image generation task. In a similar motivation to theirs, we investigate translation equivariance in the generation task but pay special attention to the vector-quantized space and multimodal learning problem.

3. METHOD

3.1. Translation Equivariance in Quantized Space

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ be an image of size $H \times W$ with RGB channels. A CNN \mathcal{F} takes \mathbf{x} as input, and produces a feature map $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times C'}$. Let Q be a quantization operation, and $Q(\mathcal{F}(\mathbf{x})) = \mathcal{F}^q(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times C'}$ be the quantized feature map. \mathcal{F} is called *translation equivariant in the quantized space* if

$$\mathcal{F}^q(T_1(\mathbf{x})) = T_2(\mathcal{F}^q(\mathbf{x})) \quad (2)$$

where T_1 is a translation operation in the image space and T_2 is the translation operation in the quantized space corre-

sponding to T_1 . In other words, it is enough for the translation equivariance in the quantized space to have the T_2 relationship between the quantized code indices of $\mathcal{F}(T_1(\mathbf{x}))$ and $\mathcal{F}(\mathbf{x})$, even if $\mathcal{F}(T_1(\mathbf{x})) \neq T_2(\mathcal{F}(\mathbf{x}))$. Note that we can safely focus only on the relationship between the code indices, because the decoder-only Transformer [5] in the following stage takes the input image in the form of code indices, not image feature maps or codebook embedding vectors.

3.2. Image Quantizer (TE-VQGAN)

VQGAN [6] consists of a CNN-based encoder \mathcal{F} , a CNN-based decoder \mathcal{G} , and codebook embeddings $e \in \mathbb{R}^{C' \times K}$, where K is the codebook size and C' is the number of feature channels. \mathcal{F} gets an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ and produces a feature map $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times C'}$. Let a fiber of the feature map whose spatial coordinate is (h', w') be $\mathcal{F}(\mathbf{x})_{(h', w')} \in \mathbb{R}^{C'}$, where $(h', w') \in [1, H'] \times [1, W']$ and $h', w' \in \mathbb{Z}$. $\mathcal{F}(\mathbf{x})_{(h', w')}$ is assigned to the closest embedding vector $e_k \in \mathbb{R}^{C'}$ based on L_2 distance:

$$\mathcal{F}^q(\mathbf{x})_{(h', w')} = e_k, k = \arg \min_j \|\mathcal{F}(\mathbf{x})_{(h', w')} - e_j\|_2^2. \quad (3)$$

The decoder \mathcal{G} gets $\mathcal{F}^q(\mathbf{x})$ and produces the reconstructed image, $\hat{\mathbf{x}} = \mathcal{G}(\mathcal{F}^q(\mathbf{x})) \in \mathbb{R}^{H \times W \times 3}$. \mathcal{F} , \mathcal{G} , and e are jointly trained by minimizing $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$.

As mentioned in Section 1, if we interpret alias as a noise due to downsampling, the following relationship may hold:

$$\mathcal{F}(T_1(\mathbf{x}))_{(h', w')} = T_2(\mathcal{F}(\mathbf{x}))_{(h', w')} + \epsilon. \quad (4)$$

Here, it is a non-trivial task to make ϵ zero. However, as mentioned in Section 3.1, as long as we are only interested in the image code indices, we can focus only on assigning the same code e_k to both $\mathcal{F}(T_1(\mathbf{x}))_{(h', w')}$ and $T_2(\mathcal{F}(\mathbf{x}))_{(h', w')}$ rather than trying to actually make $\epsilon = 0$. Surprisingly, this can be achieved by simply enforcing the orthogonal structure in the quantized space; that is, by regularizing the embedding vectors e_k to be orthogonal to each other. The intuition behind this is well illustrated in Figure 1. In the unregularized quantized space, feature maps z_1 and z_2 that slightly differ due to alias are mapped to respective embedding vectors e_2 and e_8 , respectively. In the orthogonal quantized space, however, every embedding vector is orthogonal to one another, and it is easier to ignore small noise due to alias when z_1 and z_2 are mapped to the same embedding vector e_2 .

We add the following regularization term to the loss function of VQGAN to enforce orthogonality among the codebook embeddings:

$$\mathcal{L}_{\text{REG}}(e) = \lambda \frac{1}{K^2} \|\ell_2(e)^\top \ell_2(e) - I_K\|_F^2 \quad (5)$$

where $I_K \in \mathbb{R}^{K \times K}$ and $\ell_2(e) \in \mathbb{R}^{C' \times K}$ denotes the identity matrix and an L_2 -normalized code embedding along the first dimension, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. We set λ to 10 in all experiments provided in Section 4.

(a) Examples of the test set.

(b) Constraints on the dataset.

Fig. 2: Dataset examples and constraints.

4. EXPERIMENTS

Data Construction. To carefully explore the advantages of the translation-equivariant image quantizer, we created MNIST-based image-caption pairs, as shown in Figure 2a. First, we sample one to four digits from the original dataset, and randomly place them in four quadrants of the 64×64 space. Note that the original samples have the 28×28 resolution, so we pad every sample with 2 pixels. Then, we created image captions with syntactic variations. To perform zero-shot-like generation, we applied different constraints between the train and test set, as shown in Figure 2b. For example, white 0 on the upper left does not exist in the train set, while all 0's in the test set are white and placed in upper left. In total, we generated 300K samples for train, 3,000 for validation, and 3,000 for test set.

Experiment Setting. Since aliasing occurs during downsampling, we train image quantizers with four different downsampling methods as baselines, including with Blurpool [8], as blurring before subsampling might be sufficient to anti-alias: StridedConv (SC), ConvBlurpool (CB), MaxPool (MP), and MaxBlurpool (MB). Then in Stage 2, similar to DALL-E, The decoder-only Transformer is trained by next-token prediction changing the quantizers pretrained in Stage 1. We measure the performance of a generator when TE-VQGAN (Ours) is used and vanilla VQGANs with various downsampling methods are used. Experiments in Section 4 were conducted three times with random initialization.

4.1. Text-to-Image Generation

We measure the digit accuracy when the generated image has one (Quad1), two (Quad2), three (Quad3), and four digits (Quad4). Each case has approximately 600 images. As an example, in Quad2, given a text 'the lower right is blue nine, and the white two is on the upper left.', a rule-based caption parser will convert the text into $\{(LR, B, 9), (UL, W, 2)\}$. With the parsed output, we crop the corresponding locations, lower right and upper left, from the generated image. Then, we measure the accuracy of digits on these two cropped

Table 1: Semantic accuracy of text-to-image and image-to-text generation.

Model	Text-to-Image				Image-to-Text			
	Quad1	Quad2	Quad3	Quad4	Quad1	Quad2	Quad3	Quad4
MB	0.724 (0.033)	0.697 (0.066)	0.635 (0.067)	0.620 (0.077)	0.587 (0.032)	0.507 (0.021)	0.433 (0.013)	0.378 (0.020)
MP	0.713 (0.007)	0.703 (0.015)	0.669 (0.017)	0.656 (0.028)	0.561 (0.042)	0.500 (0.038)	0.412 (0.034)	0.360 (0.039)
CB	0.753 (0.037)	0.741 (0.034)	0.682 (0.043)	0.649 (0.055)	0.563 (0.041)	0.455 (0.056)	0.362 (0.046)	0.309 (0.043)
SC	0.812 (0.032)	0.806 (0.030)	0.790 (0.038)	0.761 (0.051)	0.786 (0.052)	0.692 (0.063)	0.595 (0.086)	0.542 (0.106)
Ours	0.931 (0.033)	0.913 (0.032)	0.890 (0.048)	0.868 (0.059)	0.825 (0.037)	0.713 (0.066)	0.621 (0.061)	0.554 (0.065)

images using a pretrained classifier with a accuracy of 99.5%.

We compare Ours, which uses the SC downsampling method with the orthogonality regularization, with four baselines. As shown in Table 1, our method brings significant performance improvements compared to the baselines (MB, MP, CB, SC), clearly demonstrating the importance of translation equivariance in the quantized space.

4.2. Image-to-Text Generation

Given a 64×64 image, the generator synthesizes a caption describing it. We measure its accuracy of digit identity using a simple rule-based parser. As shown in Table 1, the gap between Ours and baselines in this task ($I \rightarrow T$) is smaller than that of $T \rightarrow I$, and the actual performance of $I \rightarrow T$ is significantly lower than that of $T \rightarrow I$ for all methods. We conjecture that this difference comes from how texts and images are evaluated. Since an image is classified as a whole, miss-predicting a code index or two might have little effect. On the other hand, miss-predicting a specific text token, such as the digit token, can have a catastrophic impact.

4.3. Sample Efficiency

Table 2: Digit accuracy of generated images in Quad1 varying the size of the train set.

Size	VQGAN	TE-VQGAN (Ours)
300K	0.812 (0.032)	0.931 (0.033)
100K	0.739 (0.037)	0.919 (0.022)
Diff	0.073	0.012

We demonstrate that the translation-equivariant image quantization could improve the generator’s ability to synthesize the shifted images, which are never shown at the training phase. This is possible due to our model’s consistent use of code indices for a given digit regardless of its position in the image. From this result, we posit that the Stage 2 generator would be able to learn the relationship between image and text with an even smaller dataset. To verify this, we conduct an experiment where we train two generators on a small (100K) and a large (300K) training set. We then measure the two generators’ digit accuracy of $T \rightarrow I$ generation using the test set.

As we hypothesize, Table 2 shows that the generator with TE-VQGAN is more robust to the reduced training set size than the one with vanilla VQGAN. From this, we claim that translation-equivariance could be a potential solution to the scalability issue that modern generative models suffer from.

5. DISCUSSION

Table 3: Code usage, reconstruction loss, and perceptual loss.

Dataset	Code usage		Recon loss		Perceptual loss	
	VQGAN	TE-VQGAN	VQGAN	TE-VQGAN	VQGAN	TE-VQGAN
MNIST	252	14	0.0079	0.0084	0.0036	0.0048
FASHION	250	18	0.0150	0.0187	0.0108	0.0198

The first person to divide rainbow into 7 colors was Sir Isaac Newton. It is known that up to 207 colors of rainbows can be distinguished by human eye, but Newton expressed rainbows with only seven ‘essences’. Our methodology is similar in spirit: using the orthogonality regularization, only a few essence codes are used as described in Figure 1.

We count the number of codes that are used at least once in Stage 1. Surprisingly, TE-VQGAN uses only 14 and 18 codes to represent colored digits and colored fashion items¹, as seen in Table 3. In other words, the 14 and 18 ‘essence’ codes are sufficient to represent each dataset.

One possible limitation of using fewer essence codes is sacrifice in image fidelity. Drawing a rainbow with only 7 colors will certainly miss its finer hue. Table 3 shows the reconstruction loss $\|\mathbf{x} - \hat{\mathbf{x}}\|_1$ and the perceptual loss [21] between \mathbf{x} and $\hat{\mathbf{x}}$. This empirically confirms that, even though orthogonal regularization miss some visual information, its impact is marginal. Since our ultimate aim lies in multimodal generation rather than reconstructing images, the ability to understand the semantics of the given condition is more important than reconstructing every bit of fine details. Therefore, considering the result from Section 4, we can conclude that despite some minor drawbacks, we can achieve the more desirable goal of multimodal generation.

6. CONCLUSION

This work is a exploratory study that first explore *translation equivariance in the image quantized space* and propose a simple yet effective way to achieve it. Our proof-of-concept experiments demonstrate that our method improves image-text multimodal generation performance and sample efficiency. Future research directions may include experimentation with real world datasets and diverse set of downstream tasks.

¹We additionally conducted this experiment using Fashion-MNIST [20].

7. REFERENCES

- [1] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, “Zero-shot text-to-image generation,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.
- [3] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang, “Cogview: Mastering text-to-image generation via transformers,” in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 19822–19835.
- [4] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al., “Scaling autoregressive models for content-rich text-to-image generation,” *arXiv preprint arXiv:2206.10789*, 2022.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [6] Patrick Esser, Robin Rombach, and Bjorn Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.
- [7] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger, “Shiftable multiscale transforms,” *IEEE transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, 1992.
- [8] Richard Zhang, “Making convolutional networks shift-invariant again,” in *International conference on machine learning*. PMLR, 2019, pp. 7324–7334.
- [9] Alan V Oppenheim, *Discrete-time signal processing*, Pearson Education India, 1999.
- [10] Rafael C Gonzalez, *Digital image processing*, Pearson education india, 2009.
- [11] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, “Alias-free generative adversarial networks,” in *Advances in Neural Information Processing Systems*, 2021, vol. 34.
- [12] Li Deng, “The mnist database of handwritten digit images for machine learning research,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
- [13] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, 2020, vol. 33.
- [14] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu, “Vector-quantized image modeling with improved VQGAN,” in *International Conference on Learning Representations*, 2022.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
- [16] Harry Nyquist, “Certain topics in telegraph transmission theory,” *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.
- [17] Aharon Azulay and Yair Weiss, “Why do deep convolutional networks generalize so poorly to small image transformations?,” *Journal of Machine Learning Research*, vol. 20, pp. 1–25, 2019.
- [18] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee, “Delving deeper into anti-aliasing in convnets,” in *BMVC*, 2020.
- [19] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, and Ross Goroshin, “Impact of aliasing on generalization in deep convolutional networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10529–10538.
- [20] Han Xiao, Kashif Rasul, and Roland Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.