Causal-DFQ: Causality Guided Data-free Network Quantization

Yuzhang Shang[1,2], Bingxin Xu[1], Gaowen Liu[2], Ramana Rao Kompella[2], Yan Yan[1*]
[1]Illinois Institute of Technology, [2]Cisco Research
{yshang4, bxu21}@hawk.iit.edu, {yuzshang, gaoliu, rkompell}@cisco.com, yyan34@iit.edu

Abstract

Model quantization, which aims to compress deep neural networks and accelerate inference speed, has greatly facilitated the development of cumbersome models on mobile and edge devices. There is a common assumption in quantization methods from prior works that training data is available. In practice, however, this assumption cannot always be fulfilled due to reasons of privacy and security, rendering these methods inapplicable in real-life situations. Thus, data-free network quantization has recently received significant attention in neural network compression. Causal reasoning provides an intuitive way to model causal relationships to eliminate data-driven correlations, making causality an essential component of analyzing data-free problems. However, causal formulations of data-free quantization are inadequate in the literature. To bridge this gap, we construct a causal graph to model the data generation and discrepancy reduction between the pre-trained and quantized models. Inspired by the causal understanding, we propose the Causality-guided Data-free Network Quantization method, Causal-DFQ, to eliminate the reliance on data via approaching an equilibrium of causality-driven intervened distributions. Specifically, we design a content-style-decoupled generator, synthesizing images conditioned on the relevant and irrelevant factors; then we propose a discrepancy reduction loss to align the intervened distributions of the pre-trained and quantized models. It is worth noting that our work is the first attempt towards introducing causality to data-free quantization problem. Extensive experiments demonstrate the efficacy of Causal-DFQ. The code is available at Causal-DFQ.

1. Introduction

There have been significant advances in deep learning models in the fields of computer vision [9, 14] and natural language processing [33, 45]. To accommodate the increasing demand for equipping cumbersome models on resource-constrained edge devices, researchers have proposed several network quantization methods [18, 57], in which high-precision parameters are converted into low-precision ones. To mitigate the performance degradation induced by model quantization, fine-tuning approaches are extensively studied to optimize quantized models on the full training datasets [19,41,42,49]. However, original training data is sometimes inaccessible in real-world situations due to the privacy and security concerns. A patient's electronic health record, for instance, is typically inaccessible because the information contained is private. Hence, the fine-tuning methods requiring training data are no longer applicable in such real-life scenarios.

To address this issue, researchers have proposed data-free quantization to quantize models without requiring access to real data [1, 2, 24, 43, 51, 54]. For example, ZeroQ [2] is proposed to generate 'optimal' fake data, which learns an input data distribution to best match the batch normalization statistics of the FP32 model. Nevertheless, most data-free quantization methods attempt to reconstruct the original data from the pre-trained model utilizing prior statistical distribution information of the underlying data, such as BNS [2,51,52], Dirichlet distribution [28] and category information [3]. However, those methods ignore a powerful tool in the human cognition, *i.e.*, causal reasoning, which commonly aids humans in learning without relying upon data collection. Human cognitive systems are immune to the data deficiency because humans are more sensitive to causal relations than data-driven statistical associations [10, 55]. Using causal language, causal reasoning can extract causal relationship from the pre-trained models and ignore irrelevant factors by interventions [34].

There are two significant challenges that need to be overcome before causality can be introduced to eliminate the reliance on data during the quantized model training. First, constructing an informative causal graph is the fundamental premise for causal reasoning [30,34], but how causal graphs should be constructed in a data-free situation is still inadequate in the literature. Second, using causal language to formalize data generation and network alignment is the key to connecting causality with data-free quantization, but it also remains unsolved. These two challenges are the fundamental obstacles that prevent us from employing causality

[*] Corresponding author

in data-free quantization.

To address these challenges, we construct a causal graph to model the data-free quantization process, including data-generation and discrepancy reduction mechanisms, where the irrelevant factors in the pre-trained models are taken into consideration. Based on the causal graph, we propose a novel Causality-Guided Data Free Network Quantization method, ***Causal-DFQ***, to remove the reliance on data during quantized model training. Specifically, we design a content-style-decoupled generator, synthesizing images conditioned on the relevant and irrelevant factors (content and style variables). Then we propose a discrepancy reduction loss to align the intervened distributions of the outputs from pre-trained and quantized models.

Overall, the contributions of this paper are four-fold: **(i)** We provide a causal perspective on data-free quantization, which is the first attempt towards using causality to facilitate data-free network compression; **(ii)** To leverage causality to facilitate data-free quantization, we construct a causal graph to model data generation process and discrepancy reduction process in data-free quantization mechanism; **(iii)** We propose a novel quantization method called Causality Guided Data-free Network Quantization, *Causal-DFQ*, in which we generate fake images conditioned on style and content variables, and align style-intervened distributions of pre-trained and quantized models. **(iv)** Extensive experiments demonstrate that the proposed method can significantly improve the performance of data-free low-bit models. Importantly, it is the first method where data-free fine-tuned models outperform the models fine-tuned with data on the ImageNet.

2. Related Work

Data-free Network Compression. Although model compression has become a hot topic recently, compressing model without training data still is a challenge. As pioneers, [44] initially devise a channel pruning method without original training data. And then a large number of data-free (DF) or zero-shot compression methods were proposed, *e.g.* DF quantization [1, 2, 24, 51, 54], DF factorization [27] and DF knowledge distillation [3,8,25]. Especially for DF quantization, recent work [1, 2, 24, 51, 54] go further to data-free quantization, which requires neither training nor validation data for quantization. Most of the data-free KD methods attempt to reconstruct the original data from the pre-trained model utilizing prior information about the underlying data statistical distribution, such as BNS [2, 51, 52], Dirichlet distribution [28] and category information [3]. However, all existing methods overlook causal reasoning, a powerful tool for humans to cognize even in situations where data are inaccessible.

Causal Reasoning. One core purpose of causal reasoning is to pursue the causal effect of interventions, contributing to achieving the desired objective. Recent work shows the benefits of introducing causality into machine learning from various aspects [38]. After the deep connections of causal systems and the concept of exogeneity having been successfully implemented in social science, such as in Economics and Genetics [30], Schölkopf *et al.* [37] originally develop a technique, named independence mechanisms via introducing causal mechanisms to independently separate the exogenous and endogenous variables *w.r.t.* specific tasks in the field of machine learning [38].

However, thanks to the unique nature of data-free quantization, our data generation process is steerable, unlike previous works. Thus, we can design a content-style-decoupled generator where both content and style variables are accessible in the causal graph. Then we can easily implement do-calculus [31] for causal reasoning.

3. Method

In this section, we elaborate on the methodology of Causality-guided Data-Free Quantization, named *Causal-DFQ*. Firstly, we review the idea of network quantization and the general framework of data-free compression. Secondly, we construct the causal graph model for the data-free network compression, which is adopted as a theoretical tool to bridge causality with data-free quantization. Thirdly, based on the causal graph, we observe that there is a unique property of data-free compression, where the data variable is completely accessible; thus we design a generator to synthesize images conditioned on style and content images for training quantized models. Next, we focus on the optimization formulation that converts the causal task into an optimizable problem. Finally, we discuss the potential insights for the *Causal-DFQ*. Note that we only elaborate on the key derivations in this section due to the space limitation. Detailed discussions, technical theorems, and implementation details in Codes can be found in the supplemental materials.

3.1. Preliminary

Here, we revisit the basic ideas of network quantization and data-free network compression.

Network Quantization. Network quantization is a popular technique for compressing neural networks. The quantization function is the key to train a neural network with low-precision weights and activations. The most common quantization function is called uniform quantization function, which is pioneerly proposed in [57]. The uniform quantization function $q(\cdot)$ for k-bit quantization is defined as follows:

$$q(v) = \mathrm{round}(L \cdot (v - Z)), \tag{1}$$

where v denotes a scalar value (full-precision, float32), L is the scaling factor, and Z is the zero point in float32. According to whether the parameter Z is zero, uniform quantization can be categorized into symmetric quantization and asymmetric quantization. In our work, we use symmetric quantization, *i.e.*, $Z = 0$, and then S is written as follows:

$$S = \frac{2^{k-1} - 1}{\max(|x_f|)}, \tag{2}$$

where x_f is the full-precision numbers.

Data-Free Compression. The key of most network compression methods [11, 18] including quantization is to reduce the discrepancy $\mathcal{D}$ between the pre-trained full-precision model $f(\cdot;\theta_P)$ and the quantized model with low-precision weights $f(\cdot;\theta_Q)$ through optimizing θ_Q. The idea of discrepancy reduction can be formalized as follows:

$$f(\cdot;\theta_Q)^\star = \min_{\theta_Q} \mathcal{D}(f(\cdot;\theta_P), f(\cdot;\theta_Q)). \tag{3}$$

This discrepancy reduction module can be considered as a knowledge distillation mechanism for aligning model $f(\cdot;\theta_P)$ and $f(\cdot;\theta_Q)$. Consequently, the integral framework of the data-free quantization can be considered as the incorporation of a generator and knowledge distillation [15, 40], and the core idea is to reconstruct some samples from full-precision models to fine-tune quantized models [23, 24]. Therefore, to achieve the goal of data-free compression via causality, we modify the existing framework from the perspectives of generator and discrepancy reduction.

3.2. Causal Graph of Data-Free Compression

Humans can perform causal reasoning, an essential ability that makes humans learn differently from machine learning algorithms. The superiority of causal reasoning endows humans with the ability to identify causal relationships. This allows them to ignore irrelevant factors that are not causally related to the targeted task and removes the reliance on collecting data for learning [34, 38, 53, 55]. Contrary to this, neural networks are normally trained based on data-driven correlation. In other words, neural networks do not have the ability to distinguish causal relationships. In the absence of this ability, irrelevant factors in data are overfitted, further resulting in the overreliance of networks on data. For instance, in a car recognition case, road background is a data-driven irrelevant factor yet cannot reflect the causality $w.r.t.$ the targeted task, *i.e.*, human can recognize a car with causal reasoning even if it is not on the road. Moreover, data even are unavailable in our data-free case. Therefore, we desire to incorporate causal reasoning to remove the reliance on data, *i.e.*, the pre-trained model guides the training of the quantized model via causality in a data-free manner. Before performing causal reasoning to guide the training in a data-free manner, we need to construct a causal graph since causal graphs are the key to formulating causal reasoning [34, 55]. In the context of data-free quantization, we desire a causal graph by which the distributions of the outputs of pre-trained and quantized models can be included. Besides, the graph is expected to reflect the impact of irrelevant factors on these two output distributions, and then we are able to align the distributions. Specifically, we investigate the difference in irrelevant factors between these two distributions and enforce the quantized models to focus on the relevant factors. Consequently, this encourages the quantized model to learn via causal reasoning.

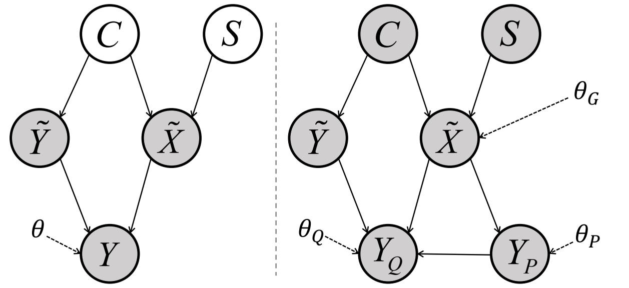


Figure 1. **Left**: Causal graph of the ideal data generation and model learning process. **Right**: Causal graph of the data-free quantization process. Each node represents a random variable, and shallow ones indicate observable variables, where C, S, $\tilde{X}$, $\tilde{Y}$, Y_P, Y_Q, θ_G, θ_P, θ_Q are content variable, style variable, generated data, generated label, distilled label, output label, parameters of the generator, parameters of the pre-trained model, parameters of the quantized, respectively.

There are two general approaches to building a causal graph of the targeted learning mechanism. One approach is to use causal structure learning to infer causal graphs [30, 34, 38], but it is challenging to apply this approach to high-dimensional data. Using external knowledge to construct causal graphs is another approach [38, 46, 55]. As automatically learning a precise causal graph is out of scope for this work, external human knowledge of the data generation process is employed to construct the causal graph. Here, we aim to construct a causal graph to model the data-free quantization process, including data-generation and discrepancy reduction mechanisms, where the irrelevant factors in the pre-trained models are also considered.

Specifically, we construct a causal graph $\mathcal{G}$ to formalize the general idea of data-free quantization process including an image generation mechanism and a discrepancy reduction mechanism to allow the pre-trained model $f(\cdot;\theta_P)$ to guide the training of the quantized model $f(\cdot;\theta_Q)$. In the previous studies [34, 38, 55], even though there is a number of different causes of natural data, researchers ideally and effectively divide all the causes into two categories for simplicity. We follow the existing work, and group content-related causes into one category, called content variable C. The rest causes, *i.e.*, irrelevant factors, are grouped into another category, called style variable S, which is content-independent, *i.e.*, $S \perp\!\!\!\perp C$. This implies that $C \rightarrow \tilde{X} \leftarrow S$

and $C \rightarrow \tilde{Y}$. Then the generated data are fed into the pre-trained model and quantized model. Under the supervision of the output of pre-trained model Y_P and the generated label $\tilde{Y}$, we obtain the output of quantized model Y_Q. The causal graph is shown in Fig. 1. Based on the causal graph, we first use a structural causal model [32] to represent the data generating mechanism:

$$\tilde{X} := \mathcal{M}(S, C, \theta_G), \tag{4}$$

where θ_G is the parameters of generator.

After formulating the process of obtaining the generated data, we expect to define valid interventions and the corresponding intervention distributions [30, 38]. Defining valid interventions is equivalent to determining which variables or mechanisms in the causal graph can be intervened. The common practice is to utilize the independence mechanism [37] to *construct probabilistic relations in causal reasoning* and *discover the irrelevant factors as the intervened variable*. This practice has been proven to be an effective way to realize causality reasoning by previous work [17, 26, 36, 37], in which the conditional (intervened) distribution does not change under the interventions on irrelevant variable (*i.e.*, S). Theoretically, for an image generation mechanism (ideally even collected natural images are included), $P(\tilde{Y} \mid C)$ is invariant to S. Therefore, we claim C as a representation of invariant content of data *w.r.t.* the $\tilde{Y}$ under interventions I on style domain $\mathcal{S}$ as shown Fig. 1(left), and the relationship can be mathematically denoted as:

$$P^{do(S=i_l)}(\tilde{Y} \mid C) = P^{do(S=i_k)}(\tilde{Y} \mid C) \quad \forall i_l, i_k \in \mathcal{I}, \tag{5}$$

where i_l and i_k form a pair of interventions in the domain of interventions $\mathcal{I}$, and $P^{do(S=i_l)}$ stands for the distribution under intervention i_l on $\mathcal{S}$ [30]. In the data-free compression literature, we also desire to access the intervened distributions of the outputs of the pre-trained model and quantized model and then derive a computationally reachable equilibrium between them. *Because our data generation process is steerable, unlike the fixed datasets collected from natural distributions, we design a content-style-decoupled generator where both content and style variables are accessible in the causal graph.*

Therefore, based on the analysis of Eq.4, 5 and the above-mentioned nature of the data-free mechanism, the desirable equilibrium in data-free quantization can be formulated as follows: $\forall l, k \in \{1, 2, \cdots, M\}$,

$$P^{do(S=i_l)}(Y_P \mid f(\tilde{X}; \theta_Q)) = P^{do(S=i_k)}(Y_P \mid f(\tilde{X}; \theta_Q)), \tag{6}$$

which can be reformed as follows: **Targeted Causal Equilibrium:**

$$\begin{aligned} &P^{do(S=i_l)}(f(\tilde{X}; \theta_P) \mid f(\tilde{X}; \theta_Q)) \\ =&P^{do(S=i_k)}(f(\tilde{X}; \theta_P) \mid f(\tilde{X}; \theta_Q)), \end{aligned} \tag{7}$$

Algorithm 1 Pseudo code of Content-Style-Decoupled Generator in a PyTorch-like style.

```
def  generator(S, C): # generator
    input = torch.mul(Embedding(C), S)
     # style & content fusion
    x = conv_blocks(input)
     # generate images via conv layers
    return x
content = torch.randint(0, class_number,
(batch_size),)) # define content
style = torch.randn(batch_size, latent_dim)
     # define style
generated_x = generator(style, content)
     # generate images based on C and S
```

where M is the number of interventions in style domain $\mathcal{S}$, $f(\cdot; \theta_P)$ and $f(\cdot; \theta_Q)$ are the pre-trained and quantized model with parameters θ_P and θ_Q, respectively. Straightforwardly, we desire the distribution, $P(f(\tilde{X}; \theta_P) \mid f(\tilde{X}; \theta_Q))$ to be invariant over style variable change.

3.3. Content-Style-Decoupled Generator

Here, we present the design of the content-style-decoupled generator, rendering accessibility to the content and style variables. Structural Equation Modeling (SEM) [32, 48] is a primary causal model, which is originally proposed to apply explicit causal interpretations to regression equations based on direct and indirect effects of observed variables in the fields of Genetics and Economics [30]. SEM has two main components: the structural model showing potential causal dependencies between endogenous and exogenous variables and the measurement model showing the relations between latent variables and their indicators. SEM aims to obtain an informative representation of some observable output.

Inspired by the concepts of SEM, we naturally introduce the above two variables into the data-free scenarios and expect to enforce the outputs of quantized models exclusively correlated with the content variable in the view of causal reasoning. To realize the goal, we generalize three fundamental assumptions of SEM into the literature on data-free compression [26, 31, 37]. The generalized assumptions can be interpreted as follows: **(i)** The data are generated from content variables C representing factors inside the model and style variables S (irrelevant factors outside the model) for targeted tasks. **(ii)** Only variable C is relevant for the model output, *i.e.*, content dominates the model performance. **(iii)** Content and style are causally independent, *i.e.*, style changes are content-preserving.

Based on the above assumptions, to achieve the goal of directly performing interventions on the style domain and building the equilibrium of the intervened distributions as

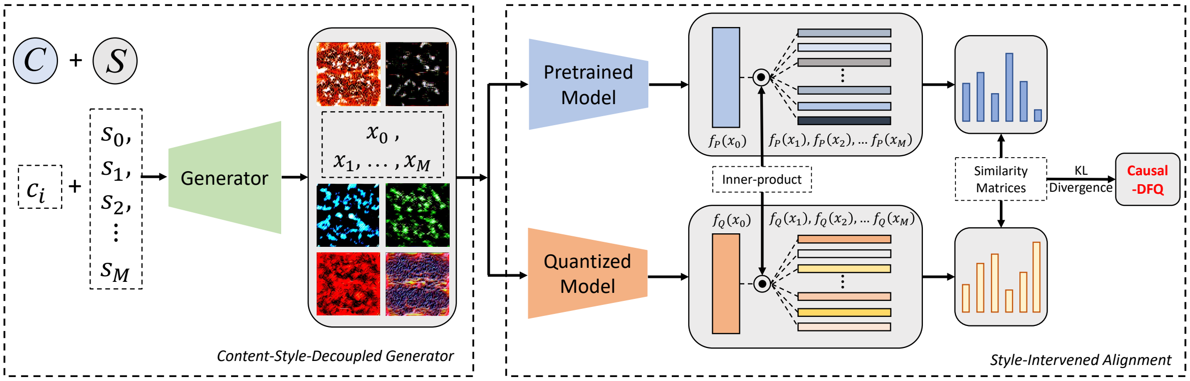


Figure 2. **Overview of the pipeline.** To eliminate the reliance on data and utilize the causality in the discrepancy reduction stage, we disentangle the invariant content C and semantics-irrelevant style S in the view of causality. We propose a **Content-Style-Decoupled Generator** to synthesize fake images conditioned on the independent content and style variables. Follow by the generator, we design ***Causal-DFQ*** loss to achieve knowledge exclusively based on content by intervening with the style variable. In particular, we use KL-divergence to minimize the distance between conditional distributions (similarity matrices, calculated in a contrastive way) of pre-trained and quantized models.

presented in Eq. 7, we design an image generator that can synthesize fake data conditioned on independent style and content variables. We call this generator a content-style-decoupled generator. Specifically, we assign every to-be-generated sample a content label and a style noise; then we feed this pair of content and style into the generator network to generate the sample. In this way, for each sample of the following discrepancy reduction process, we can access its style variable and perform interventions by keeping its content labels consistent and adjusting its style noise.

Here, we give a straightforward explanation of how our generator produces fake images conditioned on content and style variables (Algorithm 1). First, the integer function, `content = randint()` (function of randomly generating non-nagetive integer) generates the pseudo label based on the number of classes of the real dataset, which can be interpreted as a content variable for each generated image. Note that using the number of classes does not imply information leakage and is still within our data-free scenarios, as we can acquire the number of classes via accessing the pre-trained model's classification head rather than accessing the labels. And the Gaussian noise generation function, `style = randn()` can assign a Gaussian noise to style variable. By pairing the content and style and then feeding them to a generator, we can synthesize fake data conditioned on the content and style. In this way, we can directly manipulate the style variable. More details can be found in the *codes in the Supplemental Materials*.

3.4. Style-Intervened Discrepancy Reduction

After accessing the style variable in the data-free quantization mechanism, the only remaining problem is to achieve the equilibrium of the intervened distributions as derived in Eq. 7. We maintain the invariance under interventions via a regularization term to address this. The optimization problem is formalized as follows:

$$\begin{aligned} \min \mathop{\mathbb{E}}_{X \in \mathcal{D}} \mathop{\mathbb{E}}_{\{i_{lk}, i_{qt}\}} \Big[& \mathcal{L}_{i_{lk}}(f(X;\theta_P), f(X;\theta_Q)) \\ & + \mathcal{L}_{i_{qt}}(f(X;\theta_P), f(X;\theta_Q)) \Big]. \\ s.t. \quad KL\Big[& P^{do(S=i_{lk})}(f(X;\theta_P) \mid f(X;\theta_Q)), \\ & P^{do(S=i_{qt})}(f(X;\theta_P) \mid f(X;\theta_Q)) \Big] \leq \tau \end{aligned} \tag{8}$$

where $i_{lk} \triangleq i_l \times i_k \sim \mathcal{I} \times \mathcal{I}$ stands for a pair of interventions, $\mathcal{L}$ is the vanilla alignment loss, and $KL(\cdot,\cdot)$ is the KL-divergence. τ is a small threshold to adjust the similarity between two distributions. Any distance measure on distributions can be used in place of the KL divergence such as cross-entropy, since we only expect the intervened distributions $P^{do(S=i_{lk})}(f(X;\theta_P) \mid f(X;\theta_Q))$ and $P^{do(S=i_{qt})}(f(X;\theta_P) \mid f(X;\theta_Q))$ to be similar. In practice, we define the output representations of pre-trained and quantized models (*i.e.*, $f(X;\theta_P)$ and $f(X;\theta_Q)$) at the penultimate layer.

How to approach the conditional distribution under interventions $P^{do(S=i_{lk})}(f(\tilde{X};\theta_P) \mid f(\tilde{X};\theta_Q))$ becomes the key problem. To estimate the distribution, we introduce the noise-contrastive estimation (NCE) [13, 16]. Specifically, we take pairs of points (x_i, x_j) to compute similarity scores and use pairs of intervention i_{lk} to perform a style intervention. Given a batch of samples $\{x_i\}, i \in \{1, 2, \cdots, N\}$, the conditional probability of the pair can be estimated as follows:

$$\begin{aligned} & P^{do(S=i_{lk})}(f(X;\theta_P) \mid f(X;\theta_Q)) \\ & \propto h(f(x_j^{S=i_l};\theta_P), f(x_i^{S=i_k};\theta_Q)), \end{aligned} \tag{9}$$

in which h is the function to measure the similarity between the representations of the pre-trained model $f(x_j^{S=i_l};\theta_P)$

and the one from quantized model $f(x_i^{S=i_k};\theta_Q)$. Using this function to estimate the conditional distribution is originally proposed in NCE [13], also called the critic in contrastive learning [16]. It is defined as below:

$$h(\mathbf{x},\mathbf{y}) = \exp\left(\frac{<g(\mathbf{x}), g(\mathbf{y})>}{\beta}\right), \tag{10}$$

where β is the temperature to adjust degree of concentration, and g is a fully-connected network [13].

Combining all the equations, we obtain the optimizable objective function as follows: $\mathcal{L}_{\textit{Causal-DFQ}} =$

$$\begin{aligned} &\mathop{\mathbb{E}}_{X\in\mathcal{D}}\ \mathop{\mathbb{E}}_{\{i_{lk},i_{qt}\}}\Big[\mathcal{L}_{i_{lk}}(f(X;\theta_P), f(X;\theta_Q)) \\ &\qquad + \mathcal{L}_{i_{qt}}(f(X;\theta_P), f(X;\theta_Q))\Big] \\ &+ \sum_{i_{lk}}\sum_{i_{qt}} KL\Big[P^{do(S=i_{lk})}(f(X;\theta_P) \mid f(X;\theta_Q)), \\ &\qquad P^{do(S=i_{qt})}(f(X;\theta_P) \mid f(X;\theta_Q))\Big]. \end{aligned} \tag{11}$$

Concretely, the probability of a pair of samples in the conditional distribution $P^{do(S=i_{lk})}(f(X;\theta_P) \mid f(X;\theta_Q))$ can be approximated by the critic function as follows [13, 16, 26]:

$$\begin{aligned} &P^{do(S=i_{lk})}(f(x_j;\theta_P) \mid f(x_i;\theta_Q)) \\ =&\frac{h(f(x_j^{S=i_l};\theta_P), f(x_i^{S=i_k};\theta_Q))}{\sum_{i_{lk}} h(f(x_j^{S=i_l};\theta_P), f(x_i^{S=i_k};\theta_Q))}. \end{aligned} \tag{12}$$

Overall Loss Function. Taking into account all the above discussions, the *Causal-DFQ* loss can be calculated with differentiability and the overall loss function can be written as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{vanilla} + \lambda \cdot \mathcal{L}_{\textit{Causal-DFQ}}, \tag{13}$$

where $\mathcal{L}_{vanilla}$ is the objective from the vanilla data-free quantization loss, and λ is the parameter to balance the targeted task and the distillation task. In practice, we adopt and modify the codebase of GDFQ [51] to achieve our causality-based data-free quantization baseline, thus more details about the $\mathcal{L}_{vanilla}$ can be found in GDFQ.

3.5. Discussions on *Causal-DFQ*

Besides the derivation originated from the perspective of causality, we would like to give a straight-forward explanation of *Causal-DFQ*. Combining Eq.11 and Eq.12, we can observe that our method minimize the distributional distance between $P^{do(S=i_{lk})}(f(x_j;\theta_P) \mid f(x_i;\theta_Q))$ and $P^{do(S=i_{qt})}(f(x_j;\theta_P) \mid f(x_i;\theta_Q))$. Specifically, with the critic function to estimate the conditional distribution, $P^{do(S=i_{lk})}(f(x_j;\theta_P) \mid f(x_i;\theta_Q)$ acts as the similarity matrix between generated images with same content, *i.e.*, a series of differences among samples with the same content and different styles. Finally, the similarity matrices of pre-trained and quantized models are aligned with KL-divergence as shown in Fig.2.

Difference with RELIC [26]. From the perspective of causality, the most related work is RELIC [26] which acts as a regularizer in self-supervised learning via the independence mechanisms [34] to encourage networks to be invariant to different augmentations of the same instance. This self-supervised learning method also constructs a causal graph to model the data generation process. However, the focus of this work is on the content invariant property using data augmentations to stimulate inaccessible interventions [55], which varies from our data-free work, *Causal-DFQ*. Specifically, our work is different from RELIC (and most of the previous causality-guided computer vision models, such as [4, 26, 50]) for two significant reasons. Firstly, there is a unique nature in data-free scenarios where both the content and style variable are accessible, and thus we do not need to stimulate the interventions on the style domain. Secondly, the derived equilibrium is different where we focus on the distributions of outputs of pre-trained and quantized models. Detailed differences between our data-free approach and previous works are discussed in Appendix.

4. Experiments

4.1. Experimental Setup

Datasets. We validate the *Causal-DFQ* on four well-known data sets including CIFAR-10, CIFAR-100 [21], ImageNet [6] for recognition, and PASCAL VOC 2012 [7] for detection. More details about the datasets are in Supplemental Materials.

Baselines. To evaluate the effectiveness and advantages of our proposed method, we compared it with both data-free fine-tuning methods and post-training quantization methods. The baselines are presented as follows. **FP32**: the full-precision pre-trained model. **FT**: we use real training data instead of fake data to fine-tune the quantized model by minimizing L2. **ZeroQ** [2]: a data-free post-training quantization method. **DFQ** [27]: a post-training quantization method uses a weight equalization scheme to remove outliers in both weights and activations. **ZAQ** [24]: It is a fine-tuning method by optimizing the quantized models in an adversarial learning way. **DSG** [54]: It is a fine-tuning method where the diversity of generated data is enhanced. **GDFQ** [51]: It is also a fine-tuning method for recovering fake data via a conditional generator. **SQuant** [12] and **IntraQ** [56] are recently SoTA. Note that our code is modified from the code of GDFQ.

Implementation Details. On CIFAR, we optimize the generator and quantized model using Adam [20] and SGD with Nesterov [29] respectively, where the momentum term and

Table 1. **Comparisons on ImageNet**. We quantize both the weights and activations of the models to *6-bits* and report the top-1 accuracy.

Dataset	Model	Real Data		Data Free					
		FP32	FT	ZeroQ [2]	GDFQ [51]	DSG [54]	SQuant [12]	IntraQ [56]	*Causal-DFQ*
ImageNet	ResNet-18	71.47	70.76	69.84	70.13	70.46	70.74	70.60	**71.01** ± 0.06
	ResNet-50	77.74	77.70	72.93	76.59	76.07	77.05	76.90	**77.45** ± 0.13
	Inception-v3	78.80	78.80	74.94	77.20	-	78.30	77.46	**78.40** ± 0.02
	SqueezeNext	69.38	68.78	16.54	65.46	-	67.34	67.45	**67.87** ± 0.11
	ShuffleNet	65.07	64.55	35.21	60.12	-	60.25	60.18	**60.83** ± 0.06

Table 2. **Comparisons on CIFAR-10/100 and ImageNet with 4W4A quantization setting**.

Dataset	Model	Real Data		Data Free						
		FP32	FT	DFQ [27]	ZeroQ [2]	GDFQ [51]	DSG [54]	SQuant [12]	IntraQ [56]	*Causal-DFQ*
CIFAR-10	ResNet-20	94.03	93.11	89.03	79.30	90.25	78.99	-	91.49	**92.30** ± 0.08
CIFAR-100	ResNet-20	70.33	68.34	63.21	45.20	63.58	46.03	-	64.98	**65.67** ± 0.28
ImageNet	BN-VGG16	74.28	68.83	45.56	1.15	67.10	31.06	68.32	68.73	**71.09** ± 0.30
	ResNet-18	71.47	67.84	55.78	26.04	60.60	34.53	66.14	66.47	**68.11** ± 0.17
	ResNet-50	77.74	72.89	47.34	-	70.23	-	70.80	70.65	**72.49** ± 0.22
	Inception-v3	78.80	73.80	49.62	26.84	70.39	34.89	73.26	73.12	**73.35** ± 0.42
	SqueezeNext	69.38	65.78	-	-	39.18	-	43.45	42.78	**45.99** ± 0.13

weight decay in Nesterov are set to 0.9 and 1×10^{-4}. Moreover, the learning rates of quantized models and generators are initialized to $1 \times 10^{-4}$and 1×10^{-3} respectively. Both of them are decayed by 0.1 for every 100 epochs. In addition, we train the generator and quantized model for 400 epochs with 200 iterations per epoch. On ImageNet, we set the initial learning rate of the quantized model as 1×10^{-6}. Other training settings are the same as those on CIFAR. More details can be found in *Supplemental Materials*.

4.2. Comparison to SoTA

Image Classification. We quantize both weights and activations to 6-bit, and report the comparison results in Table 1. We also quantize them to 4-bit, and report the results in Table 2. In all three classification datasets, our method *Causal-DFQ* outperforms other existing state-of-the-art methods with various network architectures. In particular, when the number of categories increases in CIFAR-100, our method suffers a much smaller accuracy degradation than other methods. The main reason is that our method based on causality gains more prior knowledge from the full-precision model. These results demonstrate the superiority of our method. Especially for the large-scale dataset, ImageNet, existing data-free quantization methods suffer from severe performance degradation. However, our generated images contain style-irrelevant information and satisfy the similar distribution of real data. As a result, our method recovers the accuracy of quantized models significantly with the help of the content-style-decoupled generator and style-intervened discrepancy reduction on three commonly-used networks.

Importantly, *Causal-DFQ* comprehensively outperforms recent SoTA [12] and [56] as shown in Fig. 3. There is a breakthrough where the data-free quantized models fine-tuned by *Causal-DFQ* outperform the (quantized) ones re-trained with real data w.r.t. accuracy on the ImageNet dataset. In addition, data-free quantization is more efficient in terms of training time, *e.g.*, fine-tuning 4-bit ResNet via our data-free quantization method costs 8.4 GPU hours while re-training in a data-given manner costs 29.6 hours. These experimental results demonstrate that data-free quantization can empirically replace the method of re-training low-bit networks.

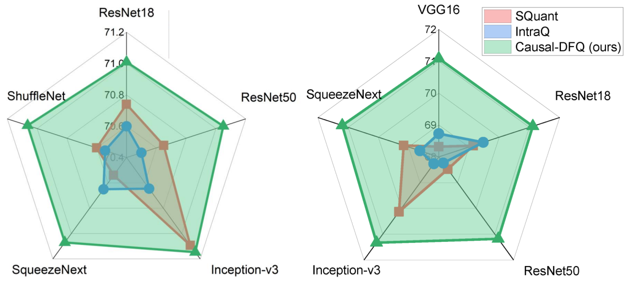


Figure 3. Overall performance on 6-bit (Left, corresponding to Tab. 1) and 4-bit (Right, corresponding to Tab. 2) settings.

Table 3. **Comparisons on VOC 2012 for object detection**. mAP is the metric, and higher is better.

Bits Method	W8A8	W4A8	W4A4	W2A2
FT	70.35	68.24	64.28	57.12
DFQ [27]	69.16	64.57	13.15	2.65
ZeroQ [2]	69.04	67.53	62.72	56.96
ZAQ [23]	70.02	68.12	64.44	56.96
Ours	**70.63**	**68.45**	**66.10**	**57.26**

Object Detection. To demonstrate the application on object detection, we apply *Causal-DFQ* to the model MobileNetV2 SSD [22] and evaluate it on VOC2012. Table 3

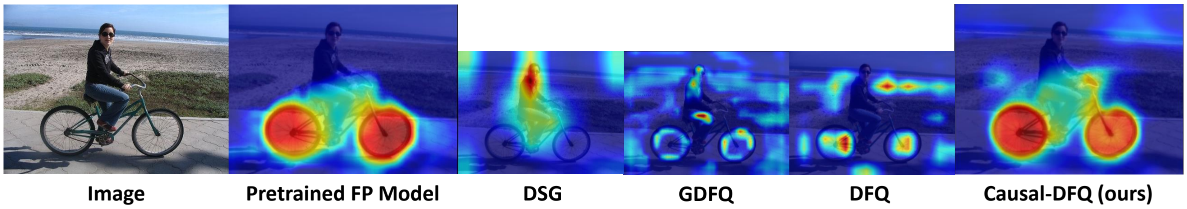
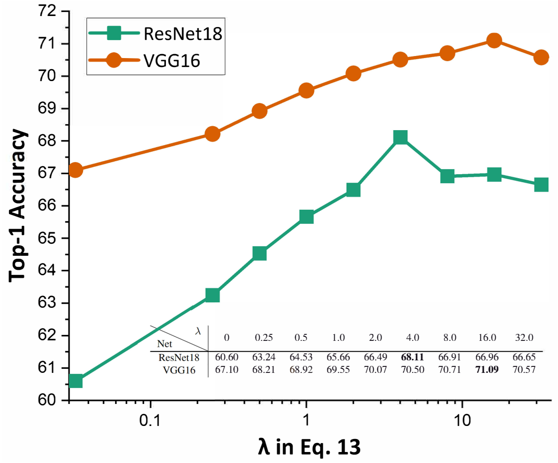


Figure 4. What makes the data-free quantized network for detection on VOC think the pixel label is 'bicycle', visualized via Grad-Cam [39]. We can see that model quantized by *Cuasal-DFQ* is able to focus on task-specific Content.



Net \ λ	0	0.25	0.5	1.0	2.0	4.0	8.0	16.0	32.0
ResNet18	60.60	63.24	64.53	65.66	66.49	**68.11**	66.91	66.96	66.65
VGG16	67.10	68.21	68.92	69.55	70.07	70.50	70.71	**71.09**	70.57

Figure 5. Ablation Study: Effect of λ. Note that $\lambda = 0$ equals to no *Causal-DFQ* as our baseline (*i.e*., GDFQ [51]).

demonstrates the advantages of our method compared to other quantization methods. In particular, *Causal-DFQ* also outperforms FT that utilizes the original training dataset.

4.3. Ablative Studies and analyses

Ablation Study.

We conducted a series of ablative studies of our proposed method on ImageNet with the ResNet18 and VGG16 architectures. By adjusting the coefficient λ in the loss function (Eq.13), where $\lambda = 0$ equals to no *Causal-DFQ* as our baseline (*i.e*., GDFQ [51]). The results are shown in Fig.5. With λ increasing, the performance improvements show the effectiveness of our method. However, when the ratio of $\mathcal{L}_{\textit{Causal-DFQ}}$ in $\mathcal{L}_{overall}$ (Eq. 13) is greater than 10% (on average), data-free quantization performance drops. A well-trained quantized network should have both the ability to align low-level feature maps (*i.e*., aligning as GDFQ [51]) and learn from causality (*i.e*., *Causal-DFQ*).

Network Similarity between FP and Quantized Networks.

Centered kernel alignment (CKA) [5,47] analyzing (hidden) layer representations of neural networks, enabling quantitative comparisons of representations within and across networks. It is a widely acknowledged tool for measuring the similarity between two networks [35]. Higher similar score between two layers' output representations mean those two layers share more similarity. The visualization of CKA analysis is presented in Fig. 6. More details about CKA for metricing network similarity are in *Supplemental Materials*.

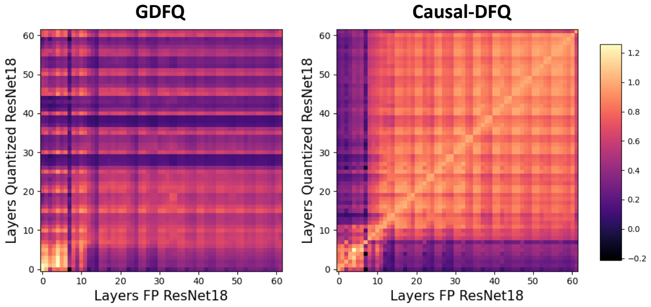


Figure 6. Cross model CKA [5, 47] heatmaps between FP and quantized networks. The lighter the dot, the more similar of the two corresponding layers learned from different datasets. We can conclude that quantized network trained by *Causal-DFQ* is more similar to the FP network.

Attention of Quantized Model Analysis via Grad-Cam [39] Visualization.

We analyze the attentions of several quantized models w.r.t. targeted task. The results are presented in 4. We can see that the quantized model created by our method behaves more similarly to the pre-trained model. Thus, we conclude that *Causal-DFQ* can quantized pre-trained FP model in a content-preserving manner.

5. Conclusion

In this paper, we introduce causal reasoning into data-free quantization. We first formalize a causal graph to model the data-free quantization mechanism. Based on the causal graph, we propose the Causality-guided Data-free Network Quantization method to eliminate the reliance on data while training a quantized model. Specifically, we design a generator which can generate images conditioned on the content and style variables in the view of causality, and then we devise a discrepancy reduction loss to align the intervened distributions of the outputs of pre-trained and quantized models.

Acknowledgement. This research was supported by the gift donation from Cisco. This article solely reflects the opinions of its authors and not the funding agent.

References

[1] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Aciq: analytical clipping for integer quantization of neural networks. In *ICLR*, 2018. 1, 2

[2] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *CVPR*, 2020. 1, 2, 6, 7

[3] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *ICCV*, 2019. 1, 2

[4] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *CVPR*, 2021. 6

[5] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *JMLR*, 2012. 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[7] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 6

[8] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. In *CVPR*, 2020. 2

[9] Ross Girshick. Fast r-cnn. In *CVPR*, 2015. 1

[10] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 2004. 1

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 2021. 3

[12] Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. Squant: On-the-fly data-free quantization via diagonal hessian approximation. In *ICML*, 2022. 6, 7

[13] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 5, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, 2014. 3

[16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 5, 6

[17] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *JMLR*, 2020. 4

[18] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NeurIPS*, 2016. 1, 3

[19] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, 2018. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 7

[23] Yuang Liu, Wei Zhang, and Jun Wang. Zero-shot adversarial quantization. In *CVPR*, 2021. 3, 7

[24] Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. Data-free knowledge transfer: A survey. *arXiv preprint arXiv:2112.15278*, 2021. 1, 2, 3, 6

[25] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. In *NeurIPS*, 2017. 2

[26] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021. 4, 6

[27] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, 2019. 2, 6, 7

[28] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, 2019. 1, 2

[29] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, 1983. 6

[30] Judea Pearl. *Causality*. Cambridge University Press, 2009. 1, 2, 3, 4

[31] Judea Pearl. The do-calculus revisited. In *UAI*, 2012. 2, 4

[32] Judea Pearl et al. Models, reasoning and inference. *Cambridge University Press*, 19, 2000. 4

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 1

[34] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 1, 3, 6

[35] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021. 8

[36] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019. 4

[37] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *ICML*, 2012. 2, 4

[38] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. 2, 3, 4

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 8

[40] Yuzhang Shang, Bin Duan, Ziliang Zong, Liqiang Nie, and Yan Yan. Lipschitz continuity guided knowledge distillation. In *ICCV*, 2021. 3

[41] Yuzhang Shang, Dan Xu, Bin Duan, Ziliang Zong, Liqiang Nie, and Yan Yan. Lipschitz continuity retained binary neural network. In *ECCV*, 2022. 1

[42] Yuzhang Shang, Dan Xu, Ziliang Zong, Liqiang Nie, and Yan Yan. Network binarization via contrastive learning. In *ECCV*, 2022. 1

[43] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *CVPR*, 2023. 1

[44] Suraj Srinivas and R Venkatesh Babu. Data-free parameter pruning for deep neural networks. *arXiv preprint arXiv:1507.06149*, 2015. 2

[45] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014. 1

[46] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[47] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *CVPR*, 2019. 8

[48] Sewall Wright. The genetical structure of populations. *Annals of eugenics*, 1949. 4

[49] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *CVPR*, 2016. 1

[50] Shaoan Xie, Mingming Gong, Yanwu Xu, and Kun Zhang. Unaligned image-to-image translation by learning to reweight. In *ICCV*, 2021. 6

[51] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *ECCV*, 2020. 1, 2, 6, 7, 8

[52] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *CVPR*, 2020. 1, 2

[53] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 3

[54] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *CVPR*, 2021. 1, 2, 6, 7

[55] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. In *ICLR*, 2022. 1, 3, 6

[56] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, 2022. 6, 7

[57] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 1, 2