Interpretable Measurement of CNN Deep Feature Density using Copula and the Generalized Characteristic Function

Anonymous authors
Paper under double-blind review

Abstract

We present a novel empirical approach toward measuring the Probability Density Function (PDF) of the deep features of Convolutional Neural Networks (CNNs). Measurement of the deep feature PDF is a valuable problem for several reasons. Notably, *a.* Understanding the deep feature PDF yields new insight into deep representations. *b.* Feature density methods are important for tasks such as *anomaly detection* which can improve the robustness of deep learning models in the wild. Interpretable measurement of the deep feature PDF is challenging due to the Curse of Dimensionality (CoD), and the Spatial intuition Limitation. Our novel measurement technique combines copula analysis with the Method of Orthogonal Moments (MOM), in order to directly measure the Generalized Characteristic Function (GCF) of the multivariate deep feature PDF. We find that, surprisingly, the one-dimensional marginals of non-negative deep CNN features after major blocks are not well approximated by a Gaussian distribution, and that these features increasingly approximate an exponential distribution with increasing network depth. Furthermore, we observe that deep features become increasingly independent with increasing network depth within their typical ranges. However, we surprisingly also observe that many deep features exhibit strong dependence (either correlation or anti-correlation) with other extremely strong detections, even if these features are independent within typical ranges. We elaborate on these findings in our discussion, where we propose a new hypothesis that exponentially infrequent large valued features correspond to strong computer vision detections of semantic targets, which would imply that these large-valued features are not outliers but rather an important detection signal.

1 Introduction

Convolutional Neural Networks (CNN) have revolutionized the performance of image analysis tasks including image classification, semantic segmentation, object detection, and image and video synthesis (Yuan & Zhang, 2016; Goodfellow et al., 2020; Hao et al., 2020; Xing et al., 2023). At the time of writing, CNNs continue to play a domenent role as image feature encoders for state-of-the-art techniques including several prominent Vision-Language Models (VLMs) (Long, 2024; Radford et al., 2021; Li et al., 2021) as well as diffusion models for image generation (Rombach et al., 2022; Yang et al., 2023). Nevertheless, the extraordinary complexity of CNNs has coined the nick-name of *black-box*, that learns an uninterpretable and high-dimensional feature representation. This work contributes toward improving our understanding of the learned representation by measuring the statistical characteristics of deep CNN features through high-dimensional statistical analysis techniques including copula analysis, and the Method of Orthogonal Moments (MOM) to obtain the Generalized Characteristic Function (GCF). Our approach is unique because it provides a novel and interpretable probability density estimate of CNN features, without making any rigid parametric assumptions that may be unjustified, and without altering the CNN feature representation which may reduce accuracy. Furthermore, as an empirical technique, we can gain greater insight by plotting and analyzing the marginal and interdependence components of the feature copula density PDF which can lead to improved understanding of the statistical behavior of the native CNN feature space. We want to measure and observe the probability density of deep features for popular CNN architectures, as we believe that this will lay the groundwork for future feature-density analysis methods that can identify stronger parametric assumptions, thereby leading

to improved understanding of the feature space as well as improved probability density estimation of the distribution of deep CNN features.

Measuring the probability density of deep CNN features has many practical applications including Out-of-Distribution detection (Lee et al., 2018; Liu et al., 2020; Zhu et al., 2022; Jiang et al., 2023; Le Lan & Dinh, 2021), adversarial detection (Lee et al., 2018; Le Lan & Dinh, 2021), domain generalization (Chen et al., 2024), and federated learning (Sun et al., 2023). It is anticipated that if one knows the density distribution of deep features, that it would be possible to use outlier detection to statistically distinguish legitimate inlier data from anomalous outlier data, thereby providing a greater level of model robustness to data that is out of model scope (Guérin et al., 2023; Lee et al., 2018; Liu et al., 2020). Furthermore, methods that can model both the conditional as well as the joint distribution meet the technical definition of *generative methods* rather than purely *discrimination methods* (Kingma, 2013; Caterini et al., 2021; Kingma et al., 2016; Rezende & Mohamed, 2015). Generative methods have many theoretical advantages because they enable complex inferencing tasks regarding the distribution of unlabeled samples in addition to performing basic discrimination tasks.

Several works have attempted to construct deep generative architectures with explicit feature representations that enable exact or approximate PDF estimates (Kingma, 2013; Caterini et al., 2021; Kingma et al., 2016; Rezende & Mohamed, 2015). Such techniques include variational auto-encoders (Kingma, 2013), as well as autoregressive and normalizing flows (Caterini et al., 2021; Kingma et al., 2016; Rezende & Mohamed, 2015). Nevertheless, open questions remain regarding the applicability of deep generative PDFs for anomaly detection for several reasons including the often unaccounted for effects of reparameterization (Le Lan & Dinh, 2021) and class-imbalance (Jiang et al., 2023), as well as questions regarding the empirical validity of the underlying 'typical set' hypothesis Zhang et al. (2021) stating that there is a discernible difference between the statistical properties of *typical valued* in-distribution and out-of-distribution features (Lee et al., 2018; Liu et al., 2020; Zhu et al., 2022). We believe that further empirical analysis, including alternate ways of obtaining density estimates will be valuable toward studying these observed phenomena.

We take a purely empirical approach toward accurate estimation of deep CNN feature density. An important contribution of our approach is that we do so without unverified parametric or linearity assumptions, and without altering the CNN architecture or feature representation which may reduce accuracy. Several prior works have assumed the CNN features follow a Multi-Variate Gaussian (MVG) (Majurski et al., 2024; Zhu et al., 2022; Lee et al., 2018; Rippel et al., 2021; Zhu et al., 2022). But the motivation for MVG traces back to a handful of histogram plots of the penultimate critic features in the supplemental materials of Lee et al. (2018). There has never been a verification that MVG is suitable to the deeper intermediate features (non-penultimate), which is especially important because in most architectures, deep features undergo ReLU activation which deactivates (zeros out) any negative valued features, thereby enforcing non-negativity. Max-pooling moreover retains the strongest positive features, throwing away weaker less-positive detections. Thus it is important to look at and measure the distribution of non-negative features as these are the only features that have an impact on subsequent activations. We observe that these non-negative deep features are highly non-Gaussian, and increasingly approximate an exponential distribution with increasing network depth. We furthermore show that at adequate network depth, many deep features exhibit non-linear dependence, with typical feature values showing statistical independence for typical (non-extreme) values, but strong statistical dependence of extreme values. In other words, we observe that two deep features may be uncorrelated within their typical ranges, but if one feature is extremely large or small, then other features may show strong correlation or anti-correlation in their extreme values.

This work also contributes to an active and growing body of ongoing empirical work toward improved CNN feature understanding (Allen-Zhu & Li, 2023; Chen et al., 2024; Giraldo & Schwartz, 2019; Hermann & Lampinen, 2020; Qiu et al., 2024; Shwartz-Ziv & Tishby, 2017). Several recent works have attempted to visualize CNN features, in order to determine under what circumstances a model may learn or retain features that are spurious (correlated) versus invariant (causal) to semantic meaning (Chen et al., 2024; Qiu et al., 2024; Shwartz-Ziv & Tishby, 2017). We do not distinguish spurious versus invariant features in this work, however we do observe strong correspondence between extreme-valued features, suggesting that many deep features within intermediate representations may be learning correlated (or anti-correlated) semantic concepts with a statistical backdrop of uncorrelated detection noise. Allen-Zhu & Li (2023) propose a multi-

view hypothesis that intermediate deep CNN features may learn to detect specific *views* of target objects. Based on our results, we propose a follow-on hypothesis which we coin the *exponential view* hypothesis. That is, at adequate network depth, the distribution of feature detections for any given *semantic target view* follows exponential distribution. That is, prevalent small-value features correspond to a lack of detection of their respective *target view*, whereas rare large-value features correspond to a strong detection of the *target view*. As the actual *target* often corresponds to only a small portion of the image frame, the strong-valued detections are statistically rare occurrences but represent important computer vision signal, whereas the lack of detection, although more statistically prevalent, is primarily a noise signal. Our *exponential view* hypothesis offers a new perspective on the possible empirical shortcomings of the *typical set* hypothesis. It logically follows that the *typical set* may be accidentally eliminating the most important computer vision signal by excluding the extreme-value features from consideration in the *typical set*. We discuss this hypothesis further in the discussion section.

2 Methodology

Our primary goal is to measure, assess, and interpret the statistical distribution of CNN features without making unnecessary and unverified assumptions such as parametric or independence assumptions. The Curse of Dimensionality, as well as the Spatial intuition Limitation complicate this evaluation. In order to measure the distribution under this context, we present a novel approach that combines copula analysis with the Method of Orthogonal Moments. To the best of our knowledge neither of these techniques has been previously applied toward measuring the distribution of CNN features, but they provide a general assumption free way of looking at the high-dimensional interdependence of features and separating this interdepedence from the univariate marginal feature distributions.

2.1 Copula Analysis

In high-dimensional statistics, copula analysis is a powerful method that allows one to completely separate the marginal distribution of the random variables from their interdependence. Given a set of random variables $(X_1, X_2, \ldots, X_D)$ and marginal cumulative distribution functions $(F_1, F_2, \ldots, F_D)$, one can perform a probability integral transform as follows,

$$(Y_1,\ Y_2,\ \ldots\ ,\ Y_D) \quad = \quad (F_1(X_1),\ F_2(X_2),\ \ldots\ , F_D(X_D)) \tag{1}$$

The copula C is defined as the cumulative distribution function of the probability integral transform of the random variables as follows,

$$C\,(y_1,\ y_2,\ \ldots\ ,\ y_D) \quad = \quad Pr\,[Y_1 \leq y_1,\ Y_2 \leq y_2,\ \ldots\ ,\ Y_D \leq y_D] \tag{2}$$

Moreover the copula density is the probability density function $c\,(y_1,\ y_2,\ \ldots\ , y_D\ )$ associated with cumulative copula distribution $C\,(y_1,\ y_2,\ \ldots\ , y_D\ )$ as follows,

$$c\,(y) = c\,(y_1,\ y_2,\ \ldots\ , y_D\ ) = \frac{\partial\ C\,(y_1,\ y_2,\ \ldots\ , y_D)}{\partial y_1\ \partial y_2\ \ldots\ \partial y_D} \tag{3}$$

Typically the probability integral transform converts a marginal distribution into a uniform distribution on the interval $(0, 1)$. However, in our approach, we carry out our analysis using a re-scaled version of the probability integral transform to a uniform the interval $(-1, 1)$. This is because many well-known orthogonal functions are designed for this interval, and these orthogonal functions allow us to measure the copula density term in greater detail without parametric assumptions. This modified probability integral transform is defined as follows,

$$F_i(x) = 2\ Pr[X_i \leq x] - 1 \tag{4}$$

2.2 Method of Orthogonal Moments (MOM)

The method of moments are a way of fully describing the shape of a probability distribution through the use of *consistent estimators*, which asymptotically share sample and population statistics. Assume that x is a finite sample of n elements drawn from infinite population X, then a series of well-behaved sample statistics $\hat{\mu}_t$ should very closely approximate their population statistics μ_t as follows,

$$\hat{\mu}_t \approx \mu_t \quad \text{where} \quad \mu_t = E(\phi_t(x)) \quad \text{and} \quad \hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n} \phi_t(x_i) \tag{5}$$

The original method of moments simply defines basis functions as ϕ_t as the power functions $\phi_t(x) = x^t$ for $t = 1, 2, 3, 4 \ldots$. In this case, the zero-mean samples would correspond to the *mean*, *variance*, *skewness*, *kurtosis*, etc. But these basis functions have the disadvantage of being non-orthogonal, and thereby duplicating some shape information. It is more powerful to choose ϕ_t to represent an orthogonal basis set, such as the Fourier series $\phi_t = e^{i\pi t x}$, or one of the orthogonal polynomial sets such as the Legendre or Chebyshev polynomials thereby representing the Method of Orthogonal Moments (MOM).

We use the MOM to analyze the copula interdependence $Y_1, \ \ldots \ , Y_D$ of deep features separately from the marginals $f(X_1), \ \ldots \ , f(X_D)$. As such, we must measure the interdependence between multiple random variables by defining multivariate moments in terms of the expected product of univariate basis statistics in the copula space. If we define T as a D dimensional vector of integers where T_d represent the desired moment of the d^{th} random variable Y_d, then the joint moment μ_T corresponds to inner product of basis vectors as follows,

$$\mu_T = E\left(\prod_{d=1}^{D} \phi_{T_d}(y_d)\right) \qquad \hat{\mu}_T = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{d=1}^{D} \phi_{T_d}(y_d)\right) \tag{6}$$

For our analysis, we specifically define ϕ_t as either the real-valued Fourier series or the normalized Legendre polynomial series because in addition to being orthogonal over the rescaled copula interval of $(-1, 1)$, these basis functions also have the additional property that all non-constant basis terms exhibit zero integral over $(-1, 1)$, which is useful for our Generalized Characteristic Independence (GCI) metric. The specific forms of these series that we propose also exhibit the property of having unit L_2 norm over the target interval $(-1, 1)$. The Legendre polynomials also have the advantage that, like the power series, *mean* and *covariance* are part of the basic shape descriptors, which are highly familiar concepts thereby aiding in practical interpretation.

2.3 Generalized Characteristic Function

A useful property of the MOM, is that it allows one to recover the actual copula probability density function $c(y)$ in high-dimensions completely non-parametrically, and without any overly-rigid assumptions on the shape of this distribution by means of the Generalized Characteristic Function (GCF). The original Characteristic Function refers to the observation that if one defines the basis set as the Fourier series, then the population moments resembles the Fourier transform of the PDF as follows,

$$\mu_p = E\left(\phi_p(y)\right) = \int_{-\infty}^{\infty} c(y)\ e^{i\pi t y} dy \tag{7}$$

As such, one can recover the copula density by means of an inverse discrete Fourier transform of the population moments μ. If this process is performed using the sample moments $\hat{\mu}$ then one recovers the sample estimate of the copula density function $\hat{c}$ as follows,

$$c(y) = \sum_{t=1}^{K} \mu_t\ \phi_t(y) \qquad \hat{c}(y) = \sum_{t=1}^{K} \hat{\mu}_t\ \phi_t(y) \tag{8}$$

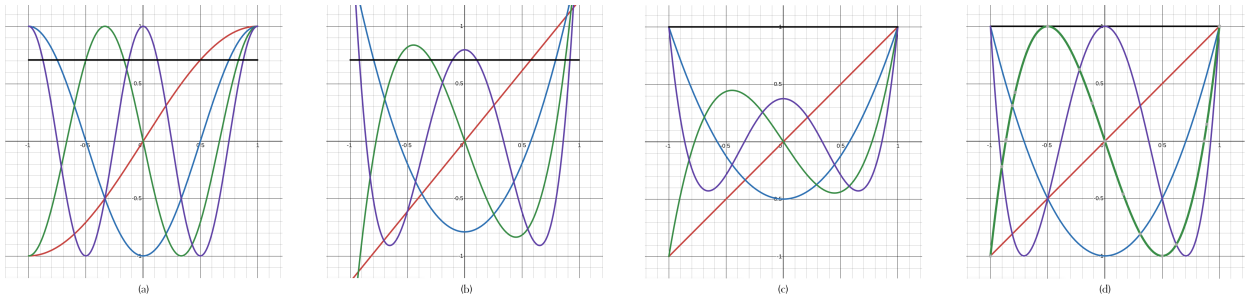

Figure 1: Comparison of orthogonal basis functions over the uniform interval $(-1, 1)$. (a) Real-valued Fourier series (b) Normalized Legendre polynomials (c) Legendre polynomials without normalization (d) Chebyshev polynomials.

The GCF refers to the straightforward extension of this technique to non-Fourier orthogonal moments. Analogously, one can recover the full copula density using an discrete inverse Fourier-like transform for any orthogonal basis set. Nevertheless, it is highly desirable to select basis functions that exhibit the following properties,

- Orthogonal over unit interval $(-1, 1)$
- Real valued, exhibiting even and odd harmonics
- Unit length L_2 norm over interval $(-1, 1)$
- All non-constant moments exhibit zero integral over $(-1, 1)$

In order to adhere to these properties, we propose to employ a specific normalized form of the Legendre polynomials, as well as a real-valued form of the Fourier series as we describe in further detail.

2.4 Normalized Legendre Polynomials

The Legendre Polynomials (figure 1b,c) are a set of real-valued orthogonal basis functions over the target interval $(-1, 1)$ with several desirable properties. Unlike the Chebyshev polynomials (figure 1d), the Legendre polynomials (figure 1c) exhibit zero integral over the interval $(-1, 1)$, except trivially for the constant polynomial P_0. This zero-integral property is highly-desirable for our resultant GCF. The polynomials can be generated efficiently using Bonnet's recurrence as follows,

$$\begin{aligned} P_0(y) &= 1 \\ P_1(y) &= y \\ P_{n+1}(y) &= \frac{2n+1}{n+1} x P_n(y) - \frac{n}{n+1} P_{n-1}(y) \end{aligned} \tag{9}$$

The Legendre polynomials in this form do not exhibit unit-length L_2 norm over the interval $(-1, 1)$, as such, we propose normalizing the Legendre polynomials based on their L_2 norm in order to obtain unit-length orthogonal moments as in (figure 1b). This normalized form is obtained as follows,

$$\phi_t(y) = \frac{P_t(y)}{||P_t||_2} \quad \text{where} \quad ||P_t||_2 = \sqrt{\int_{-1}^{1} P_t^2(y)\, dy} \tag{10}$$

2.4.1 Real-valued Fourier Series

As our sample is real-valued, one can equivalently represent the Fourier series as a sum of real-valued *cos* (even) and *sin* (odd) harmonic terms. Moreover, it is possible to simplify this to only *cos* terms if one makes use of the trigonometric phase identity as follows,

$$sin(y) = cos(y - \frac{\pi}{2} + 2\pi t) \qquad \text{for integer } t \tag{11}$$

As such, one can present both even and odd real-valued Fourier basis functions using an elegant and simplified form as follows,

$$\begin{aligned} \phi_0(y) &= \frac{\sqrt{2}}{2} \\ \phi_t(y) &= cos\left(t\frac{\pi}{2}(y-1)\right) \end{aligned} \tag{12}$$

The real-valued Fourier basis functions in this form are shown in (figure 1a). One can see that these basis functions correspond analogously one-to-one with the Legendre and Chebyshev polynomials, with ϕ_t exhibiting exactly t roots over the target interval $(-1, 1)$. This form of the Fourier series also exhibits our desired properties, as the non-constant basis functions have zero integral over the target interval $(-1, 1)$. Moreover, for all basis functions the L_2 norm over the target interval is exactly equal to 1 when presented in this form.

2.5 Generalized Characteristic Distance and Independence

Given a set of orthogonal population moments μ_t and ν_t for probability distributions P and Q respectively, the Generalized Characteristic Distance (GCD) fully describes the difference in shape between probability distribution as the Manhattan distance of the Fourier-like transforms of the PDFs. This is calculated by taking the Manhattan distance between the moments as follows,

$$D_{char}(P, Q) = \sum_{t=1}^{K} |\mu_t - \nu_t| \tag{13}$$

In the event that a set of features are completely independent, then the copula density $c\,(y_1,\ \ldots\ y_D)$ corresponds to the uniform distribution on the hypercube $y_i \in (-1, 1)$. As such, we new define a Generalized Characteristic Interdependence (GCI) metric $H_{char}(c)$ as the GCD between the copula distribution c and the ideal uniform copula density Q_{unif}. H_{char} is exactly zero when variables $y_1,\ \ldots\ , y_D$ are statistically independent, and nonzero when these variables show some statistical dependence along one or more orthogonal moments. $H_{char}(c)$ is defined as follows,

$$H_{char}(c) = D_{char}(c, Q_{unif}) \tag{14}$$

The real-valued Fourier and normalized Legendre series have the convenient property that all basis functions integrate to zero over $(-1, 1)$ (except trivially the constant basis function ϕ_0). As such, it is straightforward to show that the ideal uniform distribution Q_{unif} has all zero moments. Therefore, for the Fourier and Legendre moments, the GCI simplifies to L1 norm as follows,

$$H_{char}(c) = ||\mu||_1 \qquad \text{for Fourier and Legendre moments} \tag{15}$$

3 Experimental Setup

For our experiments we evaluate three CNN architectures resnet18, resnet50, and vgg19 across four image classification datasets mnist, cifar10, cifar100, and imagenette2 (Deng, 2012; Krizhevsky et al., 2009; Deng et al., 2009; Howard, 2019; He et al., 2016; Simonyan & Zisserman, 2014). Imagenette2 is a subset of Imagenet exhibiting 10 classes with the full resolution images that are center-cropped to 224×224 pixels.

Figure 2 shows the extracted feature spaces from the resnet and vgg architectures. For each image, after every major convolutional block, we obtain a tensor of size $[N_{img}, \text{filters}, \text{rows}, \text{cols}]$. Due to the self-similarity of features in each row and column, we evaluate the distribution for each of the filters over the entire training or test set. Thus we have a sample of $N = N_{img} \times \text{rows} \times \text{cols}$ for every filter for each block.

As we perform copula analysis, we separately analyze the marginal and the copula interdependence terms. This means that we have separate results and separate analysis for the univariate marginal density, versus the multivariate interdependence. For the marginals, we evaluate 1D probability density functions for each feature. For the copula we analyze the independence of pairs and groups of features. Our exploratory plots of copula interdependence are evaluated for pairs of features. Nevertheless, our method is capable of measuring the density of higher dimensional groups of features which we validate by a task of predicting the distribution of groups of test features given the set of training features.

For imagenette2, we used the pre-trained versions of resnet18, resnet50, and vgg19 as included with PyTorch, with a custom trained final linear classification layer. For mnist, cifar10, and cifar100 the standard versions of resnet and vgg are not designed to work with such small resolution images, and thus we used the small-image optimized versions of these architectures by Kuang (2017). This small-image optimized version is widely used, often without attribution, in papers that achieve high accuracy on these datasets. The small-image resnet and vgg architectures were trained from scratch using a learning rate of 0.01, momentum 0.09, and the SGD optimizer.

Rectified Linear (ReLU) is a very common activation function in CNNs, and both resnet and vgg employ ReLU in order to introduce non-linearity. ReLU also has the additional effect of forcing all-features to be non-negative. As such, the CNN features after major convolutional blocks such as those shown in figure 2 are always non-negative and entirely reside in the positive quadrant of the feature space. For the 1D analysis, we measure the marginal distribution by first measuring the percentage of non-negative features, as well as fitting the observed univariate distribution of non-negative features. The combination of these measurements fully describes the univariate marginal term. For the multivariate copula interdependence term, we add an infinitesimal random jitter to the zero-valued features in order to ensure a statistically independent ordering of zero valued features for the copula interdependence. We present and describe plots of the copula density in two dimensions, and further analyze the goodness of fit of high-dimensional feature copula through a KL-divergence task.

4 Results

As we are performing copula analysis, we present separate analysis of the univariate marginal distribution, and of the multivariate copula interdependence terms. The analysis of marginals shows the distribution of feature density including the percentage of non-zero features as well as the distribution of non-zero features. The copula density interdependence term shows the interdependence of the features with the marginal distribution removed by means of a probability integral transform. The interdependence is modeled by measuring the GCF using the MOM. We present results showing qualitative description of the feature distribution as well as quantitative goodness of fit using KL-divergence and/or cross entropy loss.

4.1 Analysis of Copula Marginals

The marginals describe the univariate distribution of CNN features, and are an important component of the copula analysis. ReLU has the effect of zeroing out (deactivating) any negative-valued features such that they have no further impact on the intermediate calculations. In order to fully and adequately describe the 1D marginal distribution post-activation, we must separately model the zero-valued and non-zero features

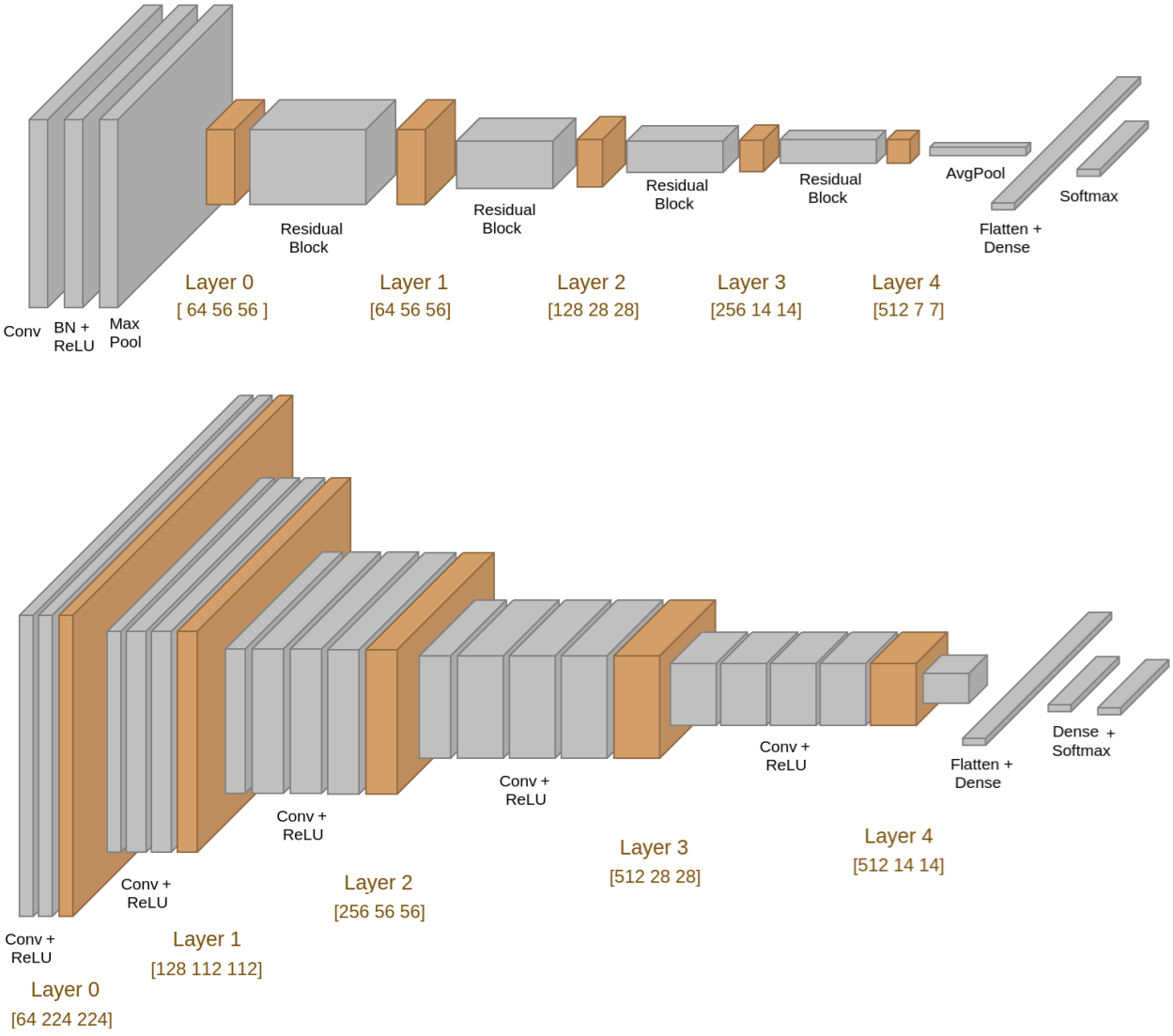

Figure 2: Illustration of resnet18 (top) and vgg19 (bottom) deep feature layers selected for probability density analysis using copula. Orange shaded regions represent deep feature layers after major architectural blocks that were selected for density analysis. Resnet50 diagram is not shown, but is similar to the resnet18 diagram shown (top).

Table 1: Percent of nonzero features per layer.

	resnet18				resnet50				vgg19			
	INET	CF10	CF100	MNST	INET	CF10	CF100	MNST	INET	CF10	CF100	MNST
Layer 0	87.7	66.4	67.2	74.8	91.0	61.3	74.8	86.0	43.2	41.9	43.2	58.0
Layer 1	77.0	68.1	79.6	66.7	80.1	83.9	81.6	90.8	29.6	21.1	23.8	32.4
Layer 2	50.3	43.7	50.9	52.9	55.8	68.0	85.2	90.8	14.8	22.0	13.2	23.3
Layer 3	46.1	24.0	35.6	38.4	29.9	21.7	49.6	58.5	9.9	79.1	51.3	52.1
Layer 4	52.2	46.8	55.0	41.3	53.9	84.9	57.9	89.6	9.0	30.6	25.8	21.5

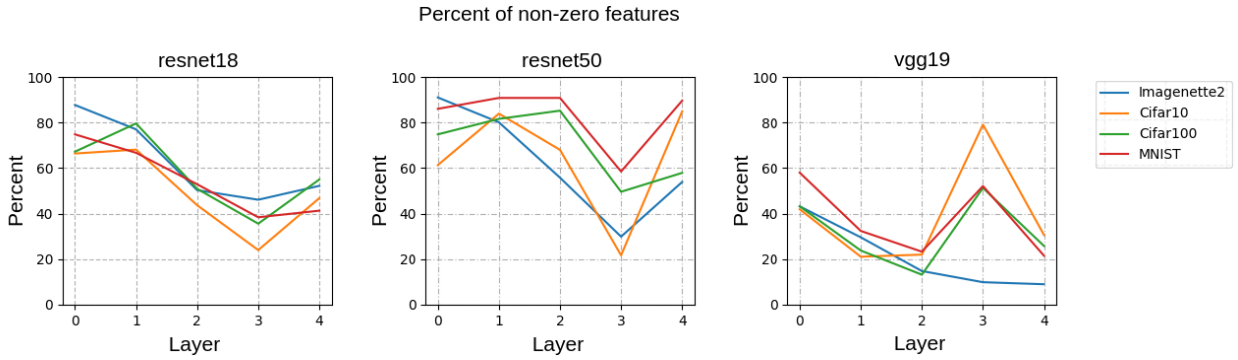

Figure 3: Percent of nonzero features per layer.

as the combination of these distributions describes the overall marginal density. The percentage of non-zero features is shown in figure 3 and also in tablular form in table 1. The percent of nonzero features varies for each architecture (resnet18, resnet50, vgg19) as well as for each dataset (imagenette2, cifar10, cifar100, and mnist). We see however a general trend amongst all models that the percent of nonzero features is quite high in layer 0, and tends to decrease in subsequent layers until layer 3 (resnet18, resnet50) or layer 2 (vgg19) before increasing slightly in subsequent layers. One primary exception to this overall shape is shown in vgg19 over imagenette2, in which the percent of nonzero features appears to decrease monotonically all the way to layer 4. The resnet models tend to start with $70-90\%$ of nonzero features in layer 0, decreasing to around 50% of nonzero features by layer 4 with the exception of resnet50 mnist, which exhibits 89.6% nonzero features in layer 4. The vgg19 architectures exhibit greater sparsity than the resnet18 architectures with $40-60\%$ non-zero features in layer 0 decreasing to $9-31\%$ non-zero features in layer 4. Overall these results exhibit an increase in feature sparsity (percent zeros) with network depth, although this trend is somewhat noisy often showing an uptick in nonzero percentage in the deepest layers. The magnitude and scale of these percentages are dataset dependent with vgg19 showing greater sparsity than resnet architectures.

Figure 4 shows a histogram of the non-zero portion of the marginal density for resnet18 on imagenette2 for filters 0, 1, and 2. Additional plots of the non-zero marginals for resnet18, resnet50, and vgg19 are available in supplemental materials. We see in figure 4 that filters 0 (left), 1 (middle), and 2 (right) show similar shape characteristics that depend on the network depth. Importantly, we observe an interesting phenomenon, where the early layers show a more complicated univariate shape, whereas the later layers appear to more closely resemble an exponential distribution shape. The complicated shape seen in layers 0 and 1 is most likely due to contamination of the input pixel distribution into the marginal feature distribution, as these are early shallow layers in the network. In the early layers, the model is unable to transform the input pixel space very much, so we still observe remnants of this pixel distribution. However, we see in the deeper layers such as layer 3 and 4 a very clear exponential distribution of the non-zero features. We observed a similar

exponential distribution consistently across other architectures and datasets with additional similar figures in supplemental materials.

We now quantitatively compare the goodness of fit of four parametric distributions to the marginals of the non-zero features. The distributions that we compare are the uniform distribution, the Gaussian distribution, the gamma distribution, and the Weibull distribution. The optimal parameters of these distributions are determined using the method of simulated annealing, and the goodness of fit of each of the distributions is compared using KL-divergence. Figure 5 shows the plots (with 1σ uncertainty) of each of the distributions tested over all of the architectures and datasets in our analysis. As these are 1σ confidence intervals, overlapping intervals represent an insignificant difference, whereas non-overlapping intervals represent a statistically significant difference in the goodness-of-fit. In order to prevent overfitting, the parametric distribution is fit to the training data features, whereas the KL-divergence measures the goodness of fit to the test distribution.

As we see from figure 5, the gamma and Weibull distributions show a significantly better fit to the feature marginals than the uniform and Gaussian distributions. The exponential distribution can be seen as having performance in-between the Gaussian and Weibull distribution. Notably, we observe that the exponential distribution increasingly approximates the observed features as we look at deeper network depth. It is important to note that the gamma and Weibull distributions generalize the exponential distribution, and thus it is not possible for the exponential to achieve a statistically better fit than the gamma and Weibull distributions. Nevertheless, the exponential distribution is highly interpretable and exhibits a very good fit to the non-zero features that significantly outperforms the Gaussian and uniform distributions, while coming close to the gamma and Weibull goodness of fit particularly for the deeper layers.

4.2 Analysis of Copula Interdependence

Figure 6 shows the copula interdependence term of the probability density for resnet18 over imagenette2 for select pairwise features. Additional plots for other architectures and datasets are available in supplemental materials. Figure 6 also shows the reconstructed copula interdependence using three different methods, the GCF with normalized Legendre polynomials (left), the GCF with real-valued Fourier series (middle) and 2D histograms as a control (right). We see that all three functions appear to show very good agreement with very similar pairwise density plots for all of the feature pairs shown in figure 6.

We observe that the early layers show much more complicated interdependence with greater variation, whereas later layers tend to show a similar structure for most feature pairs that we observed. Similar to our analysis of the marginals, we believe that the variation in the pairwise interdependence seen in the shallow layers (layer 0 and layer 1) is very likely due to the contamination of the input pixel and texture distribution into the shallow features. As these are early layers in the network, the network architecture is unable to fully remove the contribution of the initial pixel distribution. The later layers however show very similar pairwise interdependence. We observe a very interesting phenomenon, in which the deep features in layers 2,3 and 4 (rows 3-5 in figure 6) appear to be uncorrelated for typical values, but show a strong statistical dependence for extremely large valued features. This strong statistical dependence can be observed by the prevalence of very high density in the upper right corner of the interdependence plot (yellow), with much lower density along the top and right edges of the plot (blue). The apparent statistical independence of the terms over typical values (green) is also clearly apparent in the pairwise interdependence plots. We also observed (shown in supplemental materials) the prevalence of deep features showing a strong anti-correlation of the extreme values, with high density (yellow) along the top left and bottom right, but low density in the upper right hand corner (blue), again with independence for the typical valued features (green). This prevalence of uncorrelated features, except in the event of extreme-valued detections, is a new observation with profound implications regarding the distribution of deep CNN features, we discuss this result further in our discussion section.

In order to evaluate the goodness-of-fit of the copula independence term, we perform a supervised evaluation using cross entropy loss. Our task is to fit the copula interdependence to the training features using our method, and then we evaluate how well this describes the interdependence of the test features. For this experiment, we calculated the goodness of fit for resnet18 features for random groups of four features at a

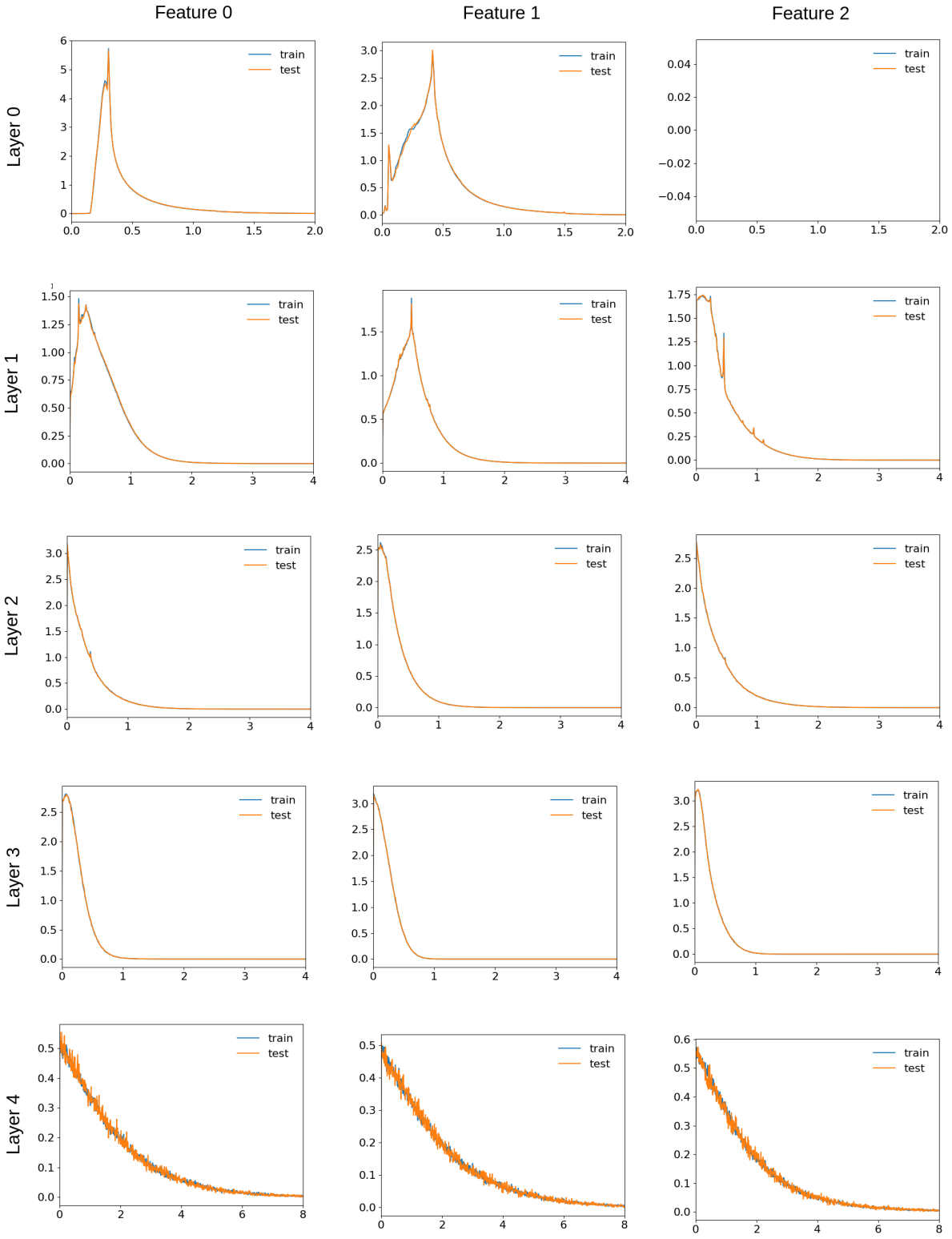

Figure 4: Histogram of marginal density for pre-trained resnet18 on imagenette2 for features 0,1,2, for each of five convolutional layers. In early layers, features show some influence of original pixel distribution. Layer 0 also exhibits a few dead features (e.x. Feature 2). Deeper layers appear to exhibit an exponential distribution.

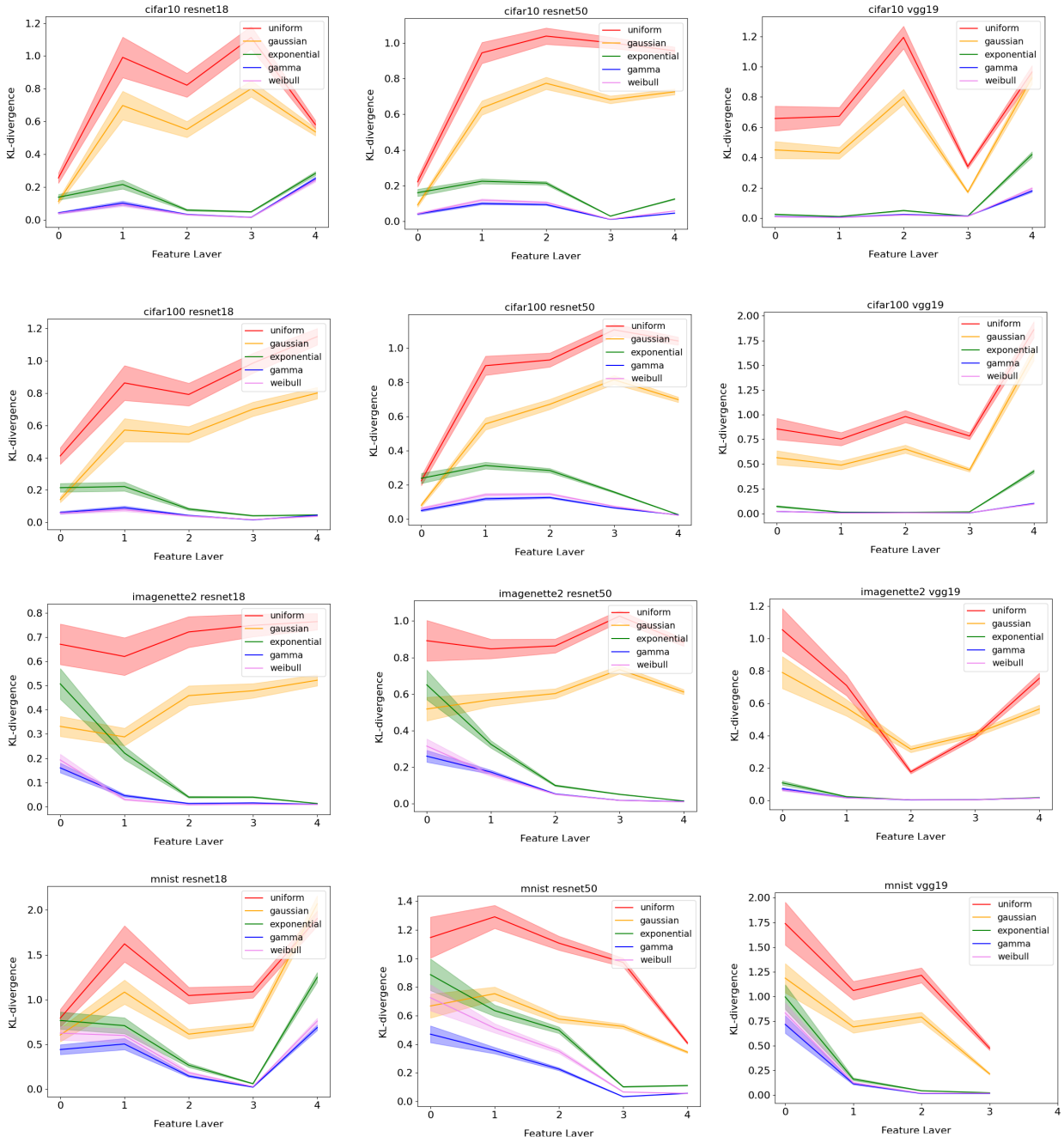

Figure 5: Quantitative goodness-of-fit of five standard distributions to the feature marginals for cifar10, cifar100, imagenette2, and mnist, across three models resnet18, resnet50, and vgg19. Shaded region shows 1σ confidence interval for goodness of fit. We see that exponential, gamma and Weibull are substantially and significantly better fit than uniform and Gaussian for most layers across all models and datasets, with gamma and Weibull showing the best fit but these distributions generalize the exponential distribution which shows increasingly good fit in the deeper layers of the network.

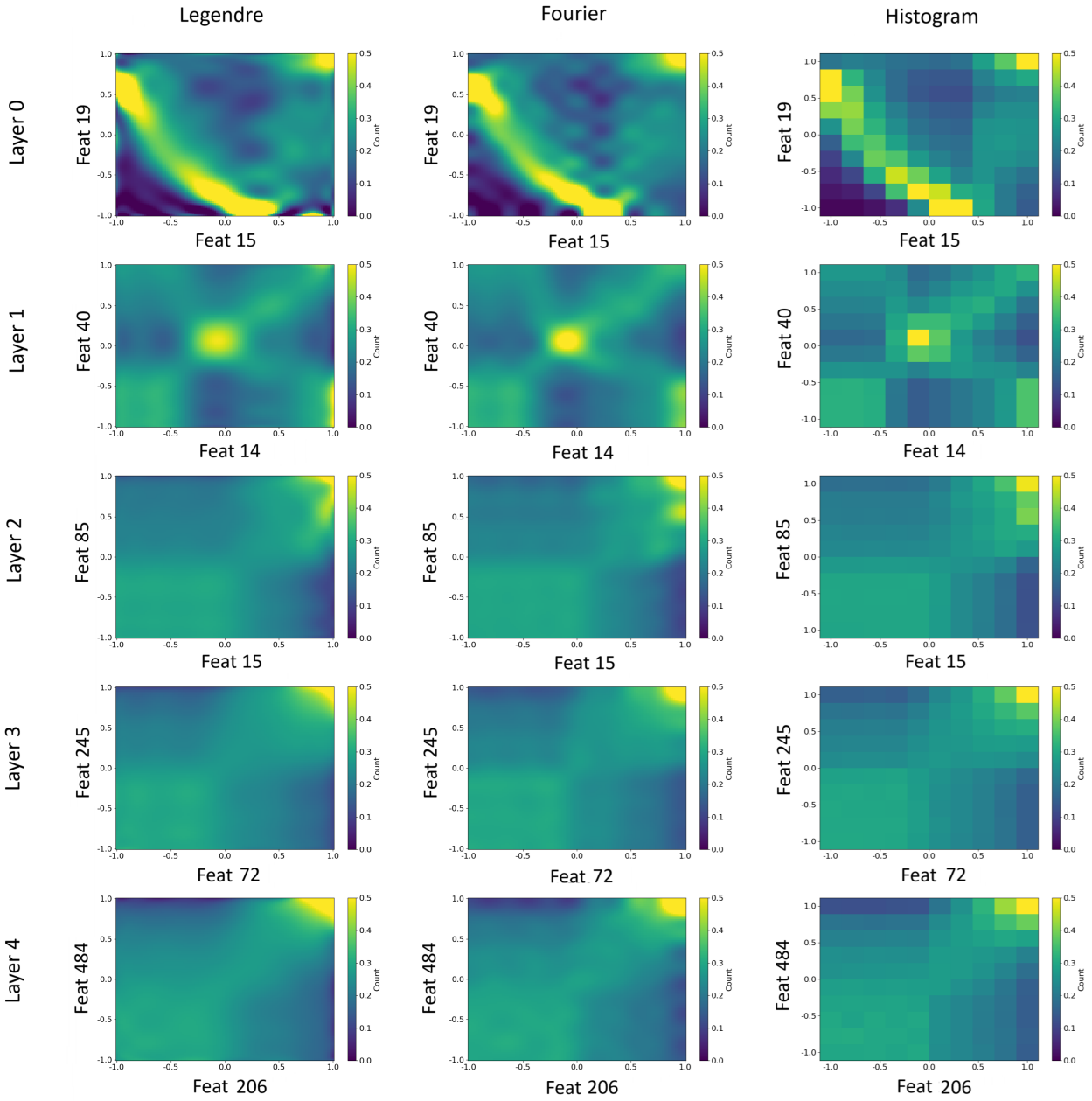

Figure 6: Select copula interdependence for pairwise features for 5 layers of resnet18 over Imagenett2. Top to bottom: Layers 0 through 4. Left: Legendre pdf. Middle: Fourier pdf. Right: Histogram pdf.

Table 2: Goodness of fit of copula interdependence term for random subsets of four features on imagenette2 with resnet18. Evaluation criteria is cross entropy loss, 1σ confidence intervals are shown in parentheses. Bold shows significantly best method for each layer task.

	Legendre	Fourier	Histogram
Layer 0	1.9793 (1.9587, 1.9998)	2.0036 (1.9864, 2.0207)	**1.7521 (1.7209, 1.7833)**
Layer 1	2.6303 (2.6299, 2.6308)	2.6290 (2.6285, 2.6295)	**2.6172 (2.6164, 2.6179)**
Layer 2	**2.7183 (2.7182, 2.7184)**	2.7192 (2.7192, 2.7193)	2.7215 (2.7214, 2.7216)
Layer 3	2.7489 (2.7486, 2.7492)	**2.7435 (2.7434, 2.7435)**	2.7440 (2.7439, 2.7440)
Layer 4	**2.7170 (2.7170, 2.7171)**	2.7192 (2.7191, 2.7192)	2.7323 (2.7322, 2.7323)

Table 3: Goodness of fit of copula interdependence term for random subsets of four features on cifar10 with resnet18. Evaluation criteria is cross entropy loss, 1σ confidence intervals are shown in parentheses. Bold shows significantly best method for each layer task.

	Legendre	Fourier	Histogram
Layer 0	2.2306 (2.2174, 2.2438)	2.2359 (2.2229, 2.2490)	**2.1850 (2.1712, 2.1989)**
Layer 1	2.5844 (2.5830, 2.5859)	2.5812 (2.5797, 2.5827)	**2.5578 (2.5548, 2.5608)**
Layer 2	**2.6834 (2.6828, 2.6841)**	2.6852 (2.6846, 2.6858)	2.6865 (2.6858, 2.6873)
Layer 3	**2.7442 (2.7440, 2.7444)**	2.7461 (2.7459, 2.7463)	2.7493 (2.7491, 2.7495)
Layer 4	2.1568 (2.1502, 2.1634)	2.1268 (2.1204, 2.1332)	**2.0399 (2.0339, 2.0459)**

Table 4: Goodness of fit of copula interdependence term for random subsets of four features on cifar100 with resnet18. Evaluation criteria is cross entropy loss, 1σ confidence intervals are shown in parentheses. Bold shows significantly best method for each layer task.

	Legendre	Fourier	Histogram
Layer 0	**2.2020 (2.1812, 2.2227)**	2.2085 (2.1883, 2.2288)	2.0857 (2.0644, 2.1070)
Layer 1	2.5588 (2.5571, 2.5605)	2.5600 (2.5583, 2.5618)	**2.5242 (2.5219, 2.5264)**
Layer 2	**2.6794 (2.6788, 2.6800)**	2.6806 (2.6800, 2.6812)	2.6829 (2.6822, 2.6835)
Layer 3	**2.7414 (2.7414, 2.7414)**	2.7415 (2.7415, 2.7415)	2.7449 (2.7449, 2.7450)
Layer 4	**2.7200 (2.7198, 2.7202)**	2.7220 (2.7218, 2.7222)	2.7303 (2.7301, 2.7304)

Table 5: Goodness of fit of copula interdependence term for random subsets of four features on mnist with resnet18. Evaluation criteria is cross entropy loss, 1σ confidence intervals are shown in parentheses. Bold shows significantly best method for each layer task.

	Legendre	Fourier	Histogram
Layer 0	-3.3750 (-3.6474, -3.1025)	-2.7168 (-2.8623, -2.5714)	**-3.6498 (-3.9232, -3.3764)**
Layer 1	1.3320 (1.2504, 1.4137)	1.2019 (1.1019, 1.3019)	**0.5817 (0.3903, 0.7730)**
Layer 2	2.1437 (2.1295, 2.1579)	2.1298 (2.1136, 2.1460)	**1.8885 (1.8526, 1.9244)**
Layer 3	**2.6451 (2.6444, 2.6458)**	2.6551 (2.6545, 2.6557)	2.6481 (2.6472, 2.6489)
Layer 4	2.1646 (2.1467, 2.1824)	2.1572 (2.1376, 2.1768)	**2.0449 (2.0147, 2.0751)**

time for the imagenette2, cifar10, cifar100 and mnist dataset. We performed 30 rounds per experiment, and calculated the copula interdependence using 3 methods, Legendre GCF, Fourier GCF, and Histograms. The interdependence of four randomly selected features within the same layer is calculated, and compared with the sample interdependence of the test features using cross entropy loss as the evaluation criteria.

For resnet18 on cifar10, Legendre GCF works better for layers 2 and 3; however, for the latest layer, the Histogram method performs the best. Table 4 demonstrates that on cifar100, Legendre GCF performs well across all layers except layer 1, where the Histogram approach outperforms due to pixels still being contaminated by the input image. On mnist (table 5), the Histogram method is the best for all layers except layer 3, where Legendre GCF outperforms. This is because mnist is essentially a binary dataset with most pixels being either black or white. This binary input pixel distribution contaminates the deep feature distributions, causing a discreet quantization which can be better modeled by the discreet Histogram approach. For the other datasets (imagenette2, cifar10, cifar100) with more continuous input pixel distributions, we observe in tables 2-4 that the GCF methods significantly outperform the Histogram methods for layers 2,3 and 4, with the two exceptions being layer 3 for imagenette2 dataset, and layer 4 for the cifar10 dataset. The Histogram approach is particularly sensitive to the CoD, because with D dimensions and B bins per dimension, the average number of samples per bin is N/B^D. Due to max-pooling, the deeper layers exhibit fewer samples relative to the shallow layers, and one would expect the GCF methods to outperform the histogram methods particularly in the situations of higher dimensionality as well as smaller sample sizes, which is consistent with our results in tables 2-4.

5 Conclusion

We present an empirical analysis of the density distribution of deep CNN features through direct measurement of the GCF using a novel non-parametric approach that combines copula analysis with the MOM. We demonstrate that our approach is able to model the marginal and interdependence terms of feature density after each major Conv+ReLU block of resnet18, resnet50, and vgg19. Moreover, as a non-parametric technique, we do not introduce overly-restrictive assumptions as to the shape of the marginal or interdependence terms. We report empirical findings on the observed marginal distributions and copula density interdependence terms as a function of network depth. In our analysis of marginals, we observe that features after major Conv+ReLU blocks exhibit both zero and non-zero features. Furthermore, we demonstrate through hypothesis testing that the non-zero features for the deeper layers of the network more closely fit the Weibull or gamma distribution versus the Gaussian or uniform distribution. Weibull and gamma distributions generalize the exponential distribution which also significantly outperforms the Gaussian and uniform distributions in the deeper layers. Our analysis of the copula interdependence shows that pairs of features in the deeper layers of the network also exhibit an unusual form of statistical dependence, for which these features are highly independent throughout their typical value ranges, yet become strongly dependent (either correlated or anti-correlated) for extremely large feature values. We observe that the Legendre and Fourier GCF techniques are able to better model the copula interdependence density relative to the histogram technique for groups of four features with resnet18 over imagenette2, cifar10 and cifar100 for the deeper layers of the network. Our approach is the first purely empirical technique to model the multivariate probability density distribution of deep CNN features. We believe that empirical analysis of the feature density distribution will lead to a better understanding of CNN feature representations. Moreover empirical feature density has the potential to lead to a new branch of generative methods that model the full joint distribution of the native CNN feature space, thereby enabling off-the-shelf CNN architectures to attain all of the benefits of generative techniques.

6 Discussion

One of our key findings is that the exponential distribution is a surprisingly good fit for the non-zero features of the marginals for the deeper layers of the network. This result has many implications for applications of feature density techniques, and also leads to several new questions regarding the reason *why* an exponential distribution fits these deeper non-zero marginals so well.

One possible explanation is that the exponential distribution is the maximum entropy distribution for random variables on the interval $[0, \infty)$. As such, the complex composition of functions with the restriction of non-negativity (as enforced by ReLU) are likely to tend toward the exponential distribution. This explanation would also explain why the exponential distribution does not well fit the feature marginals that arise in the shallow early layers, because the shallow features will have direct more influence from the input pixel distribution and will not have sufficiently deep composition to adequately tend toward the exponential distribution.

An additional explanation for the exponential distribution of marginals in the deeper layers arises from the computer vision perspective. In the deeper layers, features correspond to increasingly expressive semantic concepts. If the features represent presence of a discriminative target of interest, then most of the scene does not usually contain the target. For example, given a face in a cluttered scene, the vast majority of the scene does not contain the face (and would exhibit a small-value for face detection). A small portion of the scene however does present a face, and a face detector would present a strong finding. It is possible that in the deeper layers, features correspond to detectors of semantic concepts, in which case each large valued feature would correspond to a strong concept detection.

The observation that deep CNN features are particularly sensitive to specific *views* of known targets is a relatively new finding (Allen-Zhu & Li, 2023). As such we believe that the observation of an exponential distribution of strong non-zero feature detections is highly compatible with this observation that features correspond to specific *target views*, because a given *target view*, if it is present in an image, will only typically take up a small portion of the image thereby leading to a large number of non-detections, with a small number of actual strong detections of the *target view*.

We observed a very peculiar phenomenon where pairs of deep CNN features at adequate network depth show very little dependence for *typical values*, whereas they exhibit very strong dependence (either correlation or anticorrelation) for extremely strong detections. We conjecture that this distribution that we see is highly related to our *exponential view* hypothesis, in that the *typical values* of the features correspond to a lack of detection of the *target view*, whereas the extremely large values correspond to a strong detection of the *target view*. As such, it may be the case that the *typical values* are uncorrelated because they do not represent discriminative foreground features, but instead represent background variability especially as most images have a large number of background pixels relative to foreground pixels. The rarity of discriminative foreground features would therefore be a reason for CNN models to produce very large detections in order to overcome the more prevalent background signal.

If one believes in this *exponential view* hypothesis, then the strong statistical dependence of extremely large feature values would imply that in fact deep CNN features are more correlated than they appear in the presence of actual foreground target detection signal, even if they are uncorrelated over the background. This is very possible and we plan to investigate this hypothesis further as part of future work.

Moreover, the uncorrelated nature of *typical value* features, yet strong correlation of *extreme value* features, would suggest that future work should revisit the assumption that the *typical set* is in fact the most descriptive set of features for a given image, as modeling of feature density has overwhelmingly emphasized evaluation of the density distribution of *typical valued* features. We believe that modeling of the distribution of *extreme valued* features may also be highly important. Moreover, given the assumption of an exponential distribution, the removal of outlier features under the assumption of an MVG has the potential to accidentally eliminate strong computer vision signals of foreground target detection which would be an unwanted side effect.

Our approach toward modeling the distribution of features by measuring the GCF with the MOM has a notable advantage that the entire training set is used to measure each sample moment, leading to high certainty in estimation. This is a major inherent advantage in modeling the density distribution of high-dimensional feature spaces, versus other techniques that run into low-sample sizes in high dimensions due to the CoD. Nevertheless, this technique also has limitations. Notably, the number of moments increases exponentially with the number of features under consideration at once. Not all moments are inherently important, and future work would involve the extension of this method to find a set of sparse moments, in order to reduce the computational burden of modeling the copula interdependence for a large number of features within one subset group. An additional limitation of this methodology is the error-of-approximation

that comes from reconstructing the density using sample moments rather than population moments. We infer that this error-of-approximation is reasonably small by looking at the cross entropy goodness of fit between the train and test distributions. However, in future work, we would like to statistically model the uncertainty of the moment estimates, as this would allow us to quantify the uncertainty of the empirical probability density estimate.

Acknowledgments

Anonymous Acknowledgment

References

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Anthony Caterini, Rob Cornish, Dino Sejdinovic, and Arnaud Doucet. Variational inference with continuously-indexed normalizing flows. In *Uncertainty in Artificial Intelligence*, pp. 44–53. PMLR, 2021.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Luis Gonzalo Sánchez Giraldo and Odelia Schwartz. Integrating flexible normalization into midlevel representations of deep convolutional neural networks. *Neural computation*, 31(11):2138–2176, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Joris Guérin, Kevin Delmas, Raul Ferreira, and Jérémie Guiochet. Out-of-distribution detection is not all you need. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14829–14837, 2023.

Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL `https://github.com/fastai/imagenette`.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*, pp. 15067–15088. PMLR, 2023.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

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

Liu Kuang. Pytorch-cifar: optimized cnn aarchitectures for cifar10, 2017. URL `https://github.com/kuangliu/pytorch-cifar`.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf`.

Hanhua Long. Hybrid design of cnn and vision transformer: A review. In *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*, pp. 121–127, 2024.

Michael Majurski, Sumeet Menon, Parniyan Favardin, and David Chapman. A method of moments embedding constraint and its application to semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7809–7818, 2024.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2403.03375*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6726–6733. IEEE, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yuwei Sun, Ng Chong, and Hideya Ochiai. Feature distribution matching for federated domain generalization. In *Asian Conference on Machine Learning*, pp. 942–957. PMLR, 2023.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Zheng-Wu Yuan and Jun Zhang. Feature extraction and image retrieval based on alexnet. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, volume 10033, pp. 65–69. SPIE, 2016.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pp. 12427–12436. PMLR, 2021.

Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue', Xiang Tian, bolun zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20758–20769. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/82b0c1b954b6ef9f3cfb664a82b201bb-Paper-Conference.pdf`.