Why CNN Features Are not Gaussian: A Statistical Anatomy of Deep Representations

David Chapman Parniyan Farvardin
University of Miami
dchapman@cs.miami.edu, pxf291@miami.edu

Abstract

Deep convolutional neural networks (CNNs) are commonly analyzed through geometric and linear–algebraic perspectives, yet the statistical distribution of their internal feature activations remains poorly understood. In many applications, deep features are implicitly treated as Gaussian when modeling densities. In this work, we empirically examine this assumption and show that it does not accurately describe the distribution of CNN feature activations. Through a systematic study across multiple architectures and datasets, we find that the feature activations deviate substantially from Gaussian and are better characterized by Weibull and related long-tailed distributions. We further introduce a novel Discretized Characteristic Function Copula (DCF-Copula) method to model multivariate feature dependencies. We find that tail-length increases with network depth and that upper-tail dependence emerges between feature pairs. These statistical findings are not consistent with the Central Limit Theorem, and are instead indicative of a Matthew process that progressively concentrates semantic signal within the tails. These statistical findings suggest that CNNs are excellent at noise reduction, yet poor at outlier removal tasks. We recommend the use of long-tailed upper-tail-dependent priors as opposed to Gaussian priors for accurately CNN deep feature density. Code available at https://github.com/dchapman-prof/DCF-Copula

1. Introduction

Understanding the statistical distribution of CNN deep feature activations remains an important yet largely unexplored problem. Mechanistic interpretability aims to reverse engineer deep learning models in order to understand how their internal representations are formed [6, 7, 32]. While significant progress has been made in understanding the *semantic meaning* of deep vision features, the *statistical structure* of these activations has received far less attention [2, 7–9, 26, 30, 33, 34]. Our goal is to reverse engineer these feature distributions, including marginal and interdependence behavior using exploratory analysis and confirmatory tests.

Understanding this distribution matters because many methods implicitly assume Gaussian priors [18, 19, 22, 24, 28, 36]. If these assumptions do not hold, learning methods based upon them may be unpredictable or sub-optimal.

If Gaussian priors were realistic for activations, then Anomaly Detection (AD) would be inherently easy; it is not [16, 21–23, 25, 33, 36]. Gaussian priors motivate simple density and outlier-removal strategies (e.g., ellipsoidal level sets). Just detect the outliers, call them anomalies. Clearly things are not so simple, and successful AD methods have often avoided this approach [13, 15, 31, 35], but why?

Our empirical analysis provides new clues about this discrepancy. In particular, long-tailed distributions with upper-tail dependence are poorly matched to classical *outlier detection* assumptions. Under Gaussian assumptions, outlier removal should improve performance. We find the opposite, outlier removal catastrophically destroys performance.

Our explanation is that deep learning models attempt to learn the long-tailed distribution of the natural underlying image statistics. The distribution of object parts in imagery naturally follows a power law [8, 9]. We find that CNN must *learn* this distribution, starting with an uninformed prior, and layer-by-layer increasing the tail length.

We propose the use of a Weibull marginal to represent deep feature marginals with a range of tail lengths. We analyze feature interdependence using a novel DCF-Copula approach that is highly expressive, and we observe upper tail dependence. We argue that most of the semantic signal in deep features is represented by the tail, and thus we recommend the consideration of long-tailed upper-tail dependent priors. Our contributions are as follows.

- Systematic exploratory and confirmatory analysis of the CNN deep feature marginals and interdependence.
- Novel DCF-Copula analysis framework that is more expressive than Archimedean copulas.
- Observe long-tailed, and upper-tail-dependent behavior with tail length increasing with depth.
- Outlier removal destroys performance, but noise suppression improves performance due to feature distribution.

2. DCF-Copula Methodology

We now define the DCF-Copula methodology that is used to empirically estimate the statistical behavior of deep CNN features. Copula analysis separates the marginal distributions of random variables from their dependence structure. Given random variables $(X_1, \ldots, X_D)$ with cumulative distributions $(F_1, \ldots, F_D)$. The copula C is defined as the joint cumulative distribution of these transformed variables

$$C(y_1, \ldots, y_D) = Pr[Y_1 \leq y_1, \ldots, Y_D \leq y_D], \quad (1)$$

Typical Archimedean copulas are inflexible, and represent density using parametric methods. Our DCF-Copula method is highly expressive because we apply the Method of Orthogonal Moments (MOM) to represent copula density, thereby capturing all multivariate non-linear statistical dependencies. Given basis functions $\phi_t(\cdot)$, population and sample moments are

$$\mu_t = E[\phi_t(x)], \qquad \hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N}\phi_t(x_i). \quad (2)$$

For multivariate variables $Y_1, \ldots, Y_D$, joint moments capture cross-feature dependence

$$\mu_T = E\left(\prod_{d=1}^{D}\phi_{T_d}(Y_d)\right), \quad \hat{\mu}_T = \frac{1}{N}\sum_{i=1}^{N}\prod_{d=1}^{D}\phi_{T_d}(Y_{d,i}). \quad (3)$$

DCF-Copula reconstructs density using a discretized version of the Generalized Characteristic Function which requires far fewer terms than the traditional characteristic function techniques. If ϕ_t is an orthogonal basis, the moments correspond to transform coefficients

$$\mu_t = E[\phi_t(y)] = \int c(y)\phi_t(y)dy. \quad (4)$$

Using a finite expansion, the copula density can be obtained as

$$c(y) = \sum_{t=1}^{K}\mu_t\phi_t(y), \qquad \hat{c}(y) = \sum_{t=1}^{K}\hat{\mu}_t\phi_t(y). \quad (5)$$

We find the normalized Legendre polynomials and real-valued Fourier harmonics to be a suitable basis function series. We discuss these functions and their desirable properties in greater detail in Appendix A. The multivariate basis series is defined as the product of the univariate basis function ϕ_{T_d} as follows.

$$\Phi_T(\vec{y}) = \prod_{d=1}^{D}\phi_{T_d}(y_d). \quad (6)$$

The empirical copula density is then estimated as

$$\hat{c}(\vec{y}) = \sum_{T\in\mathbb{Z}_K^D}\hat{\mu}_T\Phi_T(\vec{y}). \quad (7)$$

3. Tail-Length Parameterization

We perform tail analysis in order to quantify the long-tailed nature of the CNN deep features. Toward this aim we use the Weibull distribution, as this allows for a range of tail lengths from sub-Gaussian to long-tailed. The PDF and CDF of the Weibull distribution are defined as the following.

$$f(x;\theta,k) = \begin{cases}\theta k\,(\theta x)^{k-1}\,e^{-(\theta x)^k} & x \geq 0\\ 0 & x < 0\end{cases} \quad (8)$$

$$F(x;\theta,k) = \begin{cases}1 - e^{-(\theta x)^k} & x \geq 0\\ 0 & x < 0\end{cases} \quad (9)$$

Where θ is the tail parameter, and k is the shape parameter. In particular the tail parameter θ is indicative of the tail-length with $\theta \leq 0.5$ indicating sub-Gaussian, $\theta \leq 1$ indicating sub-exponential, and $\theta > 1$ indicating long-tailed.

For tail analysis, we only consider samples that exceed the 99^{th} percentile which is a common choice for extreme value theory. We define $\ddot{x} \subseteq X$ as the high value subset as follows,

$$\ddot{x} = \{\, \ddot{x}_i : \ddot{x}_i \in X \text{ and } \ddot{x}_i > u \,\} \quad \text{where} \quad Pr[X \leq u] = 0.99 \quad (10)$$

Our loss function for fitting the tail is the 1-Wasserstein distance, because it is a well-behaved distance metric for comparing partial distributions. The 1-Wasserstein distance can be calculated by comparing the CDF of the observed samples $\ddot{y}$ with that of the theoretical Weibull distribution $F(\ddot{x};\theta;k)$ in order to determine the optimal Weibull tail parameter of the deep feature marginals.

$$\underset{\theta,\,k}{argmin}\ \sum_{i=1}^{n}\left|\left|\,\ddot{y}_i - F(\ddot{x}_i;\theta,k)\,\right|\right|_1 \quad (11)$$

4. Results

We analyze the deep features for three CNNs, ResNet-18 [12], ResNet-50 [12], and VGG-19 [29] across four image classification datasets: MNIST [5], CIFAR-10 [20], CIFAR-100 [20], and Imagenette2 [4, 14]. Figure 3 shows the extracted features ResNet-18 spanning multiple layer depths; similar features were extracted for other models.

Figure 2 shows examples of univariate histograms of ResNet-18 on Imagenette2. Plots for other datasets are

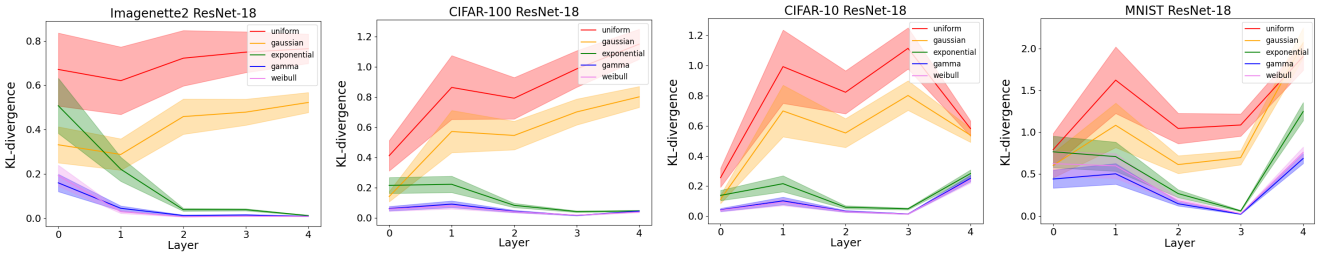

Figure 1. Overall goodness of fit of parametric marginal distributions, with shaded region showing 2σ error bars.

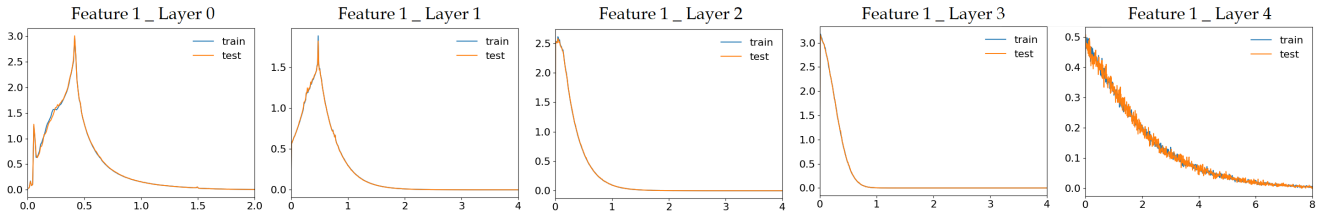

Figure 2. Example histograms of ResNet-18 feature density for extracted layers on Imagenette2.

shown in Appendix D. All plots show a prominent tail-shape. The early layers show some remnants of a bell-curve due to the lingering distribution of input colors. But by the third layer onward, there is no more bell-curve, and the entire histogram appears tail-like. By the final layer, the tail is very long and prominent. We see similar behavior on the other datasets D. These visual results already suggest that the deep semantic features form a prominent tail-like distribution that becomes increasing prominent with depth. Instead of the Central Limit Theorem (increasing bell-shape), this is a Matthew effect (increasing tail-shape) indicating a winners-compounding process (i.e. semantic signals become more semantic with depth).

The second step is confirmatory analysis to obtain a suitable distribution. In Figure 1 we analyze five standard distributions: Uniform, Gaussian, Exponential, Gamma, and Weibull. Similar tests are performed for ResNet-50 and VGG-19 in Appendix D. Numeric details of this test are given in Appendix B. This hypothesis test confirms with high significance that for all datasets and models, the Uniform and Gaussian distributions fit very poorly after the first layer, whereas the Exponential, Gamma and Weibull distributions are a much better fit. Moreover, it is highly significant that the Gamma and Weibull distributions outperform

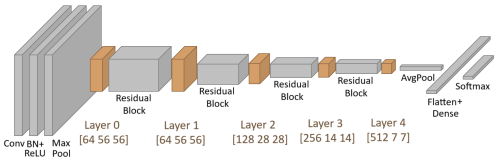

Figure 3. ResNet-18, orange shows extracted features.

the Exponential distribution. From here onward we assume a Weibull distribution, because it is not only a good distributional fit, but it can also model a range of tail behaviors by adjusting its θ parameter.

Our next goal is to determine just how *long-tailed* the feature distribution really is. Figure 5 shows the estimated Weibull θ parameter across all layers for the upper 99^{th} percentile of observations for ResNet-18. Additional plots for ResNet-50 and VGG-19 are shown in Appendix D. We see a Matthew effect, where the length of the θ parameter increases with each successive layer. Although for the *easy* datasets (MNIST and CIFAR-10), θ sharply drops off in the last layer. But θ does not drop in the *hard* datasets (Imagenette2 and CIFAR-100).

Also notable, is that although the tail length increases with depth until long-tailed $\theta > 1$ it never reaches a power-law $\theta \to \infty$. Natural image statistics follow power-laws, so why do deep CNNs exhibit a more modest long-tail? Our explanation is that the neural network never fully learns the full spectrum of semantic concepts within the image, only enough concepts for the subtasks of classification, and only after many layers. After each layer the network represents more and more semantic signal leading to a longer and longer tail (Matthew effect). But it never learns the full power-law spectrum within the natural image. Moreover, if the classification task is *easy* (CIFAR-10, MNIST), it abruptly shortens the tail just before classification.

Our next step is to perform non-parametric Copula analysis of the feature inter-dependence. Table 1 shows that our proposed DCF-Copula method greatly outperforms a range of comparable methods to model statistical interdependence including the Archimedean copulas (AMH, Clayton, Frank, Gumbel, Joe) as well as characteristic functions (ECF). Ad-

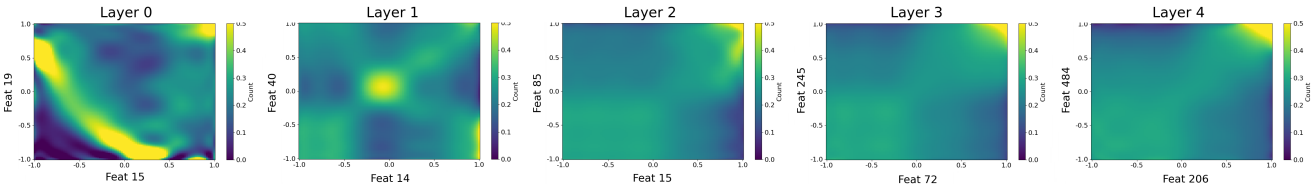

Figure 4. DCF-Copula density between random feature pairs for ResNet-18 on Imagenette2.

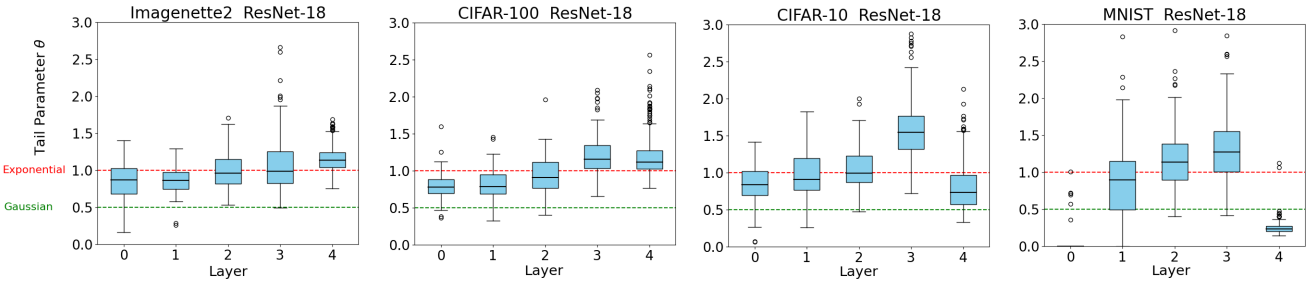

Figure 5. Estimated Weibull tail parameter θ for ResNet-18 features across layers. Larger values correspond to heavier-tailed activation distributions. Deeper layers exhibit stronger tail behavior, particularly for more complex datasets.

ditional details of the experiment are given in Appendix C additional plots in Appendix D.

Figure 4 shows example bivariate copula plots for random feature pairs. We see that the first two layers show remnants of the input pixel distribution, but all subsequent layers show a bright spot at (1,1) (upper right). This indicates an upper tail dependence, that features are uncorrelated within their normal ranges, but highly correlated in the tail regime. We believe this is due to *polysemanticity* [7], the observation that a single neuron feature will fire due to multiple semantic concepts. Similarly, a single highly semantic region of an image (i.e. object part) may cause multiple neurons to strongly fire simultaneously, thereby creating an upper tail dependence.

We conclude our analysis with Figure 6, showing what happens if one unintentionally confuses the tail of the distribution with outliers. Removing the high-percentile features (green), catastrophically destroys the accuracy. But

Table 1. Copula density loss for ResNet-18 on Imagenette2.

	L0	L1	L2	L3	L4
DCF-Legendre	1.1883	1.3757	**1.3794**	**1.3823**	**1.3711**
DCF-Fourier	**1.1878**	**1.3748**	1.3796	1.3825	1.3716
ECF [27]	2.1029	2.0450	2.1529	2.0635	1.4537
AMH [1]	1.5329	1.4933	1.4953	1.4969	1.5002
Clayton [3]	1.4469	1.4756	1.4755	1.4806	1.4631
Frank [10]	1.4779	1.4818	1.4825	1.4869	1.4762
Gumbel [11]	1.4749	1.4836	1.4845	1.4888	1.4808
Joe [17]	1.5020	1.4867	1.4883	1.4917	1.4890

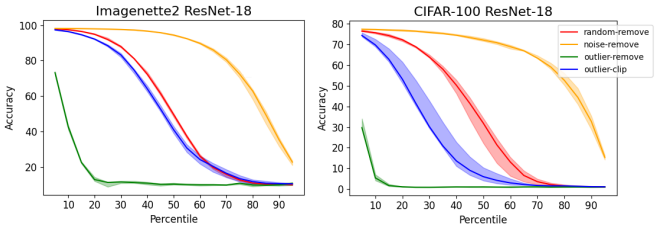

Figure 6. Accuracy vs removal of noise and outlier signals.

the model can tolerate removal of low percentile features (yellow). Even clipping the high percentile features (blue) underperforms zeroing an equal number of random features (red). Clearly the tail exhibits an important signal, whereas the head of the distribution is irrelevant for classification.

5. Conclusion

Our results bring a new understanding to the statistical distribution of deep CNN features. We clearly show that the tail of the distribution contains the most important signal. We observe that deep CNN features follow a Matthew process, as the tail length increases layer-by-layer. We believe this is due to the models increasing ability to represent natural image statistics with increasing depth. Upper-tail dependence may be due to polysemanticity, or natural image statistics, or both. Many advanced learning methods incorporate statistical feature priors, with the Gaussian prior being a common choice. Matthew process behavior is not consistent with the Central Limit Theorem. This is why the CNN features are not Gaussian, and we recommend the careful consideration of long-tailed distributions instead.

Acknowledgement

This work is supported by the Frost Institute for Data Science and Computing.

References

[1] Mir M Ali, NN Mikhail, and M Safiul Haq. A class of bivariate distributions including the bivariate logistic. *Journal of multivariate analysis*, 8(3):405–412, 1978. 4, 8

[2] 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. 1

[3] David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1): 141–151, 1978. 4, 8

[4] 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*, pages 248–255, 2009. 2

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

[6] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021. 1

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 1, 4

[8] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020. 1

[9] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 1

[10] Maurice J Frank. On the simultaneous associativity of f (x, y) and x+ y- f (x, y). *Aequationes mathematicae*, 19:194–226, 1979. 4, 8

[11] Emil J Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960. 4, 8

[12] 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*, pages 770–778, 2016. 2

[13] Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard. Self-supervised anomaly detection in computer vision and beyond: A survey and outlook. *Neural Networks*, 172: 106106, 2024. 1

[14] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. 2

[15] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19606–19616, 2023. 1

[16] 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*, pages 15067–15088. PMLR, 2023. 1

[17] Harry Joe. Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22(1):47–64, 1994. 4, 8

[18] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[19] 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. 1

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[21] Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021. 1

[22] 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. 1

[23] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475. Curran Associates, Inc., 2020. 1

[24] 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*, pages 7809–7818, 2024. 1

[25] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018. 1

[26] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 1

[27] John Nolan. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28 (5):2067–2089, 2013. 4

[28] 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)*, pages 6726–6733. IEEE, 2021. 1

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[30] Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pages 6458–6467. PMLR, 2019. 1

[31] Yan Wan, Yingqi Lang, and Li Yao. Dcs: A zero-shot anomaly detection framework with dino-clip-sam integration. *Applied Sciences*, 16(4):1836, 2026. 1

[32] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022. 1

[33] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021. 1

[34] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018. 1

[35] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 1

[36] 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 *Advances in Neural Information Processing Systems*, pages 20758–20769. Curran Associates, Inc., 2022. 1

A. Appendix: Orthogonal Basis Functions

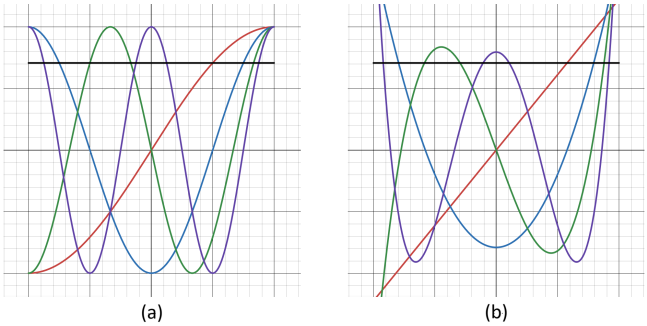

Figure 7. a. Real-valued Fourier Series b. Normalized Legendre Polynomials

DCF-Copula requires orthogonal basis functions with 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 $(-1, 1)$

We propose a specific normalized form of the Legendre polynomials, as well as a real-valued form of the Fourier series as seen in figure 7.

A.1. Normalized Legendre Polynomials

The Legendre Polynomials (figure 7b) are a set of real-valued orthogonal basis functions over the target interval $(-1, 1)$ with several desirable properties. Unlike the Chebyshev polynomials, the Legendre polynomials exhibit zero integral over the interval $(-1, 1)$, except trivially for the constant polynomial P_0. The following polynomials can be generated efficiently using Bonnet's recurrence. The Legendre polynomials in this form do not exhibit unit-length L_2 norm over the interval $(-1, 1)$, as such normalize these Legendre polynomials based on their L_2 as follows.

$$\phi_t(y) = \frac{P_t(y)}{||P_t||_2} \quad \text{where} \quad ||P_t||_2 = \sqrt{\int_{-1}^{1} P_t^2(y)\, dy} \tag{12}$$

A.1.1. 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, which is given by the following.

$$\begin{aligned} \phi_0(y) &= \frac{\sqrt{2}}{2} \\ \phi_t(y) &= cos\left(t\frac{\pi}{2}(y-1)\right) \end{aligned} \tag{13}$$

The real-valued Fourier basis functions in this form are shown in (figure 7a). These basis functions also exhibit all of our required properties.

B. Appendix: Details of Goodness of Fit

After fitting, the parametric models are evaluated using the KL-divergence between the parametric fit and the test histogram. As such, the task presented is to determine how well each of the parametric models fit to the training histogram can approximate the empirical distribution of test samples as measured by a histogram. The shaded regions present 95% confidence intervals of the layer-by-layer KL-divergence as estimated using Student's t-test. A breakdown of the steps involved with this testing procedures are as follows.

1. Compute the KL-divergence for the non-zeros samples of each filter d within the target layer of D filters, We denote this KL-divergence as KL_d. This value measures how well the trained parametric model explains the test histogram of filter d.
2. Compute the sample mean of KL-divergence values across all non-zero features in the layer.

$$\overline{\mathrm{KL}} = \frac{1}{D}\sum_{d=1}^{D} \mathrm{KL}_d \tag{14}$$

3. Estimate the standard error of the mean (SE), which quantifies the uncertainty in the estimated average KL-divergence where s is the sample standard deviation of KL-divergences,

$$s = \frac{\sigma}{\sqrt{D}}, \quad \text{where } \sigma = \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}(\mathrm{KL}_d - \overline{\mathrm{KL}})^2} \tag{15}$$

4. Construct a 95% confidence interval (CI) around the mean. Since $N \geq 64$, we approximate the t-distribution with the standard normal distribution.

$$\mathrm{CI} = \overline{\mathrm{KL}} \pm z_{0.975} \cdot s, \quad \text{with } z_{0.975} \approx 1.96 \tag{16}$$

These intervals are visualized as shaded bands around the mean KL-divergence values in Figures 1 and 8. They indicate the uncertainty in the average KL-divergence for each fitted distribution and enable statistical comparison across distributions and layers. Overlapping intervals suggest no significant difference, while non-overlapping intervals indicate a statistically significant difference in goodness-of-fit.

C. Appendix: Experimental Design for Copula Inter-comparison

This appendix provides a detailed description of the methodology used in analysis of copula interdependence.

Separate Processing for Training and Testing

We have extracted training features from the training data and testing features from the testing data. First, we obtain the empirical marginal and interdependence terms strictly from the training data. Once we obtain these terms, we evaluate our model of copula interdependence by determining how well it fits the probability density of the withheld test features by using the criteria of cross entropy loss.

Probability Integral Transform

In our implementation, the empirical PIT is calculated by sorting all of the training features in the range $[0, n-1]$ in order to obtain a set of n ordered ranks. Typically, the probability integral transform converts a marginal distribution into a uniform distribution over the interval $(0, 1)$. However, in our approach, we carry out the analysis using a rescaled version of the probability integral transform that maps the feature values to the interval $(-1, 1)$. This rescaling is motivated by the fact that many standard orthogonal functions are defined on this interval, allowing us to represent the copula density in a richer and more flexible way without parametric assumptions. The modified probability integral transform is defined in the following equation.

$$F_i(x) = 2 \cdot \Pr[X_i \leq x] - 1 \tag{17}$$

Copula Density and Its Evaluation

The empirical moments $\hat{\mu}$ and copula density $\hat{c}$ are calculated in a C program that takes as input the entire training sample for the specified D features, and outputs a set of K^D empirical moments known as $\hat{\mu}$. As such the empirical moments are computed entirely from the training set. This set of moments further fully defines a model of the copula interdependence $\hat{c} : \mathbb{R}^D \rightarrow \mathbb{R}$ which is the dot product of the moments and the set of orthogonal functions.

$$\hat{c}(\vec{y}) \quad = \quad \hat{\mu} \cdot \Phi(y) \quad = \sum_{T \in \mathbb{Z}_K^D} \hat{\mu}_T \; \Phi(\vec{y}) \tag{18}$$

Now that our copula interdependence model is $\hat{c}$ is estimated from the training data, our task is to evaluate how well it models the probability density of the features from the test set. Cross entropy loss is used for the evaluation criteria.

For a given set of test features X_{test}, the transformed test features Y_{test} are calculated using the probability integral transform. Then, we use our model $\hat{c}$ to determine the predicted probability of the test features. This predicted probability is compared against the true probability of $1/N_{test}$ because empirically each test feature is equally likelihood. Therefore, the overall cross entropy evaluation is calculated using the following summation over the test set.

$$\text{Cross-Entropy} \; = \; -\frac{1}{N_{\text{test}}} \sum_{y \in Y_{test}} \log\left(\hat{c}(y)\right) \tag{19}$$

Confidence Intervals

This training and testing process is repeated 30 times for each model, dataset, and layer using a different subset of D features. For this analysis we used $D = 4$. By repeating this process 30 times, we straightforwardly calculate 95% confidence intervals using Student's t-test for the reported cross entropy loss statistics.

Archimedean copulas

Archimedean copulas make use of a Generator function $\Psi(y; \theta)$ that is invertible as follows.

$$C(y_1, y_2, \theta) \quad = \quad \Psi^{-1}\big(\; \Psi(y_1; \theta) \; + \; \Psi(y_2; \theta) \; ; \; \theta \; \big) \tag{20}$$

The Gumbel [11], Frank [10], Clayton [3], Ali Mikhail and Haq (AMH) [1], and Joe [17] generator functions are the most popular choices. The hyperparameter θ was determined using a formula based on either Spearman's ρ or Kendall's τ of the bivariate series. For Gumbel, Clayton and AMH, this formula is of closed form. For Joe and Frank the inverse formula is closed form. For the other methods, the inverses were calculated to high precision using binary search of the following equations.

$$\begin{aligned}
\text{Gumbel:} \quad & \theta = \frac{1}{1-\tau} \\
\text{Clayton:} \quad & \theta = \frac{2\tau}{1-\tau} \\
\text{AMH:} \quad & \theta = \frac{3\rho}{3+\rho} \\
\text{Joe:} \quad & \tau = 1 - 4\sum_{k-1}^{\infty} \frac{1}{k\big(\theta k + 2\big)\big(\theta(k-1)+2\big)} \\
\text{Frank:} \quad & \tau = 1 - \frac{4}{\theta}\left(1 - \int_0^{\theta} \frac{t}{e^t - 1} dt\right)
\end{aligned} \tag{21}$$

D. Appendix: Additional Analysis Plots

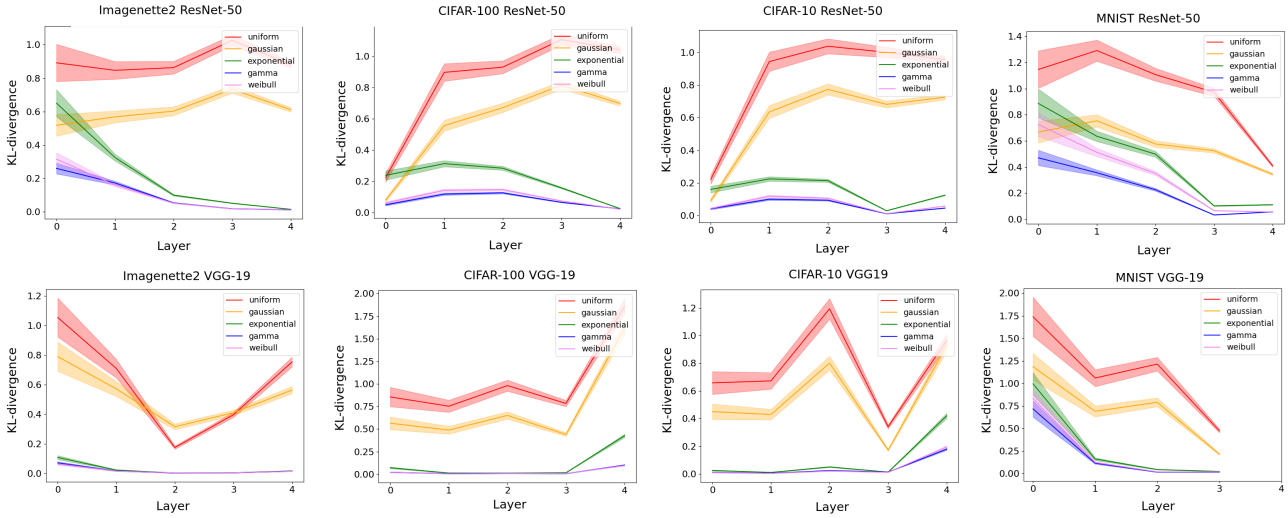

Figure 8. Overall goodness of fit of parametric marginal distributions, with shaded regions showing 2σ error bars.

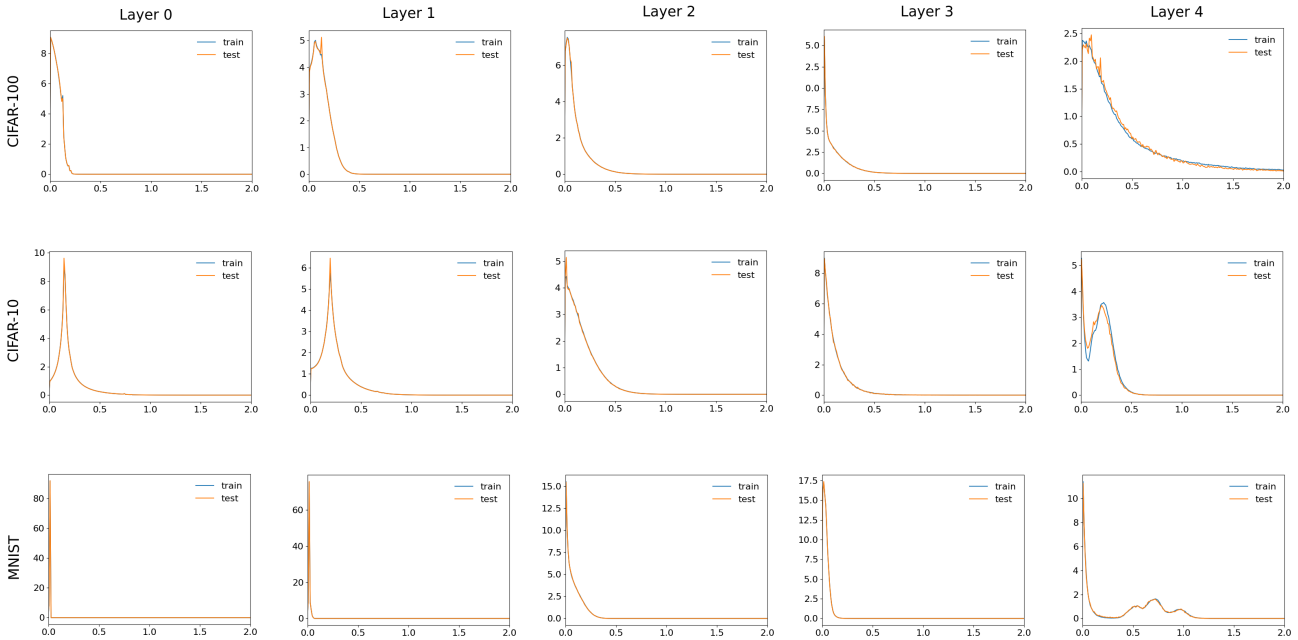

Figure 9. Supplemental examples of univariate histograms for ResNet-18 for feature 1 across CIFAR-100, CIFAR-10, and MNIST

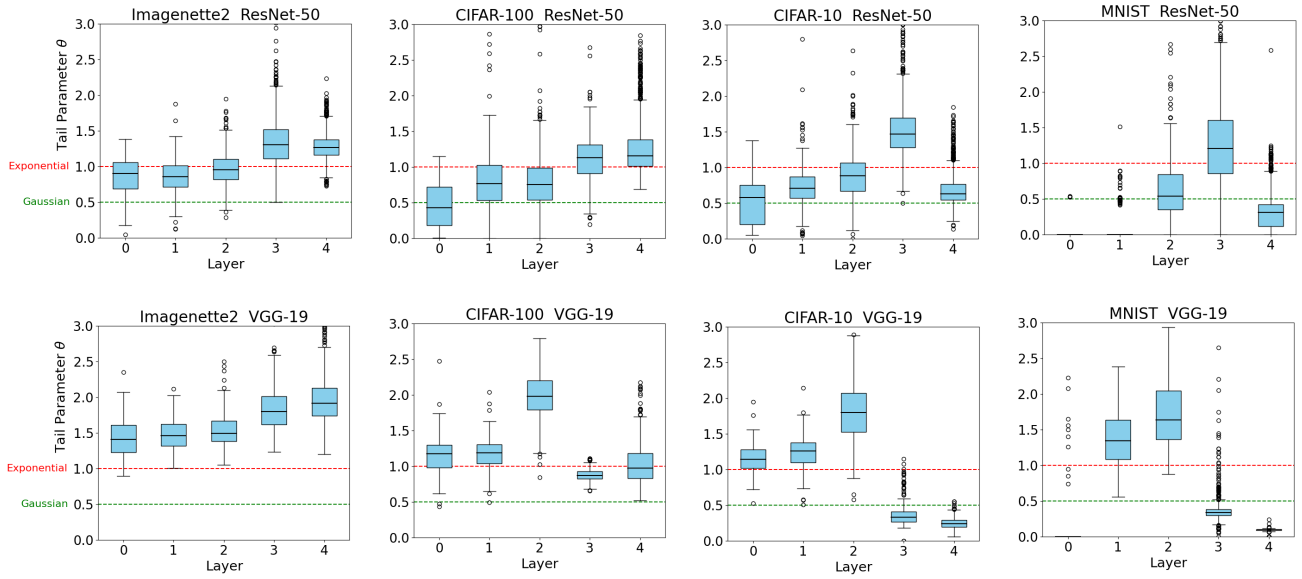

Figure 10. Estimated Weibull tail parameter θ for ResNet-50 and VGG-19 features across layers. Larger values correspond to heavier-tailed activation distributions.

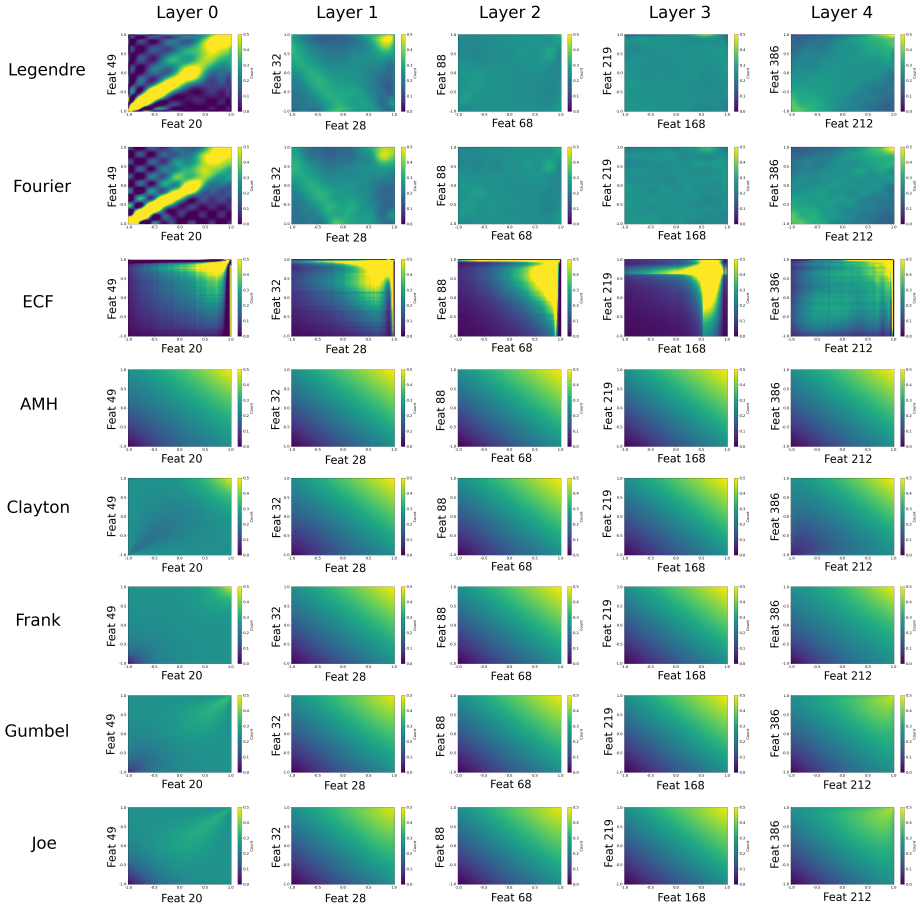

Figure 11. Comparison of copula density over random bivariate features for all methods using ResNet-18 on Imagenette2.