Neural Echos: Depthwise Convolutional Filters
Replicate Biological Receptive Fields

Zahra Babaiee
TU Vienna, MIT
zbabaiee@mit.edu

Peyman M. Kiasari
University of Waterloo
p2mohsen@uwaterloo.ca

Daniela Rus
MIT
rus@mit.edu

Radu Grosu
TU Vienna
radu.grosu@tuwien.ac.at

Abstract

In this study, we present evidence suggesting that depthwise convolutional kernels are effectively replicating the structural intricacies of the biological receptive fields observed in the mammalian retina. We provide analytics of trained kernels from various state-of-the-art models substantiating this evidence. Inspired by this intriguing discovery, we propose an initialization scheme that draws inspiration from the biological receptive fields. Experimental analysis of the ImageNet dataset with multiple CNN architectures featuring depthwise convolutions reveals a marked enhancement in the accuracy of the learned model when initialized with biologically derived weights. This underlies the potential for biologically inspired computational models to further our understanding of vision processing systems and to improve the efficacy of convolutional networks.

1. Introduction

Convolutional Neural Networks (CNNs) [31], a mainstay of modern artificial intelligence (AI), owe their fundamental design principles to insights drawn from neuroscience (NS) [24], particularly our understanding of receptive fields. A receptive field is the specific region of sensory space eliciting a response from a neuron when stimulated [14, 24]. The concept is deeply ingrained in the architecture of the mammalian visual system, starting from the retina. CNNs mimic this structure through their use of 'kernels' capable of responding to a specific part of the image. This convolution process mirrors the hierarchal, spatially invariant nature of biological vision systems, underscoring the deep connections between the fields of NS and AI.

The realm of convolutional neural networks (CNNs) has witnessed remarkable evolutionary phases since its inception. Initial architectures, such as AlexNet [27], introduced in 2012, focused on varying kernel sizes to capture image features. As the field matured, architectures like VGG net-
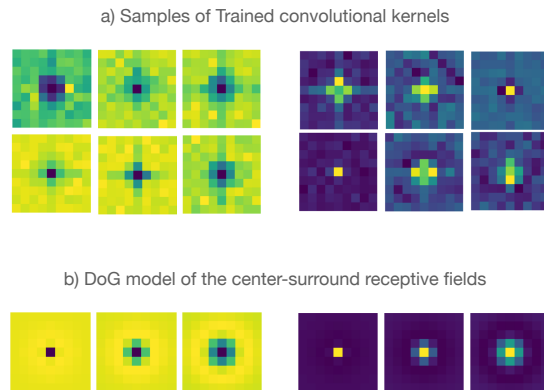

Figure 1. a) Depthwise Convolutional kernels trained on ImageNet dataset, and b) the DoG model of the biological center-surround receptive fields with different center-to-surround ratios, and with excitatory center (right) and inhibitory center (left), respectively. Artificial kernels mimic biological center-surround patterns.

works [39] and Residual Networks [18] standardized the use of 3x3 kernels, optimizing for efficiency and training speed. However, a pivotal shift emerged with the introduction and popularization of depthwise convolutions.

Depthwise convolutions introduced a novel approach to feature extraction, where each input channel is individually convolved with its own filter, as opposed to standard convolutions that aggregate information across multiple channels. This technique, exemplified by architectures like MobileNet with its 3x3 depthwise convolutions, offers significant reductions in computational overhead without markedly sacrificing model accuracy. The advent of vision transformers and their patch-centric designs [8] further accentuated the exploration into depthwise convolution behaviors with larger kernel sizes, reinforcing their unique ability to manifest structured patterns.

The cornerstone of visual processing in numerous retinal cell types, including the intricate network of ganglion neurons, is the principle of center-surround antagonism, a mechanism established in the receptive field of neurons, as
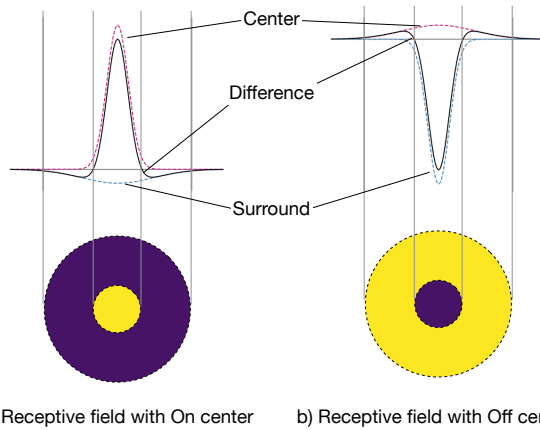
Figure 2. A "difference-of-Gaussians" is used to model a neuron's sensitivity to light at various positions on the retina. This model comprises two Gaussian functions - a narrow, positive one, representing the stimulatory center, and a wide, negative one, indicating the suppressive surround, for the neurons with an excitatory center, and the other way around for the ones with an inhibitory center.

early as in the retina [9, 24, 29]. This mechanism stems from lateral inhibitory connections and is perpetuated by neurons in higher visual processing centers, namely the lateral geniculate nucleus and the visual cortex [22].

The center-surround antagonism plays a vital role in the primate visual system, assisting in complex tasks such as edge detection, figure-background segregation, depth, and object perception, that remain consistent across various visual cues. Importantly, this architecture has two key configurations: excitatory- and inhibitory-center receptive fields, respectively [38, 44]. In the former configuration, ganglion cells are excited by light falling on the center of the receptive field and inhibited by light falling on the surrounding area. Conversely, in the latter configuration, cells are inhibited by light at the center and excited by light in the surrounding area. This design enhances contrast and aids in edge detection [9, 24].

Classical NS models frequently employ receptive fields featuring center-surround antagonism, typically realized through a Difference of Gaussians (DoG) function, which creates an excitatory peak at the receptive field's center counterbalanced by an inhibitory surround [10, 23].

In our investigations, we unearthed a remarkable parallel between the trained kernels of depthwise convolutions in various models and biological receptive fields: a significant quantity of them echoed the center-surround pattern seen in biological receptive fields. Intriguingly, such patterns were exclusively observed in depthwise convolutions, eluding their regular convolution counterparts. In Figure 1 we provide a comparative demonstration of the trained depthwise kernels and the NS-based model of center-surround kernels, highlighting their noteworthy similarities. This dis-
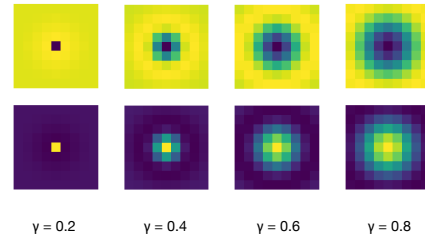

Figure 3. Size 9 DoG kernels with inhibitory (top) and excitatory (buttom) centers, with different ratios of the center-surround radiuses (γ).

covery underscores not only the computational advantages of depthwise convolutions but also their potential to mirror biologically-inspired patterns, reaffirming the value of re-thinking standard convolution operations in modern deep-learning paradigms.

Taking cues from these resemblances, we suggest a center-surround initialization procedure for depthwise convolutional kernels. Our experiments on ImageNet dataset with different models revealed that networks when initialized using our biologically-inspired methodology, display a marked increase in accuracy. Specifically, models initialized by our method gain up to more than two percent accuracy on the ImageNet dataset. Despite these notable improvements, the primary purpose of this paper is not solely to underscore performance enhancements. Rather, we aim to emphasize the intriguing discovery that artificial kernels emulate their biological counterparts without explicit supervision. Our results demonstrate the significant potential of biologically inspired computational models in enriching our comprehension of vision processing systems and enhancing the performance of artificial neural networks.

2. Related Work

Depth-wise Convolutions. The evolution of convolutional neural networks has been marked by the introduction and adaptation of diverse convolution operations. Notably, depthwise convolutions, where each input channel is convolved with its distinct filter, have gained traction. With the recent surge of modern enhanced CNN architectures, especially in the wake of the transformative impact of vision transformers, many models are now favoring depthwise convolutions with large kernels over traditional regular convolutions. Depthwise convolutions gained prominence with the introduction of the MobileNet architecture [20], which showcased their efficacy in crafting lightweight models tailored for mobile and embedded vision applications. With the resurgence of the modern CNN architectures after the introduction of vision transformers, many models use depthwise convolutions in their blocks [32, 35, 37, 42].

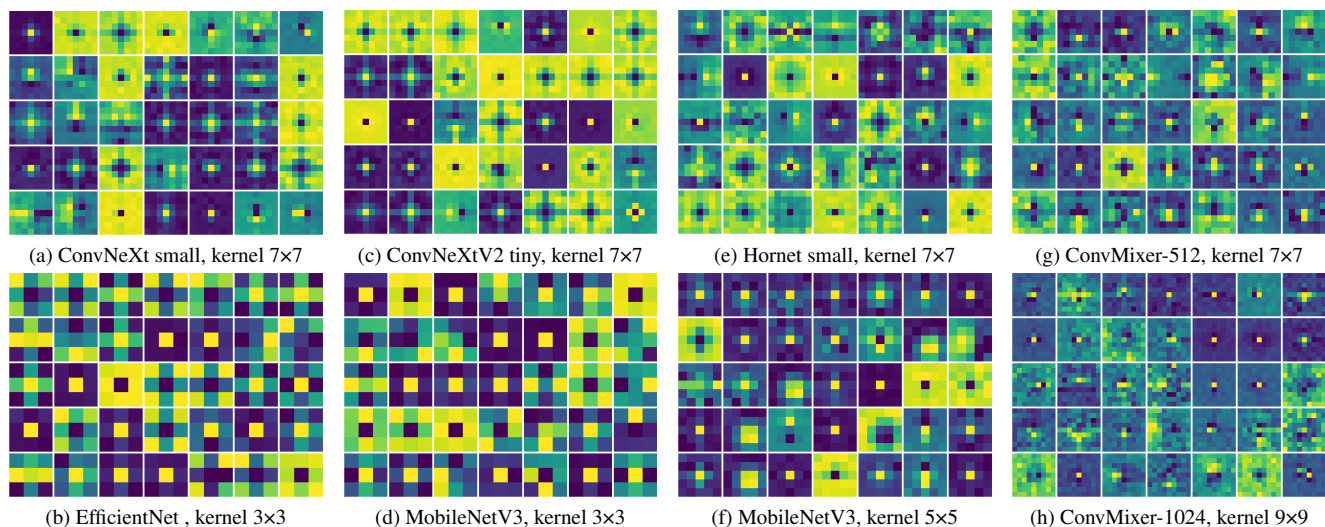**Bio-inspired Models.** A considerable volume of re-

Figure 4. Random samples from depth-wise convolutions of various models with different kernel sizes, trained on the ImageNet dataset. Trained kernels show considerable repeating patterns, many of them featuring a center-surround structure.

search has aimed to incorporate insights from NS into computer vision systems [25, 30, 47]. Initial vision models were significantly influenced by NS and psychology. In recent times, there have been substantial advances in both NS and AI, especially in computer vision. However, the majority of contemporary networks, are only loosely based on the visual system, and cross-fertilization between the two fields is less frequent as in the early days of AI. This is in spite of the fact that NS continues to be a vital source of innovative ideas that fuel advancements in AI [4, 16].

The development of AI models that closely resemble their biological counterparts and that incorporate advances in NS offers two primary advantages. Firstly, NS can be a fertile source of inspiration for designing new models and enhancing existing ones. This holds true both in isolation and in tandem with the computational and mathematical advancements underpinning new models. Secondly, NS can offer validation for existing models and methodologies in the AI domain. An example of this is the residual connections found in pyramidal cells within the cerebral cortex. These connections enable input from layer I to reach cortical layer VI neurons, bypassing intermediary layers [16, 41].

Center-Surround Receptive Fields. The Neocognitron model, proposed by Fukushima, holds a key position in the history of neural networks and machine learning, as one of the earliest examples of a CNN. Inspired by the pioneering work of Hubel and Wiesel on the visual cortex of cats [22], the Neocognitron model was designed to mimic the hierarchical structure of the visual system in mammals. It included a contrast-extracting preprocessing layer, reminiscent of the On-Off ganglion neurons, as well as inhibitory surround connections, mirroring the surround modulation observed in the visual cortex [10]. More recent studies have

tried to incorporate center-surround receptive fields into CNNs by integrating convolutional layers equipped with fixed kernels into the input feature maps. Evidence indicates that this modification enhances the network's performance and resilience, particularly with respect to variations in lighting conditions and input noise [2, 15].

Initialization Methods. Kernel initialization methods are crucial in training deep convolutional neural networks (CNNs) and have been the focus of significant research. The initialization of convolutional kernels directly impacts the convergence speed and the final performance of CNNs. Traditional initialization methods include Glorot and Bengio's uniform initialization [11], He et al.'s Kaiming initialization [17], and LeCun's Normal initialization. Mishkin and Matas introduced the LSUV initialization, optimized for deep architectures [33]. Hanin and Rolnick identified and addressed initialization failure modes in deep ReLU networks [12]. Arpit et al. proposed a robust initialization for weight normalized and ResNets, demonstrating enhanced generalization in deeper structures [1]. These methods usually generate weights from a Gaussian or uniform distribution with zero mean and a certain standard deviation. These initialization methods aim to maintain a reasonable activation variance across layers to avoid the issue of vanishing or exploding gradients.

Kaiming initialization, also known as He initialization [17], is a widely adopted technique for initializing weights in convolutional neural networks. Kaiming initialization addresses the vanishing/exploding gradients problem by initializing weights in a way that the variance of the outputs of each convolutional layer is approximately the same as the variance of its inputs. Specifically, the weights are initialized from a Gaussian distribution with a mean of

(a) ConvNeXt small, kernel 7×7 (b) ConvNeXtV2 tiny, kernel 7×7 (c) Hornet small, kernel 7×7 (d) ConvMixer-512, kernel 7×7
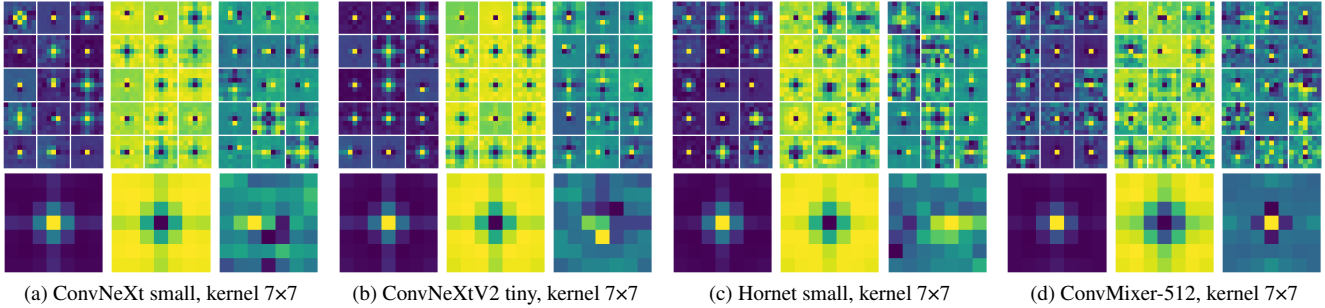
Figure 5. Kernels randomly selected from each K-Means cluster (top) and their respective cluster averages (bottom). Right clusters resemble excitatory-centered fields and middle clusters resemble inhibitory-centered. Clusters on the left contain all other patterns, resulting in their average being cluttered, implying the dominance in the first two cluster patterns.

zero and a standard deviation of $\sqrt{2/n}$, where n is the number of inputs to the neuron. This technique has been shown to significantly improve the speed of convergence in deep neural networks and to stabilize the training process.

3. Methods

In the following section, we delve into the specifics of our proposed methodology. We begin by conducting a comprehensive analysis of the trained kernels of various state-of-the-art (SOTA) models on the ImageNet dataset. We provide detailed visual illustrations coupled with quantitative results, to validate the presence of the center-surround antagonism in a considerable portion of the kernels.

Next, we move on to discuss the Difference of Gaussian (DoG) function. This function serves as a mathematical model of the center-surround receptive fields found in biological visual systems. This mathematical model provides us with a foundation for designing an initialization scheme that mimics these biological structures.

Finally, we describe our novel kernel initialization approach, which leverages the DoG function. We detail the process of applying this function to generate kernel weights that resemble the center-surround antagonism of biological vision systems. The ultimate goal is to provide the model with a starting point that is already attuned to the kind of spatial feature mappings it would otherwise have to learn through many epochs of training.

3.1. Inspecting the Trained Kernels

In this section, we conduct a detailed exploration of the trained kernels of the regular and depthwise convolutions in different models. Specifically, we examine VGG16 [39], ResNet50 [18], DenseNet201 [21], MobilenetV3 [19], EfficientNet [40], ConvNeXt [32], ConvNeXtV2 [37], Hornet [35], and ConvMixer [42]. For our analysis, we utilized the pre-trained versions of these models, sourced directly from Pytorch or their respective official code repositories.

In order to visually inspect the learned patterns within these filters, we have included Figures 4 and 6 in our paper.

Figure 4 provides a representative selection of randomly chosen samples from the trained kernels of each of the models utilizing depth-wise convolutions, and Figure 6 shows the same for models with regular convolutions. Upon inspection of these kernels, we can identify recurring patterns among the depthwise convolutions. We observed similar patterns in other variants of these models trained on ImageNet, too. However, kernels of regular convolutions do not show such observable patterns
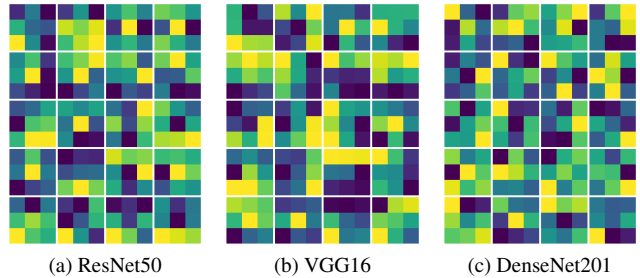

(a) ResNet50 (b) VGG16 (c) DenseNet201

Figure 6. Random samples from regular convolutions of popular models, trained on the ImageNet dataset. Unlike depthwise convolutions, kernels from regular convolutions do not have visually observable repeated patterns.

One particularly notable pattern in depthwise kernels is the center-focused structure of many of them. Interestingly, these center-focused kernels can be broadly divided into two categories. The first category includes kernels with larger weight values concentrated in their center. Conversely, the second category consists of kernels with larger weight values populating their surround.

These discovered patterns bear striking resemblance to the well-studied 'center-surround antagonism' found in the mammalian visual system which we discussed previously.

Nevertheless, it's worth noting that not all filters from our models exhibit this center-surround pattern. We observed other filters possessing different, non-center-surround patterns, but their occurrence was less frequent compared to the center-surround ones. This differential frequency points to the significance and prevalence of the
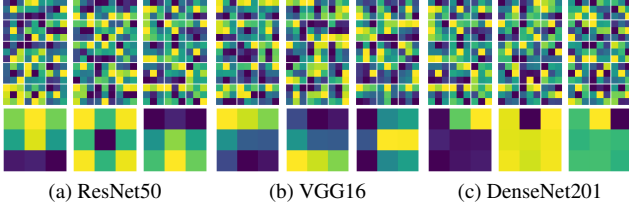
(a) ResNet50 (b) VGG16 (c) DenseNet201

Figure 7. Clusters of 3×3 kernels of regular convolutions

center-surround pattern in the learned representations of depthwise kernels in these models.

To attain a more analytic comprehension of the varying kernel patterns, we leverage a straightforward clustering algorithm to group all the kernels. Our hypothesis suggests that the two center-surround groups are the most significant patterns, thus we set the number of clusters to three in the clustering algorithm. This choice is to discern whether the algorithm can effectively categorize the kernels into two distinct center-surround clusters, and a third cluster comprising the less common patterns.

For a successful execution of clustering, we took some preparatory steps. First, we normalized all kernels to have their weight values lie within the range of 0 and 1, utilizing min-max encoding. This step is essential to ensure the numerical stability and effectiveness of the clustering algorithm. Following normalization, we flattened each kernel into a vector, to fit the input requirement of the k-means algorithm. With the transformed data, we were finally able to run the k-means algorithm [13] with a k-value of 3.

After clustering, we visually inspected the kernels belonging to each cluster, by presenting randomly selected samples in Figure 5. For each cluster, we also depict the average of all kernels belonging to it. This helps in better seeing the prominent pattern of each cluster. The first cluster predominantly contained kernels akin to the excitatory-center receptive fields, while the second cluster closely resembled inhibitory-center receptive fields. As for the third cluster, the kernels exhibited some degree of center-focused structures but lacked the precise characteristics of center-surround receptive fields. Examination of the average kernel within each cluster reveals a pronounced center-surround structure in the first two clusters, whereas the third cluster exhibits a more dispersed pattern. This observation further underscores that even an unbiased clustering approach distinctly recognizes the prominence of center-surround patterns relative to alternative patterns.

Figures 7 and 8 show the discovered clusters of the models with 3 × 3 kernels with regular and depthwise convolutions, respectively. As one can see, the models with small depthwise kernels still have prominent center-surround clusters, in contrast to the ones with regular convolutions. Only in Resnet50, the average kernels of one cluster is similar to the center-surround pattern. However,

once inspecting the kernels themselves, we can not distinguish the patterns (See Figure 6).
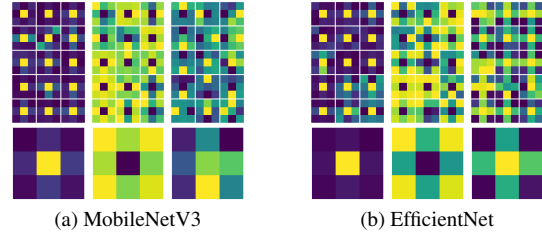

(a) MobileNetV3 (b) EfficientNet

Figure 8. Clusters of 3×3 kernels of depthwise convolutions

The distribution of the kernels across each cluster of ConvNeXt compared to ConvNeXtV2 variants is demonstrated by the histograms shown in Figure 9. The ConvNeXt model faced challenges with feature collapse, manifesting as redundant activations across channels. In response, the ConvNeXt V2 architecture introduces the Global Response Normalization (GRN) layer, promoting feature diversity. This enhancement, coupled with advanced self-supervised techniques, positions ConvNeXtV2 as a marked improvement over its predecessor in visual recognition tasks. We note a significant reduction in the number of kernels within the third cluster of the improved ConvNeXtV2 when contrasted with its predecessor, ConvNeXtV1. This is an interesting observation that underscores the pivotal role of center-surround in enhancing the model's performance.

As detailed in Figure 10, we observed a remarkable consistency in the proportions of filter clusters across various models, despite changes in model sizes, kernel sizes, and dataset sizes. This trend was evident in models such as ConvNeXt, ConvNeXtV2, and Hornet, where different model sizes were analyzed. For MobileNet and ConvMixer, we extended this analysis to include variations in kernel sizes. Additionally, the training of MobileNet on both ImageNet 1K and 21K datasets did not significantly alter the proportion of filter clusters. These findings suggest an inherent stability in the distribution of filter types within each model category, indicating that the architectural design of these models plays a more critical role in determining filter distribution than the scale of the model or the size of the dataset. This insight could have implications for understanding the scalability and adaptability of these models to different sizes and types of datasets.

3.2. Formulation of Center-Surround Kernels

The computation of center and surround weights can be accomplished using a difference of two Gaussian functions (DoG). Represented in Cartesian coordinates (CC), with the CC origin designated as the receptive field's center, the DoG can be formulated as presented in Rodieck's work [36]:

$$DoG(x,y) = K_1 e^{-\frac{x^2+y^2}{\sigma_1}} - K_2 e^{-\frac{x^2+y^2}{\sigma_2}} \quad (1)$$
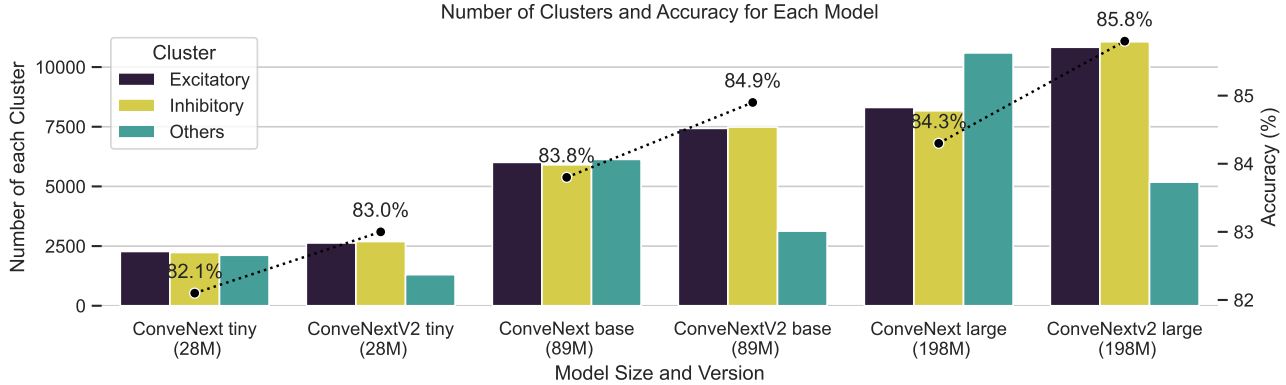
Figure 9. Histograms of the clusters discovered in ConvNeXt model variants, alongside their V2 counterpart, including the test accuracy of each model. The improved V2 versions have a considerably lower number of kernels in their "others" cluster.

where it holds true that $K_1, >, K_2$ and $\sigma_2, >, \sigma_1$ [3].

We use the DoG model proposed by Petkov and Kruizinga [28, 34], which defines the difference of Gaussians for the center and surround kernels. This model enables us to calculate the variances analytically, given the kernel size and the ratio of the center to surround [34]:

$$DoG_{\sigma,\gamma}(x,y) = \frac{A_c}{\gamma^2} e^{-\frac{x^2+y^2}{2\gamma^2\sigma^2}} - A_s e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

In this formula, $\gamma < 1$ stipulates the ratio between the center radius r and the surround. The coefficients A_c and A_s are determined by requiring the sum of all positive values in Equation 2 to be equivalent to the negative values. These are then normalized such that their sum equals 0.5 and -0.5, respectively. While in the continuous infinite case, the coefficients A_c and A_s are equal, in the discrete finite case, the values of A_c and A_s remain remarkably similar.

By setting $DoG_{\sigma,\gamma}(x,y) = 0$, σ can be calculated immediately as shown in Equation 3 below, where k is the kernel size, for any arbitrary values of k and γ:

$$\sigma \approx \frac{k}{4}\sqrt{\frac{1-\gamma^2}{-\ln\gamma}} \quad (3)$$

In Figure 2, we show the DoG functions used to model the center-surround receptive fields, with either excitatory or inhibitory centers, respectively.

3.3. Center-Surround Initialization

Observing the repetitive patterns exhibited in the depth-wise kernels trained on ImageNet, and noting their resemblance to center-surround receptive fields, we propose a novel methodology for kernel-weight initialization. Our hypothesis is based on the assumption that by offering the model kernels an initialization that aligns with patterns not only found in nature but also in fully-trained models, we

can enhance the performance of these models. Additionally, this approach might streamline the convergence process during training. This method potentially serves as a bridge, linking biological vision models with their artificial counterparts, thereby enabling the latter to benefit from the intrinsic efficiency of the former.

In our approach, we begin by initializing the weights of the depth-wise convolution kernels in the model architectures with the weights derived from the previously discussed DoG function. This is a crucial step that enables us to effectively incorporate the center-surround receptive field structure into the model. To achieve a balanced representation of both inhibitory and excitatory centers, each kernel is assigned a center type – either inhibitory or excitatory – with an equal probability of 50%. This ensures that both types of centers are represented in approximately equal proportions across the kernels, thus maintaining a balanced interaction of these opposing neural behaviors in the model.

Additionally, we introduce variability in the ratio of the center to the surround for each kernel. To do this, we select the ratio from a uniform distribution. This introduces an element of randomness to the model, ensuring that a wide variety of center-to-surround ratios are represented in the kernels. Consequently, this design enables the model to accommodate and respond to a broad range of spatial scales in the input data and adds a wider variety of weight values to the initialization.

Figure 3 shows the DoG kernels of size 9 with different ratios of the center to surround varying from 0.2 to 0.8. As the ratio increases, the center gets larger.

4. Experiments

Here, we detail our implementation, covering model selection and training. We then showcase results from testing our initialization on ConvNeXt, HorNet, and ConvMixer models using Cifar10 [26] and ImageNet [6]. Lastly, we provide an ablation study on our approach's facets.
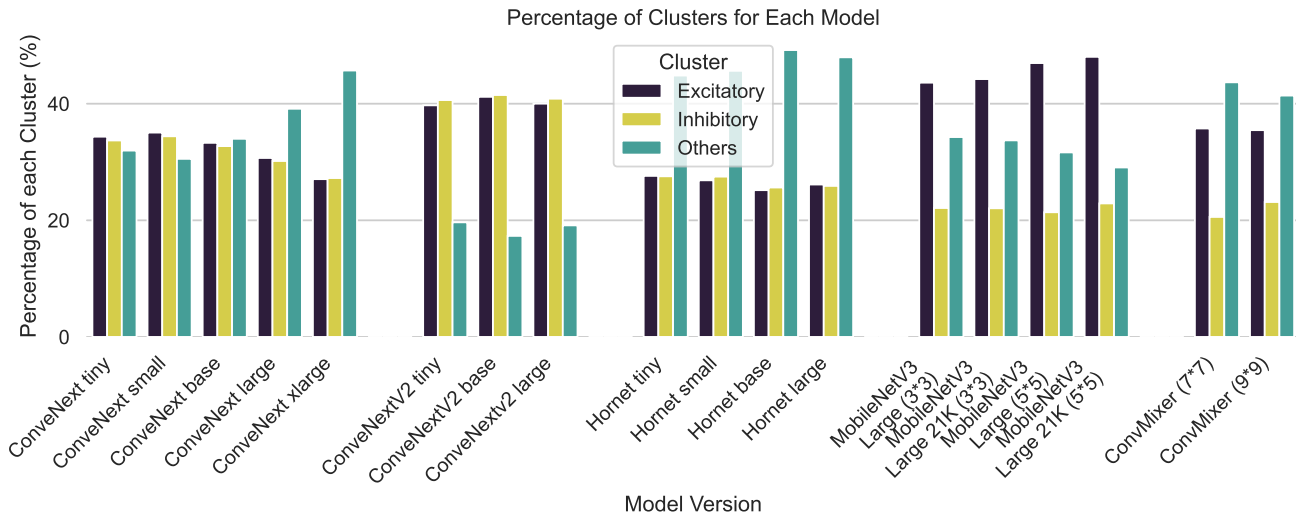
Figure 10. Filter Proportions by Cluster in Various Models: We observe almost consistent proportions of filter clusters across models of different sizes, kernel sizes, and dataset sizes within each model category.

4.1. Implementation Details

For the ImageNet evaluations, we employed ConvNeXt tiny, HorNet Tiny, and ConvMixer-512. ConvNext tiny and HorNet tiny contain 18 and 25 blocks respectively, each composed of one depthwise and two pointwise convolutions. The ConvMixer models incorporate 512 filters, within each depthwise convolutional layer and comprise mixer blocks composed of depthwise and pointwise convolutions. For each model, we used the training settings proposed in the original paper. Our training regimen included the use of a suite of data augmentation techniques, namely RandAugment [5], mixup [45], CutMix [43], and random erasing [46], in addition to gradient norm clipping. We employed the Adam optimizer [7] for the training process.

Across all experimental evaluations, we adhered to a balanced strategy for our kernel initialization, assigning half of the kernels with excitatory centers and the other half with inhibitory centers. Moreover, to determine the value of γ, representing the ratio of center to surround, we utilized a uniform distribution inside [0,0.5]. This choice is motivated by our observations derived from the trained filters, which showed us that the centers are usually quite small.

4.2. Results

In the following, we describe our experimental results on ImageNet and Cifar10 datasets. We compare our results to the Kaiming initialization, which is the default initialization method used in most of the models.

ImageNet. In Table 1 we present the empirical results of our experiments on ImageNet. Across all configurations, our initialization consistently outperforms the Kaiming initialization, with improvements ranging from marginal to

substantial. This was particularly evident in the improvement exceeding 2% for the ConvMixer-512 with kernel-size 9×9.

First, we present the performance metrics for the ConvNeXt tiny model, utilizing a 7×7 kernel size and trained over 50 epochs. Employing the conventional Kaiming initialization, the model achieved an accuracy of 76.17%. However, when initialized with our proposed method, the accuracy exhibited a slight enhancement, reaching 76.74%.

The subsequent row provides the results for the HorNet tiny model under analogous conditions: a kernel size of 7×7 and a training duration of 50 epochs. The performance with the Kaiming initialization stood at 76.06%. In contrast, our innovative initialization method yielded a superior result, registering an accuracy of 76.40%.

Finally, we explored the effectiveness of our method with larger kernel sizes using the ConvMixer-512 model. Specifically, we employed a kernel size of 9 in the ConvMixer architecture. This approach resulted in a significant improvement, with an accuracy increase of 2.34%, thereby affirming even better efficacy of our method when applied to models with larger kernel configurations.

Cifar10. We evaluated our initialization on Cifar10 with models with kernel sizes of 5, 7, and 9. However, across different runs, we observed little to no improvements. This may be primarily attributed to the lack of discernible patterns in filters trained on Cifar10, in contrast to their ImageNet counterparts. This disparity is likely rooted in the significant size difference between the datasets, both in number of classes and image sizes. Clusters of kernels trained on Cifar10 are depicted in Figure 11.

Table 1. Results of Depthwise Convolutional Models on ImageNet with different settings and initializations.

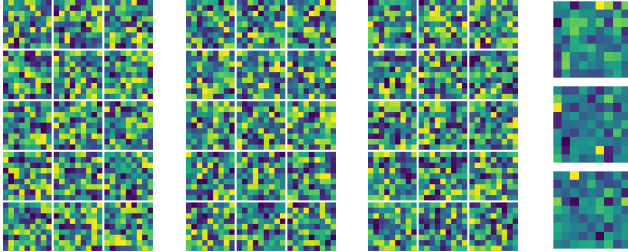| Model	kernel Size	Kaiming Initialization	Our Initialization
ConvNeXt tiny	7×7	76.17	**76.74**
HorNet tiny	7×7	76.06	**76.40**
ConvMixer-512	9×9	64.00	**66.34**

Figure 11. Clusters from kernels trained on the Cifar10 dataset (left) and the average of each cluster (right).

Table 2. Ablation on initialization settings with ConvMixer-512 with kernel size 9×9 on ImageNet.

Initialization	Accuracy
Kaiming	64.00
Ours, $\gamma \in (0,1)$, On and Off Centeres	65.20
Ours, $\gamma \in (0,0.5)$, Only On Centers	64.78
Ours, $\gamma \in (0,0.5)$, On and Off Centeres	**66.34**

4.3. Ablation Study

To assess our initialization's impact, we conducted an ablation study on the ConvMixer model, as this model exhibited the most significant improvement from our initialization method. The study, performed on the ImageNet dataset, is detailed in Table 2.

Our baseline evaluation using Kaiming initialization achieved 64.00% accuracy. We applied our initialization, adjusting the DoG function parameters and the excitatory/inhibitory center arrangement. First, we sampled γ from a uniform distribution between (0,1) and used both On (excitatory) and Off (inhibitory) centers. This yielded an improved accuracy of 65.20%.

Next, we narrowed the range of γ to (0,0.5), while only utilizing On centers. The resulting accuracy, though slightly lower at 64.78 percent, still exceeded the baseline Kaiming initialization. This suggests that the selection of γ and center types both play significant roles in the performance.

Finally, maintaining γ in the (0,0.5) range, we reintroduced both On and Off centers into our model. This resulted in the highest observed accuracy of 66.34%. It is clear from this ablation study, that the selection of γ and the type of centers (On or Off) significantly influence the model performance.

5. Conclusions and Future Work

Conclusions. This paper has delved into an intriguing discovery of center-surround patterns in depthwise convolutional kernels, highlighting a fascinating interplay between artificial neural networks and natural vision systems. We capitalized on this finding by introducing a novel initialization strategy for depthwise kernels, incorporating the principles of the DoG method, typically utilized in bio-inspired vision models. This unique approach taps into the center-surround antagonism property of retinal ganglion cells, offering enhanced contrast sensitivity, mirroring the proficiency of biological vision systems.

The empirical evidence from our extensive experiments on ImageNet firmly backs the efficacy of our proposed method. Compared to the widely used Kaiming initialization, our technique demonstrated a notable improvement in the accuracy of the models. As illustrated in Figure 11, it is also interesting to observe that on the Cifar10 dataset, models do not seem to be able to learn biological kernels so effectively. In our opinion, this is a result of the small size of their images (32×32 compared to 224×224 in ImageNet), and the very limited number of their classes (10 compared to 1000 in ImageNet).

Future Work. While the results obtained are promising, there's still scope for further exploration. A promising direction is to test our initialization method across a broader range of CNNs, potentially advancing a ubiquitous biology-inspired initialization approach.

Furthermore, the primary aim of this paper was to highlight the resemblance between trained kernels and their biological counterparts. We have not yet embarked on any form of hyperparameter search to fine-tune the parameters of our initialization method. Parameters like the range for γ, the proportion of excitatory and inhibitory kernels, initialization specific to each layer, and the balance of positive and negative values in the Difference of Gaussians function have been left unexplored. Potentially, these factors could be tweaked for optimal performance.

Finally, this paper has not explored the patterns in the "Others" cluster of Figure 9. It is very likely that these patterns are linked to biological neural processing, too.

6. Acknowledgements

References

[1] Devansh Arpit, Víctor Campos, and Yoshua Bengio. How to initialize your network? robust initialization for weight-norm & resnets. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[2] Zahra Babaiee, Ramin Hasani, Mathias Lechner, Daniela Rus, and Radu Grosu. On-off center-surround receptive fields for accurate and robust image classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 478–489. PMLR, 18–24 Jul 2021. 3

[3] M.R. Blackburn. A Simple Computational Model of Center-Surround Receptive Fields in the Retina. Technical Report 2454, Ocean Surveillance Center, Feb 1993. 6

[4] Rodney Brooks, Demis Hassabis, Dennis Bray, and Amnon Shashua. Turing centenary: Is the brain a good model for machine intelligence? *Nature*, 482:462–3, 02 2012. 3

[5] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019. 7

[6] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and F. Li. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, Florida, USA, June 2009. IEEE Computer Society. 6

[7] P.K. Diederik and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015. 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 1

[9] C. Enroth-Cugell and L. H. Pinto. Properties of the Surround Response Mechanism of Cat Retinal Ganglion Cells and Centre-Surround Interaction. *The Journal of Physiology*, 220(2):403–439, Jan 1972. 2

[10] K. Fukushima. Neocognitron for Handwritten Digit Recognition. *Journal of Neurocomputing*, 51:161–180, 2003. 2, 3

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 3

[12] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture, 2018. 3

[13] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979. 5

[14] H. K. Hartline. The receptive fields of optic nerve fibers. *American Journal of Physiology-Legacy Content*, 130(4):690–699, 1940. 1

[15] Hosein Hasani, Mahdieh Soleymani, and Hamid Aghajan. Surround modulation: A bio-inspired connectivity structure for convolutional neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[16] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016. 1, 4

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[20] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 2

[21] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017. 4

[22] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968. 2, 3

[23] Jorn-Henrik Jacobsen, Jan Van Gemert, Zhongyu Lou, and Arnold WM Smeulders. Structured receptive fields in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2610–2619, 2016. 2

[24] E.R. Kandel, T.M. Jessell, J.H. Schwartz, S.A. Siegelbaum, and A.J. Hudspeth. *Principles of Neural Science*. Fifth Edition. McGraw-Hill Medical / Education, 2013. 1, 2

[25] Jonghong Kim, O Sangjun, Yoonnyun Kim, and Minho Lee. Convolutional neural network with biologically inspired retinal structure. *Procedia Computer Science*, 88:145–154, 2016. 3

[26] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 6

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[28] P. Kruizinga and N. Petkov. Computational Model of Dot-Pattern Selective Cells. *Biological Cybernetics*, 83(4):313–325, Jun 2000. 6

[29] S.W. Kuffler. Discharge Patterns and Functional Organization of Mammalian Retina. *Journal of Neurophysiology*, 16(1):37–68, 1953. 2

[30] Md Nasir Uddin Laskar, Luis G Sanchez Giraldo, and Odelia Schwartz. Correspondence of deep neural networks and the brain for visual textures. *arXiv preprint arXiv:1806.02888*, 2018. 3

[31] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 1

[32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[33] Dmytro Mishkin and Jiri Matas. All you need is a good init, 2016. 3

[34] N. Petkov and W. Visser. Modifications of Center-Surround, Spot Detection and Dot-Pattern Selective Operators. Technical Report 2005-9-01, Institute of Mathematics and Computing Science, University of Groningen, Netherlands, 2005. 6

[35] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser-Lam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 4

[36] R. Rodieck. Quantitative Analysis of Cat Retinal Ganglion Cell Response to Visual Stimuli. *Vision Research*, 5(12):583–601, 1965. 5

[37] Ronghang Hu Xinlei Chen Zhuang Liu In So Kweon Sanghyun Woo, Shoubhik Debnath and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 2, 4

[38] R. Shapley and V.H. Perry. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciences*, 9:229 – 235, 1986. 2

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 4

[40] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 4

[41] Alex Thomson. Neocortical layer 6, a review. *Frontiers in Neuroanatomy*, 4:13, 2010. 3

[42] Asher Trockman and J. Zico Kolter. Patches are all you need? *CoRR*, abs/2201.09792, 2022. 2, 4

[43] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7

[44] K.A. Zaghloul, K. Boahen, and J.B. Demb. Different circuits for on and off retinal ganglion cells cause different contrast sensitivities. *Journal of Neuroscience*, 23(7):2645–2654, 2003. 2

[45] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 7

[46] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, Apr. 2020. 7

[47] Georgios Zoumpourlis, Alexandros Doumanoglou, Nicholas Vretos, and Petros Daras. Non-linear convolution filters for cnn-based learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4761–4769, 2017. 3