Neuromanifold-Regularized KANs for Shape-fair Feature Representations

Mazlum Ferhat Arslan[1,2] Weihong Guo[1] Shuo Li[2,3*]
[1]Dept. of Mathematics, Applied Mathematics and Statistics [2]Dept. of Biomedical Engineering
[3]Dept. of Computer and Data Sciences
Case Western Reserve University
{mxa1328,wx49,shuo.li11}@case.edu

Abstract

*Traditional deep networks struggle to acquire shape-fair representations due to their high expressivity. Kolmogorov-Arnold Networks (KANs) are promising candidates as they learn nonlinearities directly, a property that makes them more adaptive. However, KANs perform suboptimally in terms of shape-fairness because of unconstrained nonlinearities, a limitation we demonstrate for the first time. On the other hand, shape-fair networks reside on a neuromanifold of low-degree. Motivated by this, we investigate neuromanifold regularization of KANs to enable learning of shape-fair feature representations. The proposed method, **N**euro**M**anifold **R**egularized-**KAN**s, is a novel regularization that addresses failure modes during the acquisition of local and global shape cues, separately. This is done by constraining the degree of the neuromanifolds of two jointly trained feature extractors. Additionally, we propose a novel Style Decorrelation Loss that promotes decorrelation of intermediate representations. Our experiments demonstrate that NMR-KAN improves shape bias over baseline convolutional KANs by 14.8% while also providing robustness under image corruptions and adversarial attacks. Code is avaliable at:* ***http://www.github.com/kaptres/NMR-KAN/***.

1. Introduction

Image shape extraction is the process of identifying structural object features —such as contours, boundaries, and part relationships—while minimizing reliance on fine-scale texture details that can mislead recognition systems. It is typically modelled as a two-stage nonlinear process [3, 26, 36, 43]. Local shape cues, such as contours and object parts, are first extracted. These are then aggregated into structured global representations. Given the domain-invariance of object shapes, correct shape processing potentially implies robustness of the networks under image-space corruptions and adversarial attacks, and generalizability of the networks to domains unseen during training [20, 44].

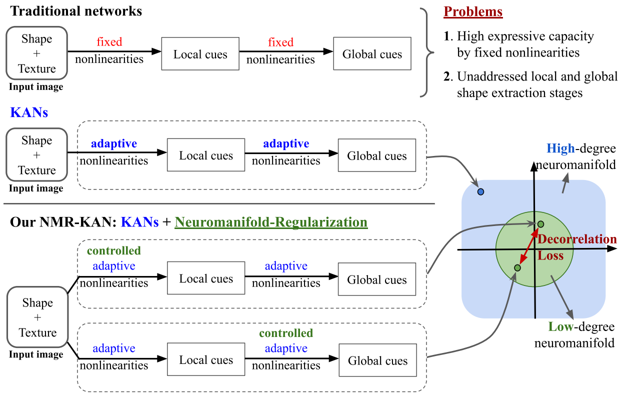

Figure 1. NMR-KAN employs two KAN feature extractors with constrained adaptive nonlinearities at different stages. The feature extractors lie in a lower degree neuromanifold, leading to better shape-models due to the emphasis on learning *simpler* functions in the extraction of *both local and global* cues.

Existing image processing methods (Figure 1, top) struggle to acquire shape-fair representations. High expressivity in traditional networks enables the capture of complex features but also leads to learning non-generalizable, dataset-specific details [1, 21, 49]. In computer vision, this often results in texture-biased representations, where models rely on high-frequency details rather than structural form [3, 14, 19]. Data augmentation techniques mitigate this to some extent, but they do not address the root cause: unconstrained network expressivity. This leads to overly complex models that fall short of robustness expectations [39]. Other approaches, such as sparsity-based methods [31], require large datasets, making them impractical in low-data

[*]Corresponding author: shuo.li11@case.edu

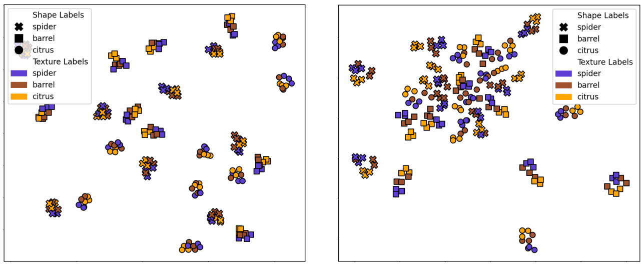

Figure 2. Representations of cue conflict images have lower shape spread for our NMR-KAN (left) than the traditional networks (right). Representation mapping my UMAP [37].

scenarios where synthetic augmentation is insufficient.

Thanks to their advantage of learning nonlinearities directly, Kolmogorov-Arnold Networks (KANs) [33] are promising candidates for modeling nonlinear processes in computer vision tasks [2, 22, 47, 48]. However, KANs perform suboptimally in terms of shape-fairness, a limitation we demonstrate for the first time. When unconstrained, the adaptive nonlinearities have excessive expressivity, resulting in similar issues in traditional networks (Figure 1, middle). Furthermore, KANs do not inherently distinguish between local and global shape processing stages, limiting their ability to form structured representations.

Neuromanifolds provide a means to regulate network expressivity, offering several advantages for shape-fairness. Low-degree neuromanifolds, which have lower expressive capacity, contain shape-fair networks [45] that generalize across minor variations. Additionally, neuromanifold regularization enables targeted control over local and global shape processing stages, aligning network design with human-like perception. Beyond empirical benefits, this approach also provides a principled mathematical framework for enforcing shape-fair representations.

The power of combining KANs and neuromanifold regularization (NMR) lies in their complementary strengths. KANs learn what features to extract by adapting nonlinearities to shape-relevant information, overcoming the limitations of fixed nonlinearities in CNNs. NMR controls where these nonlinearities reside, constraining the search space to prevent capture of texture. This, along with our novel strategy of specializing network stages for local and global shape cues via distinct NMR constraints, yields robust, shape-biased representations. Traditional methods lack this targeted, hierarchical control, hindering their ability to achieve similar shape-fairness (Figure 2).

We propose **N**euro**M**anifold-**R**egularized KANs (NMR-KANs) for shape-feature representation (Figure 1, bottom). Our contributions are *i)* for the first time shape-fairness in KANs is investigated, *ii)* for the first time the strengths of neuromanifolds and KANs are combined together via a novel architectural regularization strategy (Sec. 3.1) and a novel loss function (Sec. 3.2), *iii)* for the first time the shape-fairness problem in neural networks is addressed explicitly in terms of the local and global cue extraction stages (Sec. 3.3).

2. Related Work

Why Neuromanifolds? Neuromanifolds provide a principled way to control network expressivity. In particular, a low-degree neuromanifold defines a constrained function space, where shape-fair networks reside [45], limiting expressivity to prioritize structural shape cues over texture.

A neuromanifold, distinct from the data manifold in manifold learning, is the *function space* containing functions a given neural network can represent [9, 24, 27, 41]. Given an architecture, f, and the set of parameters Φ, it is the space $\mathcal{M} = \{f_\phi \mid \phi \in \Phi\}$. For a convolutional network with polynomial nonlinearities of depth L, kernel sizes $\mathbf{k}$, stride s, layer dimensions $\mathbf{d}$, and polynomial degree r, we denote the neuromanifold by $\mathcal{M}_{\mathbf{d},\mathbf{k},s,r}$.

Neuromanifold degree and dimensionality govern expressivity. Higher degrees enable complex functions but risk overfitting and thus reducing overall robustness. Formal definitions for polynomial CNNs are [41]:

$$\dim(\mathcal{M}_{\mathbf{d},\mathbf{k},s,r}) = |\mathbf{k}| - L + 1, \tag{1}$$

$$\deg(\mathcal{M}_{\mathbf{d},\mathbf{k},s,r}) = (|\mathbf{k}| - L)! \prod_{0 \leq j < L} \frac{r^{(L-j-1)(k_j-1)}}{(k_j - 1)!}. \tag{2}$$

Standard regularization techniques, such as weight decay, indirectly influence expressivity by penalizing parameter magnitudes. In contrast, neuromanifold-based regularization *directly constrains* the function space itself. Rather than merely discouraging certain parameter sets, NMR prevents the network from learning functions outside the neuromanifold, ensuring structured expressivity.

Kolmogorov-Arnold Networks KANs [34] aim to learn nonlinearities directly, based on the Kolmogorov-Arnold representation theorem (KART) which states a function can be exactly represented as a sum of single-variable functions,

$$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \varphi_{p,q}(x_p) \right) \tag{3}$$

where n is the dimensionality of the input space. While KART provides an upper bound on the computational units, it is at the cost of finding possibly more complex functions φ and Φ than the fixed nonlinearities in traditional NNs.

In [34], parameterization of the nonlinearities Φs and φs via B-splines is proposed. Similar to how fully connected layers, and their deeper counterparts that extend UAT, can be conveniently written via matrix notation as

$$f \approx \sigma_n \left(A_n \left(\ldots \sigma_2 \left(A_2 \, \sigma_1 \left(A_1 x + b_1 \right) + b_2 \right) \ldots \right) + b_n \right),$$

treating subscripts p, q as the indices of a matrix of nonlinear functions, $\Psi^{(j)}$, the stacked model can be written as

$$f \approx \Psi^{(n)} \circ \cdots \circ \Psi^{(2)} \circ \Psi^{(1)} x$$

where $\Psi^{(j)} = \left[{\varphi_{p,q}}^{(j)} \right]$ Subsequent works investigated basis functions other than B-splines in order to describe the univariate functions φ_{p,q_j}. For example, in [7] wavelet basis is considered, while others proposed polynomial bases [5, 23], or RBF basis [33].

In computer vision applications convolutional operations are commonplace thanks to their parameter sharing and translation equivariance properties [4]. The convolutional KAN [6] operation, denoted $*_{KAN}$, can be defined with similar motivations as

$$(I *_{KAN} f)(x, y) = \sum_{h_x} \sum_{h_y} \Psi(h_x, h_y) \left(I(x + h_x, y + h_y) \right).$$

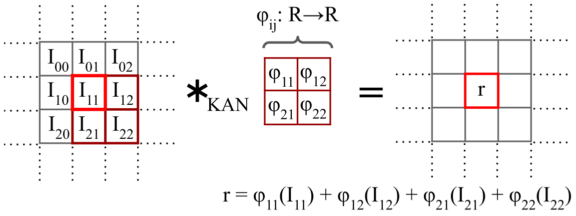

Figure 3. KANs can be implemented as convolutions, allowing the Convolutional KAN architectures.

Shape-fair feature representations in CNNs Human vision exhibits a strong shape bias, developed from an early age [14, 28, 40]. This bias allows humans to recognize objects based primarily on their shape, even with variations in texture or color. In contrast, CNNs have been shown to exhibit a texture bias [14, 19].

Data augmentation strategies may not fully address the underlying issue of the model's preference for texture, and they are not suitable in applications where real data is scarce and synthetic data is not reliable. Despite this, they are among the most popular strategies for mitigating shape bias. Geirhos et al. [14] leveraged stylized images for training, aiming to reduce the influence of texture cues, while Li et al. used style-transferred images during training [32]. Others explored different data augmentation strategies, including human-like augmentations [19] and shape-based augmentations [16, 30].

Beyond data augmentation, other factors have been shown to influence shape bias in CNNs. Müller et al. [38] investigated the relationship between shape bias and foreground-to-background ratio, finding that shape bias varies systematically with it. Interestingly, the development of adversarially robust models has also been linked to increased shape bias, suggesting a potential connection between robustness and shape-based perception [10] where it is also reported, particularly in early layers, adversarially robust networks develop smoother convolution filters.

Recently, [31] proposed an architectural solution inspired by the sparse encoding of the brain. Specifically, they reported that Top-K activations encode shape and structural information, even when such activations are used only during the inference stage.

3. Methodology

Our NMR-KAN (Fig. 4) is a regularization method leveraging neuromanifold properties and innovatively addressing the shape-feature representation problem. It consists of two novel parts: Implicit neuromanifold regularization enhances local and global shape representations by separately controlling the expressivity of the corresponding network layers; the style decorrelation loss encourages the two feature extractors to explore complementary features.

3.1. Implicit neuromanifold regularization (INMR)

INMR is an architectural regularization method that constrains the candidate set of activation functions while separately accounting for local and global shape feature extraction. On one hand, low-degree neuromanifolds naturally favor shape features; on the other hand, they still need sufficient expressive capacity to learn those representations. A direct application of Eq. (2) is insufficient because it assumes a uniform activation degree across all layers. This leads to *i)* abrupt jumps in the neuromanifold degree due to its exponential dependence on polynomial degree per layer, and *ii)* a lack of distinction between local and global shape extraction stages.

To address these limitations, we introduce heterogeneous activation degrees, allowing finer control over expressivity across layers. First, Corollary 1 extends Eq. (2) to networks with layer-wise varying degrees, enabling the construction of neuromanifolds with intermediate expressivity.

Corollary 1. *The degree of the neuromanifold with polynomial degrees* $\mathbf{r} = (r_0, r_1, \ldots, r_{L-1})$ *is*

$$deg(\mathcal{M}_{\mathbf{d},\mathbf{k},s,\mathbf{r}}) = (|\mathbf{k}| - L)! \prod_{0 \le j < L} \frac{r_j^{(L-j-1)(k_j-1)}}{(k_j - 1)!}.$$

Note that although this extension is straightforward, its implications are important. By allowing structured control over expressivity, this formulation enables targeted regulation of shape extraction stages, which was not possible with uniform-degree neuromanifolds.

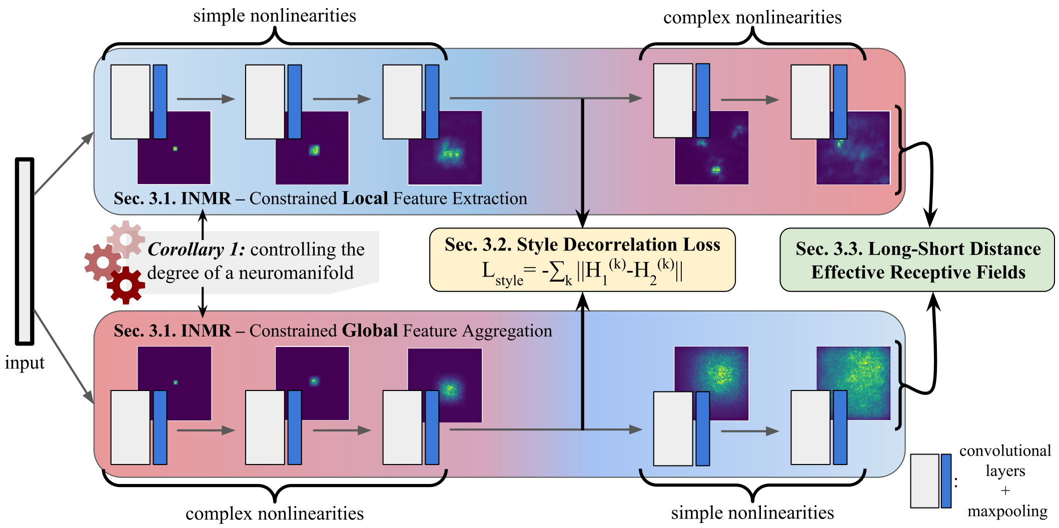
simple nonlinearities
complex nonlinearities
Sec. 3.1. INMR – Constrained Local Feature Extraction
Corollary 1: controlling the degree of a neuromanifold
Sec. 3.2. Style Decorrelation Loss
L_style= -∑_k ||H_1^(k)-H_2^(k)||
Sec. 3.3. Long-Short Distance Effective Receptive Fields
Sec. 3.1. INMR – Constrained Global Feature Aggregation
input
complex nonlinearities
simple nonlinearities
convolutional layers + maxpooling

Figure 4. **Neuromanifold-regularized KANs (NMR-KAN)** enable shape-fair feature representations through two novel components. **(a) Implicit neuromanifold regularization (INMR, Sec. 3.1):** INMR, based on **Corollary 1**, separately handles *local* and *global* shape cue extraction. One feature extractor uses simple activations in earlier layers to promote robust, shape-related features, while the other applies them in later layers to prevent over-reliance on texture-like combinations. **(b) Style decorrelation loss (SDL, Sec. 3.2):** SDL enriches features by disentangling the style representations of the two extractors, enhancing shape-fair representations. The two components lead to **Long-short distance effective receptive fields (Sec. 3.3):** the receptive field of one branch has Gaussian spread, capturing short distance correlations, while the other branch is non-Gaussian, accounting for long distance correlations.

Next, leveraging the hierarchical learning property of neural networks, we associate earlier layers with local shape extraction and later layers with global shape aggregation. Using Corollary 1, we define a structured neuromanifold:

$$r_j = \begin{cases} r_1, & 0 \le j < L_1 \quad \text{(local shape cues)} \\ r_2, & L_1 \le j < L \quad \text{(global shape cues)} \end{cases}$$

where the transition at L_1 reflects the shift from local to global shape cue extraction. Keeping r_i constant within each stage stabilizes the joint optimization. Below, we apply this structure in two dedicated feature extractors.

As a final note, INMR offers a scalable implicit regularization, a key advantage over the analogous implicit methods for traditional CNNs, such as reducing the number of neurons, which are not scalable.

Constrained Local Feature Extraction To prevent texture-driven feature learning in early-stage processing, we impose a lower polynomial degree, $r_1 < r_2$, in view of Corollary (1). This ensures that local shape cues, such as edges and contours, are extracted using smoother, lower-complexity functions, limiting the network's ability to capture high-frequency, texture-like patterns. Without this constraint, the early-stage feature extractor may overfit to fine-grained details, reducing generalization. The increased nonlinearity in later layers compensates for this restriction, allowing a more expressive encoding of global structures.

Constrained Global Feature Aggregation In contrast, the second feature extractor follows the inverse constraint, $r_1 > r_2$, mirroring the first branch but shifting the role of complexity control. Here, the lower-degree nonlinearities in deeper layers prevent the network from overfitting to complex texture compositions when integrating local cues into a global shape representation. The higher expressivity in earlier layers enables richer local feature extraction, capturing necessary variations without introducing texture bias at the global aggregation stage. This dual-branch strategy ensures a structured separation of local and global shape processing, reinforcing shape-fair learning.

Summary of the advantage: NMR introduces a structured regularization that balances Constrained Local Feature Extraction, which prevents texture bias in earlier layers, and Constrained Global Feature Aggregation, which ensures structured shape representations in later layers. Unlike traditional implicit regularization, INMR is scalable and provides targeted control over neuromanifold degrees.

3.2. Style decorrelation loss (SDL) to enrich features

SDL aims to decorrelate the styles of intermediate representations of the two feature extractors. The two feature extractors of INMR can still extract similar features, despite being constrained to a lower degree neuromanifold and aligns the design of the model architecture with that of the problem, This prevents the branches from converging to similar feature extractors, enriching the learned representations and improving shape-fairness.

A style space is defined in [12] in terms of the Grammian of lth layer in the network as

$$G_{ij}^{(l)} = \frac{1}{N_l M_l} \sum_n F_{in}^{(l)} F_{jn}^{(l)} \tag{4}$$

where N_l is the number of channels and M_l is the number of spatial dimensions. Noting that the matrix is symmetric and its diagonal corresponds to self-correlation of channels, we define $H_{ij,n}^{(l)} = G_{ij,n}^{(l)}$ if $i < j$ and 0 otherwise, and optimize based on H matrices.

Then, SDL promotes style dissimilarity between the two branches, as measured by the Frobenius norm:

$$\mathcal{L}_{decorr} = -\sum_l \left(\sum_{i,j} \|H_{ij,1}^{(l)} - H_{ij,2}^{(l)}\|_F \right) \tag{5}$$

where $H_{ij,n}^{(l)}$ denotes the H matrix from nth branch, leading to the total loss

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \mathcal{L}_{decorr}.$$

In practice, since the Grammian can be costly to calculate when N_l is high, SDL only uses the outputs of the layers where the basis changes, that is, $l = L_1$.

Summary of the advantage: SDL aims at feature enrichment by disentangling the style representations of the two feature-extractors.

3.3. Long-short distance effective receptive fields

The combination of INMR and SDL leads to an important property of NMR-KAN: the effective receptive field (ERF) [35] has Gaussian spread in one of the branches, indicating short distance correlations are leveraged; while the other branch has a non-Gaussian distribution, accounting for long distance correlations. This is because the design of the two components of INMR encourages the learning of local and global shape cues, while SDL actualizes the intent of the design by decorrelating the styles of intermediate representations. Fig. 5, following [35], demonstrates our ERF is significantly better for addressing local and global shape cues thanks to the long and short distance ERFs when compared to the ERF of a traditional CNN and a convolutional KAN with a single feature extractor. This *i)* provides insight for the increased shape-fairness of NMR-KAN and *ii)* the complementarity of the learned features.

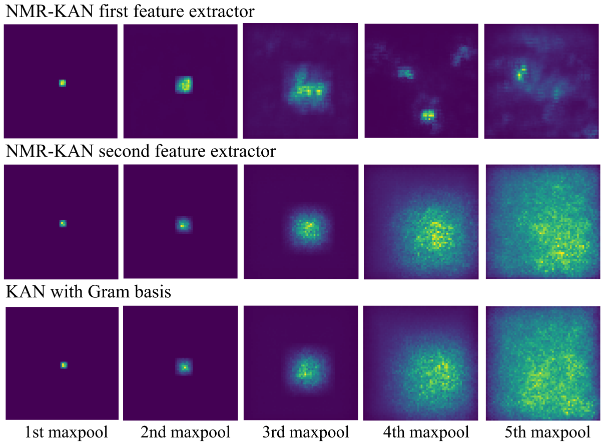

Figure 5. Integrated framework yields non-Gaussian and Gaussian effective receptive fields on the first and second features extractors of the network, respectively. ERFs are acquired using random inputs and at every max pool layer.

4. Experiments and Results

4.1. Experimental Setting

Datasets. Our novel cue conflict dataset, Tiny ImageNet [29], Tiny ImageNet-C [18], and CIFAR-10 are used. Since downsampling the original dataset from [14] loses the texture, the new cue conflict dataset is created using 64 × 64 images from Tiny ImageNet as detailed in the Appendix. The dataset will be made publicly available. Sample images are shown in Fig 6.

Evaluation metric. Shape bias (s) is defined as the ratio of correct shape decisions to the total number of correct guesses. A "correct shape decision" occurs when the model classifies an image into the same category as its source image, while a "correct guess" includes either a correct shape or texture classification.

Models and training details. Vanilla convolutional KANs (CKANs) are implemented using ResNet [17] and VGG-like [42] architectures, following [11], with 3 different bases: B-spline, Gram polynomial, and wavelet. For B-spline, the spline order is 3 and the number of grid points is 5, for Gram polynomial basis the polynomial degree is set to 3, and for wavelet basis the mexican hat wavelets are used. When the number of parameters is too high, bottleneck convolutional layers are utilized. All models are trained on Tiny ImageNet's training split, and reported Top-1 scores are acquired on the validation split. The novel cue conflict dataset is used only for shape bias tests, and not during training of the models. In our implementations of NMR-KAN, the modules up to the third pooling layer are considered as the early stage layers so that the receptive field is large enough to capture shape information in the later stages. For the implementation of [31], the original paper is followed by ap-

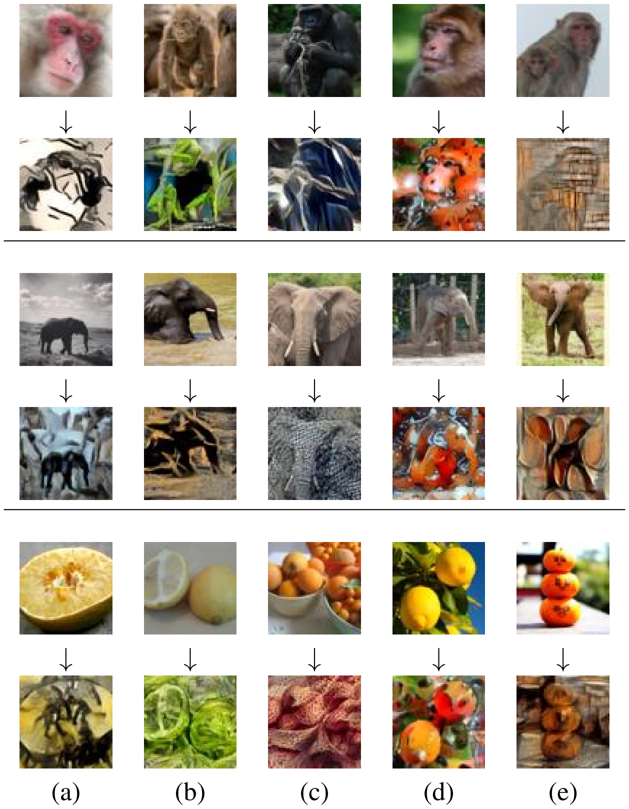

Figure 6. Content shapes in odd-numbered rows followed up by cue conflict images acquired from them. Style images are (a) spider (b) orthopetra (c) garment (d) ladybug (e) barrel/chest

plying the Top-k operation after every max-pooling layer. However, applying a sparsity of 5%, as reported in the original work, led to poorly trained models. Instead, a sparsity of 50% yielded best-shape bias results over a search with a granularity of 5%. All models are trained with cross-entropy loss as $\mathcal{L}_{task}$, (when applicable) loss weight $\alpha = 1$, Adam optimizer, and cosine learning scheduling.

4.2. First demonstration of shape bias in CKANs

For the first time, we show the shape bias in CKANs in Table 1. The evaluations indicate that Gram polynomials achieve the highest Top-1 accuracy on Tiny ImageNet for both ResNet and VGG architectures, with the VGG model demonstrating the highest shape bias. This is in alignment with our mathematical analysis suggesting the potential of KANs for shape fairness in the supplementary material. In contrast, traditional CNNs, such as ResNet, exhibit better shape bias compared to VGG. Given that VGG11 contains 132M parameters, while ResNet18 has only 18M, this suggests that CKANs may leverage neurons more efficiently.

Basis	Architecture	Conv.	Shape bias	Top-1 acc. (val)
UAT	ResNet18	full	31.76	**59.9**
UAT	ResNet18-ft	full	35.79	**65.4**
UAT	VGG11	full	28.02	48.6
UAT	VGG11-ft	full	28.52	61.2
Gram	VGG	BotNeck	**42.53**	57.4
B-spline	VGG	BotNeck	42.11	50.7
Wavelet	VGG	BotNeck	29.86	46.4
Gram	ResNet	BotNeck	25.00	55.8
Gram	ResNet	full	20.80	58.0
B-spline	ResNet	full	26.35	52.0
Wavelet	ResNet	full	18.18	51.8

Table 1. Shape bias and Top-1 validation accuracy for various CKAN implementations (ResNet-18, VGG-11 with Gram, B-spline, and Wavelet bases) and traditional CNNs. “BotNeck” indicates bottleneck convolutional layers, “ft” indicates ImageNet pre-training, included as a reference.

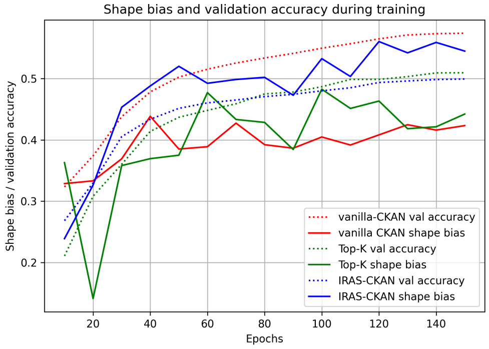

Figure 7. Shape bias and validation accuracy during training for a VGG-like CKAN with a degree 3 Gram polynomial basis (baseline), Top-K activations [31], and the proposed NMR-KAN. NMR-KAN demonstrates consistently higher shape bias throughout training while maintaining reasonable validation accuracy.

Based on both shape bias and Top-1 accuracy scores, VGG-Gram is adopted as the baseline architecture. Table 2 shows that the variations in degree, where all layers of the network share the same basis, do not exhibit a clear correlation with the shape bias. However, both the averaged shape bias and the Top-1 accuracy scores slightly favor the degree 3 results.

degree	2	3	4	5	6	7
s	41.5	**42.5**	41.8	41.1	41.9	42.0

Table 2. Uniformly changing activation degrees does not significantly affect acquiring shape-fair feature representations.

4.3. NMR-KAN enables shape-fairness

Fig. 7 demonstrates that NMR-KAN improves the shape bias to **57.3%**, an improvement of **14.8%** over the baseline, demonstrating the effectiveness of our implicit regularization strategy in promoting shape-fairness. It is seen that failing to improve in the later stages. In contrast to vanilla CKAN which reach its peak shape-fairness during early stages of the training, NMR-KAN reaches its peak near the end of the training, correlating more strongly with validation accuracy.

Interestingly, Top-K operation [31] only slightly improves the shape bias of the baseline. The performance drop observed in the Top-K operation compared with the values reported in the original work could be due to the nature of the method: sparsity operations might require more data to confidently learn representations.

We have also tested DuFeNet [46], which uses additional edge information to enhance shape bias, on our cue conflict dataset which achieved 41.6% top-1 accuracy and 39.7% shape bias.

While shape bias increased, the Top-1 accuracy drops by 7.3% requiring consideration. The "strong negative correlation" between accuracy and shape bias is reported systematically in [13] (Fig.4). This trade-off aligns with previous work [8], where texture-based models achieve surprisingly high ImageNet accuracy. Standard CNN training often inadvertently encourages this bias [14, 15], exploiting unintended features for faster learning. Thus, imposing shape-fairness might require unlearning these shortcuts, potentially reducing accuracy while promoting reliance on more robust shape features.

4.4. NMR-KAN strengthens adversarial defense

Table 3 compares adversarial robustness of NMR-KAN with that of ResNet-18 and Top-K under PGD ($\alpha = 10$, $\epsilon = 8/255$), DeepFool ($\alpha = 20$, $\epsilon = 8/255$) and AutoAttack. All models are finetuned on CIFAR-10 and tested following the implementation in [25]. The results demonstrate that NMR-KAN is **consistently** well-performing under adversarial attacks.

	ResNet-18	Top-K	Ours
PGD*	32.8%	46.4%	**56.2%**
DeepFool†	**83.0%**	69.4%	**79.6%**
AutoAttack	23.8%	**50.5%**	**50.7%**

Table 3. NMR-KAN is consistently well-performing under adversarial attacks. *,†: $\epsilon = \frac{8}{255}, \alpha = 10^{*}/20^{\dagger}$ steps.

4.5. NMR-KAN improves robustness to corruptions

Table 4 presents the relative mean corruption error (mCE) on Tiny ImageNet-C, assessing the robustness of our models under various image corruptions. Each component of NMR-KAN progressively contributes to lowering the relative mCE. The introduction of the hybrid basis reduces the relative error from 105.7 to 100.1, suggesting that constraining activation complexity enhances robustness. Fusing the features from both branches further improves robustness, lowering the mCE to 95.1. The complete NMR-KAN method achieves the lowest relative mCE of 92.5, demonstrating the effectiveness of combining the hybrid basis, feature fusion, and style decorrelation loss in enhancing resilience to image corruptions. The improvement in robustness aligns with the observed increase in shape bias, suggesting a potential correlation between shape-fairness and resistance to corruptions.

4.6. Qualitative results

Figure 8 shows GradCAM visualizations for VGG11 (traditional CNN), VGG-Gram (CKAN), and our NMR-KAN model's two branches. Traditional CNNs prioritize textures, while VGG-Gram, though somewhat sensitive to shape, remains texture-sensitive. In contrast, NMR-KAN consistently emphasizes object contours and edges, indicating a stronger focus on shape.

For example, with the fish images, VGG11 focuses on the textured tail fin, while VGG-Gram expands attention to the body, but remains influenced by texture. NMR-KAN highlights the fish's outline, demonstrating a clear preference for shape. Similarly, for the cat and bird images, NMR-KAN focuses on key shape features (ears, face, whiskers, paws, tail, etc.) while traditional methods are distracted by textures or background elements. These results demonstrate that NMR-KAN promotes shape-fairness in CKANs, focusing on regions richer with shape information.

5. Ablation studies

Table 5 presents the ablation study to dissect the individual contributions of each component by incrementally introducing them to the baseline CKAN model, VGG11-like CKAN architecture with Gram basis of degree 3 (*cf.* Section 4.2). For the models with heterogeneous degrees, earlier layers use degree 2 and later layers use degree 3 polynomials.

Hybrid basis. Introducing the hybrid bases, *i.e.* using different bases in the earlier and later layers of the network, already yields a noticeable improvement in shape bias (+5.1%), albeit with a slight drop in Top-1 accuracy (-3.8%). This suggests that constraining the activation search space, even without fusing features from two different branches or style decorrelation, encourages the network to prioritize more robust shape-related features. The accuracy drop likely reflects the reduced expressivity of the constrained branch, as a result of which, the ability to capture fine-grained details, which can benefit classification performance on the in-distribution data, are limited.

Fused features. Fusing the features from the two branches results in the most significant increase in shape bias (+6.0%) but also the largest decrease in accuracy (-4.4%). The fused features provide a richer representation of shape, however, this might be leading to leveraging textural cues.

Models	rel. mCE	Blur				Noise			Weather				Digital			
		Zoom	Defocus	Motion	Glass	Shot	Impulse	Gauss	Fog	Frost	Snow	Brightness	Pixel	Elastic	JPEG	Contrast
VGG-Gram	105.7	108.9	105.2	108.2	105.9	74.6	77.5	84.3	96.5	70.6	74.2	185.7	85.8	147.0	139.6	122.0
+ hybrid basis	100.1	97.8	95.0	96.8	97.1	70.5	73.2	78.1	92.7	69.7	72.7	185.9	83.1	139.6	135.3	114.7
+ fused features	95.1	96.7	93.0	96.3	**91.9**	64.9	68.7	72.1	83.3	66.4	**64.7**	183.9	81.0	135.3	132.3	102.3
Top-K [31]	96.9	100.4	94.2	96.0	94.5	**63.5**	68.8	**71.6**	88.9	72.6	69.1	201.4	82.4	137.0	128.9	**84.8**
NMR-KAN (ours)	**92.5**	**93.7**	**89.7**	**92.6**	92.3	65.4	**67.8**	72.9	**82.8**	**63.1**	65.4	**168.9**	**77.5**	**129.3**	**123.2**	103.4

Table 4. Relative mean corruption error (mCE) on Tiny ImageNet-C for the baseline VGG-Gram CKAN, incremental components of NMR-KAN, the Top-K method [31], and the full NMR-KAN. Lower rel. mCE indicates better robustness to image corruptions. Best results are in bold.

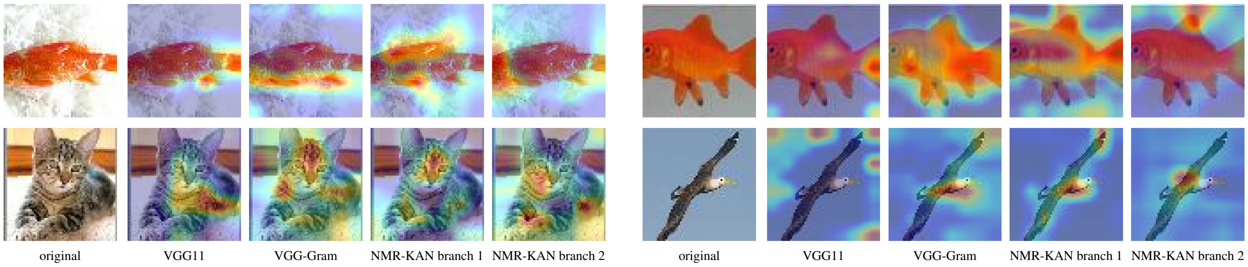

Figure 8. Grad-CAM results comparing VGG11 (traditional), VGG-Gram (degree 3), and the two branches of the proposed NMR-KAN model. For the fish examples, VGG11 primarily focuses on the tail, while VGG-Gram highlights the fin and the overall object, including the textured interior. In contrast, NMR-KAN exhibits attention primarily around the object's edges, with the two branches offering mostly complementary information. In the cat example, VGG11 emphasizes the face, texture around the neck, and rear limbs, while VGG-Gram attends to characteristic shape features, such as the ears and eyes, along with textured regions such as the top of the head. NMR-KAN, on the other hand, focuses more on distinct features, including the ears, face, whiskers, paws, and rear limbs, with the two branches again offering complementary attention. For the bird example, VGG11 attends to the background, whereas the other models focus on similar regions of the bird itself.

	Shape bias	Top-1 acc. (val)
baseline	42.5	57.4
+ hybrid basis	47.6	53.6
+ fused features	53.6	49.2
+ style decorr. (NMR-KAN)	57.3	50.1

Table 5. Contribution of each component of NMR-KAN towards shape bias.

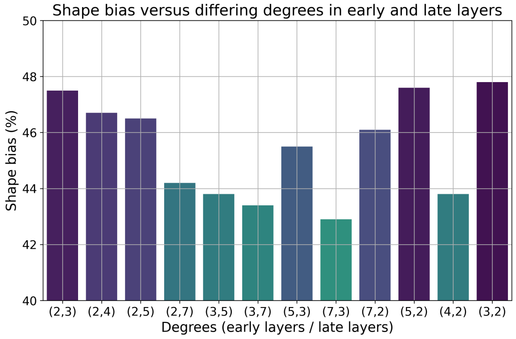

Figure 9. The choice of the heterogeneous degrees of a single KAN feature extractor consistently demonstrates higher shape-fairness, though the lower degree ones yields the better results.

Style decorrelation. SDL further boosts the shape bias (+3.7%) with a small recovery in accuracy (+0.9%), suggesting that encouraging stylistic diversity between the two branches prevents them from collapsing into similar feature extractors. The improvement in accuracy indicates that SDL might help capture complementary information.

5.1. Robustness under hyperparameters

In Fig. 9 we report the shape biases for differing degrees for hybrid bases, each averaged over 3 separately trained models. In the figure, degrees denoted as (n, m) refers to a model with a basis of degree n in the early layers, and degree m in the later layers.

All evaluated models exhibit a higher shape bias than the baseline model, with the lowest being the (7,3)-model at 42.9% while the $(2, 3)$ and $(3, 2)$ models form the best symmetric pair (47.65% shape bias). Notably, the general trend suggests the increase in the total degree leads to a decrease in shape bias. With a higher total degree, the models have more parameters and greater capacity, thus, potentially relying more on texture-related features.

6. Conclusions

Our work and the newly proposed NMR-KAN contributes *i)* the first time investigation of shape-fairness in KANs, *ii)* leveraging the neuromanifold and KAN in a unified framework for the first time, and *iii)* for the first time, explicitly addressing the shape-fairness problem in terms of local and global cue extraction stages. The results demonstrate that NMR-KAN improves shape bias over baseline CKANs by 14.8% while also providing robustness under image corruptions and adversarial attacks.

References

[1] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[2] Basim Azam and Naveed Akhtar. Suitability of kans for computer vision: A preliminary investigation. *arXiv preprint arXiv:2406.09087*, 2024. 2

[3] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, 172:46–61, 2020. 1

[4] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*. MIT press Cambridge, MA, USA, 2017. 3

[5] Subhransu S. Bhattacharjee. Torchkan: Simplified kan model with variations. `https://github.com/1ssb/torchkan/`, 2024. 3

[6] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau. Convolutional kolmogorov-arnold networks. *arXiv preprint arXiv:2406.13155*, 2024. 3

[7] Zavareh Bozorgasl and Hao Chen. Wav-kan: Wavelet kolmogorov-arnold networks. *arXiv preprint arXiv:2405.12832*, 2024. 3

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 7

[9] Ovidiu Calin. *Deep learning architectures*. Springer, 2020. 2

[10] Peijie Chen, Chirag Agarwal, and Anh Nguyen. The shape and simplicity biases of adversarially robust imagenet-trained cnns. *arXiv preprint arXiv:2006.09373*, 2020. 3

[11] Ivan Drokin. Kolmogorov-arnold convolutions: Design principles and empirical studies. *arXiv preprint arXiv:2407.01092*, 2024. 5

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 5

[13] Paul Gavrikov and Janis Keuper. Can biases in imagenet models explain generalization? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22184–22194, 2024. 7

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1, 3, 5, 7

[15] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 7

[16] Shruthi Gowda, Bahram Zonooz, and Elahe Arani. Inbiased: Inductive bias distillation to improve generalization and robustness through shape-awareness. *arXiv preprint arXiv:2206.05846*, 2022. 3

[17] 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. 5

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 5

[19] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020. 1, 3

[20] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing shape bias property of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1923–1931, 2018. 1

[21] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 1

[22] Ali Jamali, Swalpa Kumar Roy, Danfeng Hong, Bing Lu, and Pedram Ghamisi. How to learn more? exploring kolmogorov–arnold networks for hyperspectral image classification. *Remote Sensing*, 16(21):4015, 2024. 2

[23] Unknown Khochawongwat. Gram: Kan meets gram polynomials. `https://github.com/Khochawongwat/GRAMKAN/tree/main`, 2024. Accessed: 2024-08-16. 3

[24] Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. *Advances in neural information processing systems*, 32, 2019. 2

[25] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020. 7

[26] Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013. 1

[27] Kaie Kubjas, Jiayi Li, and Maximilian Wiesmann. Geometry of polynomial neural networks. *Algebraic Statistics*, 15(2): 295–328, 2024. 2

[28] Barbara Landau, Linda B Smith, and Susan Jones. Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, 31(6):807–825, 1992. 3

[29] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[30] Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4323–4331, 2022. 3

[31] Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of shape bias in convolutional neural networks through activation sparsity. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 5, 6, 7, 8

[32] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020. 3

[33] Ziyao Li. Kolmogorov-arnold networks are radial basis function networks. 2024. 2, 3

[34] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024. 2

[35] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 5

[36] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 1

[37] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2

[38] Niklas Mueller, Cees GM Snoek, Iris Isabelle Anna Groen, and H Steven Scholte. Shape-biased learning by thinking inside the box. *bioRxiv*, pages 2024–05, 2024. 3

[39] Xinkuan Qiu, Meina Kan, Yongbin Zhou, Yanchao Bi, and Shiguang Shan. Shape-biased cnns are not always superior in out-of-distribution robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2326–2335, 2024. 1

[40] Larissa K Samuelson and Linda B Smith. They call it like they see it: Spontaneous naming and attention to shape. *Developmental science*, 8(2):182–198, 2005. 3

[41] Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. On the geometry and optimization of polynomial convolutional networks. *arXiv preprint arXiv:2410.00722*, 2024. 2

[42] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[43] Ramanujan Srinath, Alexandriya Emonds, Qingyang Wang, Augusto A Lempel, Erika Dunn-Weiss, Charles E Connor, and Kristina J Nielsen. Early emergence of solid shape coding in natural and deep network vision. *Current Biology*, 31 (1):51–65, 2021. 1

[44] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[45] Mohsen Yavartanoo, Shih-Hsuan Hung, Reyhaneh Neshatavar, Yue Zhang, and Kyoung Mu Lee. Polynet: Polynomial neural network for 3d shape recognition with polyshape representation. In *2021 international conference on 3D vision (3DV)*, pages 1014–1023. IEEE, 2021. 2

[46] Zecong Ye, Zhiqiang Gao, Xiaolong Cui, Yaojie Wang, and Nanliang Shan. Dufenet: Improve the accuracy and increase shape bias of neural network models. *Signal, Image and Video Processing*, 16(5):1153–1160, 2022. 7

[47] Simin Zhan, Jiajun Su, Pudu Liu, Jianqing Zhu, and Huanqiang Zeng. Kaan: Kolmogorov-arnold attention networks for object re-identification. In *Chinese Conference on Biometric Recognition*, pages 13–24. Springer, 2024. 2

[48] Boheng Zhang, Haorui Huang, Yi Shen, and Mingjian Sun. Mm-ukan++: A novel kolmogorov-arnold network based u-shaped network for ultrasound image segmentation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2025. 2

[49] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. 1