Improving Shape Bias in Learnable Geometric Moment Representations

Anonymous WACV Algorithms Track submission

Paper ID *****

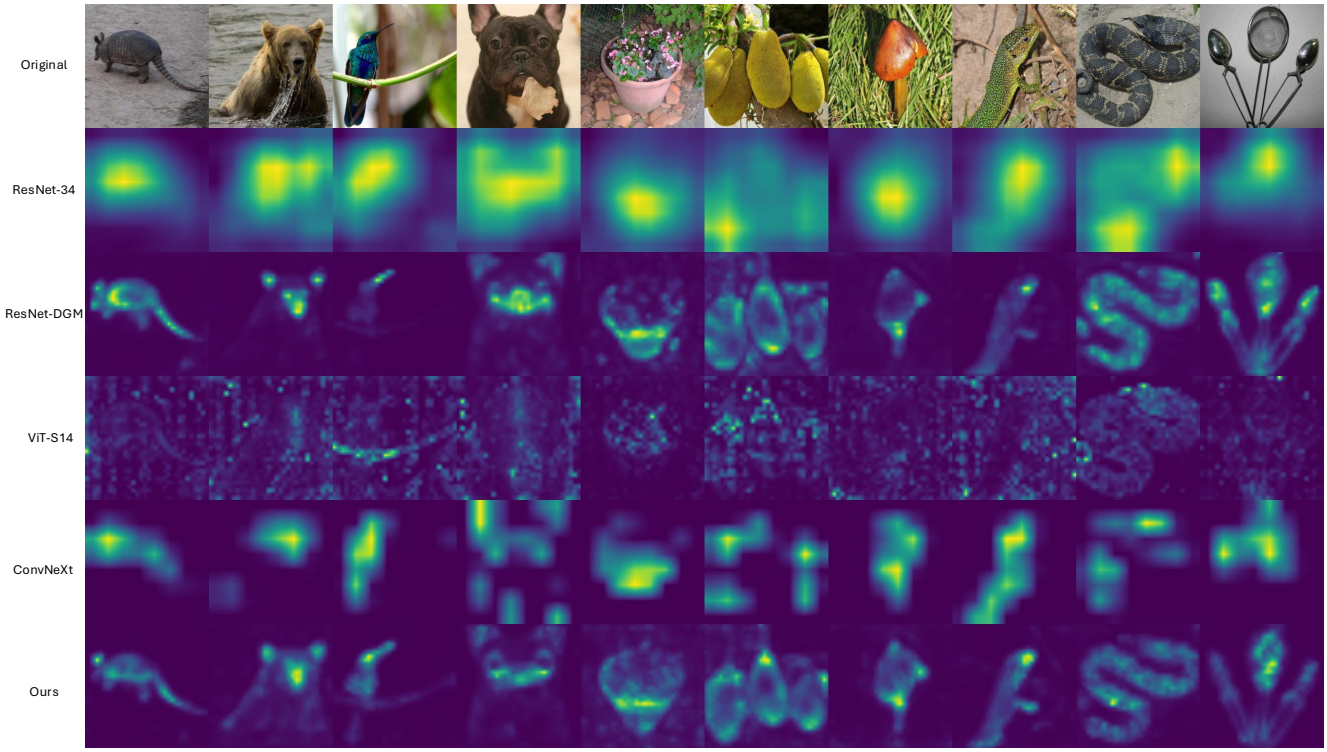

Figure 1. Feature visualization comparison across ResNet-34, ResNet-DGM, ViT-s14 (DINOv2), ConvNeXt-Base, and our method (ConvNeXt-DGM). Standard baselines (ResNet, ViT, ConvNeXt) tend to produce diffuse activations that fail to capture accurate object boundaries. In contrast, the DGM variations demonstrate superior shape alignment and localization.

Abstract

Deep Geometric Moments (DGMs) have been shown to encode shape-aware representations in image features. In this work, we revisit the DGM framework through the lens of ConvNeXt, a modern convolutional network (ConvNet) architecture. By leveraging features extracted from ConvNeXt, we improve classification accuracy while further strengthening the geometric shape-awareness of DGMs. Our results demonstrate that features from modern ConvNet backbones serve as compatible stems for training Deep Geometric Moments, and that the learned representations remain tightly aligned with object geometry while exhibiting robustness to visual perturbations. We quantitatively characterize this shape awareness using geometric metrics such as the Hausdorff distance and the Average Symmetric Surface Distance (ASSD) complemented by Intersection over Union (IoU) to assess regional overlap. Furthermore, we conduct an extensive analysis of the invariance of the learned feature representations under diverse image perturbations, including changes in rotation, brightness, color, and scale. We posit that these shape-aligned features offer significant value not only for traditional computer vision tasks, such as object detection and image segmentation, but also for modern efficient training-free image and video edit-

ing methods.

1. Introduction

Learning visual representations that are sensitive to object shape while remaining robust to appearance variations is a long-standing goal in computer vision. Shape-aware features are particularly valuable for tasks that require precise geometric reasoning, such as image segmentation, depth estimation, and 3D understanding, where semantic consistency must be maintained despite changes in color, illumination, viewpoint, or scale. [10, 18]. Beyond traditional recognition pipelines, recent works have further demonstrated that intermediate feature representations such as attention maps or activation responses can be directly manipulated for image editing, video guidance, and controllable generation, placing additional importance on the interpretability and geometric alignment of learned features.

Deep Geometric Moments (DGM) were introduced as a mechanism for encouraging shape-awareness in deep neural networks by explicitly modeling geometric properties of feature responses [17]. The original study demonstrated that DGM-based representations tend to align more closely with object structure, enabling improved spatial correspondence and robustness to certain transformations. However, this foundational work relied exclusively on ResNet [8, 13], a conventional backbone that predates more modern architectures. Consequently, the interaction between these geometric properties and advanced network designs remains underexplored, and it is unclear whether the observed benefits persist when utilizing stronger feature extractors.

In parallel, the landscape of visual representation learning has evolved significantly with the introduction of Vision Transformers (ViTs) [2, 4, 7, 11, 14, 19], which have demonstrated strong scalability and performance across a wide range of vision tasks. While ViTs excel at modeling long-range dependencies, ConvNets [6] retain important advantages due to strong inductive biases such as locality and translation equivariance. [6] These properties often lead to better generalization in limited-data regimes and reduced computational overhead. ConvNeXt revisits these design principles through Transformer-inspired training and architectural choices, demonstrating that ConvNets can achieve performance competitive with ViTs while maintaining their inherent inductive strengths.

Motivated by these developments, we revisit Deep Geometric Moments using ConvNeXt [12, 20] as the feature backbone. Our objective is twofold: first, to assess whether DGM remains effective when paired with a high-capacity ConvNet architecture; and second, to conduct a systematic analysis of the shape-aware features learned in this setting. By integrating DGM with ConvNeXt, we aim to investigate the specific contribution of the geometric moment formulation relative to the underlying feature extractor and to gain further insight into the properties of the representations DGM induces.

To this end, we conduct a comprehensive empirical study of ConvNeXt-based DGM models. We quantitatively evaluate shape awareness by measuring the alignment between learned feature responses and object masks using metrics such as mean Intersection-over-Union (mIoU), Hausdorff distance, and Average Symmetric Surface Distance (ASSD). To further probe the robustness and invariance of the representations, we perform controlled perturbation experiments spanning variations in color, brightness, scale, and orientation. These analyses provide insight into the stability of DGM features against common appearance changes and their ability to preserve geometric consistency under transformations.

Our results indicate that ConvNeXt serves as a compatible feature backbone for Deep Geometric Moments, yielding representations that align more closely with object geometry than the baseline model. Furthermore, the observed robustness to visual perturbations demonstrates that shape-aware mechanisms complement advanced ConvNet architectures by effectively enhancing geometric consistency.

In summary, this work provides a re-examination of Deep Geometric Moments by integrating them with advanced convolutional backbones and conducting a detailed representational analysis. Our findings confirm that DGM remains a powerful tool for inducing shape bias, bridging the gap between strong feature extractors and geometric interpretability.

2. Preliminary

2.1. Deep Geometric Moments

Geometric moments have long been used in computer vision as a compact and informative representation of shape, capturing spatial distributions of mass while facilitating invariance to geometric transformations. Classical moment formulations encode object geometry through weighted integrals of pixel coordinates, enabling descriptors that are sensitive to shape while being robust to appearance variations such as color or texture. Formally, consider a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ extracted from a backbone network. In a standard formulation, the general geometric moment of order (p, q) for a given channel is defined as:

$$M_{pq} = \sum_{x=1}^{W}\sum_{y=1}^{H} x^p y^q \, \phi(F_{x,y}) \tag{1}$$

where (x, y) represents spatial coordinates and $\phi(\cdot)$ denotes a mapping function applied to the feature activations. By selecting specific moment orders and normalization schemes, these moments can explicitly encode geometric properties

such as orientation, scale, and higher-order shape characteristics. However, determining the optimal set of orders (p, q) prior is non-trivial. Deep Geometric Moments (DGM) address this by embedding geometric structure directly into the representation learning process. Rather than manually prescribing fixed integer orders, our method fixes the dimensionality of the representation but learns the underlying basis functions end-to-end. Consequently, we simplify the notation by dropping the order subscripts p, q and indexing moments solely by the feature channel c. In this formulation, the DGM module projects extracted image features onto a learnable coordinate basis. The moment m^c for the c-th channel is defined as:

$$m^c = \frac{1}{N \times N} \sum_{x=1}^{N} \sum_{y=1}^{N} g^c(x, y) f^c(x, y) \tag{2}$$

where $N \times N$ denotes the spatial resolution of the feature map, $f^c(x, y)$ represents the feature activation at location (x, y) extracted by the backbone CNN, and $g^c(x, y)$ is a learnable 2D polynomial function serving as the coordinate basis. To further account for variations in object location, size, and pose, we explicitly learn affine parameters that dynamically deform the coordinate grid $g^c(x, y)$ during the moment computation, thereby ensuring geometric invariance. For our specific implementation, we set the spatial dimension $N = 32$ (yielding a 32×32 grid) and maintain a fixed channel dimensionality of $C = 256$.

2.2. ConvNeXt

ConvNeXt revisits the standard convolutional design by systematically integrating architectural patterns that contribute to the success of Vision Transformers [15]. As illustrated in Figure 2, the core building block shifts from the traditional ResNet bottleneck to an inverted bottleneck structure. This design decouples spatial mixing from channel mixing, employing large-kernel depthwise convolutions (increasing from 3×3 to 7×7) to significantly expand the effective receptive field. Additionally, the network adopts Layer Normalization [1] and GELU activations [9], replacing the conventional Batch Normalization and ReLU activation layers found in ResNet. These adjustments align the network's processing dynamics with Transformer-style blocks while retaining the computational efficiency of sliding-window convolutions. In the context of this study, ConvNeXt offers a distinct benefit by yielding discriminative feature representations that preserve the inherent spatial structure of standard convolutional networks. Unlike ViTs, which process token sequences that may lose local adjacency information, ConvNeXt outputs dense, grid-aligned activation maps. This preservation of spatial layout is crucial, as DGM computes geometric moments directly from the feature map coordinates to enforce shape constraints.

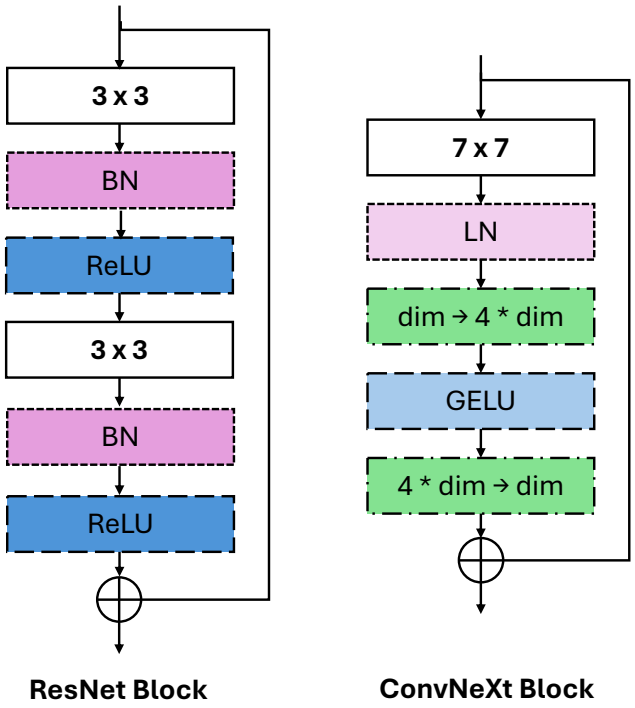

Figure 2. Composition of ResNet and ConvNeXt blocks. Red depicts normalization layers, Green is linear layer with input and output dimensions, Blue is activation functions.

2.3. Combining DGM and ConvNeXt

We combine these frameworks to evaluate the behavior of Deep Geometric Moments within a more powerful architectural setting. ConvNeXt provides a strong, spatially grounded feature substrate that is technically compatible with the moment integration process of DGM. By employing this advanced backbone, we aim to investigate the interaction between explicit geometric constraints and stronger feature representations. This approach allows us to verify how DGM adapts to improved feature extractors and whether it continues to enhance representation alignment beyond what is achieved by the backbone alone.

3. Experimental Setup

In this section, we discuss how our experiments are designed. We explain the evaluation metrics used for our analysis, following the model architecture and its variations.

3.1. Metrics

Our primary objective is to evaluate the capacity of learned features to preserve object geometry and shape structure. While the original DGM study relied primarily on qualitative visualizations to demonstrate shape awareness, we extend this analysis by establishing a rigorous quantitative

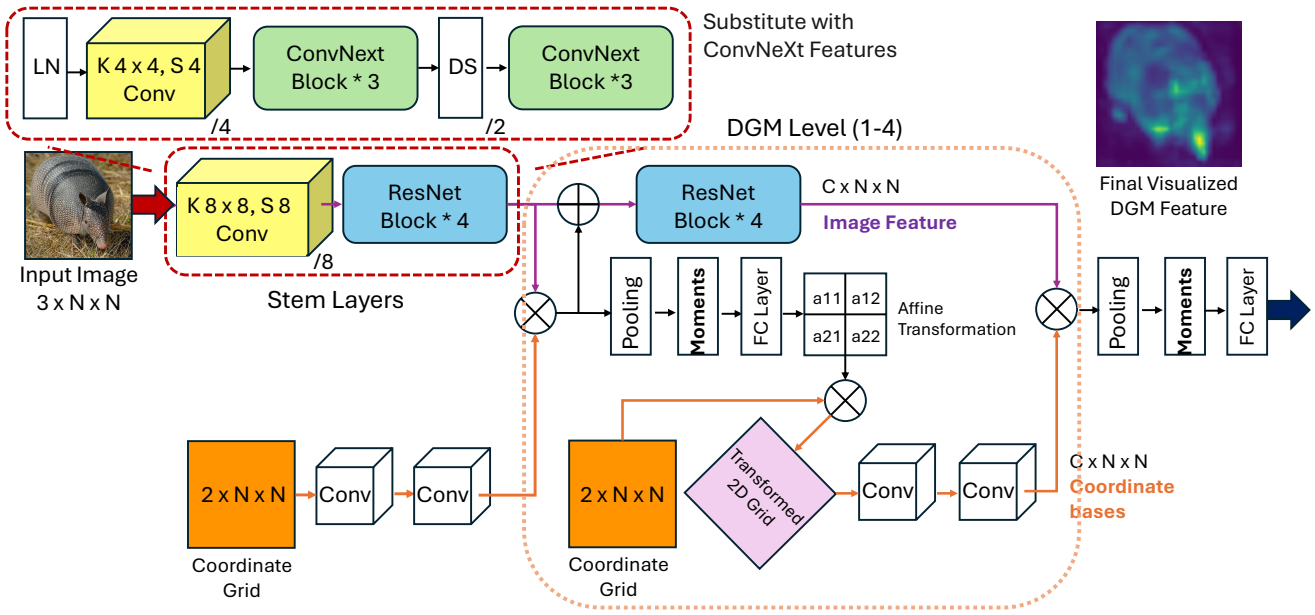

Figure 3. DGM model architecture. DGM is learned by dual path processing; upper image features, lower grid bases. We substitute the stem layers for the image features with ConvNeXt.

framework. We adopt the following three metrics to systematically measure the geometric alignment of each model's image features.

Intersection over Union (IoU)

The Intersection over Union (IoU), also known as the Jaccard Index, is a standard metric for quantifying the spatial overlap between two distinct regions. It is calculated by dividing the area of the intersection by the area of the union of the two sets. Formally, given a predicted binary mask A (derived from the feature response) and the ground truth object mask B, the IoU is defined as:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

where $|\cdot|$ denotes the cardinality (area) of the set. The resulting score ranges from 0 to 1, where 1 indicates a perfect spatial match. Unlike detection tasks that often apply IoU to bounding boxes, we compute the mean IoU (mIoU) directly on pixel-wise segmentation masks to accurately capture the alignment of the feature activation with the interior shape of the object.

Hausdorff Distance

The Hausdorff Distance (HD) measures the maximum distance from a point in one set to the nearest point in another set, effectively quantifying the worst-case mismatch between two boundaries. Formally, given two sets of boundary points A and B (representing the predicted and ground-truth contours), the Hausdorff distance is defined as:

$$H(A, B) = \max\left\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\right\} \quad (4)$$

where $d(a, b)$ denotes the Euclidean distance between points a and b. While IoU focuses on the volumetric overlap of the object's interior, the Hausdorff distance supplements this by explicitly evaluating the geometric precision of the object boundaries.

However, because the standard HD is determined by the single largest deviation, it is highly sensitive to outliers. To mitigate this, we adopt the 95th percentile Hausdorff Distance (HD95), which excludes the most extreme anomalies to provide a more robust metric of shape consistency.

Average Symmetric Surface Distance (ASSD)

The Average Symmetric Surface Distance (ASSD) is a boundary-based metric that evaluates the global alignment between two surfaces. ASSD computes the average distance between all points on the predicted and ground-truth boundaries. It is defined as:

$$\text{ASSD}(A, B) = \frac{\sum_{a \in S(A)} d(a, S(B)) + \sum_{b \in S(B)} d(b, S(A))}{|S(A)| + |S(B)|} \quad (5)$$

where $S(A)$ and $S(B)$ represent the set of surface points (contour pixels) for the ground truth A and prediction B, respectively. The term $d(x, S(Y)) = \min_{y \in S(Y)} \|x - y\|$ denotes the shortest Euclidean distance from a point x to the surface set $S(Y)$. Crucially, this metric incorporates

two directional components to ensure symmetry. By averaging these bidirectional distances, ASSD penalizes both error types equally, providing a holistic measure of boundary adherence that complements the worst-case sensitivity of the Hausdorff distance.

3.2. Model

We detail our model architecture in Figure 3. The original DGM implementation utilized a convolutional stem to downsample the input by a factor of 8, followed by a sequence of four ResNet blocks. In this work, we replace this backbone with a pretrained ConvNeXt encoder. Specifically, the image is processed through two stages, each comprising three ConvNeXt blocks, configured to match the spatial downsampling ratios and output dimensions of the original architecture. These extracted features subsequently serve as the input for the DGM module, where the geometric learning formulation remains identical to the original method. All models are trained for 100 epochs using a cosine decay scheduler with an initial learning rate of 0.1.

4. Experimental Results

In this section, we outline our experimental analysis and results, with a primary focus on the shape alignment capabilities of DGM features and their robustness to perturbations.

Table 1. ImageNet-1K classification accuracy using different ConvNeXt models as stem feature extractors. We newly train ResNet34-DGM following [17]. The first two columns denote the number of channels after each stage.

Model	Stage 1	Stage 2	Accuracy (%)	Params (M)
Tiny	96	192	76.26 (+1.34)	20.86
Base	128	256	76.76 (+1.84)	21.82
Large	192	384	77.25 (+2.33)	24.51
X-Large	256	512	77.36 (+2.44)	28.25
ResNet34-DGM	-	-	74.93	24.35

4.1. Model Variations

We experiment with ConvNeXt model variants ranging from Tiny to X-Large. The architecture is structured in four stages, beginning with a stem that reduces the input resolution by a factor of four. Between subsequent stages, downsampling layers further reduce the spatial dimension by half. To accommodate our DGM training pipeline, which operates on a fixed 32×32 grid, we fix the input resolution to 256×256 and extract features from the output of Stage 2. With this setup, the cumulative downsampling transforms the 256×256 input precisely into the required 32×32 feature map. The model variants differ only in channel width at this extraction point; consequently, we omit ConvNeXt-Small, as it is structurally identical to the Tiny variation up to Stage 2. As summarized in Table 1, we trained four DGM variants utilizing different ConvNeXt backbones. For comparison, we also retrained the ResNet34-DGM baseline [17] under the same settings. All models were trained on the ImageNet-1K classification task [5]. We observed that every ConvNeXt-based variant outperformed the ResNet-DGM baseline, with performance consistently improving as the model size increased. However, since our primary focus is on geometric analysis rather than maximizing downstream classification accuracy, we utilize the ConvNeXt-Base variant for the detailed experiments presented in the remainder of this paper.

Table 2. Evaluation of geometry-shape alignment metrics on various feature extractors. GD denotes GradCAM. (*) Denotes an outlier: visual inspection confirms that ResNet-GD produces over-expanded masks, inflating mIoU despite poor shape fidelity. Consequently, we adopt **ConvNext-Base-DGM** as our experimented model for the remainder of this study.

Model	mIoU (↑)	mHD95 (↓)	mASSD (↓)
ResNet-GD	**0.4635***	128.73	58.97
ResNet-DGM	0.3282	133.91	42.41
ConvNext-GD	0.3158	131.02	46.38
ConvNext-Tiny-DGM	**0.4118**	128.61	41.66
ConvNext-Base-DGM	0.4003	**126.70**	**40.09**
ConvNext-Large-DGM	0.3972	127.33	40.57
DINOv2 (ViT-S/14)	0.1975	146.02	48.64
DINOv2 (ViT-L/14)	0.2223	141.02	46.74

4.2. Shape Alignment Evaluation

We evaluate the geometric shape alignment of the learned feature representations. As detailed in Section 3.1, we quantify this alignment by comparing generated feature masks against ground-truth object masks using three metrics: Intersection over Union (IoU), 95th percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). Feature masks are min-max normalized and binarized using a fixed empirical threshold of 0.15. For baseline comparisons, we derive masks for the standard ConvNeXt-Base using GradCAM [16], and for DINO models [3] by averaging attention maps across all heads. Through extensive analysis, we observe that region-based metrics like IoU can be misleading when applied to coarse feature extractors (e.g., standard ResNet), which often produce diffuse, over-expanded masks. To address this limitation, we incorporate boundary-aware metrics to ensure contour precision. Among these, we find ASSD to be particularly robust. As shown in Table 2, our ConvNeXt-Base DGM significantly outperforms the baselines across all three metrics.

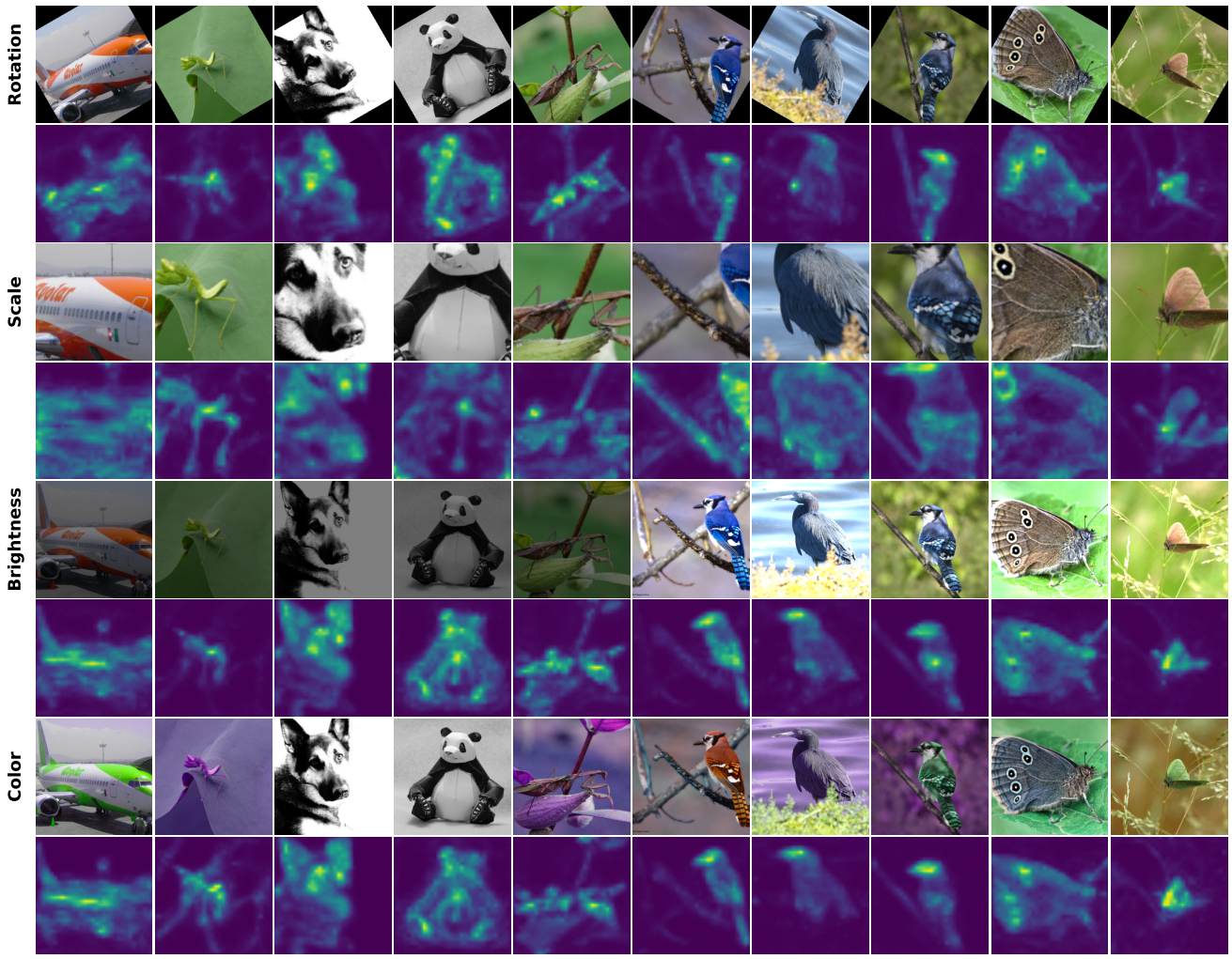

Figure 4. Qualitative comparison of model robustness against geometric (Angle, Scale) and photometric (Brightness, Color) perturbations. Our model maintains consistent shape alignment and structural integrity.

4.3. Invariance of Shape Alignment

To highlight the invariance of the learned geometric moments, we conducted experiments using diverse perturbations, including rotation, photometric shifts, and scale variations. For rotation, we evaluated six distinct angles ranging from 45° to 240°. For color, we perturbed specific RGB channel intensities (increase/decrease) and performed channel swapping. For scale, we tested performance on enlarged inputs. Table 3 demonstrates that while standard ConvNet features (ConvNeXt baseline) become inconsistent under these transformations, our model maintains high structural consistency. Figure 4 provides qualitative examples of this stability. Specifically, we observe that the baseline often suffers from catastrophic degradation when texture statistics shift, suggesting an over-reliance on local appearance. In contrast, our method's performance remains stable, confirming that the learned geometric moments effectively capture intrinsic shape properties that are independent of photometric or viewpoint variations.

4.4. Performance of Different Shape Categories

We conduct an ablation study to analyze how geometric alignment varies across different semantic categories. We selected 16 ImageNet classes that exhibit a wide range of performance gaps between the baseline and our method. As shown in Figure 5, a compelling trend emerges: our model achieves significantly higher gains on animate objects (e.g., animals) compared to inanimate objects, even for species with complex, diverse morphologies such as butterflies and crocodiles. Taken together with Figure 6, these results establish that our model offers superior shape preservation and alignment generality.

Table 3. Robustness evaluation under geometric (Angle, Scale) and photometric (Color Perturbation, Color Swap) transformations. We compare the stability of the default ConvNext-Base model versus ConvNext-DGM across mIoU, mHD95, and mASSD metrics.

Angle	mIoU (↑)		mHD95 (↓)		mASSD (↓)	
(°)	Base	Ours	Base	Ours	Base	Ours
45	0.3558	0.4364	128.32	123.08	44.72	37.68
60	0.3478	0.4468	131.60	125.65	46.25	38.70
90	0.3434	0.4751	137.32	133.53	46.58	40.05
120	0.3448	0.4413	132.31	127.98	46.35	39.59
180	0.3465	0.4734	134.27	131.31	45.95	39.77
240	0.3633	0.4419	125.31	127.66	43.94	39.57

Color Perturb	mIoU (↑)		mHD (↓)		mASSD (↓)	
	Base	Ours	Base	Ours	Base	Ours
R boost	0.3502	0.4730	133.46	127.10	45.77	38.09
R reduce	0.3532	0.4809	132.56	123.28	45.48	36.82
G boost	0.3506	0.4666	133.33	128.12	45.92	38.80
G reduce	0.3471	0.4667	134.17	127.22	46.03	38.19
B boost	0.3530	0.4811	132.49	124.82	45.27	37.31
B reduce	0.3532	0.4727	131.79	123.96	45.15	37.03

Color Swap	mIoU (↑)		mHD (↓)		mASSD (↓)	
	Base	Ours	Base	Ours	Base	Ours
R ↔ G	0.4669	0.3428	129.04	137.01	38.94	47.22
G ↔ B	0.4671	0.3455	129.11	135.47	38.82	46.69
B ↔ R	0.4730	0.3441	126.38	135.83	38.05	46.71

Scale	mIoU (↑)		mHD (↓)		mASSD (↓)	
	Base	Ours	Base	Ours	Base	Ours
1.25	0.3351	0.4801	147.34	136.62	48.09	40.39
1.5	0.3252	0.4753	156.57	144.34	50.48	42.48
2.0	0.3066	0.4614	171.47	155.91	54.59	45.62

5. Conclusion

By revisiting Deep Geometric Moments through the lens of modern convnet architectures like ConvNeXt, we have synthesized the inductive bias of geometric theory with representational power of deep learning. Our work quantitatively validates that DGMs learn representations that are both tightly aligned with object geometry through extensive evaluation using geometric metrics such as Hausdorff distance and ASSD and shows robustness to diverse visual perturbations including rotation, scaling, and photometric shifts. This establishes a distinct "shape bias" that contrasts sharply with the texture-reliance often observed in standard convnets, suggesting that DGMs can offer a robust foundation for the next generation of geometry-sensitive downstream applications, providing critical capabilities for tasks ranging from controllable image manipulation and video editing workflows.

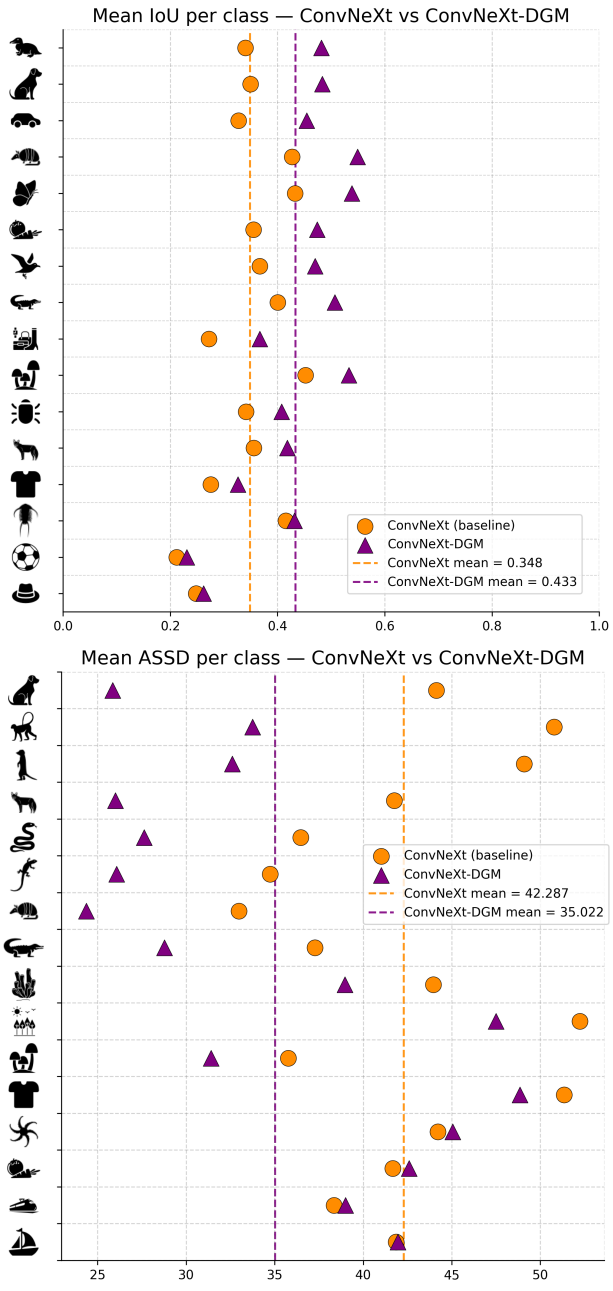

Figure 5. Per-class performance comparison on 16 curated categories from the ImageNet dataset. The chart illustrates the IoU and ASSD scores for the Base model versus DGM (Ours). While improvement margins vary depending on class complexity, DGM demonstrates consistent gains across diverse semantic categories.

6. Limitation and Future Works

In this study, we constrain our models to approximately 25M parameters to maintain computational efficiency and prioritize algorithmic interpretability. We hypothesize that scaling the input grid size offers significant potential, as

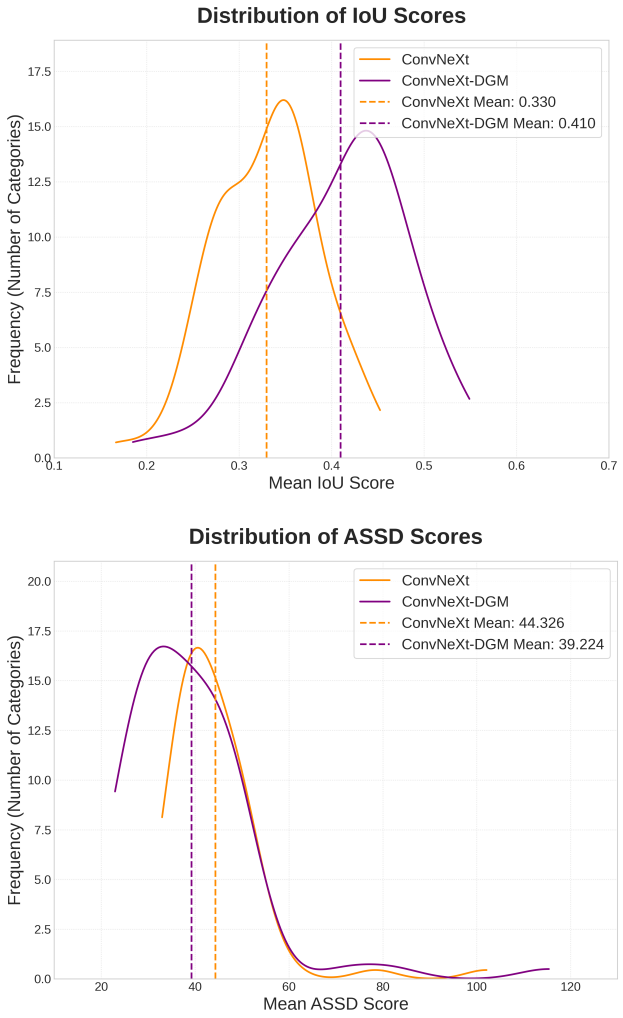

Figure 6. Distribution of IoU and ASSD scores on the ImageNet dataset. Compared to the baseline, DGM (Ours) shows a distinct distributional shift towards higher IoU and lower ASSD values, indicating superior generalization and more consistent geometric alignment across the dataset.

finer resolutions theoretically allow for the calculation of more precise geometric moments—a direction we reserve for future exploration. Additionally, we do not conduct a direct comparison with massive self-supervised ViTs (e.g., DINOv3). While we acknowledge the exceptional representational power of such foundation models, our results demonstrate that robust shape awareness is attainable in significantly more compact architectures without relying on large-scale pre-training.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[4] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023. 2

[5] 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. Ieee, 2009. 5

[6] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16458–16468, 2021. 2

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] 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

[9] D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[10] Sangmin Jung, Utkarsh Nath, Yezhou Yang, Giulia Pedrielli, Joydeep Biswas, Amy Zhang, Hassan Ghasemzadeh, and Pavan Turaga. Guiding diffusion with deep geometric moments: Balancing fidelity and variation. *arXiv preprint arXiv:2505.12486*, 2025. 2

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[13] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020. 2

[14] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 2

[15] Nataniel Ruiz, Sarah Bargal, Cihang Xie, Kate Saenko, and Stan Sclaroff. Finding differences between transformers and convnets using counterfactual simulation testing. *Advances in Neural Information Processing Systems*, 35:14403–14418, 2022. 3

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[17] Rajhans Singh, Ankita Shukla, and Pavan Turaga. Improving shape awareness and interpretability in deep networks using geometric moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4159–4168, 2023. 2, 5

[18] Rajhans Singh, Ankita Shukla, and Pavan Turaga. Polynomial implicit neural representations for large diverse datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2051, 2023. 2

[19] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 2

[20] Kirill Vishniakov, Zhiqiang Shen, and Zhuang Liu. Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy. *arXiv preprint arXiv:2311.09215*, 2023. 2