Supplementary Material For Improving Shape Awareness and Interpretability in Deep Networks Using Geometric Moments

A. Overview:

This section provides the training details used in the experiments for different datasets and presents additional quantitative and qualitative results for our proposed Deep Geometric Moment (DGM) model.

A.1. Semantic image segmentation

Training details: For PASCAL VOC dataset, we train our model with input size of 512×512 and batch size of 48, and for Cityscapes dataset we train with 768×768 images and batch size of 32. We train models on both datasets with initial learning rate of 0.01 and 'poly' learning rate policy (the learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{power}$, where $power = 0.9$). All models are trained with SGD optimizer ($weight\_decay = 1e^{-4}$ and $momentum = 0.9$), cross entropy loss and up to $30K$ iterations.

Qualitative results on Cityscapes: Figure 1 shows qualitative results on the Cityscapes dataset. We observe that the segmentation map from our DGM model is better than the standard ResNet model.

A.2. Training details for image classification

CIFAR: All the models on these datasets are trained up to 150 epochs with a batch size of 128, and SGD optimizer with $momentum = 0.9$ and $weight\_decay = 5e^{-4}$. We use cosine learning rate decay with an initial learning rate of 0.1. During training, we augment the dataset with color and affine transformations.

ImageNet: We train all our models on this dataset with a batch size of 256 and up to 100 epoch, and SGD optimizer with $momentum = 0.9$ and $weight\_decay = 1e^{-4}$. We use cosine learning rate decay with an initial learning rate of 0.1.

A.3. Performance under affine distortions

Table 1 compares the classification performance of different models under various affine distortions on CIFAR-100 dataset. In this experiment, the images are altered with rotation (R), uniformly selected between $\pm 90°$ and rotation scale and translation (RST), uniformly chosen from ($\pm 90°$), scale between $[0.7$ and $1.2]$ and translation between ($\pm 20\%$, in both x and y directions).

Table 1. Performance comparison on distorted CIFAR-100 dataset. R stands for rotation and RST stands for rotate, scale and translate

Model	Params(M)	R (%)	RST (%)
Baseline ResNet-18	9.62	72.46	69.62
ResNet-18	11.17	72.65	69.79
DGM ResNet-18	11.62	**73.45**	**71.81**
ResNet-34	21.33	73.37	70.43
DGM ResNet-34	21.06	**74.91**	**73.20**

Table 1 shows that our DGM model outperforms the baseline as well as the standard ResNet model with a similar number of parameters on both (R) and (RST) distortions. While (RST) distortion is significantly more complex than (R), the performance gain of DGM over the standard ResNet model for (RST) is higher than the (R) distortion. Figure 2 shows that our model has invariance to affine transformations and captures the object shape perfectly well while also outperforming existing classification models under such drastic distortions.

A.4. Computation cost

Table 2 compares the computation cost of the proposed DGM model with standard and baseline ResNet models. The computation cost of the DGM model is higher mainly due to the higher spatial resolution of features against the standard ResNet model, which uses pooling layers to reduce the spatial resolution. The computation cost of the baseline ResNet model (without pooling layers) is comparable to our DGM model. The computation cost can be significantly reduced using the MobileNet architecture for the image feature pipeline.

A.5. Feature visualization across different DGM models

Figure 3 shows *level-4* feature visualization of different DGM models on the ImageNet dataset. All our DGM model produces very clear object shape visualization. We also observe that the objects' shapes for DGM ResNet-34 and DGM ResNet-66* (DGM MobileNet) are slightly sharper in some cases compared to DGM-ResNet-18, which is also

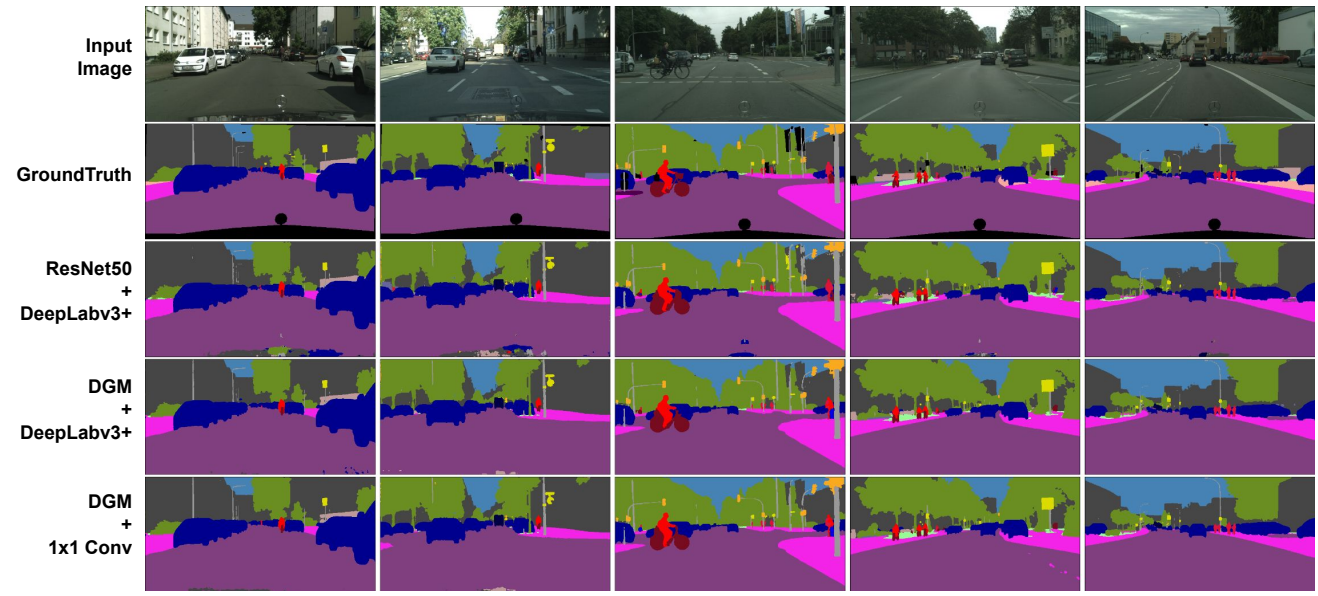

Figure 1. Cityscape segmentation results. The segmentation results from our DGM ResNet-50 model is qualitatively better than the standard ResNet-50 model.

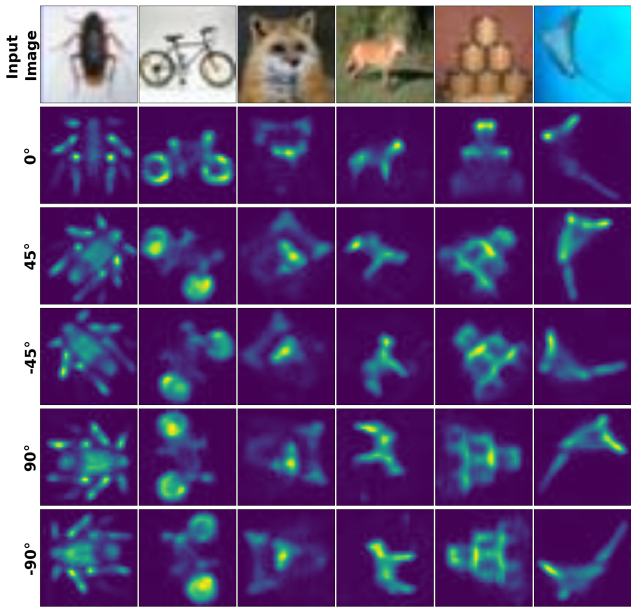

Figure 2. *Level-4* feature visualization from DGM model under different image rotations on CIFAR-100 dataset. The object shape is very well captured across different rotations of the image

reflected in the classification accuracy performance.

A.6. Bases visualization

We also observe that bases from *Level-4* of our DGM model, as shown in Figure 4 for two input images, are also indicative of the object shape. The figure compares the same set of bases sampled from 256 bases for the two images. We observe that the final bases are generated based on the input images, so each image gets a unique set of bases.

Table 2. Computation cost comparison of DGM model with standard and baseline ResNet models in terms of number of floating point operation (FLOPS) in Giga (G)

Model	Params (M)	Flops (G)
Baseline ResNet-18	9.89	9.86
ResNet-18	11.69	1.82
DGM ResNet-18	11.88	10.27
ResNet-34	21.80	3.68
DGM ResNet-34	21.32	19.94
ResNet-50	25.56	4.12
DGM ResNet-50	23.51	17.59
DGM MobileNet	4.76	2.99

A.7. Additional visualizations

In this section we provides additional visualizations. Figure 5 shows *Level-4* visualization under different color distortions on ImageNet-C dataset. Figure 6 shows *Level-4* visualization under different rotations on ImageNet dataset. Figure 7 provides additional *Level-4* visualization on ImageNet dataset.

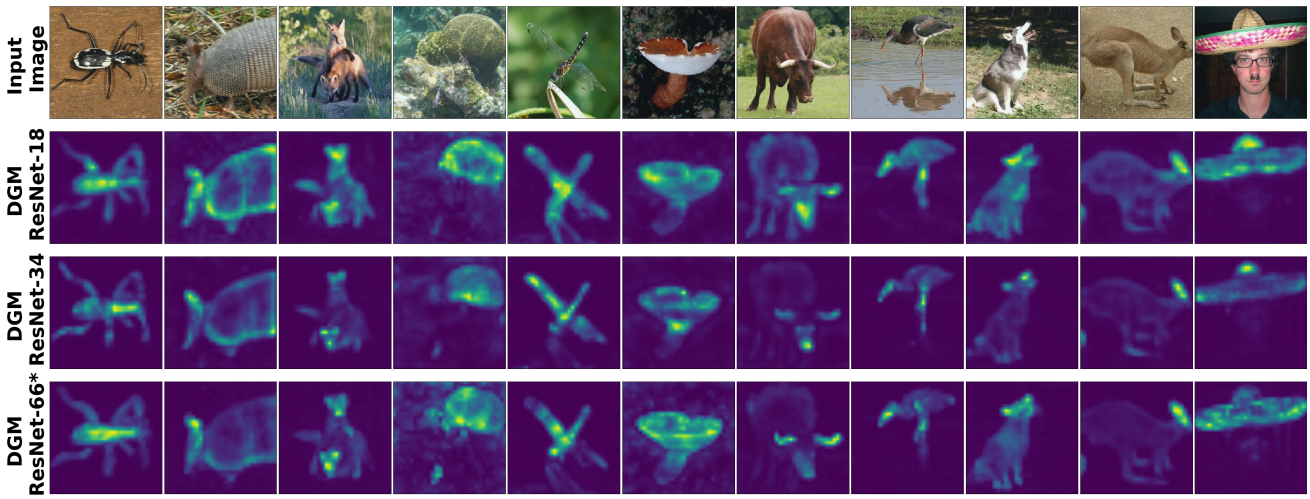

Figure 3. *Level-4* feature visualization of different DGM models on the ImageNet dataset. Note that all our DGM models produce very sharp object shape. We also observe that the objects' shape for DGM ResNet-34 and DGM ResNet-66* (DGM MobileNet) are slightly sharper in some cases compared to DGM-ResNet-18 which is also reflected in the classification accuracy performance.

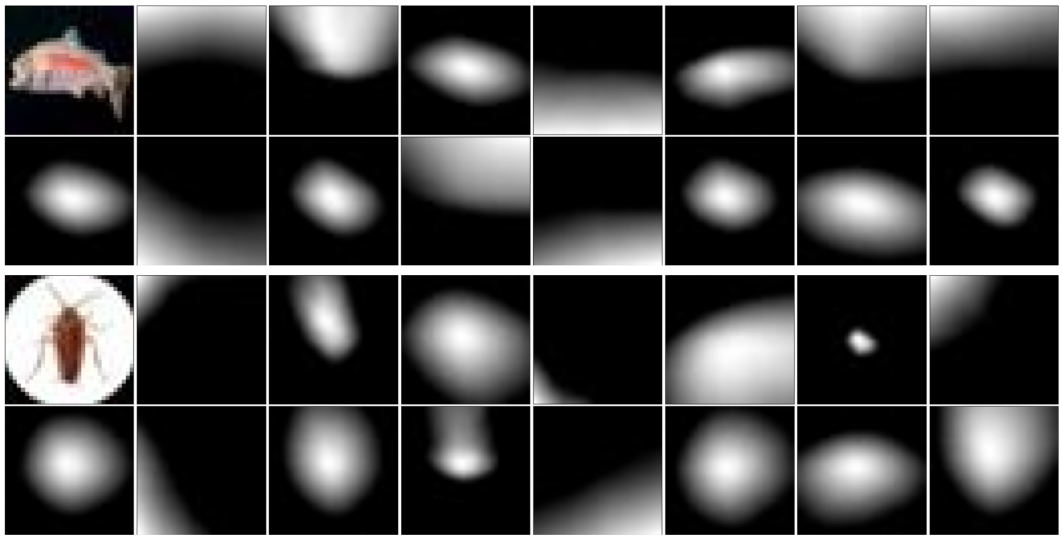

Figure 4. Comparison of bases generated from *Level-4* of our DGM model for two images from CIFAR-100 dataset (top left). Note that the bases from *Level-4* are dependent on the input image.

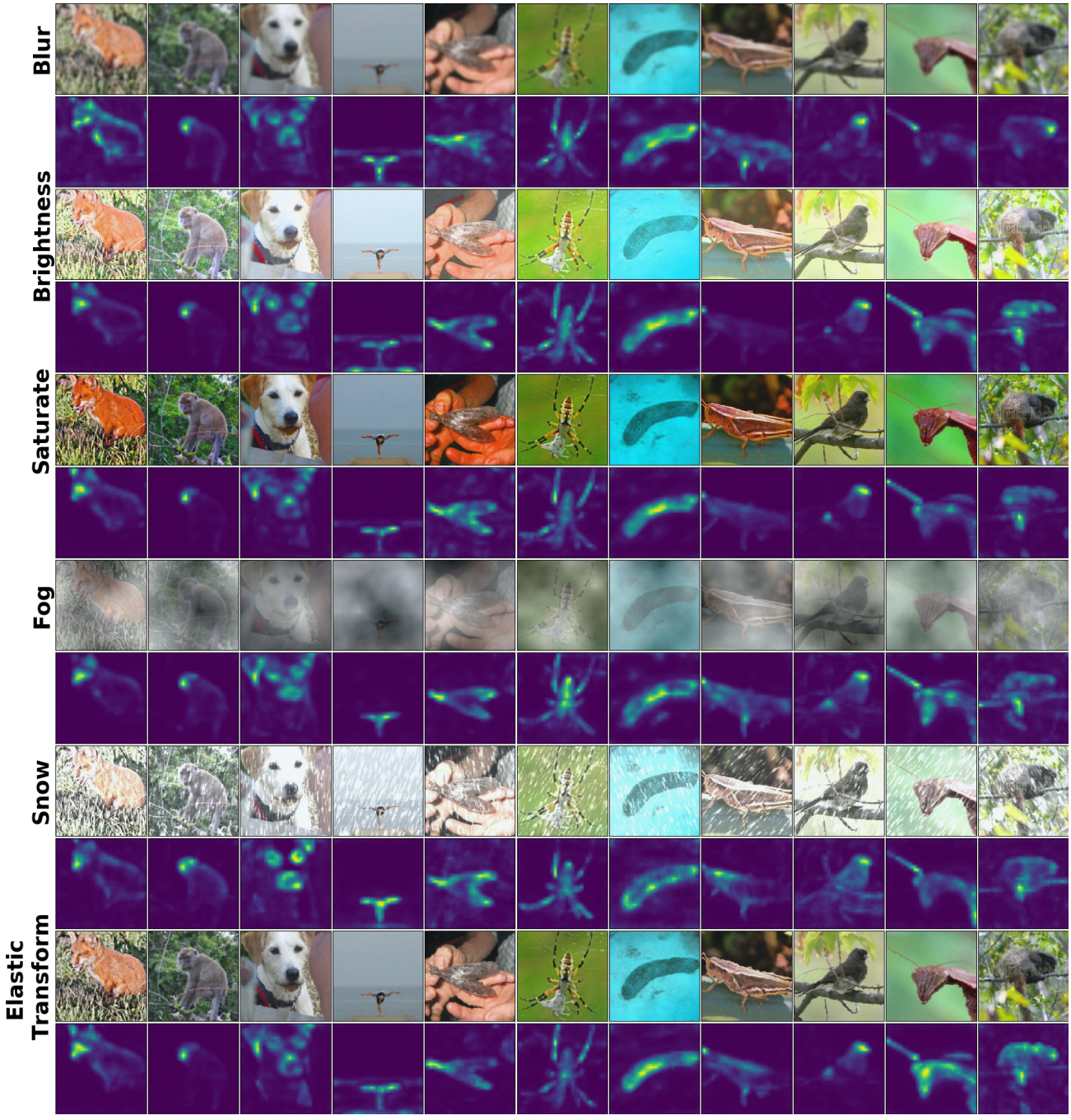

Figure 5. *Level-4* feature visualization from DGM ResNet-34 model under different color distortions (ImageNet-C). Note that our model is trained on the clean images from ImageNet dataset and is able to capture the object shape really well under challenging distortions.

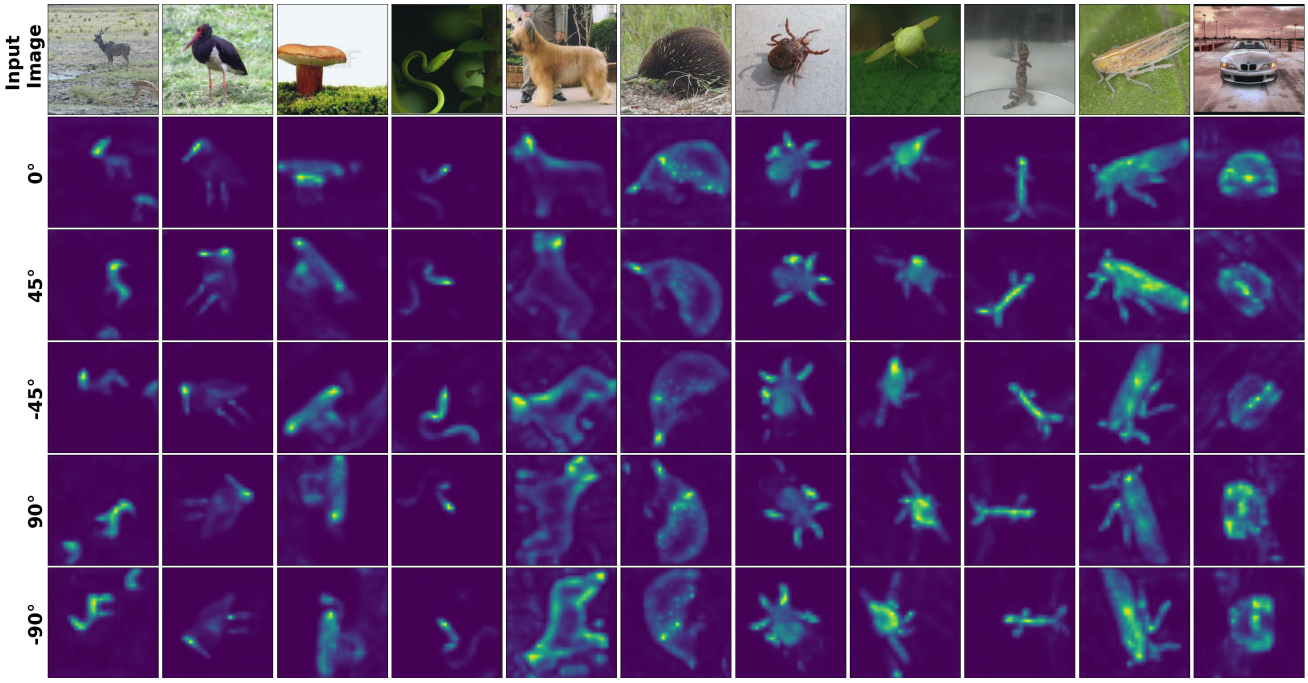

Figure 6. *Level-4* feature visualization from our DGM ResNet-34 model under different image rotation on ImageNet dataset. We do not use any affine transformation augmentation during training. Note the object shape is very well captured across different rotations of the image.

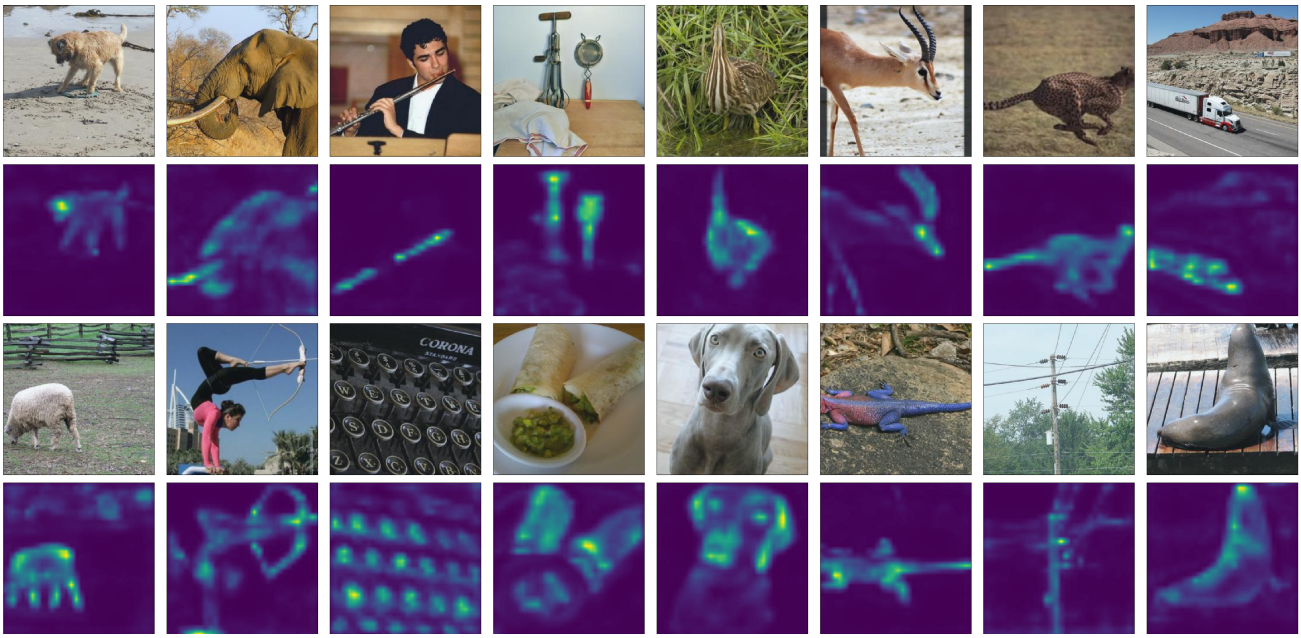

Figure 7. Few examples of *Level-4* feature visualization for DGM ResNet-34 model on the ImageNet dataset.