Supplementary Material for: HumanGuideNet: Adapter-Based Alignment of Deep Neural Networks with Human Similarity Judgments

1. Training details.

We trained HumanGuideNet (CLIP ViT-L/14, backbone frozen) with Adam (learning rate 5×10^{-4}, weight decay 1×10^{-5}), label smoothing 0.1, and gradient clipping (global norm $= 1$); batch sizes were 256 (classification) and 64 (triplet). Training used mixed precision (fp16) on an NVIDIA RTX 4070 GPU for up to 10 epochs with a cosine-annealing-with-restarts schedule (T_0=30, T_{mult}=1, $\eta_{\min}$=LR$/100$) and early stopping on validation loss (patience $= 10$); unless noted, we report the checkpoint with the lowest validation loss.

2. k-NN Purity Computation

To quantify local label consistency in the embedding space, we compute the k-nearest neighbor (k-NN) purity. Given an embedding matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ and corresponding class labels $\mathbf{y} \in \{1, \ldots, C\}^N$, we first L2-normalize each row vector to obtain $\hat{\mathbf{x}}_i = \mathbf{x}_i / (\|\mathbf{x}_i\|_2 + \epsilon)$, where $\epsilon = 10^{-12}$ ensures numerical stability. Cosine similarity is then computed between all sample pairs as the dot product of normalized embeddings.

For each sample i, we identify its k most similar neighbors, excluding itself, and compute the proportion that have the same class label:

$$\text{purity}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_i^k} \delta(y_j, y_i), \tag{1}$$

where $\delta(y_j, y_i) = 1$ if $y_j = y_i$, and 0 otherwise. The overall k-NN purity is the average across all samples:

$$\text{mean purity} = \frac{1}{N} \sum_{i=1}^{N} \text{purity}_i. \tag{2}$$

3. Linear CKA with Class Labels

To assess the global representational alignment between features and class structure, we compute linear centered kernel alignment (CKA) with a one-hot class kernel (also known as kernel–target alignment).

Given an embedding matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ and corresponding class labels $\mathbf{y} \in \{1, \ldots, C\}^N$, We first mean-center the features across samples:

$$\tilde{\mathbf{X}} = \mathbf{X} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i, \tag{3}$$

where $\mathbf{1}$ is the N-dimensional all-ones column vector. This subtracts the sample mean from each feature dimension.

We then construct a one-hot class indicator matrix $\mathbf{Y} \in \{0, 1\}^{N \times C}$ such that $Y_{ic} = 1$ if $y_i = c$ and 0 otherwise, and similarly mean-center it:

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i, \tag{4}$$

The linear CKA between $\mathbf{X}$ and $\mathbf{Y}$ is then defined as:

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\|\tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}\|_F^2}{\|\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_F \cdot \|\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}\|_F + \epsilon}, \tag{5}$$

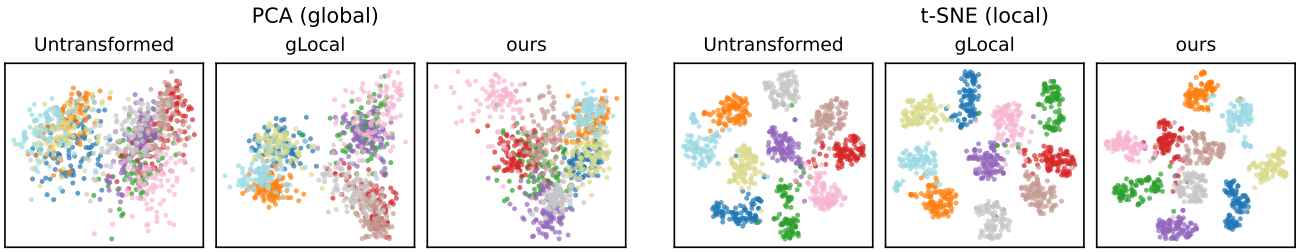

Figure 1. Comparison of representations from CLIP ViT-L/14 (WIT) using PCA (left three panels) and t-SNE (right three panels) on CIFAR-10. We show the untransformed representation, the gLocal transformation, and our fused representation. Compared to the baselines, our method produces more compact and well-separated clusters, reflecting improved global structure (PCA) and local neighborhood separability (t-SNE). Visualizations are generated from randomly selected 100 images per class, and points are colored by their class labels.

where $\|\cdot\|_F$ denotes the Frobenius norm and $\epsilon = 10^{-12}$ ensures numerical stability.

This metric lies in the interval $[0, 1]$, with higher values indicating stronger agreement between the overall geometry of the representation space and the underlying class structure.

4. Representation Visualizations on CIFAR-10 (PCA and t-SNE)

In Fig. 1, we present PCA and t-SNE visualizations of the learned representations on the CIFAR-10 dataset. Consistent with the quantitative results, the PCA visualizations reveal improved class separability and a clearer global structure in the fused representation. Similarly, the t-SNE plots show that the fused representation forms tighter and more label-consistent local neighborhoods compared to both the original and gLocal variants.

5. Effect of Sample Size on Few-Shot Performance

We evaluate the impact of the number of labeled examples per class on few-shot classification accuracy across multiple datasets. Tab. 1 summarizes the results obtained with the CLIP ViT-L/14 (WIT) backbone using three feature representations: the original frozen CLIP features, the glocal-transformed features, and our proposed HumanReg-enhanced representation.

Dataset	Method	1-shot	5-shot	10-shot	15-shot
CIFAR-100	Original	36.29	55.83	62.48	64.22
	Glocal	38.69	57.80	63.94	65.60
	Ours	**50.58**	**70.14**	**73.61**	**73.81**
CIFAR-100-Coarse	Original	34.20	55.60	63.16	66.15
	Glocal	**39.86**	59.18	67.00	69.26
	Ours	38.12	**64.50**	**75.83**	**78.60**
CIFAR-10	Original	62.80	83.80	90.52	91.00
	Glocal	**65.72**	85.48	**92.21**	91.28
	Ours	59.20	**86.44**	91.64	**93.32**
DTD	Original	35.51	54.40	59.70	62.52
	Glocal	36.41	54.47	60.00	62.55
	Ours	**38.35**	**55.80**	**60.89**	**62.63**
SUN397	Original	42.59	58.52	62.48	64.47
	Glocal	44.85	60.34	64.00	65.69
	Ours	**46.73**	**62.65**	**66.26**	**67.82**

Table 1. Few-shot classification accuracy (%) using CLIP ViT-L/14 (WIT) backbone on four datasets under different numbers of training samples per class.

Accuracy improves as the number of labeled samples per class increases, and our method generally surpasses the baselines across datasets. On CIFAR-100, SUN397, and DTD, it consistently achieves the best performance across all shot settings.

For CIFAR-10, Glocal shows slight advantages at 1- and 10-shot, whereas our method takes the lead at 5- and 15-shot. On CIFAR-100-Coarse, our approach provides the strongest results from 5-shot onward, including the highest 10- and 15-shot accuracies.

6. Anomaly Detection Performance Under Varying k

We evaluate anomaly detection performance across different neighborhood sizes k, with results summarized in Tab. 2. While the performance of the Original CLIP features decreases slightly as k grows, both the glocal transformation and our method remain stable. Our approach consistently achieves the highest AUC across all k values. These findings indicate that combining the original model representation with human-aligned features not only enhances anomaly detection but also preserves robustness to the choice of k, underscoring its reliability for practical applications.

Dataset	**Method**	$k=2$	$k=5$	$k=10$	$k=20$
CIFAR-100	Original	91.90	91.41	90.93	90.39
	Glocal	97.22	97.19	97.08	96.92
	Ours	**98.82**	**98.91**	**98.93**	**98.91**
CIFAR-100-coarse	Original	89.28	88.50	87.73	86.81
	Glocal	95.90	95.83	95.66	95.41
	Ours	**98.31**	**98.52**	**98.60**	**98.62**
CIFAR-10	Original	95.37	95.14	94.86	94.50
	Glocal	98.16	98.16	98.11	98.03
	Ours	**99.35**	**99.46**	**99.49**	**99.50**
DTD	Original	92.57	92.02	91.21	90.16
	Glocal	**95.14**	**94.90**	**94.51**	93.92
	Ours	94.74	94.67	94.43	**94.06**

Table 2. Anomaly detection AUC (%) using CLIP ViT-L/14 (WIT) across different k values. "Original" uses the pretrained features; "Glocal" applies a linear transformation; "Ours" denotes HumanGuideNet.

7. Robustness Evaluation for CLIP ViT-L/14 (L2B) and CLIP ViT-L/14 (L400M)

To further examine robustness, we evaluate our method on two additional CLIP backbones, ViT-L/14 pretrained on the L2B and L400M datasets. The results, presented in Tab. 3 and Tab. 4, show that while HumanGuideNet attains slightly lower accuracy than the gLocal method for certain corruption types, it consistently outperforms the original models. These findings indicate that the advantages of human-guided alignment are not confined to a single pretraining source but extend across different large-scale CLIP variants.

Backbone	**Method**	**ImageNet Top-1**	**Impulse Noise**		**Glass Blur**		**Fog**		**Contrast**	
			Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5
CLIP ViT-L/14 (L2B)	Original	81.50	27.93	8.93	30.65	22.83	60.17	49.23	51.12	23.39
	gLocal	**82.74**	**30.62**	10.60	**32.49**	24.62	**62.43**	**51.42**	55.18	27.69
	Ours	81.35	30.44	**10.77**	32.36	**24.78**	61.86	51.21	**55.46**	**30.15**

Table 3. CLIP ViT-L/14 (L2B): Top-1 ImageNet accuracy (%) and robustness on ImageNet-C. Results are shown under severity levels 4 and 5 for four representative corruptions.

8. Robustness Evaluation Across All Image Corruption Types and Severity Levels

We evaluate robustness under increasing corruption severity for all ImageNet-C perturbations. Tab. 5 presents top-1 accuracy across severity levels 1–5. As expected, accuracy drops as corruption severity increases, but our method consistently achieves the best performance, particularly at higher severities. For instance, under glass blur at severity 5, HumanGuideNet attains

Backbone	Method	ImageNet Top-1	Impulse Noise		Glass Blur		Fog		Contrast	
			Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5
CLIP ViT-L/14 (L400M)	Original	80.20	28.21	11.35	26.24	21.36	57.96	46.72	47.43	23.17
	gLocal	**81.75**	**30.56**	12.70	**27.38**	**22.45**	**59.33**	**47.82**	**50.11**	24.59
	Ours	79.78	29.65	**12.80**	26.21	21.66	58.30	47.30	49.54	**25.39**

Table 4. CLIP ViT-L/14 (L400M): Top-1 ImageNet accuracy (%) and robustness on ImageNet-C. Results are shown under severity levels 4 and 5 for four representative corruptions.

30.09% compared to 24.03% (Original) and 26.82% (Glocal). Comparable improvements are seen for fog and contrast, where our approach outperforms the baselines by 2–5 percentage points at the most challenging levels. These results highlight that fusing the original model representation with human-aligned features substantially enhances robustness to distributional shifts caused by common corruptions.

Category	Corruption Type	Method	1	2	3	4	5
Blur	Defocus Blur	Original	70.84	66.04	55.4	45.06	34.9
		Glocal	72.95	68.48	58.54	48.22	37.53
		Ours	**74.52**	**70.36**	**60.77**	**50.76**	**40.19**
	Glass Blur	Original	71.09	62.78	40.06	32.38	24.03
		Glocal	72.49	64.81	42.74	35.73	26.82
		Ours	**74.18**	**66.93**	**45.99**	**38.68**	**30.09**
	Motion Blur	Original	77.18	72.66	63.95	52.06	43.53
		Glocal	77.80	73.66	65.63	53.89	45.60
		Ours	**78.56**	**74.80**	**67.32**	**56.13**	**48.08**
	Zoom Blur	Original	66.80	59.46	52.73	46.44	39.16
		Glocal	68.12	61.16	54.15	47.86	40.39
		Ours	**69.47**	**62.56**	**56.14**	**50.16**	**42.47**
Noise	Gaussian Noise	Original	77.49	73.22	63.90	49.79	31.31
		Glocal	77.41	73.31	64.81	51.35	32.65
		Ours	**78.13**	**74.17**	**66.05**	**52.78**	**34.19**
	Impulse Noise	Original	71.48	65.72	60.53	48.21	33.95
		Glocal	72.45	67.27	62.05	49.75	35.33
		Ours	**74.89**	**69.93**	**65.07**	**52.46**	**37.13**
	Shot Noise	Original	76.68	70.91	62.32	45.28	33.17
		Glocal	76.72	71.53	63.30	46.69	34.42
		Ours	**77.48**	**72.50**	**64.68**	**48.00**	**35.67**
Digital	Contrast	Original	76.66	74.17	69.75	57.92	35.89
		Glocal	77.87	75.91	72.21	60.87	38.34
		Ours	**79.18**	**77.66**	**74.98**	**65.48**	**43.84**
	Elastic Transform	Original	75.22	54.43	69.68	58.10	30.30
		Glocal	75.77	55.67	70.46	59.22	31.73
		Ours	**76.28**	**56.52**	**71.26**	**60.59**	**33.02**
	JPEG Compression	Original	76.38	73.57	71.07	61.50	46.91
		Glocal	76.86	74.27	71.69	62.77	48.84
		Ours	**77.34**	**74.75**	**72.50**	**63.54**	**49.58**
	Pixelate	Original	77.29	76.18	70.53	63.28	57.96
		Glocal	77.95	76.85	72.17	66.01	60.49
		Ours	**78.66**	**77.79**	**73.86**	**67.99**	**61.85**
Weather	Brightness	Original	81.03	80.55	79.61	77.86	74.91
		Glocal	81.17	80.65	79.59	78.00	75.32
		Ours	**81.51**	**80.97**	**80.18**	**78.59**	**75.91**
	Fog	Original	76.17	74.04	69.80	65.47	55.47
		Glocal	76.52	74.57	70.80	66.52	56.63
		Ours	**77.40**	**75.91**	**72.70**	**69.51**	**61.00**
	Frost	Original	75.53	65.96	57.79	55.86	50.15
		Glocal	75.65	66.47	58.01	56.32	50.40
		Ours	**76.41**	**67.78**	**60.25**	**58.48**	**53.11**
	Snow	Original	74.45	65.80	66.67	60.11	55.90
		Glocal	74.49	66.34	66.96	60.53	56.16
		Ours	**76.20**	**68.35**	**69.31**	**62.68**	**59.15**

Table 5. Top-1 accuracy (%) of CLIP ViT-L/14 (WIT) on ImageNet under increasing severity levels (1–5) across corruption categories and types. HumanGuideNet consistently improves robustness over Original CLIP and Glocal.

9. Results for RESNET-50 backbone

We further evaluate our approach on ResNet-50, using the same training configuration as for the CLIP models, except for the learning rate, which is set to 1×10^{-4}. Tab. 6 reports the RSA results, Tab. 7 presents the few-shot performance, Tab. 8 summarizes the anomaly detection results, and Tab. 9 shows robustness to image distortions. Across all settings, our method consistently outperforms the gLocal baseline.

Model	**Method**	**Spearman ρ**
ResNet-50	Glocal	68.80
	HumReg	**78.47**

Table 6. THINGs RSA results. Spearman correlation (ρ) between model RSMs and human RSMs (higher is better).

Backbone	**Method**	**CIFAR-100**	**CIFAR-100-coarse**	**CIFAR-10**	**DTD**	**SUN397**
RESNET-50	Glocal	32.06	41.48	54.16	38.70	29.82
	Ours	**35.47**	**42.64**	**61.48**	**41.29**	**32.39**

Table 7. Few-shot classification accuracy (%) on downstream datasets using ResNet-50 backbones. "GLocal" applies the gLocal transform to the original representations, and "Ours" denotes the fused features from HumanGuideNet. Results are averaged over five runs.

Backbone	**Method**	**CIFAR-100**	**CIFAR-100-Coarse**	**CIFAR-10**	**DTD**
RESNET-50	Glocal	93.76	91.20	89.74	91.52
	Ours	**95.95**	**93.52**	**93.67**	**93.38**

Table 8. Anomaly detection AUROC (%) on four datasets using ResNet-50 backbone models. "GLocal" applies the gLocal transform to the original representations; "Ours" denotes the fused feature from the HumanGuideNet.

Backbone	**Method**	**ImageNet Top-1**	**Impulse Noise**		**Glass Blur**		**Fog**		**Contrast**	
			Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5	Sev. 4	Sev. 5
ResNet-50	gLocal	75.93	18.78	6.47	12.02	8.89	41.34	22.66	47.34	30.70
	Ours	78.00	**24.50**	**10.20**	**12.50**	**8.97**	**47.72**	**29.06**	**51.91**	**35.69**

Table 9. Top-1 ImageNet accuracy (%) and Top-1 accuracy on ImageNet-C for ResNet-50. Results are reported for severity levels 4 and 5 across four representative corruptions.