Supplementary Material: Attributional Robustness Training using Input-Gradient Spatial Alignment

Our supplementary material is organized as follows: In Section S1, we provide dataset and implementation details for training ART models. We also show ablation studies related to attributional robustness which includes testing attributional robustness of ART model on different values of ϵ. Section S2 describes dataset and implementation details for weakly supervised object localization task. Also, we provide qualitative results on this task. In Section S3, we perform additional analysis of adversarial vulnerability for ART model.

The code for our proposed methodology (ART) is available at:
https://github.com/nupurkmr9/Attributional-Robustness.

S1 Attributional Robustness: Additional Details and Results

In this section, we provide details of the datasets as mentioned in the main paper (Section 4.1), as well as some additional results on attributional robustness.

We qualitatively show in Figure S1 that attribution maps generated via ART are robust to attribution manipulation unlike *Natural* model. We also report the Top-1000 Intersection and Kendall's Correlation between original and perturbed saliency maps for ART and *Natural* models. We use target attribution attack as mentioned in [4] to perturb the attributions while keeping the predictions same. For images in Figure S1, the model predictions are correct and the attribution maps are computed using Integrated Gradient [17]. We observe that attributions of the *Natural* model are visually and quantitatively fragile as attributions are easily manipulated to resemble target attribution map that is present in the rightmost column of the figure. However, it can seen from the figure that ART models show high robustness to attribution manipulations.

S1.1 Choice of optimization objective L_{attr} and its variants

Our choice for the loss function was based on the empirical analysis as reported in table 1 on CIFAR-10. We empirically observed that instead of directly minimizing ℓ_2 distance between x and $g^y(x)$ in Equation 4 of main paper, cosine distance led to better robustness. We believe this is because cosine avoids scale mismatch issues in x and $g^y(x)$ magnitudes. The triplet loss is only introduced to improve performance on attributional robustness objective. For negative sample selection, we choose i^* as second most likely class, which represents most uncertainty, following standard principles of hard negative mining in triplet loss [7,14]. For other choices of i^* , we observed a performance drop.

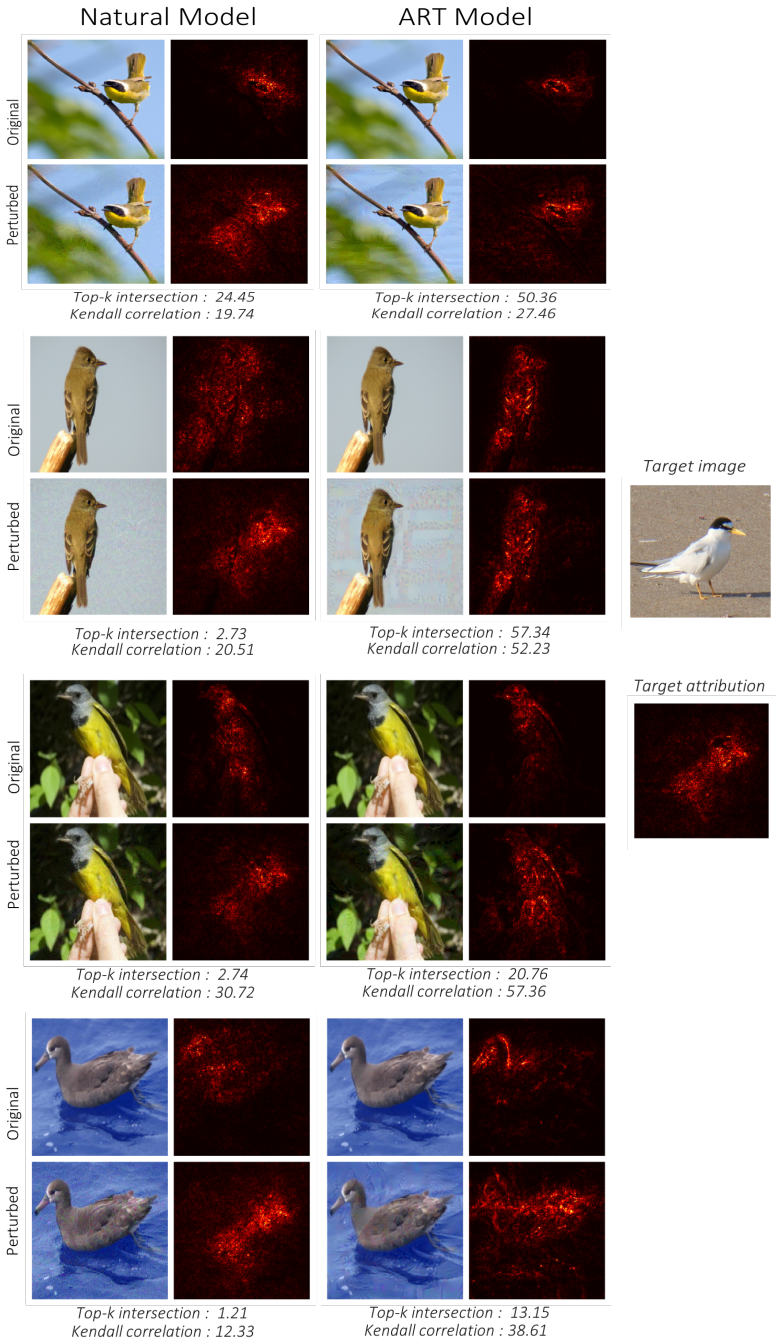

Fig. S1: Targeted attribution attack [4] using integrated gradient (IG) attribution map on *Natural* and ART trained model. Top-1000 intersection and Kendall correlation between IG attribution map of original and perturbed images is shown below each image. The target attribution manipulation uses the attribution map as depicted in the rightmost column of this figure.

Table 1: Comparison of different loss functions used as the objective function for increasing attributional robustness on CIFAR-10

Optimization Objective	Attributional Robustness		Test Accuracy	Adversarial Accuracy
	IN	K		
Equation 2	74.78	71.40	91.34	15.15
Equation 4 : ℓ_2 distance	68.41	69.75	91.66	16.64
Equation 4 : Cosine distance	91.25	89.28	89.21	35.95
Equation 5 : ART with i^*=argmin(logit)	90.75	83.32	89.94	37.93
Equation 5 : ART (ours)	92.90	91.76	89.84	37.50

S1.2 Cosine distance in L_{attr} loss

Following our discussion in Sec 3.2 of the main paper, we now elaborate on the relation of cosine distance in a unit ℓ_2-norm surface of vectors with Euclidean distance. We show below that squared Euclidean distance is proportional to the cosine distance for unit ℓ_2 norm space of vectors. Euclidean distance is a valid distance function and follows the triangle inequality which we use in Eqn 3 for obtaining the upper bound on attributional robustness as a function of the distance between an image and its attribution map.
Given two vectors x and $\tilde{x}$, with unit ℓ_2 norm i.e. $||x||_2 = 1$ and $||\tilde{x}||_2 = 1$, cosine distance between them is related to their Euclidean distance as follows:

$$\begin{aligned}(||x-\tilde{x}||_2)^2 &= (x-\tilde{x})^\top.(x-\tilde{x}) \\ &= x^\top x + \tilde{x}^\top\tilde{x} - 2.x^\top.\tilde{x} \\ &= ||x||_2 + ||\tilde{x}||_2 - 2.x^\top.\tilde{x} = 1 + 1 - 2.x^\top.\tilde{x} \\ &= 2(1 - x^\top.\tilde{x}) = 2.\mathit{CosineDistance}(x,\tilde{x})\end{aligned} \tag{1}$$

S1.3 Dataset and Implementation Details

Below, we describe the datasets and hyper-parameters used for experiments, which we could not include in the main paper owing to space constraints.

SVHN
Data and Model: SVHN dataset [12] consists of images of digits obtained from house numbers in Google Street View images, with 73257 digits for training and 26032 digits for testing over 10 classes. We perform experiments on SVHN using WideResNet-40-2 [19] architecture for training on reported approaches.

Hyperparameters for Training:
Natural: We use SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, l_2 weight decay of 2e−4 and batch size of 256. We train it for 200 epochs with a learning rate schedule decay of 0.1 at 50^{th}, 80^{th} and 0.5 at 150^{th} epoch.
PGD-7: We use the training configuration as in [10] to perform 7-step adversarial

training with $\epsilon = 8/255$ and step size 2.5/255.
ART: We use the same training configuration as mentioned for *Natural* model, $\beta = 50$ and $\lambda = 0.5$. We calculate $\tilde{x}$ using $\epsilon = 8/255$, step size 1.5/255 and number of steps $a = 3$.

CIFAR-10
Data and Model: CIFAR-10 dataset [8] consists of 50000 training images for 10 classes with resolution of $32 \times 32 \times 3$. We normalize the images with its mean and standard deviation for training. We train a WideResNet28-10 [19] model for all the experiments on this dataset.

Hyperparameters for Training:
Natural: We use SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, l_2 weight decay of 2e−4 and batch size of 256. We train it for 100 epochs with a learning rate schedule decay of 0.1 at 50^{th}, 80^{th} and 0.5 at 150^{th} epoch.
PGD-10: We use the training configuration as mentioned in [10] to perform 10-step adversarial training with $\epsilon = 8/255$ and step size 2/255.
ART: We use the same training configuration as mentioned for *Natural* model with $\beta = 50$ and $\lambda = 0.5$. We calculate $\tilde{x}$ using $\epsilon = 8/255$, step size 1.5/255 and number of steps $a = 3$.

GTSRB
Data and Model: German Traffic Signal Recognition Benchmark [16] consists of 43 classes of traffic signals with $34,799$ training images, $4,410$ validation images and $12,630$ test images. We resize the images to $32 \times 32 \times 3$ and normalize the images with its mean and standard deviation for training. To balance the number of images for each class, we use data augmentation techniques consisting of rotation, translation, and projection transforms to extend the training set to $10,000$ images per class as in [2]. We train WideResNet28-10 [19] model for carrying out experiments related to this dataset.

Hyperparameters for Training:
Natural: We use SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, l_2 weight decay of 2e−4 and batch size of 128. We train it for 12 epochs with a learning rate schedule decay of 0.1 at 4^{th}, 6^{th} and 0.5 at 10^{th} epoch.
PGD-7: We use the training configuration same as [2] to perform 7-step adversarial training with $\epsilon = 8/255$ and step size 2/255.
IG Norm and *IG-Sum Norm [2]:* We report the accuracy as mentioned in the paper [2].
ART: We use the same training configuration as mentioned for *Natural* model with $\beta = 50$ and $\lambda = 0.5$. We calculate $\tilde{x}$ using $\epsilon = 8/255$, step size 1.5/255 and number of steps $a = 3$.

Flower
Data and Model: Flower dataset [13] has 17 categories with 80 images for each class. We resize the images to $128 \times 128 \times 3$ and normalize it with its mean and

standard deviation for training. The training set consists of $1,224$ images with 72 images per class. The test set compromises of 136 images with 8 images per class. We use standard data augmentation techniques of rotation, translation, and projection transforms to extend the training data so that each class contains $1,000$ training examples as proposed in [2]. We use WideResNet28-10 [19] model for the reported approaches.

Hyperparameters for Training:
Natural: We use SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, l_2 weight decay of 2e$-$4 and batch size of 128. We train it for 68 epochs with a learning rate schedule decay of 0.1 at 15^{th}, 35^{th} and 0.5 at 50^{th} epoch.
PGD-7[9]: We use the training configuration as mentioned in [10] to perform 7-step adversarial training with $\epsilon = 8/255$ and step size 2.5/255.
IG Norm and IG-Sum Norm [2] : We report the accuracy as mentioned in the paper [2].
ART: We use the same training configuration as mentioned for *Natural* model with $\lambda = 0.5$ and $\beta = 50$. We calculate $\tilde{x}$ using $\epsilon = 8/255$, step size 1.5/255 and number of steps $a = 3$.

Attack Methodology and Evaluation
For evaluation, we perform the Top-K variant of Iterative Feautre Importance Attack (IFIA) proposed by [5]. Feature importance function is taken as Integrated Gradients [17], and dissimilarity function is Kendall Correlation. The hyperparameters used are the same as in [2] i.e. for CIFAR-10, SVHN and GTSRB datasets, k in top-k is 100, ϵ is 8/255, number of steps is 50 and step-size is 1/255. For the Flowers dataset, k is 1000, ϵ is 8/255, number of steps is 100 and step-size is 1/255. We also show the comparison by varying ϵ on CIFAR-10 dataset in Section S1.4. Evaluation is also similar to [2] using Top-k intersection and Kendall correlation measure and we report both numbers as percentage values. For Top-k intersection, k is 100 for CIFAR-10, SVHN and GTSRB datasets, and 1000 for Flowers dataset.

S1.4 Additional Analysis on CIFAR-10

Attributional Robustness: In Fig S2, we show the variance box plot of Kendall Correlation and Top-k Intersection with $\epsilon = 8/255$ for *Natural*, *ART* and *PGD-10* [9] models on CIFAR-10. *ART* has higher attributional robustness with the least variance as compared to other approaches across 1000 samples randomly selected from the test dataset. We also measure the attributional robustness of models on varying ϵ to the standard values of 2/255, 4/255, 8/255 and 12/255 in the attack methodology as explained in Section S1.3. Figure S3 shows the Top-k Intersection and Kendall correlation measure for the same. We can see that *ART* outperforms *PGD-10* and *Natural* model over all choices of ϵ.

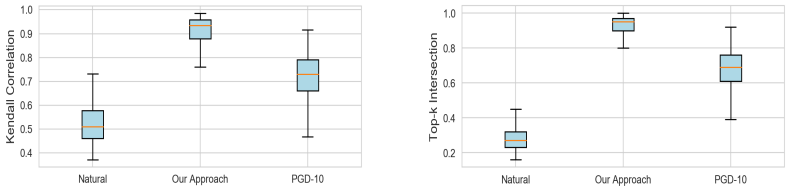

Fig. S2: Variance box plot of Attributional Robustness measure for different models on Kendall Correlation (left) and Top-k Intersection (right) for 1000 test samples of CIFAR-10

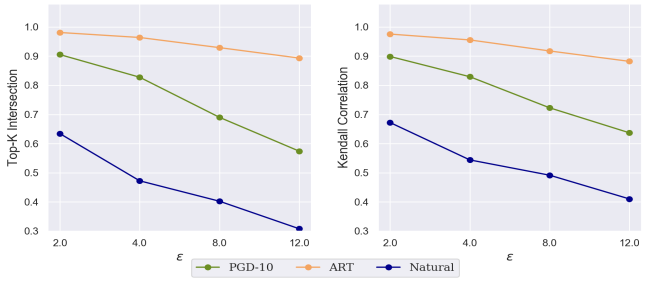

Fig. S3: Attributional robustness on varying ϵ for *ART*, *PGD-10* and *Natural* models on CIFAR-10

S2 Weakly Supervised Localization: More Details and Results

In this section, we provide more details of the dataset used for the results presented in the main paper on weakly supervised localization (Section 4.2), as well as more qualitative examples for these experiments.

S2.1 Dataset and Implementation Details

We begin by describing the dataset used in experiments for weakly supervised localization, which we could not include in the main paper owing to space constraints.

Dataset and Model: CUB-200 [18] is an image dataset of 200 different bird species (mostly North American) with 11,788 images in total. The information as a bounding box around each bird is also available. We finetune a ResNet-50 [6] model pre-trained on ImageNet for the reported approaches as in [3].

Hyper-parameters for training

Natural: We use SGD optimizer with an initial learning rate of 0.01, momentum of 0.9 and l_2 weight decay of 1e−4. We train the model for 200 epochs with batch size 128 and learning rate decay of 0.1 at every 60 epochs.

PGD-7 [9]: We use same hyper-parameters as natural training with $\epsilon = 2/255$.

Fig. S4: Examples of estimated bounding box and heatmap by ResNet50 model trained via our approach on randomly chosen images of CUB dataset; Red bounding box is ground truth and green bounding box corresponds to the estimated box

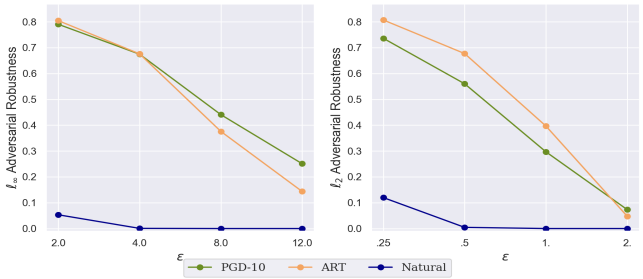

Fig. S5: ℓ_∞ and ℓ_2 adversarial robustness on varying ϵ of *ART*, *PGD-10* and *Natural* model on CIFAR-10

and $step\_size = 0.5/255$. for calculating adversarial examples.
ART: We use SGD optimizer with an initial learning rate of 0.01, momentum of 0.9 and l_2 weight decay of 1e−4. We decay the learning rate by 0.1 at every 40 epoch till 200 epochs and train with a batch size of 90. While calculating L_{attr} loss, we took mean over channels of images and gradients. Values of other hyper-parameters are $\epsilon = 2/255$, $step\_size = 1.5/255$, $a = 3$, $\lambda = 0.5$ and $\beta = 50$.

S2.2 Qualitative Analysis

Figure S4 shows the estimated bounding box and heatmap derived from gradient based attribution [15] on randomly sampled images for ResNet50 model trained via our approach. We observe that the estimated bounding box sometimes does not capture the complete object in cases where birds have extended wings, or the bird is in an occluded area with branches and twigs. Although, we observe qualitatively that this issue also exists for other models [3].

S3 Adversarial Robustness: Ablation Studies

In this section, we provide additional ablation results on adversarial robustness for the CIFAR-10 dataset.

Table 2: Comparison of Adversarial accuracy of different baseline models using transfer-based black-box attacks on CIFAR-10

Training Approach	Adversarial perturbation created using			Clean Test Accuracy
	Natural	PGD-10	ART	
Natural	0.00	80.35	49.09	95.26
PGD-10	86.44	44.07	71.34	87.32
ART	88.45	72.72	37.58	89.84

Adversarial Robustness on ℓ_∞ and ℓ_2 PGD Perturbations with Varying ϵ To analyze the adversarial robustness of *ART* model, we report and compare accuracy of the *ART* model and the *PGD-10* adversarially trained model over ℓ_∞ and ℓ_2 PGD perturbations for different values of ϵ on CIFAR-10. In Figure S5, we can observe that *ART* adversarial robustness for ℓ_∞ perturbations is similar to *PGD-10* for ϵ less than $4/255$ and better for various values of ℓ_2 perturbations.

Transfer-based black-box attacks We analyse the adversarial robustness of *ART* models on transfer-based black box attacks. Specifically, we compute the adversarial perturbations on the test set of CIFAR-10 for different baseline models and evaluate its adversarial accuracy on *ART*. We see that the transfer of adversarial perturbation from *ART* is much better than PGD-10 on Natural model. *ART* also shows higher robustness than PGD-10 for transfer attack from Natural model as reported in table 2.

Comparison with other training techniques for adversarial robustness: We consider *JARN*[1] and *CURE*[11], which are recently proposed training techniques for adversarial robustness that are different from adversarial training [9]. We compare the adversarial robustness of these techniques with *ART* on CIFAR-10 dataset using a ℓ_∞ PGD-20 adversarial perturbation with $\epsilon = 8/255$. *JARN*, *CURE* and *ART* show adversarial accuracy of 15.5%, 41.4% and 37.73% respectively and test accuracy of 93.9%, 83.1% and 89.84% respectively.

Using $L_{attr} + L_{ce}$ to Compute Perturbations $\tilde{x}$ With the motive to combine the benefits from attributional and adversarial robust models, we augment the loss function of our approach with adversarial loss [9]. We observe that the model achieves test accuracy of 85.33 and adversarial accuracy of 52.31 on PGD-40 ℓ_∞ attack with $\epsilon = 8/255$ as compared to the *PGD-10* model which has 87.32 test accuracy and 44.07 adversarial accuracy. The attributional robustness measure of Top-k intersection and kendall correlation using Integrated Gradients is 74.24 and 77.86 which is less than the attributional robustness of *ART* model but is $\sim 5\%$ better than *PGD-10* model.

References

1. Chan, A., Tay, Y., Ong, Y.S., Fu, J.: Jacobian adversarially regularized networks for robustness. ICLR (2020)
2. Chen, J., Wu, X., Rastogi, V., Liang, Y., Jha, S.: Robust attribution regularization. arXiv preprint arXiv:1905.09957 (2019)
3. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2219–2228 (2019)
4. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: Advances in Neural Information Processing Systems. pp. 13567–13578 (2019)
5. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3681–3688 (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
7. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
8. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10. URL `http://www.cs.toronto.edu/kriz/cifar.html` (2010)
9. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
10. MadryLab: cifar10_challenge. URL `https://github.com/MadryLab/cifar10_challenge` (2017)
11. Moosavi-Dezfooli, S.M., Fawzi, A., Uesato, J., Frossard, P.: Robustness via curvature regularization, and vice versa. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9078–9086 (2019)
12. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
13. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1447–1454 (2006)
14. Schroff, Florian an Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. CVPR (2015)
15. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
16. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In: IEEE International Joint Conference on Neural Networks. pp. 1453–1460 (2011)
17. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. ICML (2017)
18. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
19. Zagoruyko, S., Komodakis, N.: Wide residual networks. CoRR **abs/1605.07146** (2016), `http://arxiv.org/abs/1605.07146`