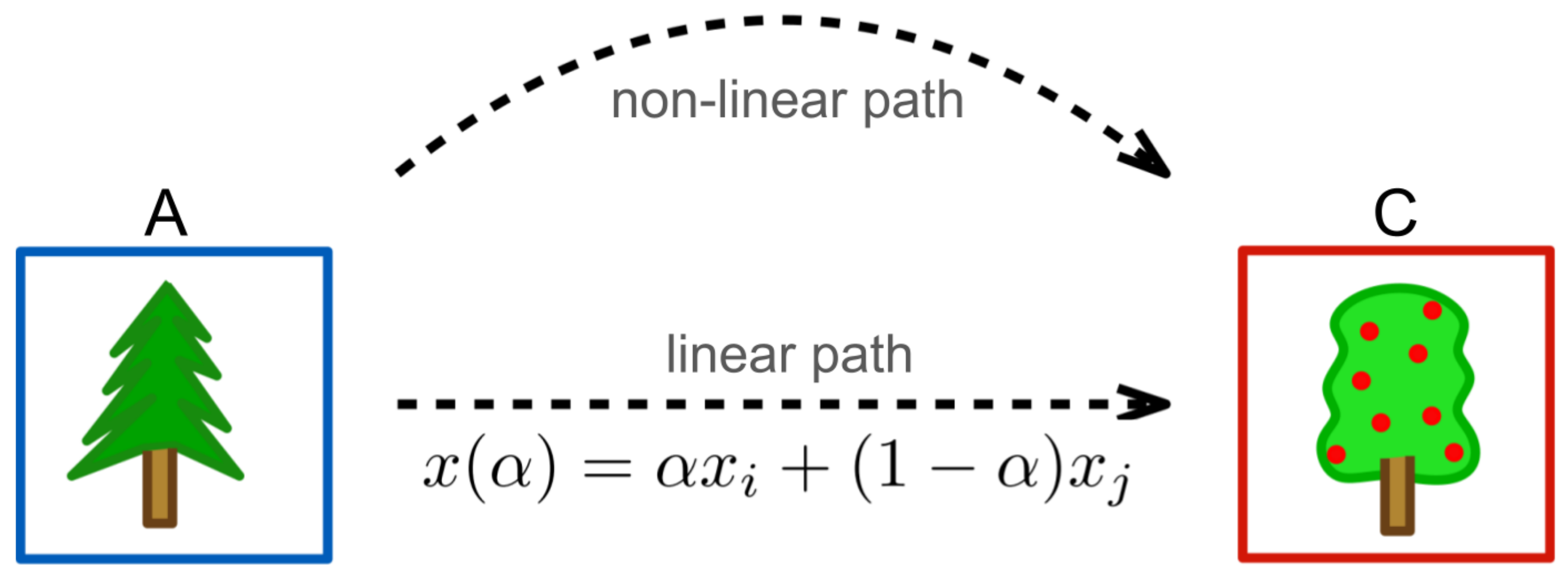
Input Space Mode Connectivity in Deep Neural Networks

Jakub Vrabel [1], Ori Shem-Ur [2] Yaron Oz [2] David Krueger [3]

[1]CEITEC, Brno University of Technology [2]Tel Aviv University [2]Mila, University of Montreal



$$\boxed{\mathcal{L}(f(x=A,\theta),y=\text{tree})} = \boxed{\mathcal{L}(f(x=C,\theta),y=\text{tree})} = 0$$

TL;DR

- We extend the concept of loss landscape mode connectivity to the input space of deep neural networks.
- We conjecture that input space mode connectivity emerges from high-dimensional geometry.
- We show that paths connecting examples of the same class are usually simple.
- We introduce a simple adversarial detection algorithm leveraging mode connectivity, which outperforms baselines under advanced attacks.
- We discuss interpretability potential through class-optimal manifold exploration.

Parameter space mode connectivity

Multiple loss minimizers (i.e., trained models) often lie along simple continuous paths in parameter space, as shown by Garipov et al. (NeurIPS 2018) and Draxler et al. (ICML 2018). Each such minimizer, or *parameter mode*, is defined as:

$$\theta_i = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(f(\mathcal{D};\theta))$$

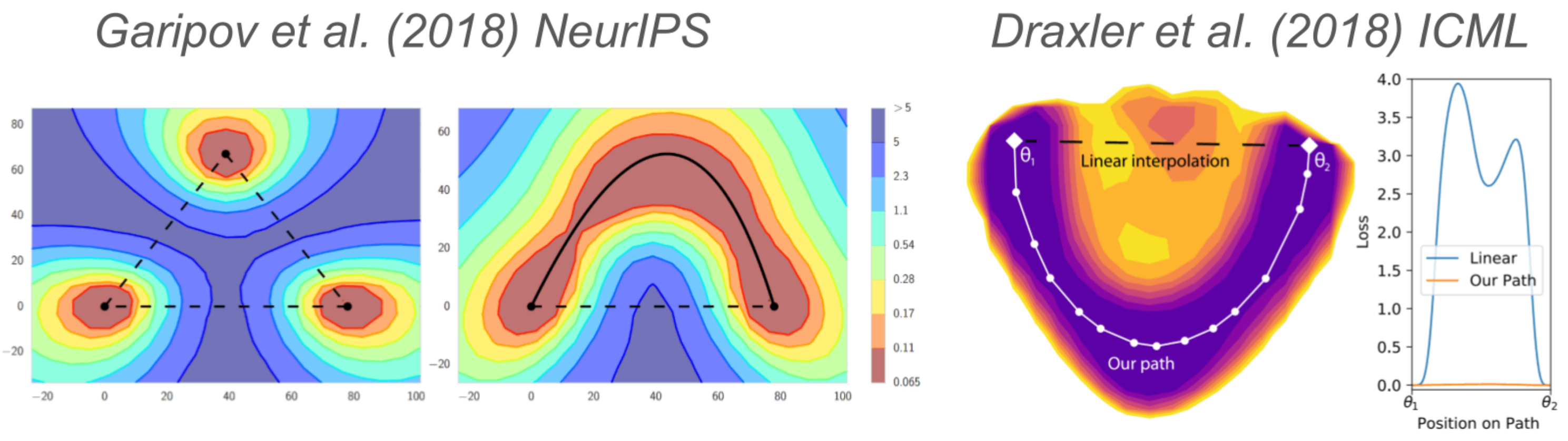


Figure 1. Early examples of mode connectivity in parameter space.

Input space modes

In the input space, multiple inputs exist that lead to confident predictions for a given class. Each such *input mode* minimizes the loss for a fixed model θ and target label y_i:

$$x_i = \underset{x}{\operatorname{argmin}}\, \mathcal{L}(f(x_i;\theta),y_i)$$

Modes are connected by simple continuous paths, where the loss remains negligible along the path. The path-finding procedure follows these steps:

1. Find two loss-minimizing inputs (images) A and C; denote $x_A \equiv A$.
2. Linearly interpolate between A and C: $x(\alpha) = \alpha x_A + (1-\alpha)x_C$.
3. Compute the loss curve by evaluating $\mathcal{L}(x(\alpha))$ for all α and locate the highest loss point B (the "barrier").
4. Perform gradient descent on the input x_B (with fixed θ) to minimize $\mathcal{L}(f(x_B;\theta),y_i)$, yielding a new mode $x_{B'}$. The path is formed by linearly connecting points A, B', and C.

Input space mode connectivity

Mode connectivity is consistently observed across a wide range of vision models (GoogLeNet, ViT, ResNet, VGG, CNN, MLP) and datasets (ImageNet, CIFAR), suggesting it is a general property of deep networks.

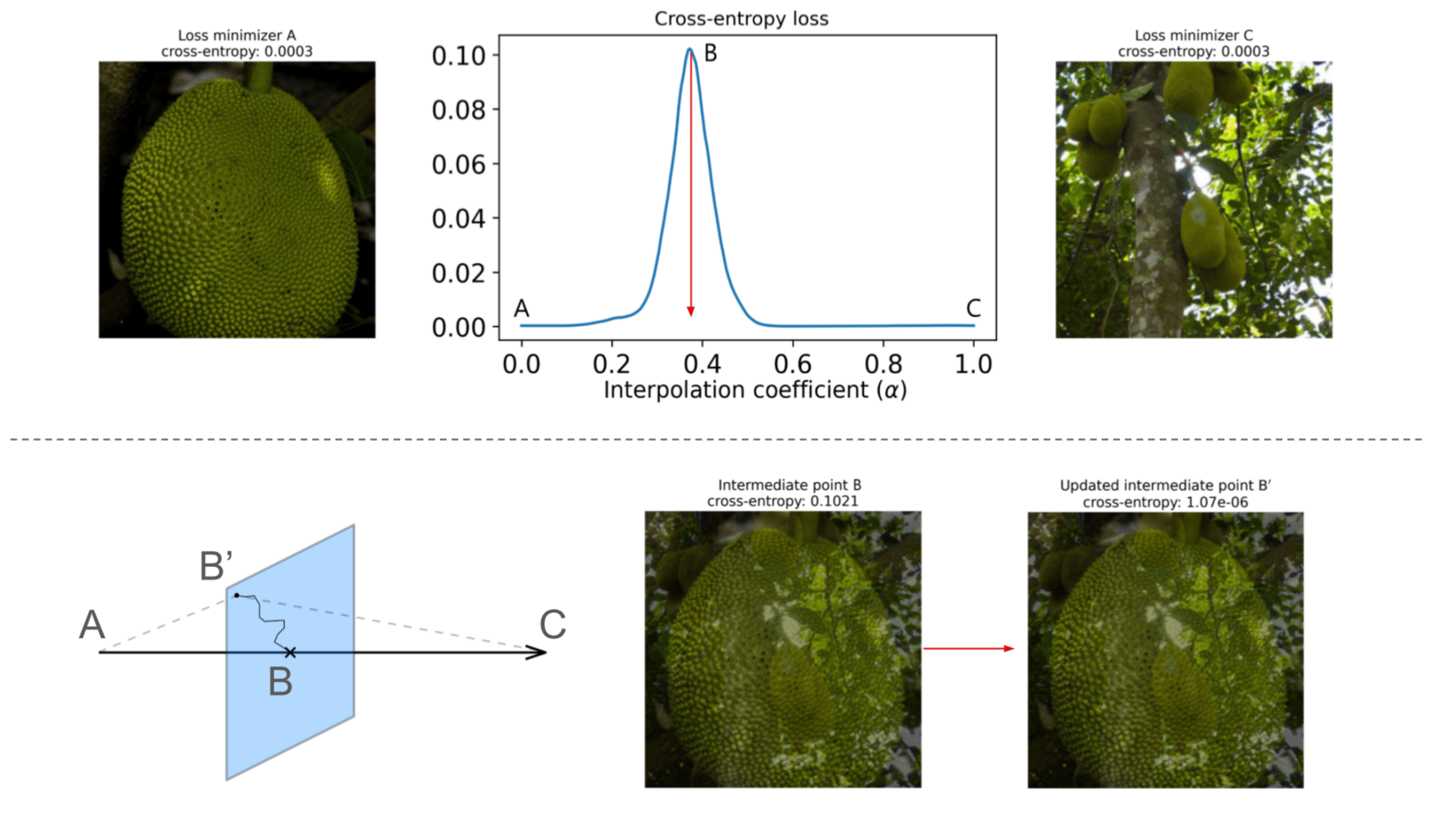


Figure 2. Input space mode connectivity example in GoogLeNet and ImageNet.

Path optimization

To find B', we optimize the intermediate point B using the Adam optimizer with an L2 norm and a high-frequency penalty. We typically run the optimization for 256 to 2048 iterations.

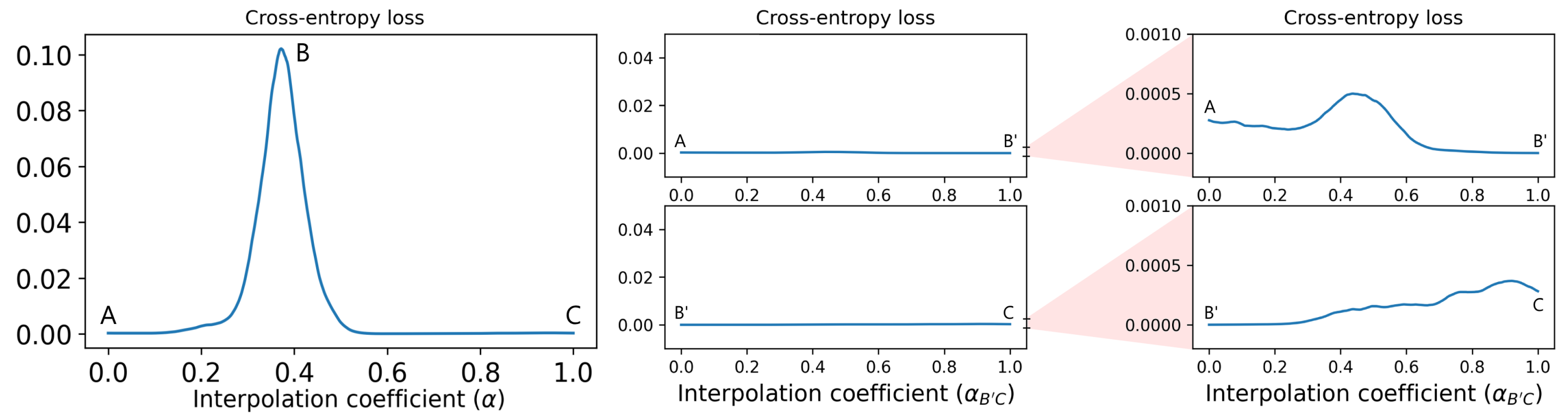
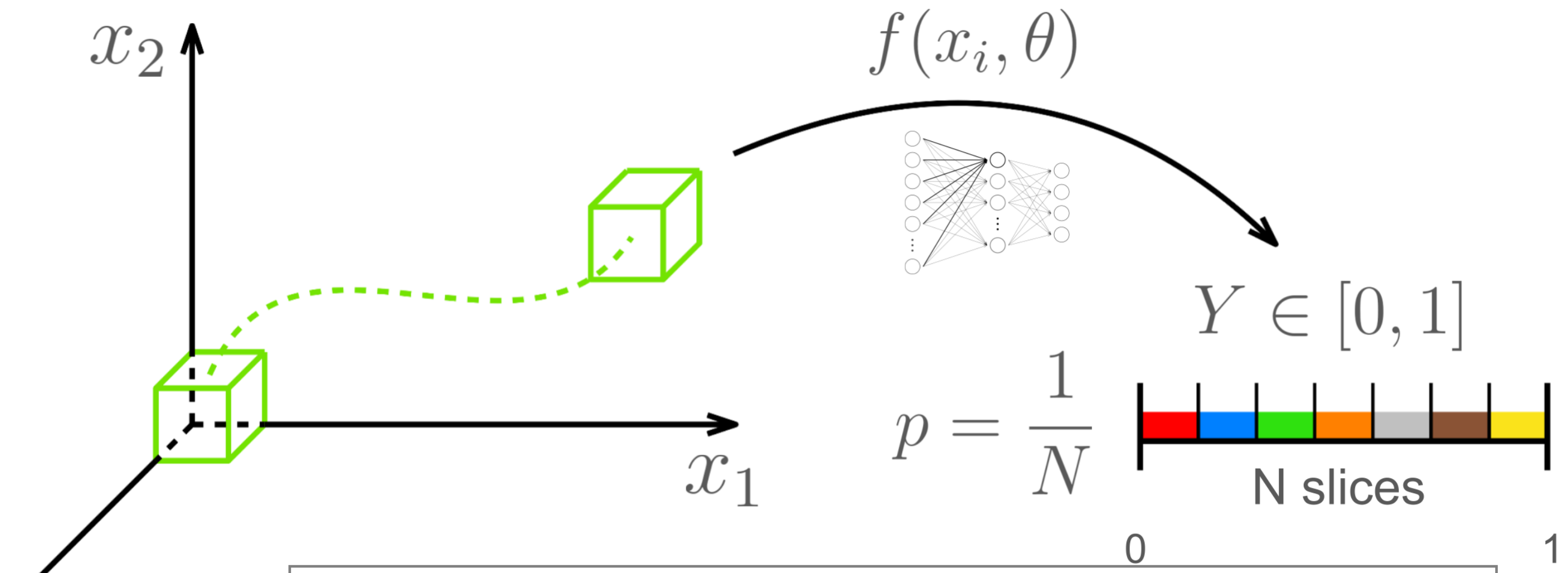


Figure 3. Barrier B is being optimized until mode B' is found. Small secondary barriers can emerge but are neglected if below a reasonable threshold ($\delta = 0.001$ considered here).

Theory: high dimensional percolation



Conjecture 5.1 (Geometric Input Space Connectivity). *Given a subset of the input space $X' \subseteq X$, a network $f(\cdot;\theta)$ at initialization, and a loss function $\mathcal{L}$, whose specifications are provided in Appendix E.1, the following holds: For arbitrarily small $0 < \delta' < \delta$, any two inputs $x_0, x_1 \in X'$, selected independently of θ and with similar predictions are almost always connected as $d_X \to \infty$:*

$$P\left(x_0, x_1 \text{ are } \delta\text{-connected} \mid \mathcal{L}\left(f\left(x_0;\theta\right), f\left(x_1;\theta\right)\right) \leq \delta'\right) = 1 - O\left(e^{-d_X \tilde{\delta}}\right), \qquad (6)$$

Figure 4. Input space mode connectivity can be studied as a percolation on a high-dimensional graph that segments the input space.

Adversarial examples

Connectivity patterns considerably differ between natural–natural and natural–adversarial image pairs. Adversarial paths typically exhibit higher barriers and more complex structure.

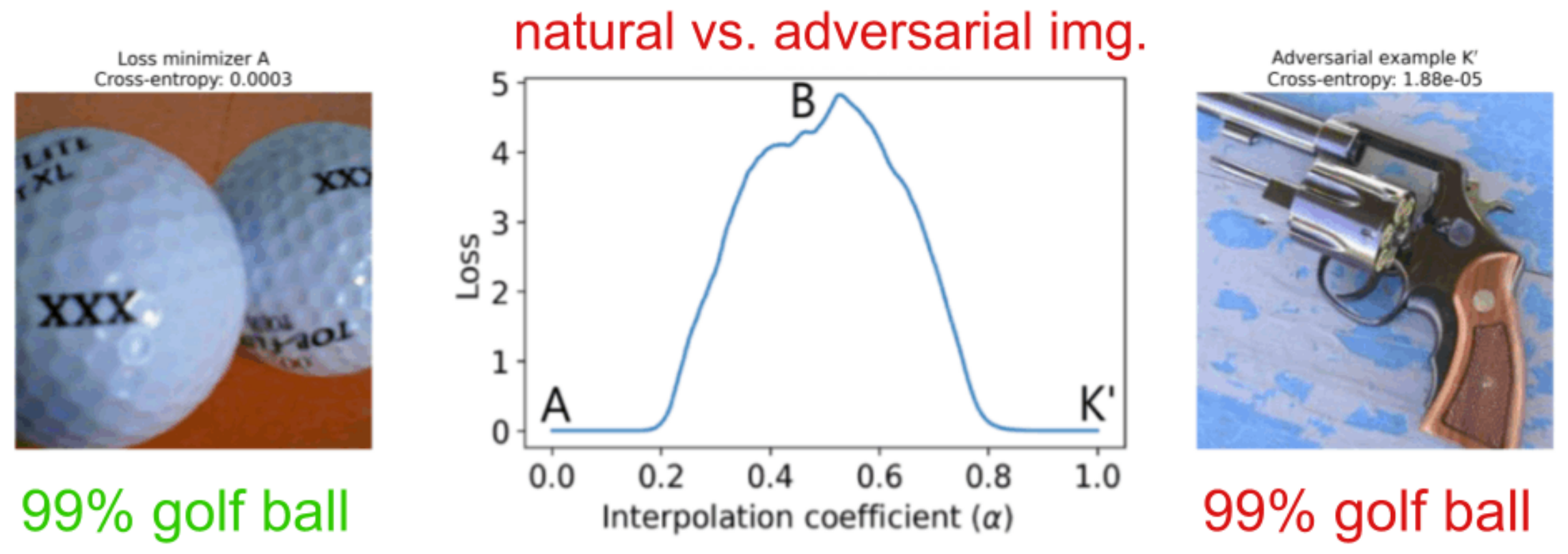


Figure 5. The loss barrier between natural–adversarial image pairs is generally higher and more complex. The path contains more segments and therefore the path-finding procedure requires more rounds to bypass all barriers.

Application: Adversarial Detection

We exploited this difference in connectivity to design a simple adversarial detection algorithm. For a given image, we compute the loss curve between it and a class-specific low-loss reference input. These loss curves are used as features for a lightweight classifier.

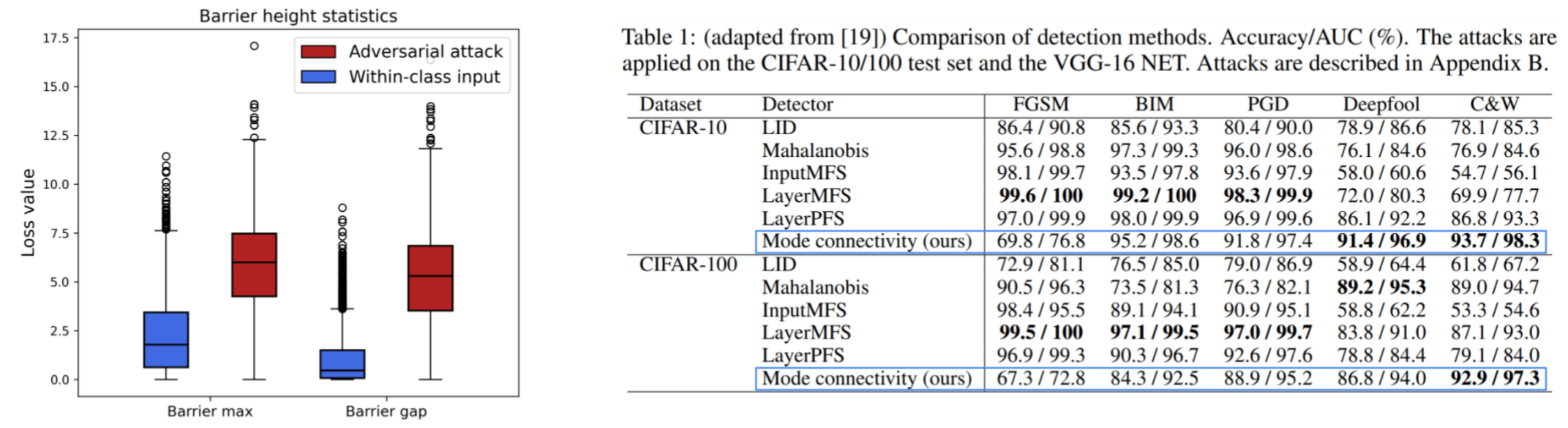


Table 1: (adapted from [19]) Comparison of detection methods. Accuracy/AUC (%). The attacks are applied on the CIFAR-10/100 test set and the VGG-16 NET. Attacks are described in Appendix B.

Dataset	Detector	FGSM	BIM	PGD	Deepfool	C&W
CIFAR-10	LID	86.4 / 90.8	85.6 / 93.3	80.4 / 90.0	78.9 / 86.6	78.1 / 85.3
	Mahalanobis	95.6 / 98.8	97.3 / 99.3	96.0 / 98.6	76.1 / 84.6	76.9 / 84.6
	InputMFS	98.1 / 99.7	93.5 / 97.8	93.6 / 97.9	58.0 / 60.6	54.7 / 56.1
	LayerMFS	**99.6 / 100**	**99.2 / 100**	**98.3 / 99.9**	72.0 / 80.3	69.9 / 77.7
	LayerPFS	97.0 / 99.9	98.0 / 99.9	96.9 / 99.6	86.1 / 92.2	86.8 / 93.3
	Mode connectivity (ours)	69.8 / 76.8	95.2 / 98.6	91.8 / 97.4	**91.4 / 96.9**	**93.7 / 98.3**
CIFAR-100	LID	72.9 / 81.1	76.5 / 85.0	79.0 / 86.9	58.9 / 64.4	61.8 / 67.2
	Mahalanobis	90.5 / 96.3	73.5 / 81.3	76.3 / 82.1	**89.2 / 95.3**	89.0 / 94.7
	InputMFS	98.4 / 95.5	89.1 / 94.1	90.9 / 95.1	58.8 / 62.2	53.3 / 54.6
	LayerMFS	**99.5 / 100**	**97.1 / 99.5**	**97.0 / 99.7**	83.8 / 91.0	87.1 / 93.0
	LayerPFS	96.9 / 99.3	90.3 / 96.7	92.6 / 97.6	78.8 / 84.4	79.1 / 84.0
	Mode connectivity (ours)	67.3 / 72.8	84.3 / 92.5	88.9 / 95.2	86.8 / 94.0	**92.9 / 97.3**

Figure 6. **Left**: Statistics over 5,000 pairs (5 per each class, ImageNet). Barrier gap ≡highest loss - max($\mathcal{L}$(A), $\mathcal{L}$(B)). **Right**: Benchmark from Harder et al. 2021.

Connectivity in random models

Connectivity also manifests in randomly-initialized (untrained) networks. Using ResNet18, we optimized class-optimal inputs from Gaussian noise. Despite random parameters, piecewise-linear paths with negligible loss were found after two rounds of the path-finding procedure.

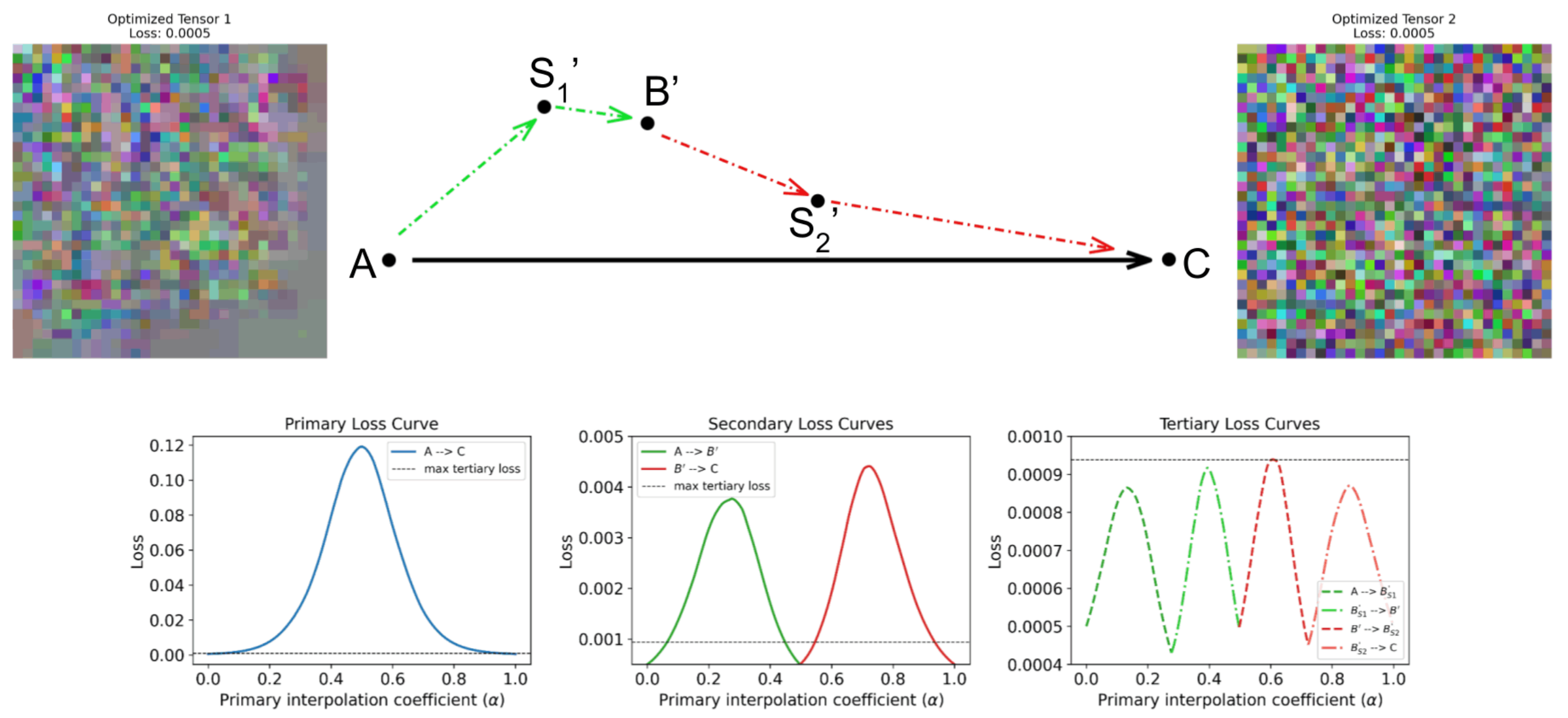


Figure 7. In untrained models, additional rounds of the path-finding procedure are often required, as secondary and higher-order barriers emerge. The procedure continues until all segments lie within the target loss level set.

Conclusion

Mode connectivity appears to be a general phenomenon arising in high-dimensional spaces, beyond just parameter landscapes. Our results show that it also takes place in input space and can be leveraged for tasks such as adversarial detection and interpretability. This perspective opens new directions for understanding and exploiting the geometry of deep networks.