

# Curved Representation Space of Vision Transformers



Kim, Juyeop, Junha Park, Songkuk Kim, and Jong-Seok Lee.  
published at AAAI, 2023

Jonas Alle

20<sup>th</sup> of March, 2025

BioMedIA Reading Group



# Motivation – Counterintuitive Findings

self-attention **vs** convolutional architectures

- Transformers are **more robust** to adversarial perturbations than CNNs
  - their citations: [3], [4], [5], [19], [20], [21], [22]
- Transformers are reported to **not** produce **overconfident** predictions unlike CNNs
  - their citations: [6], [23], [24], [25]

Kim, Juyeop, Junha Park, Songkuk Kim, and Jong-Seok Lee. “Curved Representation Space of Vision Transformers.” AAAI, 2023. <http://arxiv.org/abs/2210.05742>.

• [5]: Sayak Paul and Pin-Yu Chen. **Vision transformers are robust learners**. In AAAI, 2022

• [6]: Matthias Minderer, Josip Djolonga, Rob Romijnders, ..., and Mario Lucic. **Revisiting the calibration of modern neural networks**. In NIPS, 2021

# Expected Calibration Error (ECE) on ImageNet Validation Set

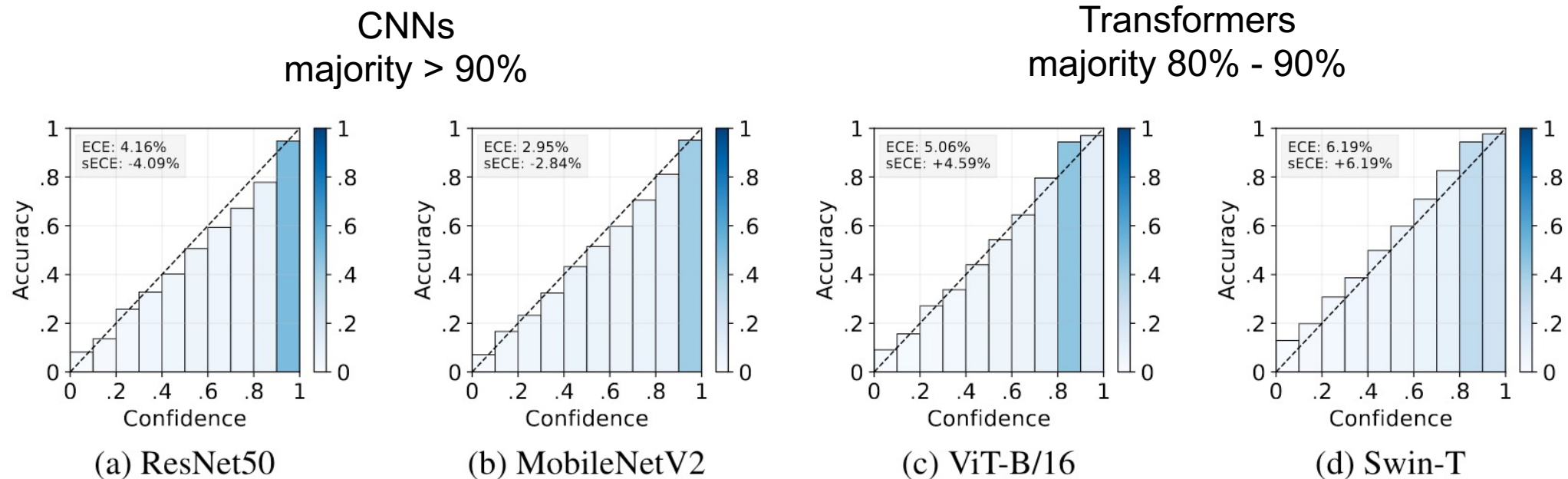


Figure 2: Reliability diagrams of CNNs and Transformers. Transparency of bars represent the ratio of the number of data in each confidence bin. ECE and sECE values are also shown in each case.

# Expected Calibration Error (ECE) on ImageNet Validation Set

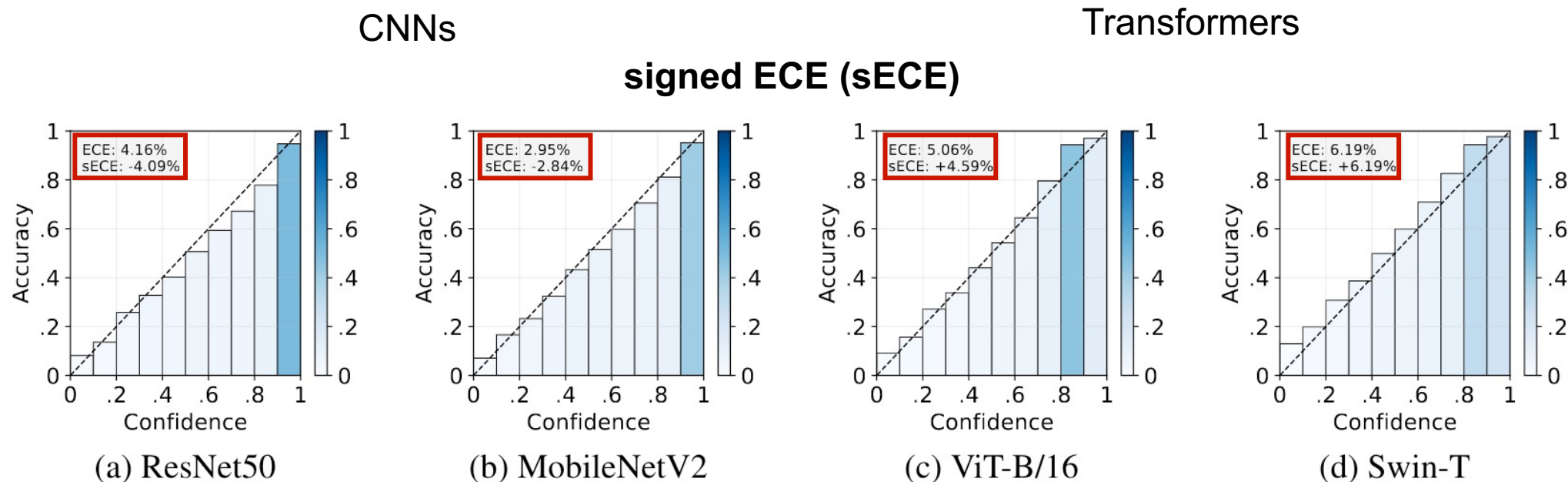
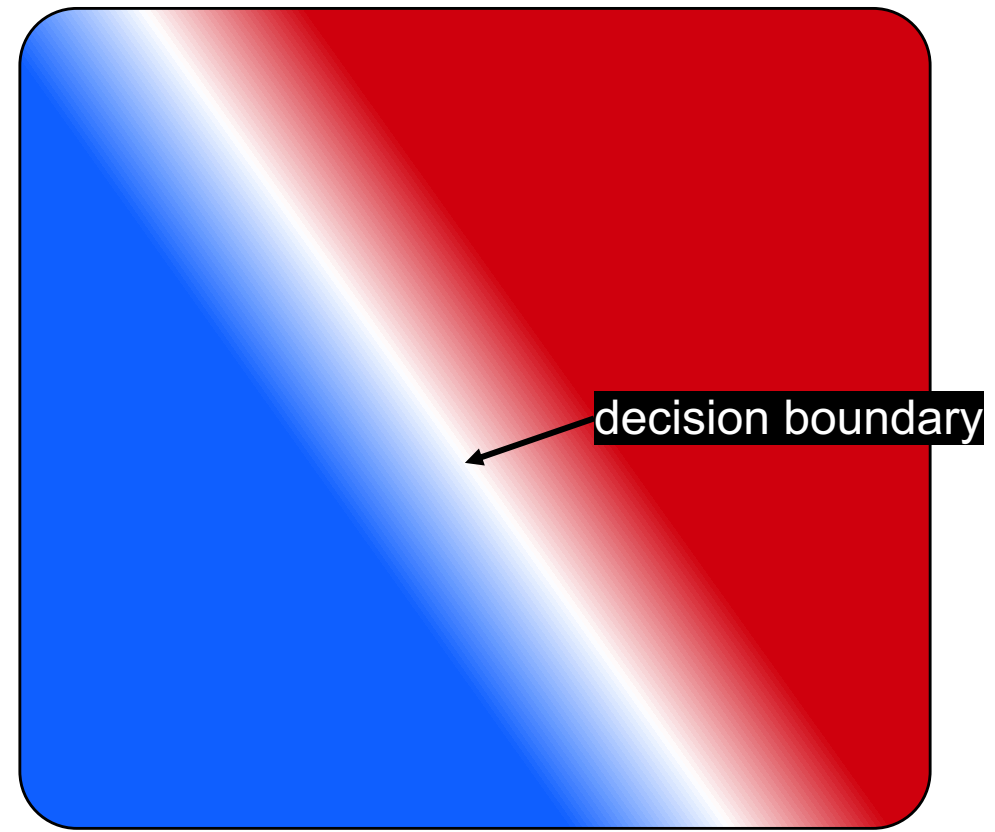


Figure 2: Reliability diagrams of CNNs and Transformers. Transparency of bars represent the ratio of the number of data in each confidence bin. ECE and sECE values are also shown in each case.

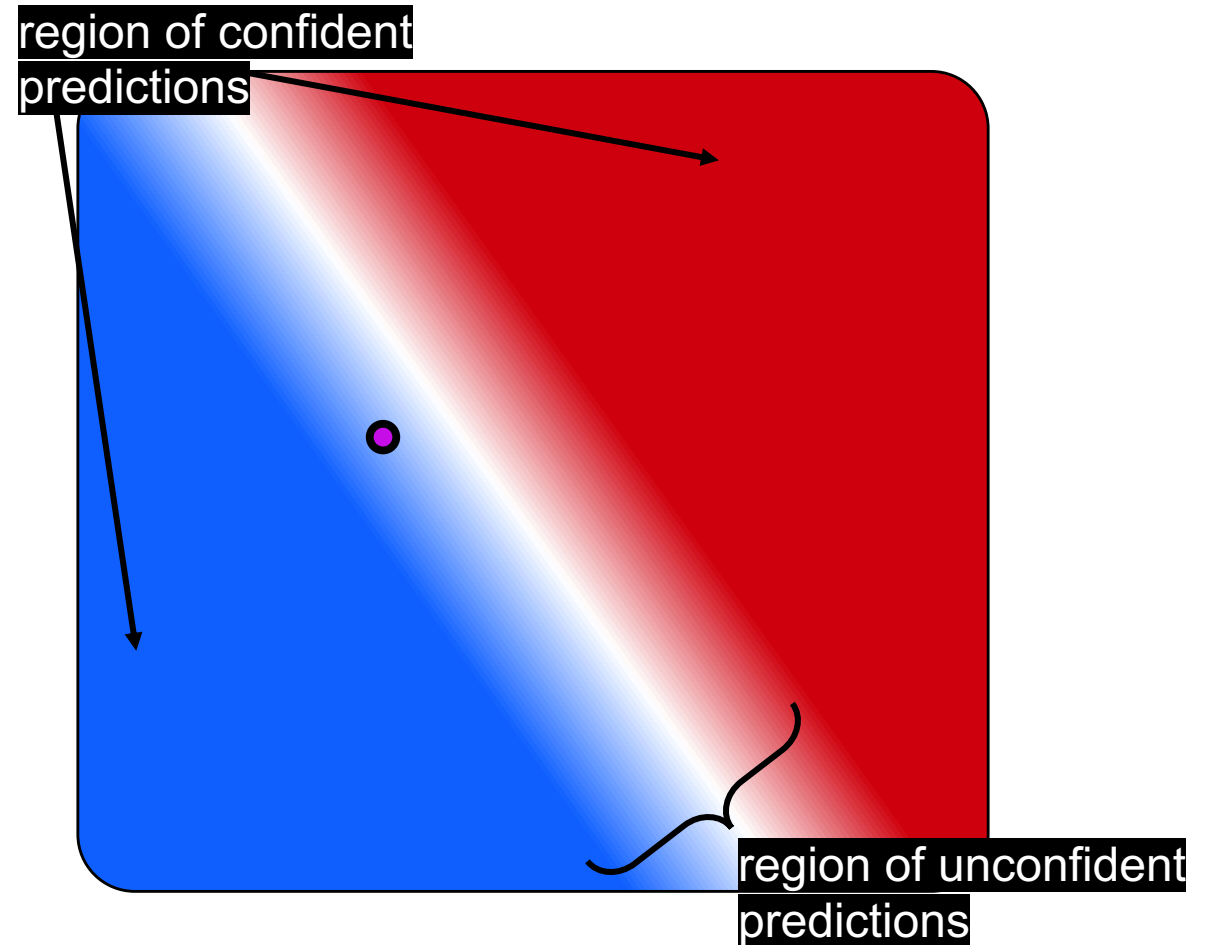
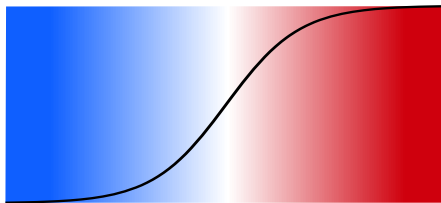
# Intuition on Robustness and Confidence

- embedding space for a two-class problem:

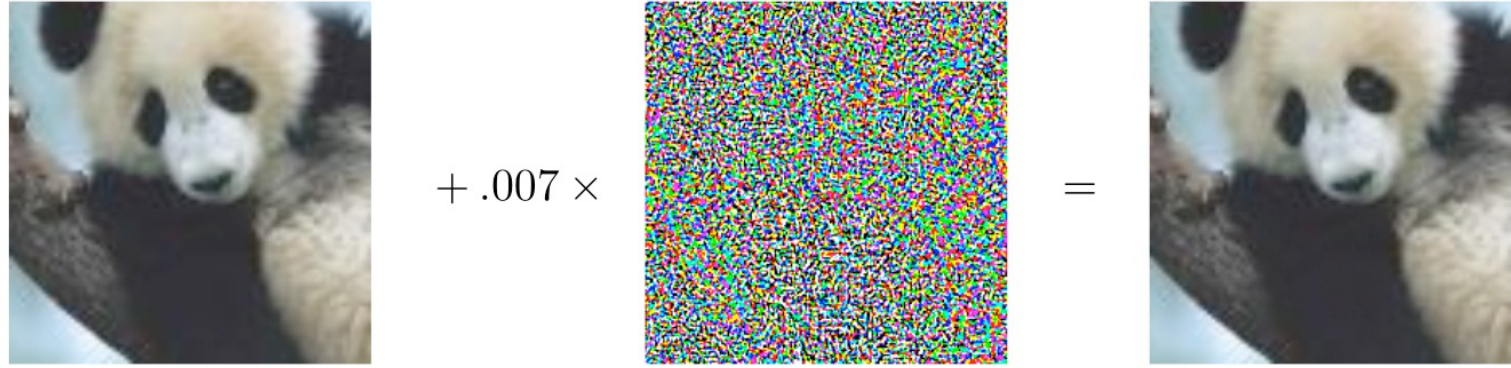


# Intuition on Robustness and Confidence

- embedding space for a two-class problem:
- samples of low confidence are expected to *lie close* to the DB



# Perturbation Direction using FGSM

$$x + \epsilon \times d = x + \epsilon d$$


$x$   $+ .007 \times$   $=$

$x$   $\text{sign}(\nabla_x J(\theta, x, y))$   $x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“panda” “nematode” “gibbon”

57.7% confidence 8.2% confidence 99.3 % confidence

Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet’s classification of the image. Here our  $\epsilon$  of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet’s conversion to real numbers.

# Linear Distance to Decision Boundary on ImageNet Validation Set

solve for each sample:

$$\operatorname{argmin}_{\varepsilon} f(\boldsymbol{x} + \varepsilon \boldsymbol{d}) \neq y$$

normalized to:

$$\|\boldsymbol{d}\| = \sqrt{D}$$



# Linear Distance to Decision Boundary on ImageNet Validation Set

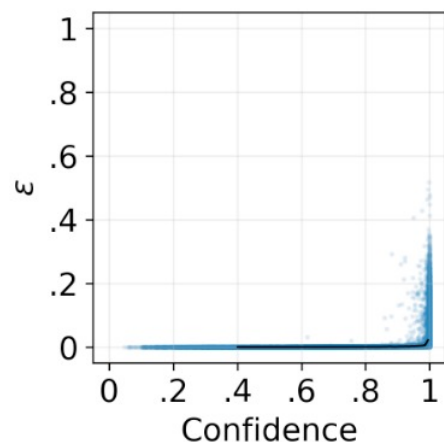
solve for each sample:

$$\operatorname{argmin}_{\epsilon} f(\mathbf{x} + \epsilon \mathbf{d}) \neq y$$

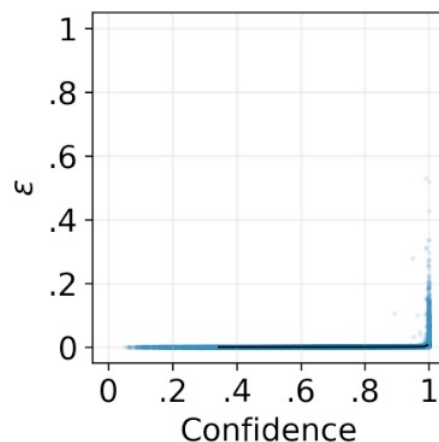
normalized to:

$$\|\mathbf{d}\| = \sqrt{D}$$

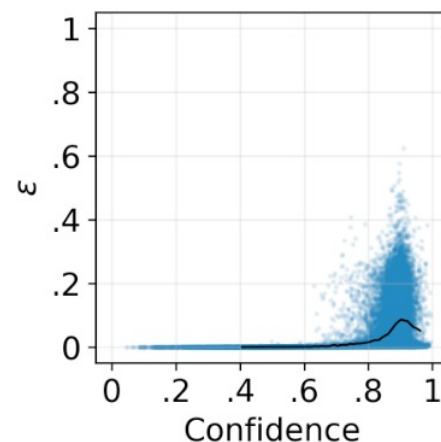
linear  
distance to DB



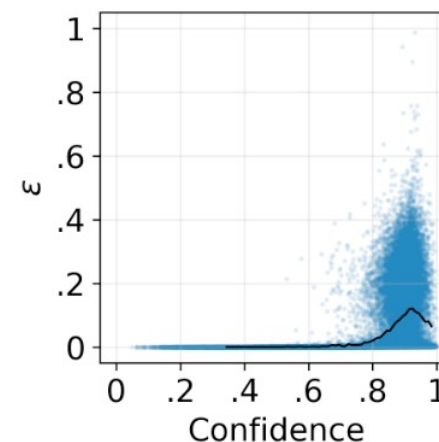
(a) ResNet50



(b) MobileNetV2



(c) ViT-B/16

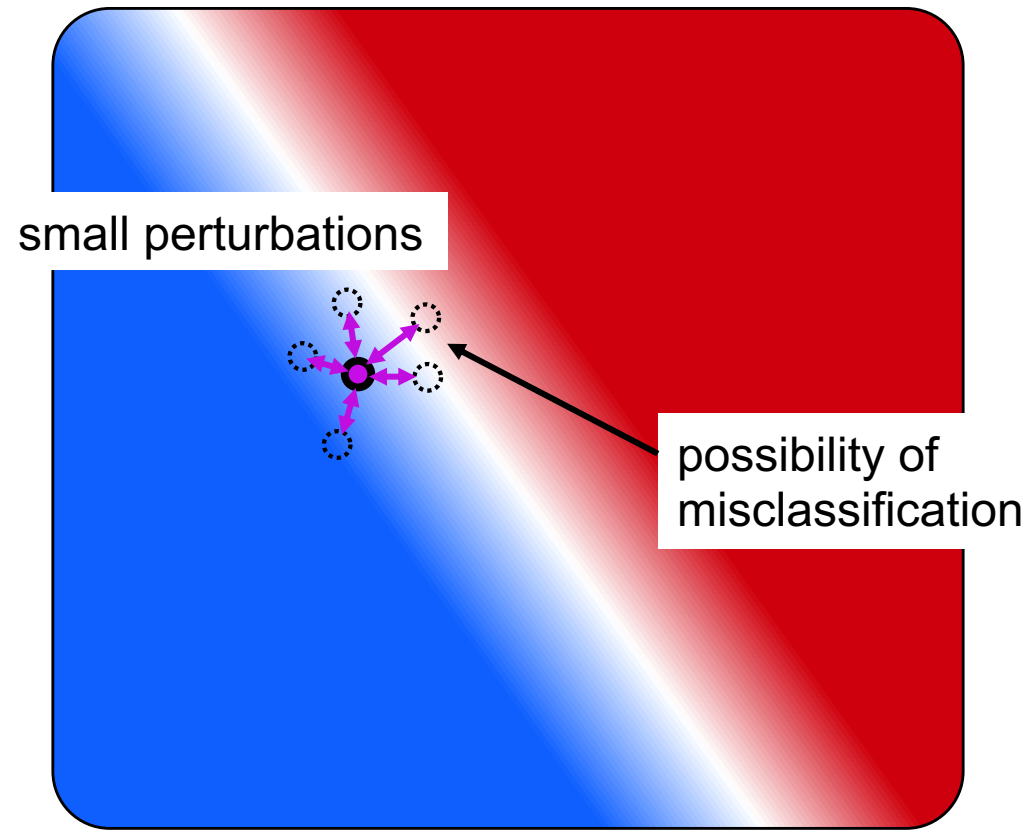


(d) Swin-T

Figure 3: Lengths ( $\epsilon$ ) of the travel to decision boundaries with respect to the confidence for the ImageNet validation data. Black lines represent average values.

# Contradicting Intuition

- embedding space for a two-class problem:
- samples of low confidence are expected to lie close to the DB
- robustness to adversarial perturbations suggests the sample lies far from DB



# Linearity Analysis

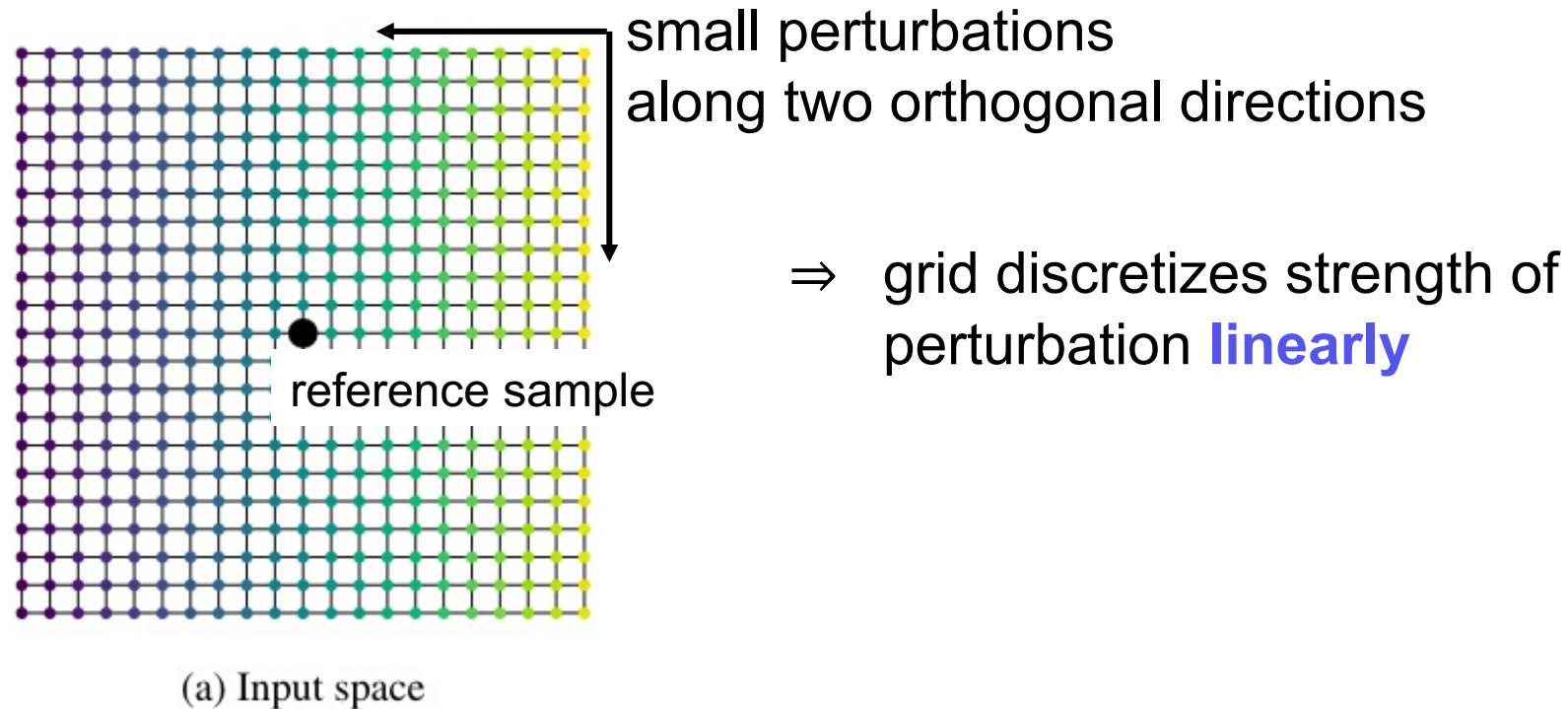
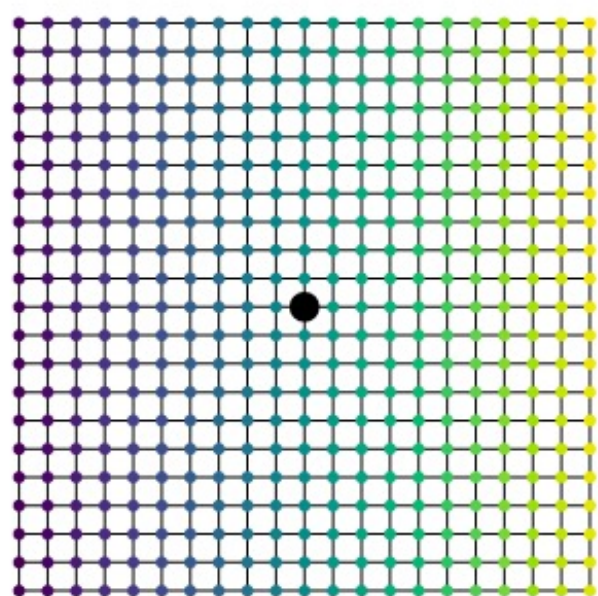
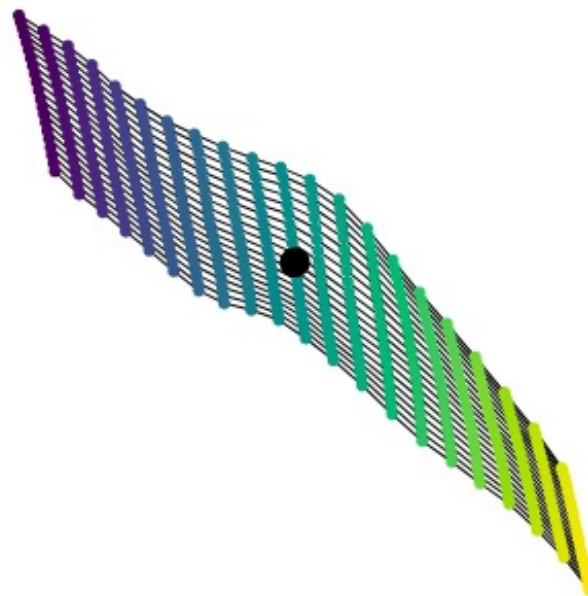


Figure 1: 2D projected movements of (a) the data (black dot) in the input space and corresponding output features in the representation space for (b) ResNet50 and (c) Swin-T.

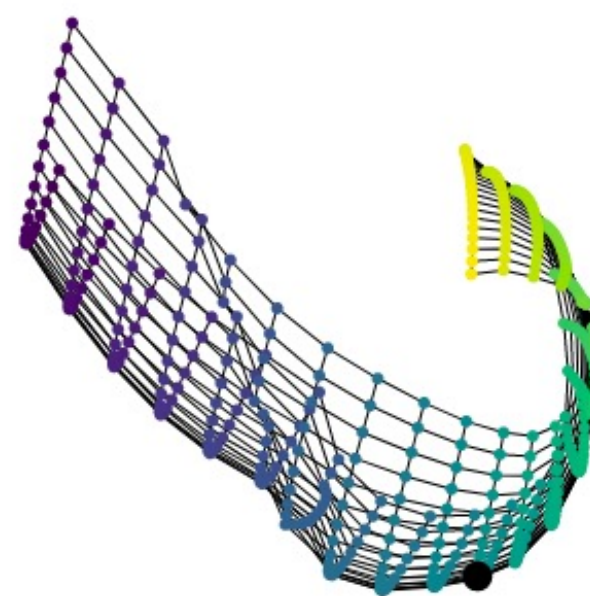
# Linearity Analysis / Local Curvature



(a) Input space



(b) Representation space  
(ResNet50)



(c) Representation space  
(Swin-T)

Figure 1: 2D projected movements of (a) the data (black dot) in the input space and corresponding output features in the representation space for (b) ResNet50 and (c) Swin-T.

# Local Curvature - Visualization

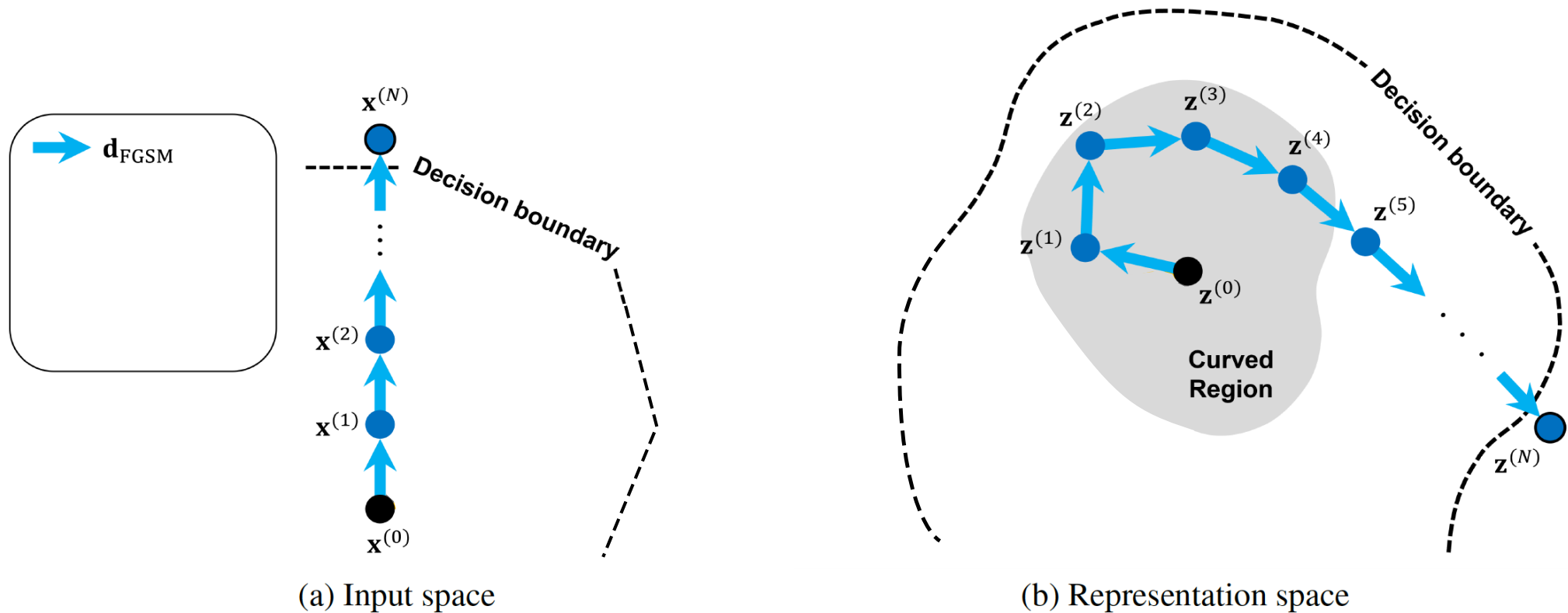
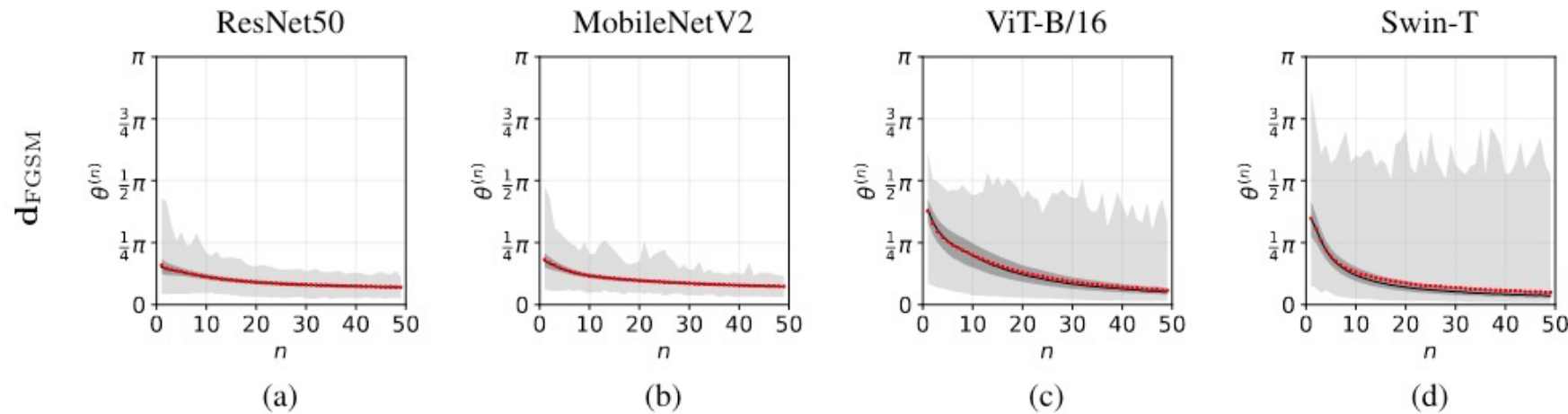


Figure 4: Illustration of the input-output relationship of Transformers in terms of the trajectories in the input space and the representation space.

# Local Curvature - Quantitative



direction-change of successive  
steps in output space:  $\theta^{(n)}$

over

steps along grid:  $n$

light-grey: min – max range

dark-grey: Q1 – Q3 range

red dots: mean values

across 1000 directions of 10 random images



# Mathematical Analysis - Linear Operators

increment  $\mathbf{P}$  to the input  $\mathbf{X}$  converts into the addition of separate responses:

CNNs

$$\text{Conv}(\mathbf{X} + \mathbf{P}) = \text{Conv}(\mathbf{X}) + \text{Conv}(\mathbf{P}). \quad (8)$$

Transformers

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{X}^\top, \quad (9)$$

$$\text{Attn}(\mathbf{X}) = \text{softmax}(\mathbf{A}/\sqrt{D_k})\mathbf{X}\mathbf{W}_v, \quad (10)$$

where  $\mathbf{W}_q$ ,  $\mathbf{W}_k$ , and  $\mathbf{W}_v$  are the projection heads for query, key, and value, respectively, and  $D_k$  is the column dimension of  $\mathbf{W}_k$ . If  $\mathbf{X}$  is moved by  $\mathbf{P}$ ,  $\mathbf{A}$  will change as follows:

$$\mathbf{A}(\mathbf{X} + \mathbf{P}) = (\mathbf{X} + \mathbf{P})\mathbf{W}_q\mathbf{W}_k^\top(\mathbf{X}^\top + \mathbf{P}^\top) \quad (11)$$

$$= \mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{X}^\top + \mathbf{P}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{P}^\top + \mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{P}^\top + \mathbf{P}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{X}^\top \quad (12)$$

$$= \mathbf{A}(\mathbf{X}) + \mathbf{A}(\mathbf{P}) + \underbrace{\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{P}^\top + \mathbf{P}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{X}^\top}_{\text{residual}} \quad (13)$$

# Curvature's Influence on Robustness and Confidence

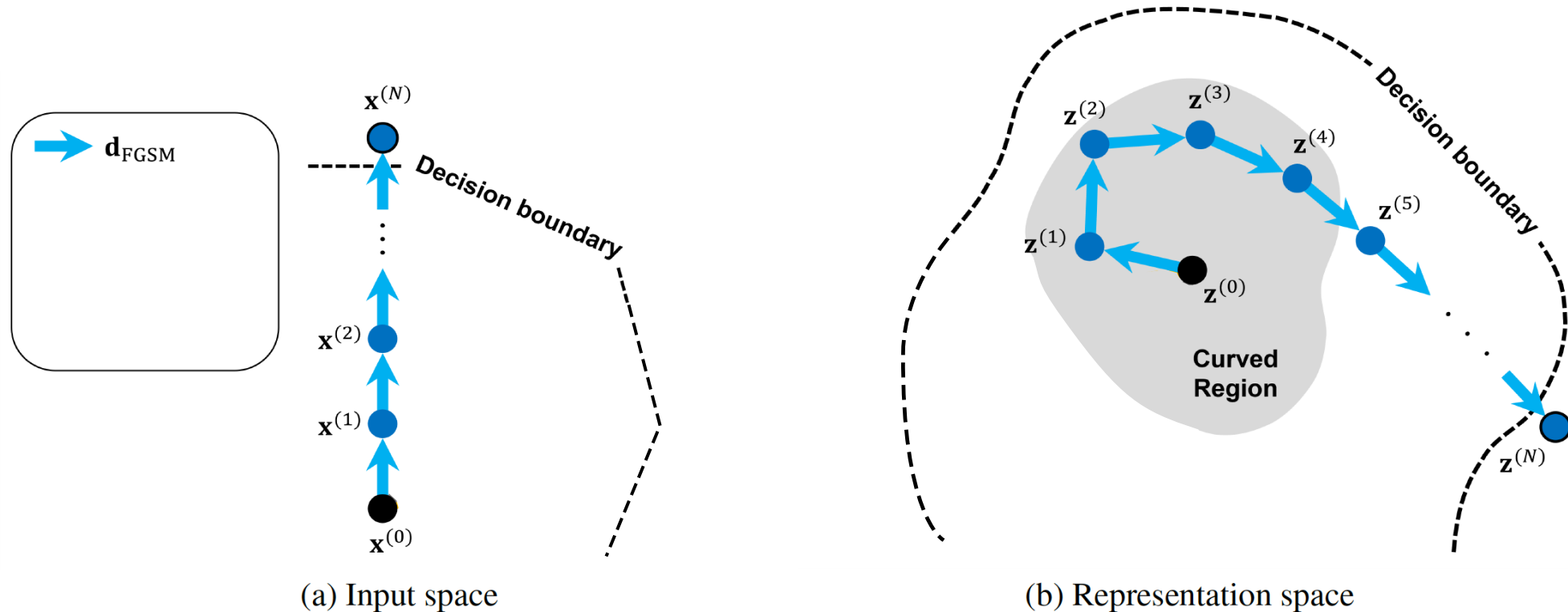


Figure 4: Illustration of the input-output relationship of Transformers in terms of the trajectories in the input space and the representation space.



# Curvature's Influence on Robustness and Confidence

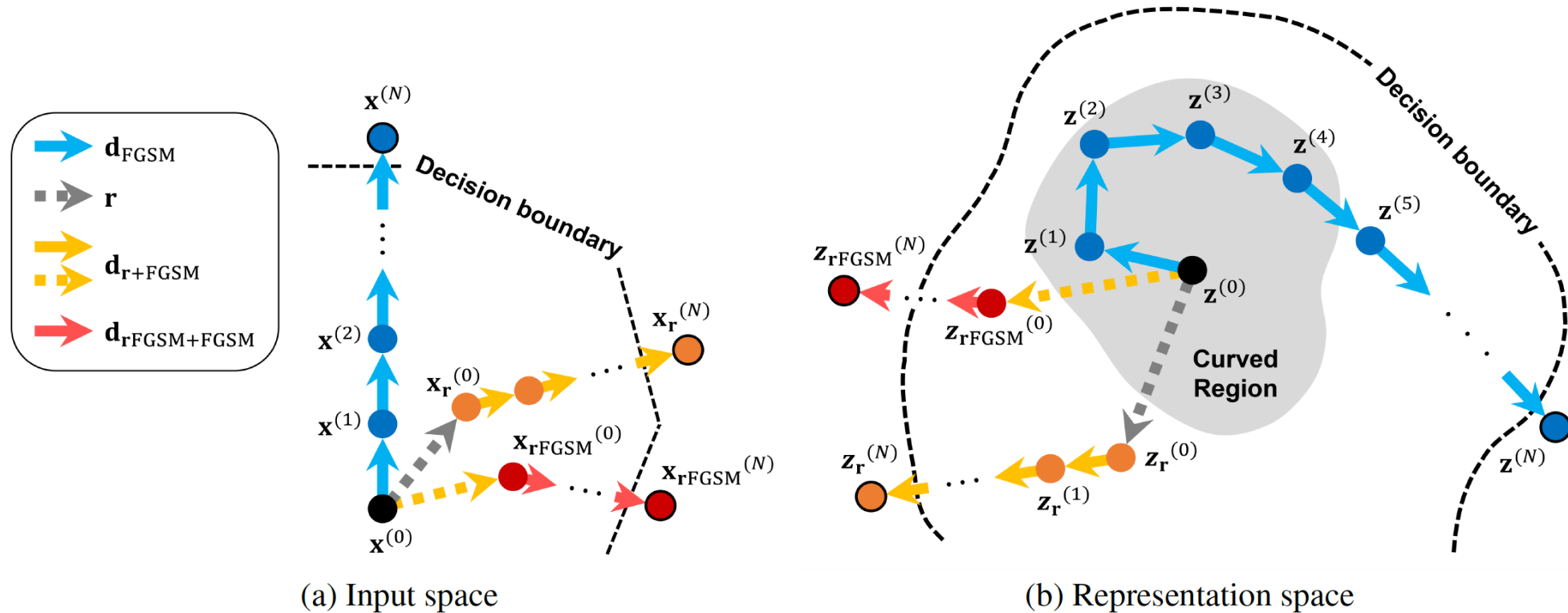


Figure 4: Illustration of the input-output relationship of Transformers in terms of the trajectories in the input space and the representation space.

# Curvature's Influence on Robustness and Confidence

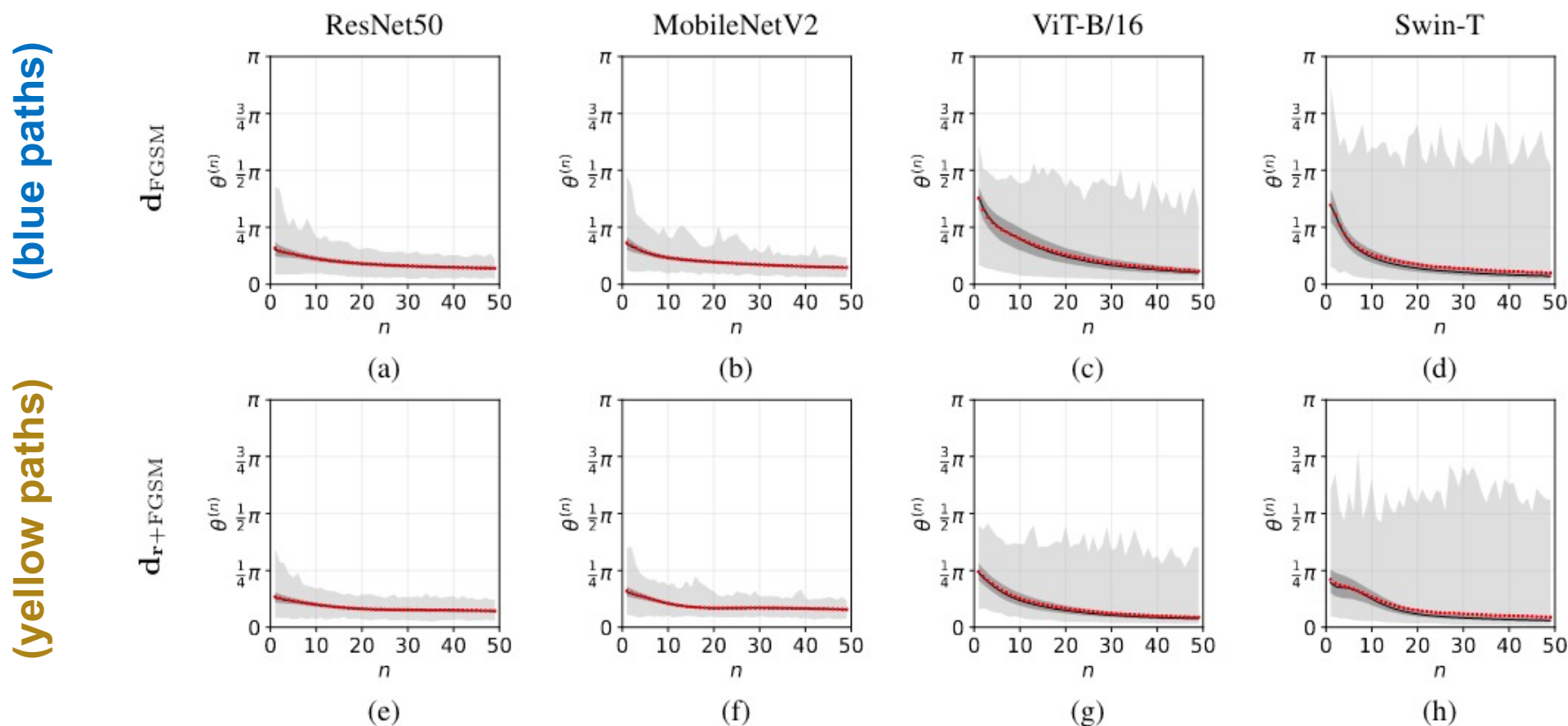


Figure 5: Direction changes of output features with respect to the travel step ( $n$ ). **Light-gray regions:** Range between the minimum and maximum values. **Dark-gray regions:** Range between the first quartile (Q1) and the third quartile (Q3). **Black lines:** Medians (Q2). **Red dots:** Mean values.

# Linearly Perturbed Example Images

- comparison of perturbation strength  $\varepsilon$  needed to reach DB
- perturbation direction  $d$  in all cases suggested by FGSM
  - perturbation  $\varepsilon$  from original output (blue path)
  - perturbation  $\varepsilon$  from randomly perturbed output ( $\varepsilon_r$  fixed) (yellow path)

# Linearly Perturbed Example Images

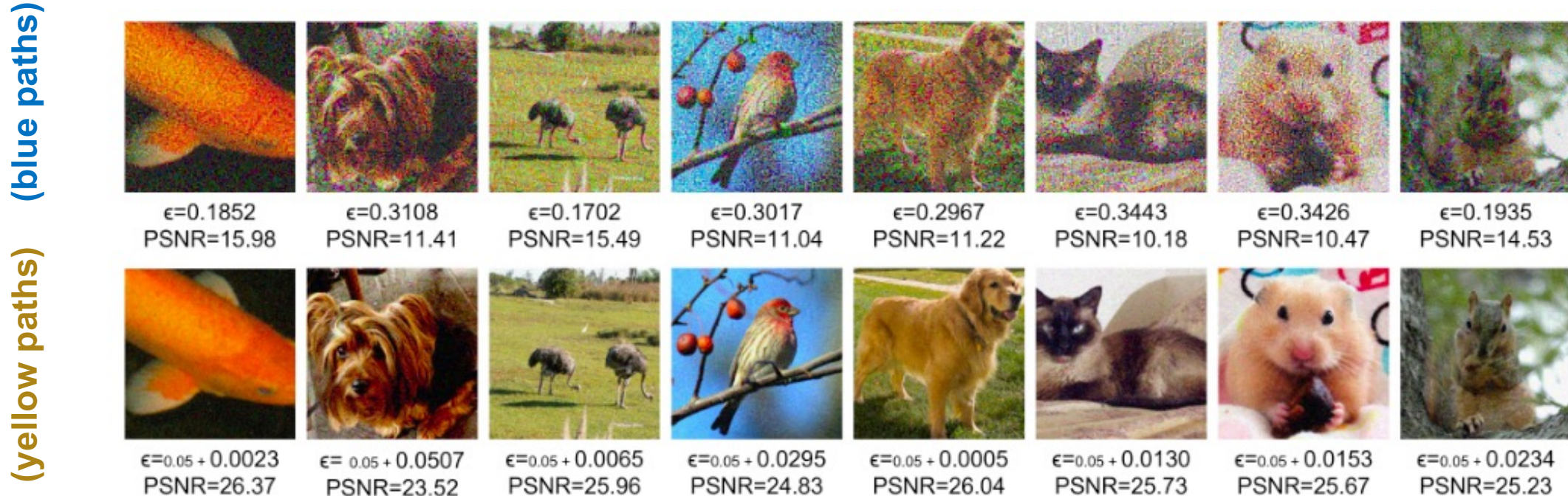


Figure 10: Example images that are perturbed by FGSM so as to reach decision boundaries and become misclassified. The total amount of perturbation ( $\epsilon$ ) and the peak signal-to-noise ratio (PSNR) in dB are also shown. **Top:** Perturbed images. **Bottom:** Images perturbed after random jump ( $\epsilon_r = .05$ ).

# Thanks for Your Attention

---

Questions ?

