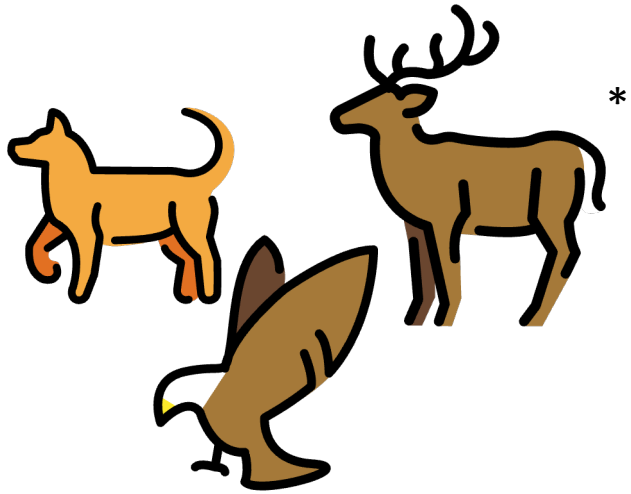
WeiPer: OOD Detection using Weight Perturbations of Class Projections

Maximilian Granz, Manuel Heurich, Tim Landgraf

Freie Universität Berlin

Out of Distribution Detection

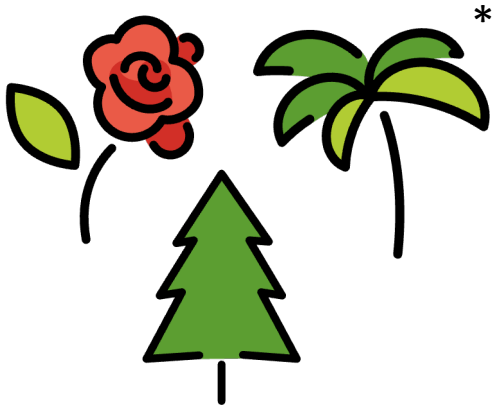
$x_{\text{ID}} \in X_{\text{ID}}$



*

Animals

$x_{\text{OOD}} \in X_{\text{OOD}}$



*

Plants

*designed by OpenMoji

Freie Universität Berlin

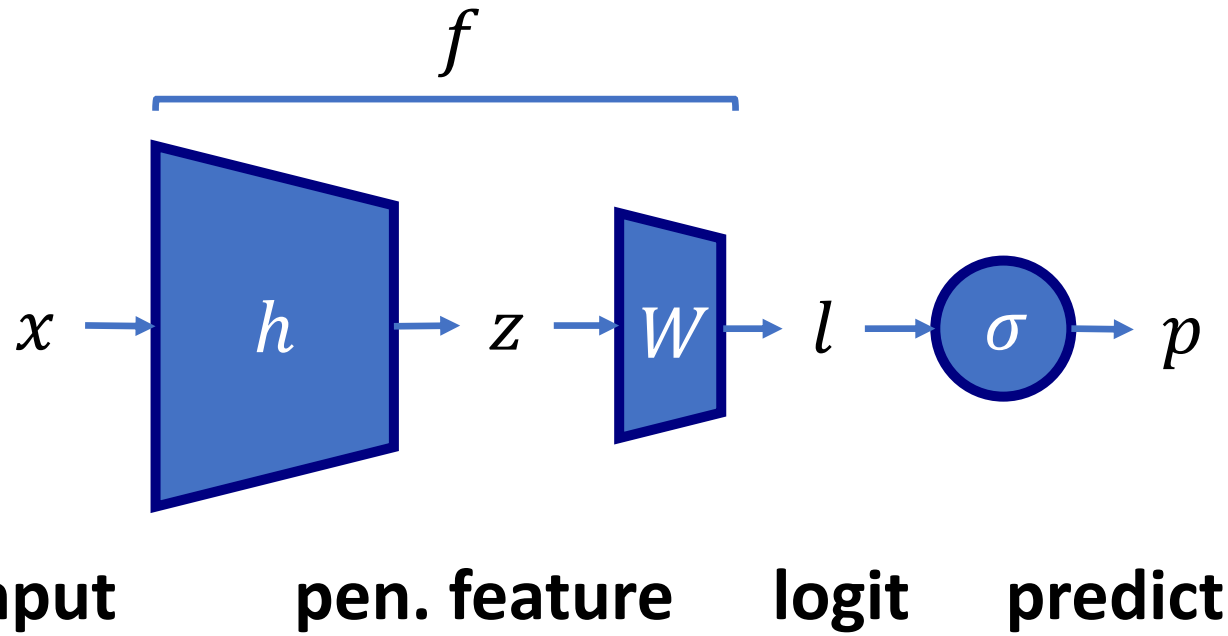
Out of Distribution Detection

Define a **score function** $S(x) \rightarrow \mathbb{R}$

$$O(x) = \begin{cases} \text{ID}, & \text{if } S(x) > \lambda \\ \text{OOD}, & \text{otherwise} \end{cases}$$

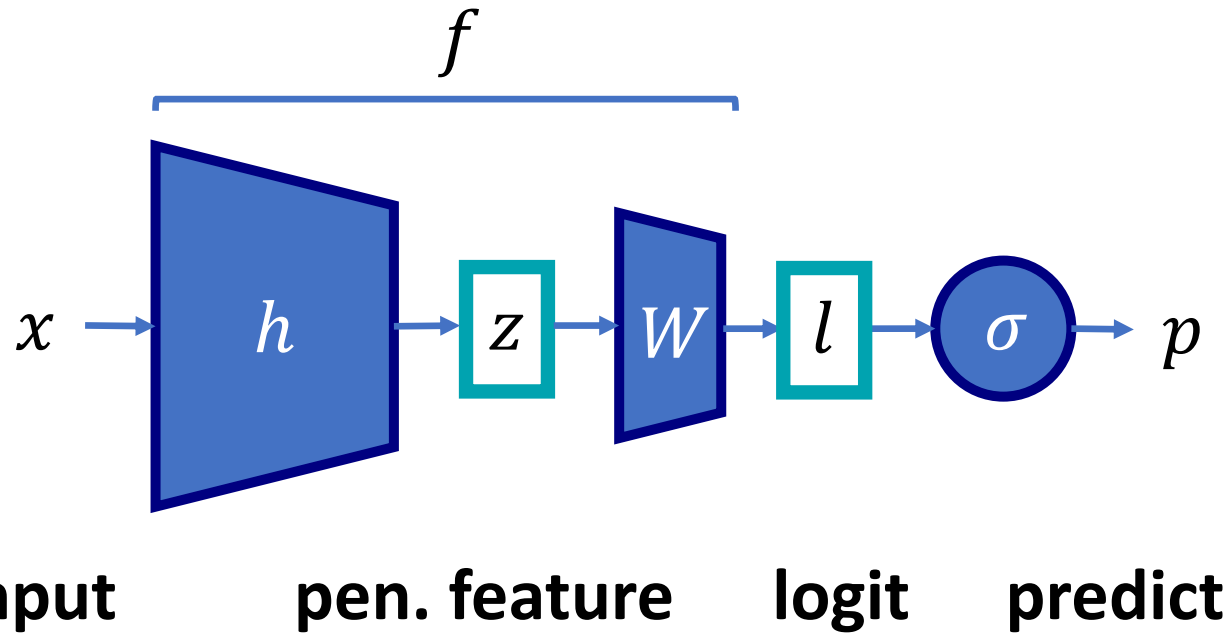
Out of Distribution Detection

Let σ be the softmax function and f be an ANN classifier



Out of Distribution Detection

Let σ be the softmax function and f be an ANN classifier



Maximum Softmax Probability (MSP)

One example for a **score function** is MSP

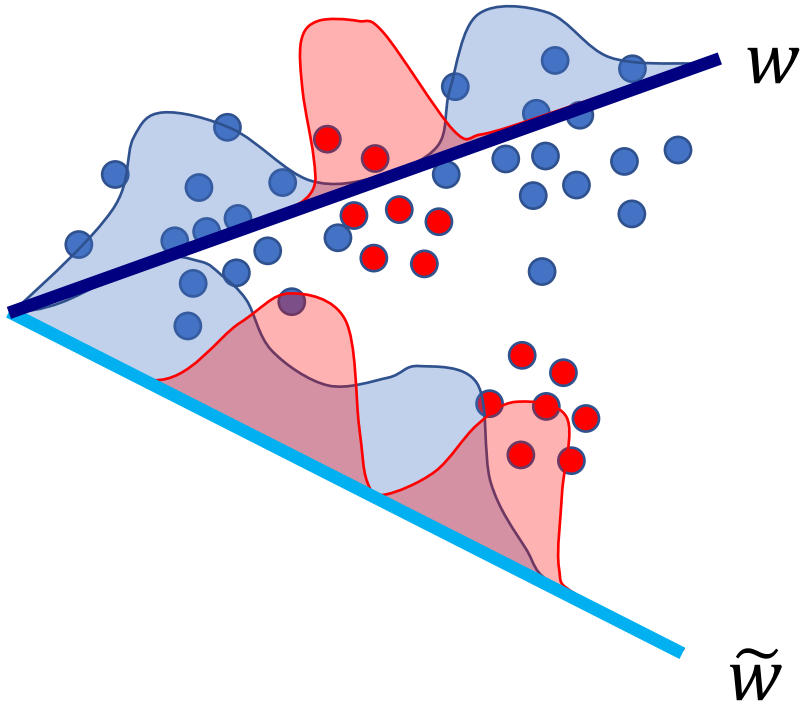
$$\mathrm{MSP}(x) = \max_{i=1,\ldots,C} \sigma(l)_i$$

How can we extract more information from
the penultimate feature space?

$$\mathrm{MSP}(x) = \max_{i=1,\ldots,C} \sigma(Wz)_i$$

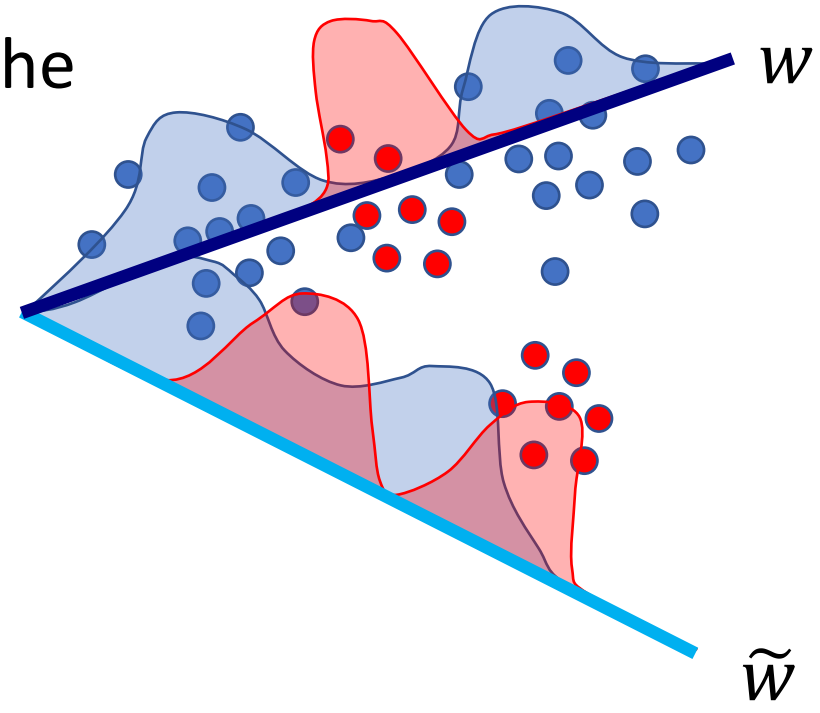
Creating a richer Representation

Cramer-Wold Thm.
If we sample all projections $\tilde{w} \in \mathbb{R}^k$, the penultimate distribution Z is fully determined by it‘s one-dimensional projections $\tilde{w}(Z)$.



Creating a richer Representation

Cramer-Wold Thm.
If we sample all projections $\widetilde{W} \in \mathbb{R}^{C \times k}$, the penultimate distribution Z is fully determined by it's **C-dimensional projections** $\widetilde{W}(Z)$.



$$S(x) \coloneqq \frac{1}{r}\sum_{j=1}^{r} \max_{i} \sigma(\widetilde{W}_j z)$$

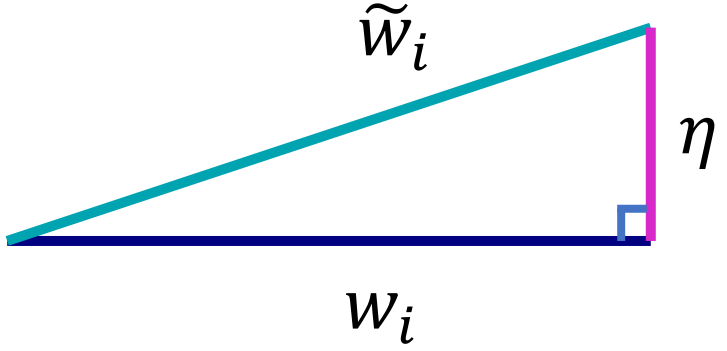
Weight Perturbations (WeiPer)

- Define projections that **correlate** with the class projections W.
- We control the **similarity** with a hyperparameter $\boldsymbol{\delta}$.

$\widetilde{w_i} \sim w_i$ row vector of W

$\widetilde{w_i} = w_i + \delta \cdot \eta,\ \eta \perp w, \eta \in S_K$

$\angle(\widetilde{w_i}, w_i) = \arctan \delta$

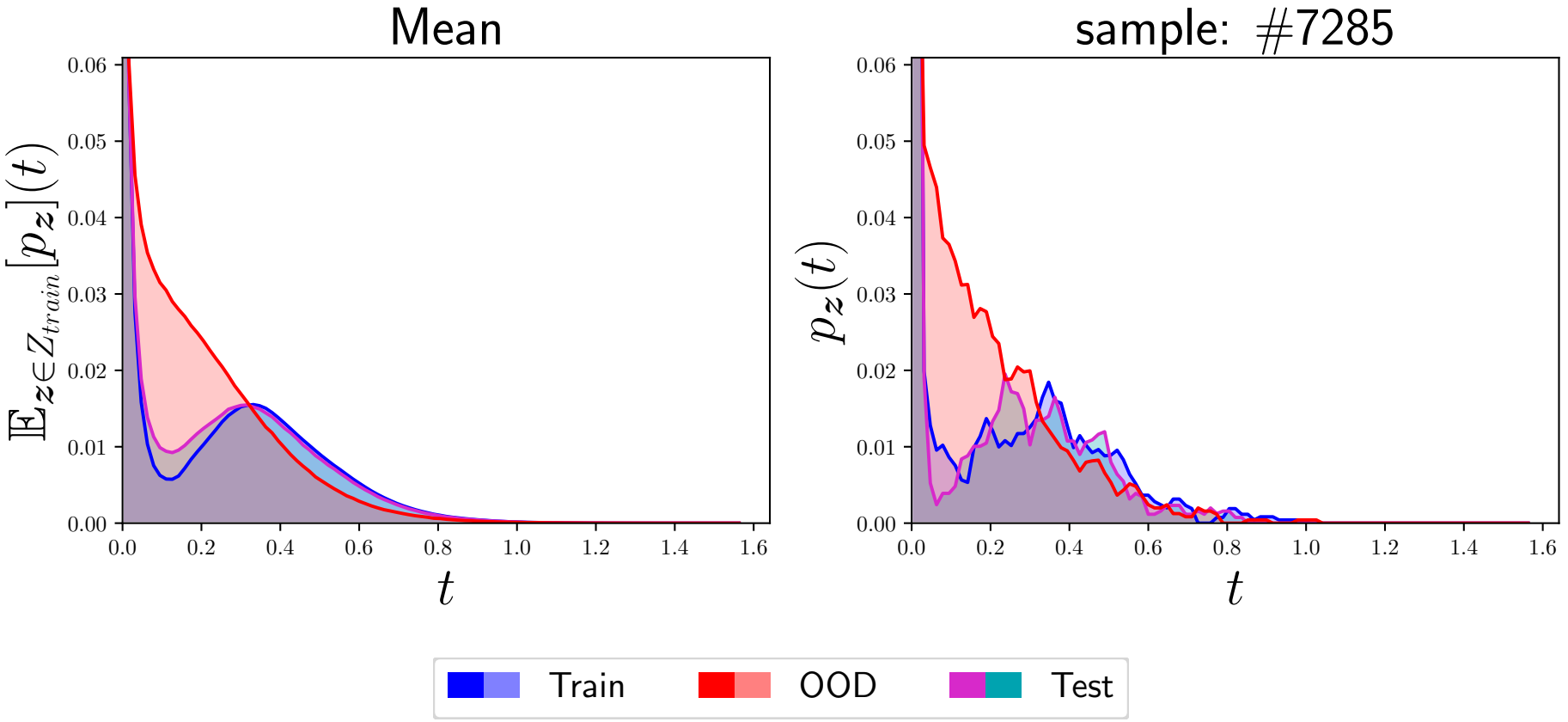


MSP on WeiPer space

Repeat perturbing each row vector r times to get $\widetilde{W}_1, \ldots, \widetilde{W}_r$

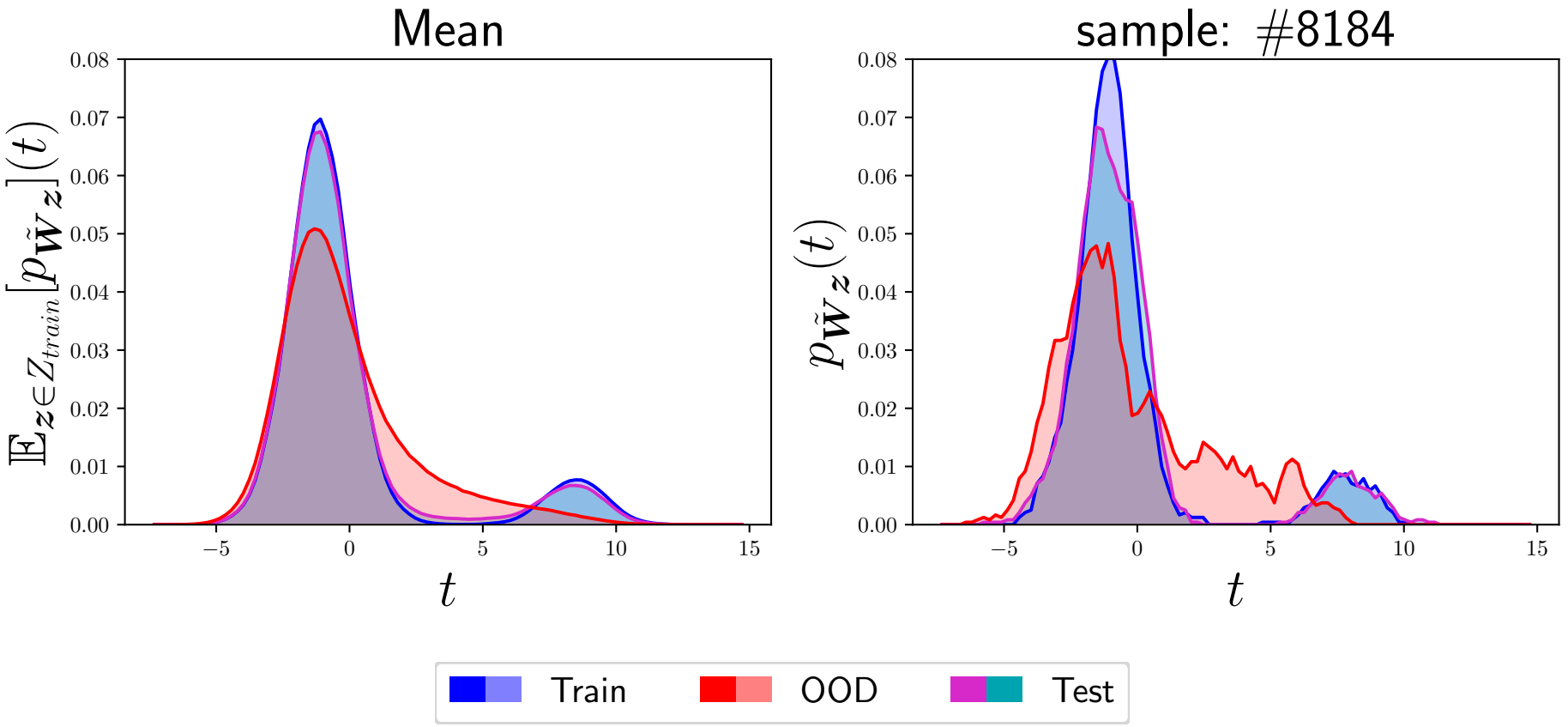
$$\mathrm{MSP}_{\mathrm{WeiPer}}(z) := \frac{1}{r}\sum_{j=1}^{r} \mathrm{MSP}(\widetilde{W}_j z)$$

Penultimate Layer Distribution



Freie Universität Berlin

WeiPer space Distribution



Freie Universität Berlin

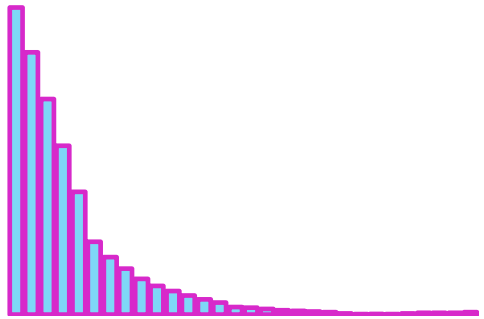
KL Divergence Score Function

1. Calculate the **binned density** p_z of the **penultimate activations** and $p_{\tilde{W}z}$ of the **WeiPer outputs** of z.

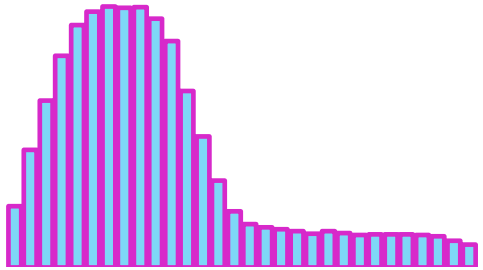
p_z $p_{\tilde{W}z}$

KL Divergence Score Function

2. Apply **normalization and smoothing** T to p_z and $p_{\tilde{w}z}$.



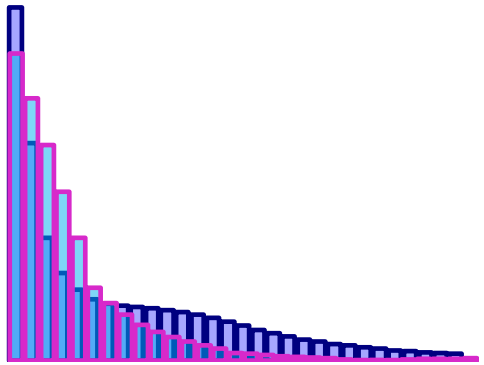
$T(p_z)$



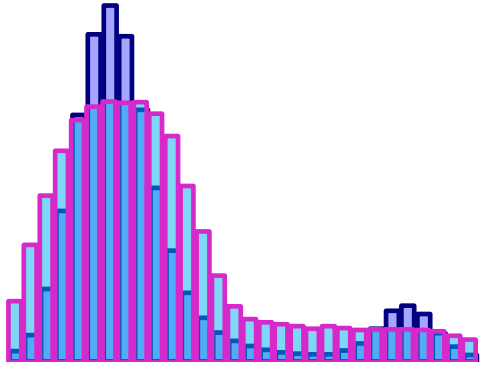
$T(p_{\tilde{w}z})$

KL Divergence Score Function

3. Calculate a **KL-Divergence based distance** D_{KL} from each sample distribution to the mean distribution.



$D_{KL}(T(p_z), \mathbb{E}(p_z))$



$D_{KL}(T(p_{\tilde{W}z}), \mathbb{E}(p_{\tilde{W}z}))$

KL Divergence Score Function

3. Combine all **scores** into one:

$$\textit{WeiPer}\text{+KLD}(x) :=$$

$$D_{KL}\big(T(p_z), \mathbb{E}(p_z)\big) + \lambda_1 D_{KL}\big(T(p_{\tilde{W}z}), \mathbb{E}(p_{\tilde{W}z})\big) - \lambda_2 \text{MSP}_{\text{WeiPer}}(z)$$

Results on OpenOOD

CIFAR10		**CIFAR100**		**ImageNet (ResNet)**		**ImageNet(ViT)**	
Near	Far	Near	Far	Near	Far	Near	Far
<u>90.54</u> (2nd)	93.12 (3rd)	**81.37** **(1st)**	79.01 (13th)	**80.05** **(1st)**	<u>95.54</u> (2nd)	75.0 (6th)	90.32 (8th)

Freie Universität Berlin

Thank You

Code: https://github.com/mgranz/weiper

Paper #96381

Freie Universität Berlin