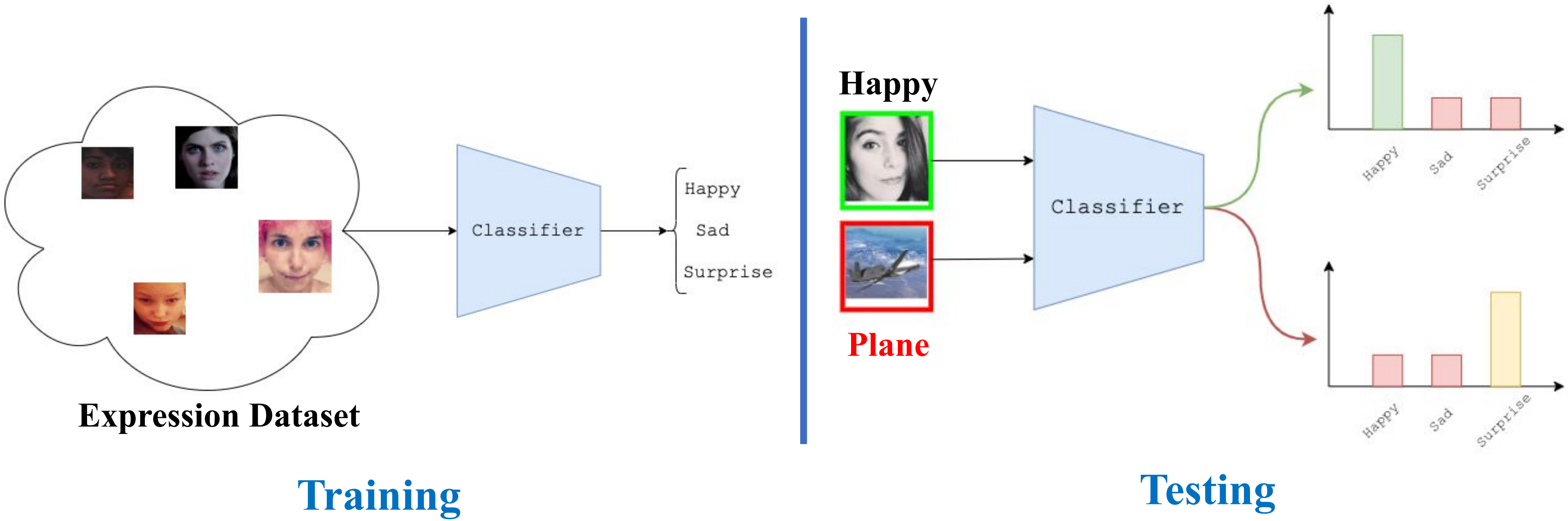
Out-of-distribution Detection using Flow-based Contrastive Learning

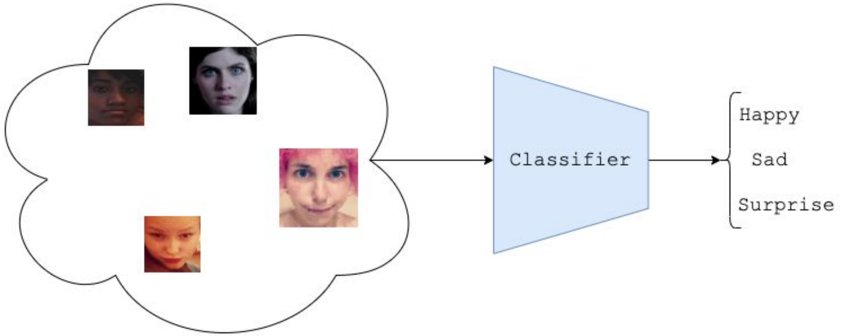
Saandeep Aathreya, Dr. Shaun Canavan

Far-OOD (Semantic Shift)

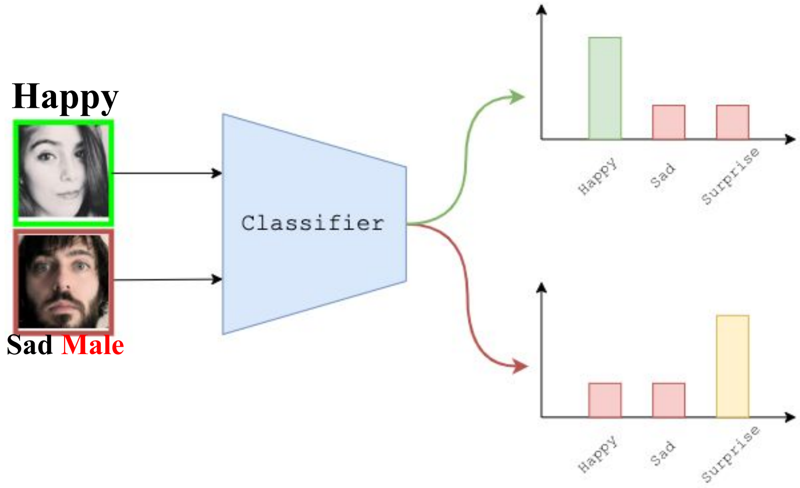


When encountering OOD data, how should the model behave?

Near-OOD (Covariate Shift)



Female **Expression Dataset**

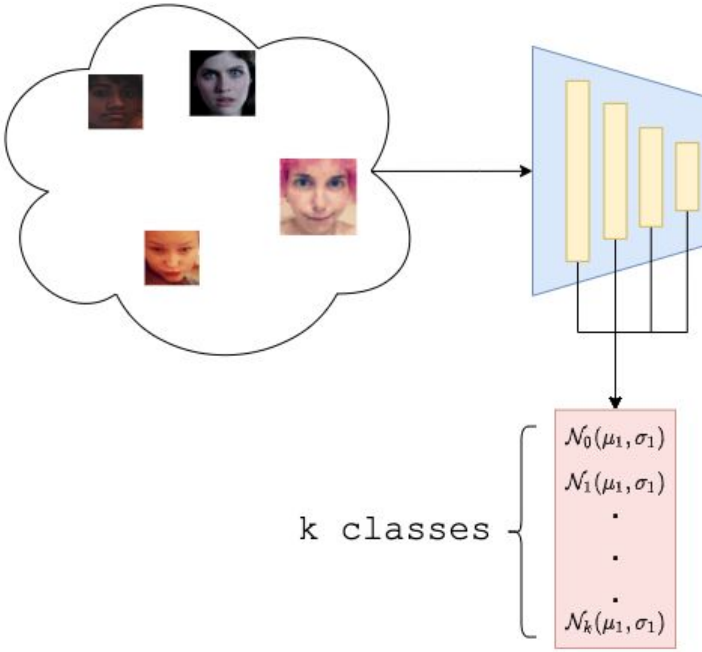


Training

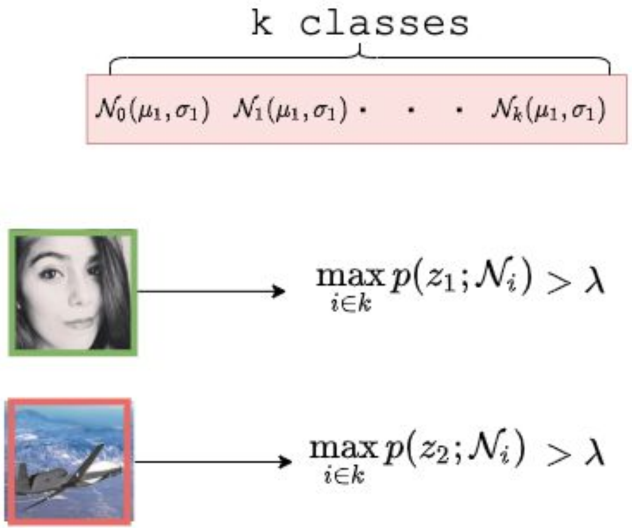
Testing

When encountering OOD data, how should the model behave?

Likelihood Based Approaches (Mahalanobis)



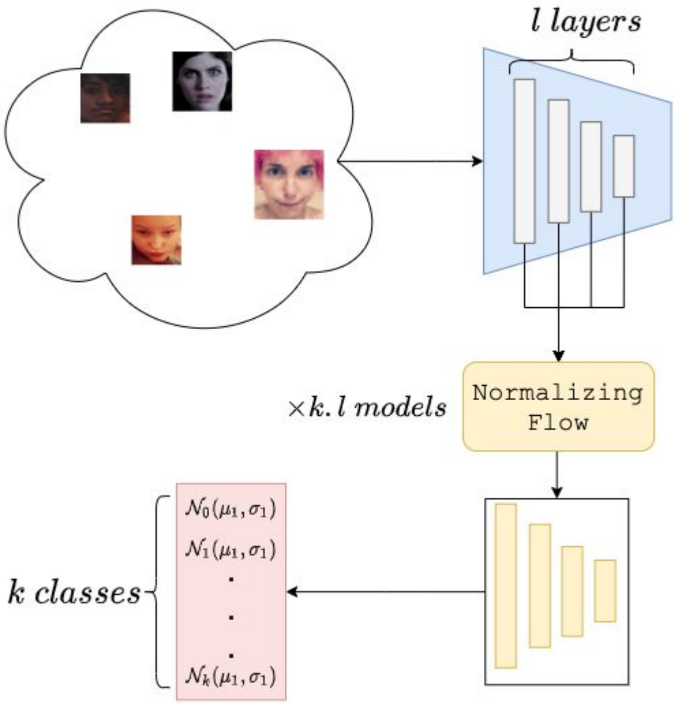
Calculating Distributions



Thresholding

Lee, Kimin, et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." *Advances in neural information processing systems* 31 (2018).

Removing Normality Assumption



Training $k \times l$ models is impractical.

Example: *Resnet18 classifier with* ***four*** *intermediate layers trained on CIFAR100 would require training 400 normalizing flow models*

Zisselman, Ev, and Aviv Tamar. "Deep residual flow for out of distribution detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Normalizing Flows (NF)

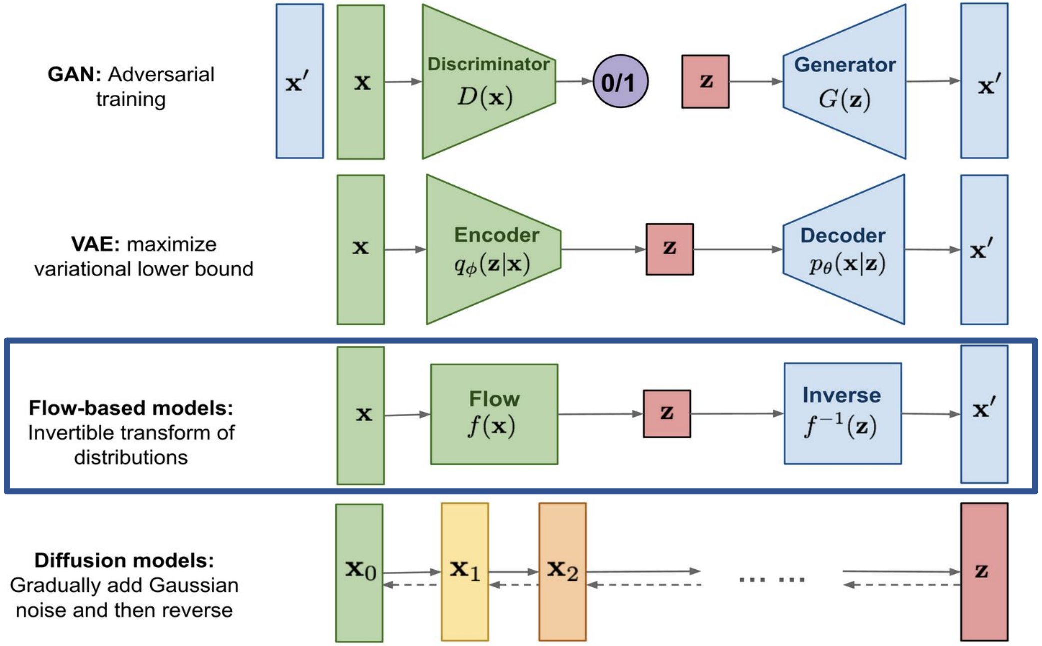


Image taken from [1]

Out-of-distribution Detection using ***Flow-based*** Contrastive Learning

[1] Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

Change of Variables

Inference

$x \sim \hat{p}_X$

$z = f(x)$

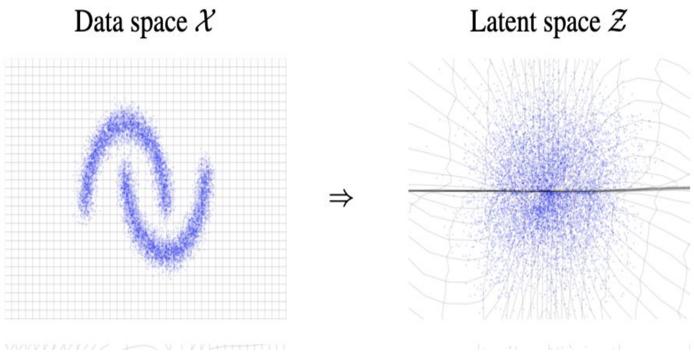


Image taken from [1]

Likelihood of the data point x relates to probability density of a simpler distribution (Z)

Data Distribution

$$p_X(x) = p_Z(z) \cdot \left| \det \frac{\partial f(x)}{\partial x} \right|$$

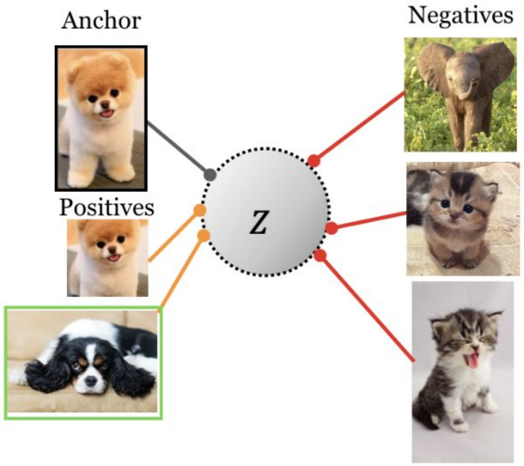
Latent Distribution (Gaussian)

Volume change in density

$$L_{flow} = -\log p_X(x)$$

[1] Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp." *arXiv preprint arXiv:1605.08803* (2016).

Supervised Contrastive Learning (SCL)



Supervised Contrastive

Image taken from [1]

Out-of-distribution Detection using Flow-based ***Contrastive Learning***

Similarity score between Anchor and Positive

$$\mathcal{L}_i = \sum_{p \in P(i)} \log \frac{S(z_i, z_p)}{\sum_{a \in A(i)} S(z_i, z_a)}$$

Sum over positives, p

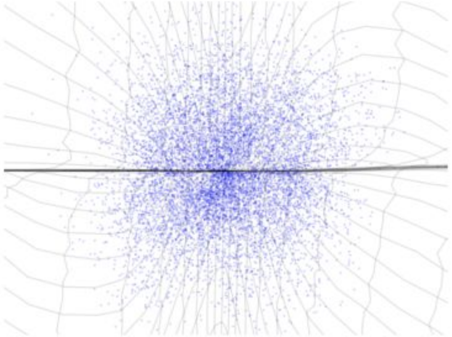
Similarity score between Anchor and All images in batch

$$S(z_i, z_j) = \exp(z_i . z_j / \tau)$$

$$L_{con} = -\Sigma_{i=I} \frac{1}{|P(i)|} L_i$$

[1] Khosla, Prannay, et al. "Supervised contrastive learning." *Advances in neural information processing systems* 33 (2020): 18661-18673.

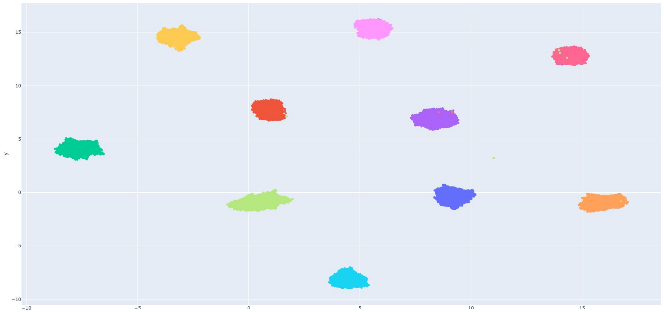
Combining NF and SCL



$$\mathcal{L}_{flow} = -p_X(x) = p_Z\left(\boxed{z_i^{flow}}\right) \cdot \left|\det \frac{\partial f(x)}{\partial x}\right|$$

Flow feature

Normalizing Flow



Similarity score between Anchor and Positive

$$\mathcal{L}_{con} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{S(\boxed{z_i^{flow}, z_p^{flow}})}{\sum_{a \in A(i)} S(\boxed{z_i^{flow}, z_a^{flow}}))}$$

Similarity score between Anchor and All images in batch

SCL

Learning Distributions Contrastively

Bhattacharyya Coefficient: Quantifies overlap between two distributions. Greater the overlap, higher the coefficient value.

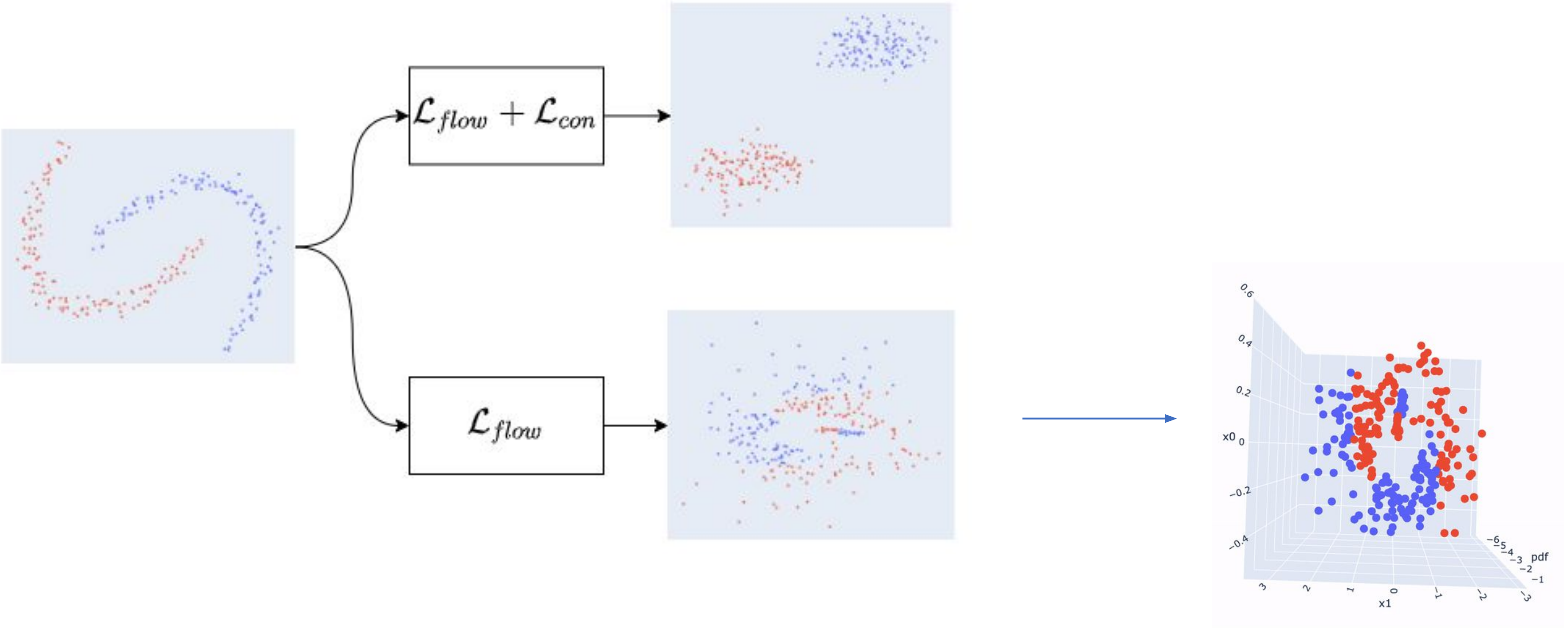
$$S(z_i, z_j) = \exp(z_i \cdot z_j / \tau)$$

$$\downarrow$$

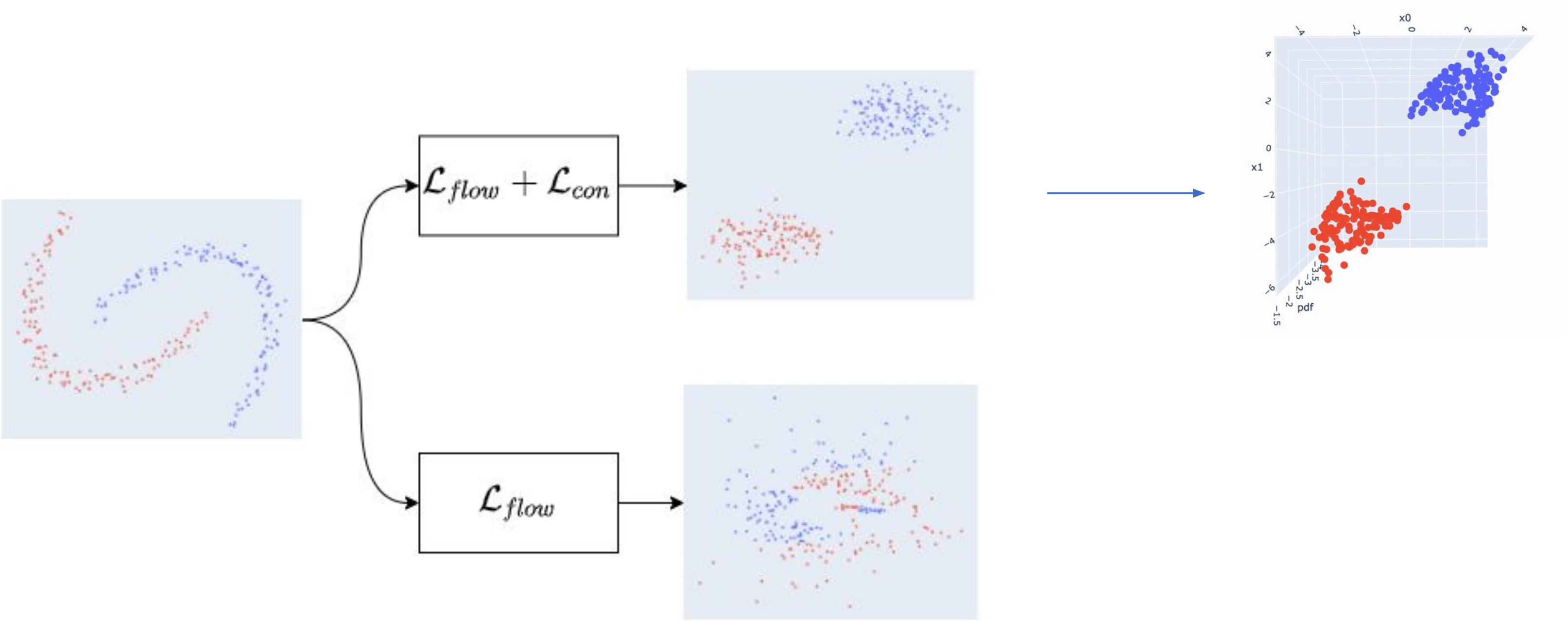
$$S_{flow}(z_i, z_j, \mathcal{N}_i) = \exp\Big(\underbrace{p_Z(z_i|\mathcal{N}_i) \cdot p_Z(z_j|\mathcal{N}_i)}_{\text{Computed from } \mathcal{L}_{flow}} \Big)^{\tau}$$

$$\boxed{\mathcal{L}_{con} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{S_{flow}(z_i, z_p, \mathcal{N}_i)}{\sum_{a \in A(i)} S_{flow}(z_i, z_a, \mathcal{N}_i)}}$$

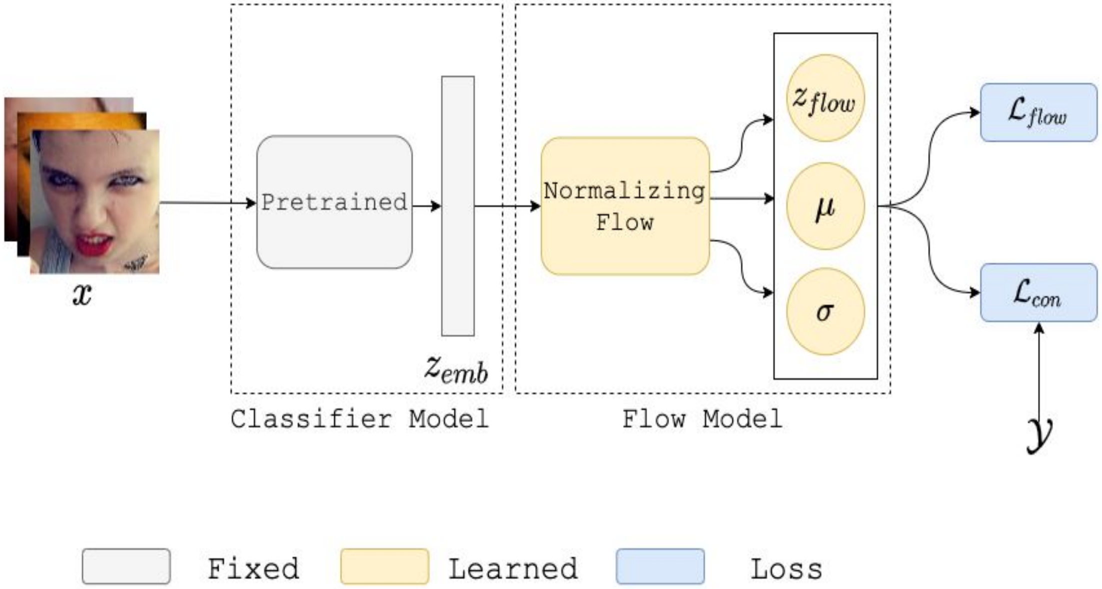
Intuition – With Flow Loss



Intuition – With Flow + Contrastive Loss



Pipeline



Training

$$\boxed{\begin{aligned} \mu_c &= \frac{1}{N_k}\sum_{i=1}^{k}\mu^i \\ \sigma_c &= \frac{1}{N_k}\sum_{i=1}^{k}\sigma^i \end{aligned}}$$

$$k \left\{ \begin{array}{c} \mathcal{N}_0(\mu_1, \sigma_1) \\ \mathcal{N}_1(\mu_1, \sigma_1) \\ \vdots \\ \mathcal{N}_k(\mu_1, \sigma_1) \end{array} \right.$$

$$\boxed{M_i = \max_{i \in \{1,\ldots,k\}} p_Z(z|\mathcal{N}_{y=i})}$$

Score Computation

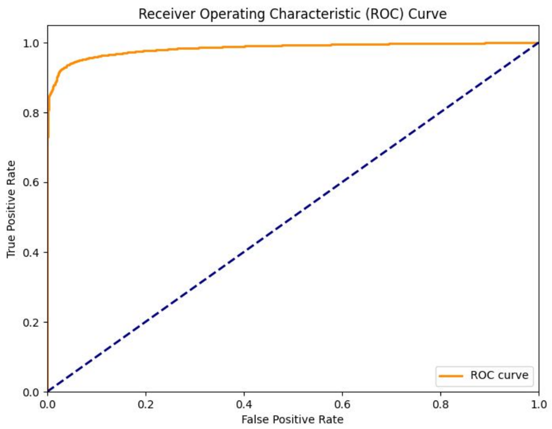
Metrics

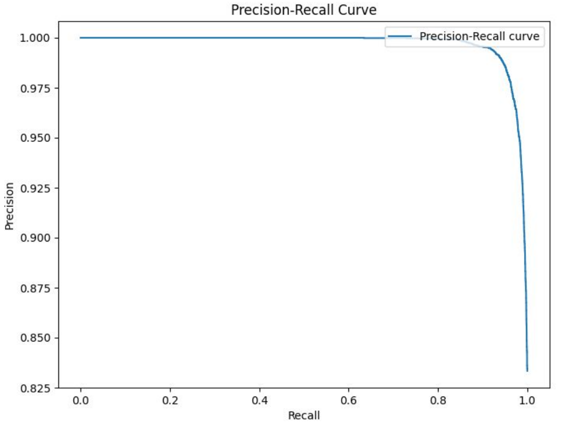
AUROC: A higher AUROC value indicates better performance in correctly classifying OOD samples across all thresholds. *Higher is better*

AUPR-Success (Area Under the Precision-Recall Curve for Success): A higher AUPR-Success indicates that the model can confidently identify OOD samples without misclassifying many ID samples as OOD. *Higher is better*

AUPR-Error: Indicates model's ability to avoid false positives (high precision) across all levels of false negative rates. *Higher is better*

FPR@TPR=95%: It represents the false positive rate when the model correctly identifies 95% of the OOD samples. *Lower is better*





Experiments

No.	Name	ID Dataset	OOD Datasets	Model	Result
1	Semantic Shift (Vision)	CIFAR10	lsun-r, lsun-c, isun, svhn, textures, places365	Resnet18 and WideResnet	Averaged
		CIFAR100			
2	Semantic Shift (Expression)	RAF-DB	lsun-r, lsun-c, isun, svhn, textures, places365		
		AffectNet			
3	Covariate Shift (Expression)	RAF-DB	AffectNet		Individual
		AffectNet	Raf-DB		

- *No additional training of pretrained classifier.*
- *No external dataset utilized as OOD.*
- *Compared with five other methods*

Semantic Shift (Common vision dataset)

D_{in} (model)	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
CIFAR-10 (ResNet)	MSP	90.90	97.94	64.11	53.99
	ODIN*	88.33	96.67	71.49	38.35
	Mahalanobis	90.09	97.04	76.92	28.07
	Energy	91.91	97.94	72.85	36.80
	ReAct	91.78	97.88	72.77	36.80
	Ours	**97.19**	**99.43**	**85.66**	**16.26**
CIFAR-10 (WideResNet)	MSP	91.79	98.27	64.09	55.45
	ODIN*	95.01	98.68	84.39	21.09
	Mahalanobis*	92.03	98.09	75.44	32.73
	Energy	**95.30**	97.87	81.89	22.5
	ReAct*	51.92	85.46	17.53	97.12
	Ours	95.19	**98.78**	**86.11**	**20.30**
CIFAR-100 (ResNet)	MSP*	79.29	95.04	40.34	76.58
	ODIN	83.28	95.96	48.74	67.96
	Mahalanobis	73.46	93.00	35.90	79.46
	Energy	82.07	95.71	43.92	74.45
	ReAct	84.22	96.27	49.08	67.78
	Ours	**88.22**	**96.85**	**67.89**	**41.85**
CIFAR-100 (WideResNet)	MSP	65.31	90.38	26.21	88.45
	ODIN	79.43	94.60	43.98	73.19
	Mahalanobis	73.99	92.58	43.80	68.45
	Energy	77.11	93.95	39.07	78.03
	ReAct	80.74	95.24	48.04	67.47
	Ours	**84.54**	**95.89**	**54.84**	**60.05**

ID Dataset: CIFAR-10, CIFAR-100
OOD Dataset: lsun-r, lsun-c, isun, svhn, textures, places365

*The results marked with * are taken from the previous work[1].*

[1] Hornauer, Julia, and Vasileios Belagiannis. "Heatmap-based Out-of-Distribution Detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

Semantic Shift (Expression dataset)

D_{in} (model)	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
RAF (ResNet)	MSP	66.72	89.15	26.16	87.8
	ODIN	65.62	88.40	28.06	82.84
	Mahalanobis	**97.92**	99.11	**88.88**	**8.97**
	Energy	75.32	93.57	37.2	79.49
	ReAct	78.89	94.7	40.85	76.12
	Ours	96.35	**99.57**	87.50	13.05
RAF (WideResNet)	MSP	66.47	89.46	25.75	89.28
	ODIN	66.21	88.92	28.92	85.01
	Mahalanobis	**98.17**	**99.62**	90.86	**9.40**
	Energy	75.32	93.57	37.2	79.49
	ReAct	71.17	93.06	30.80	84.99
	Ours	97.98	99.57	**91.97**	10.08
Aff (ResNet)	MSP	67.61	91.11	26.15	89.63
	ODIN	71.19	91.71	32.26	82.58
	Mahalanobis	**98.27**	**99.64**	**91.57**	**8.48**
	Energy	67.69	91.12	27.68	87.50
	ReAct	73.06	93.32	31.17	84.25
	Ours	76.66	90.91	69.95	30.70
Aff (WideResnet)	MSP	66.88	90.93	26.22	89.5
	ODIN	63.58	87.93	24.79	89.6
	Mahalanobis	98.33	99.76	94.13	5.2
	Energy	55.83	86.02	20.04	94.07
	ReAct	65.33	91.33	23.47	93.03
	Ours	**99.71**	**99.95**	**97.95**	**1.32**

ID Dataset: RAF-DB, AffectNet
OOD Dataset: lsun-r, lsun-c, isun, svhn, textures, places365

Only Mahalanobis and Our method show competitive performance on expression dataset.

Covariate Shifts

D_{in} (model)	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
RAF (ResNet)	Mahalanobis	**97.92**	99.11	**88.88**	**8.97**
	Ours	96.35	**99.57**	87.50	13.05
RAF (WideResNet)	Mahalanobis	**98.17**	**99.62**	90.86	**9.40**
	Ours	97.98	99.57	**91.97**	10.08
Aff (ResNet)	Mahalanobis	**98.27**	**99.64**	**91.57**	**8.48**
	Ours	76.66	90.91	69.95	30.70
Aff (WideResnet)	Mahalanobis	98.33	99.76	94.13	5.2
	Ours	**99.71**	**99.95**	**97.95**	**1.32**

Semantic Shift

OOD Dataset: lsun-r, lsun-c, isun, svhn, textures, places365

D_{in} (model)	Method	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
RAF (ResNet)	Mahalanobis	57.87	87.95	18.8	95.59
	Ours	**87.21**	**97.22**	**50.43**	**64.37**
RAF (WideResNet)	Mahalanobis	55.33	86.7	18.03	96.08
	Ours	**63.51**	**89.89**	**25.22**	**89.75**
Aff (ResNet)	Mahalanobis	75.53	93.51	35.16	79.34
	Ours	**91.4**	**98.20**	**63.40**	**48.93**
Aff (WideResnet)	Mahalanobis	77.31	94.20	36.75	77.84
	Ours	**85.23**	**96.79**	**45.42**	**72.59**

Covariate Shift

OOD Dataset: RAF-DB or AffectNet

ID Dataset: RAF-DB, AffectNet

Likelihood Plots

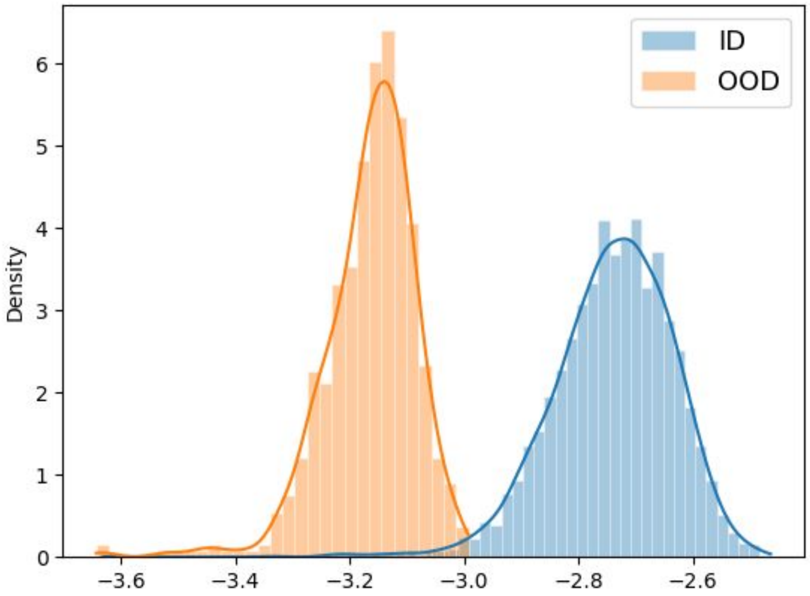
$$\mu_c = \frac{1}{N_k}\sum_{i=1}^{k}\mu^i$$

$$\sigma_c = \frac{1}{N_k}\sum_{i=1}^{k}\sigma^i$$

$$k \left\{ \begin{array}{l} \mathcal{N}_0(\mu_1, \sigma_1) \\ \mathcal{N}_1(\mu_1, \sigma_1) \\ \cdot \\ \cdot \\ \cdot \\ \mathcal{N}_k(\mu_1, \sigma_1) \end{array} \right.$$

$$M_i = \max_{i \in \{1,\ldots,k\}} p_Z(z|\mathcal{N}_{y=i})$$

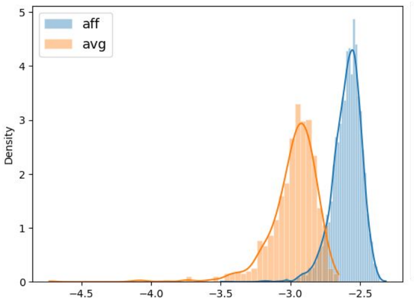
Score Computation

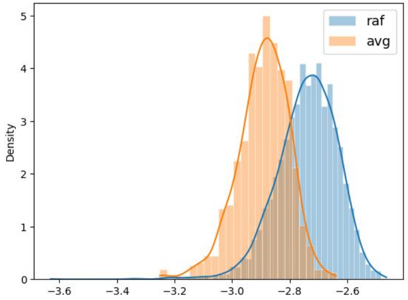


Histogram plots of M_i values on ID and OOD Test set

Semantic to Covariate Shift

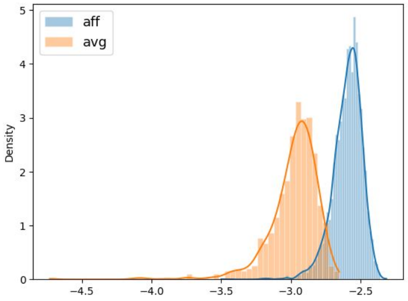
ResNet18

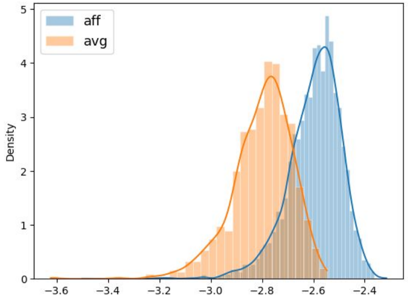




Increased Overlap

WideResnet





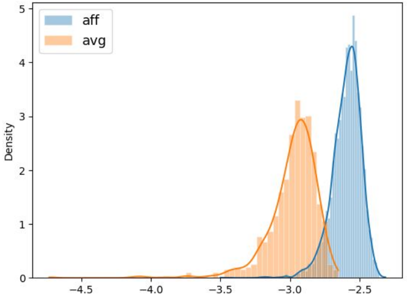
Semantic shifts

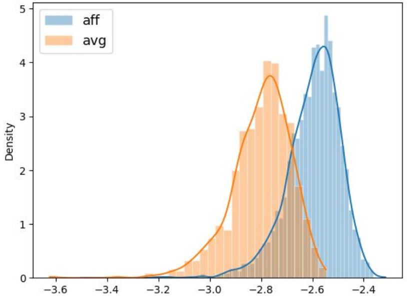
Covariate shifts

RAF-DB trained on CIFAR10

Semantic to Covariate Shift

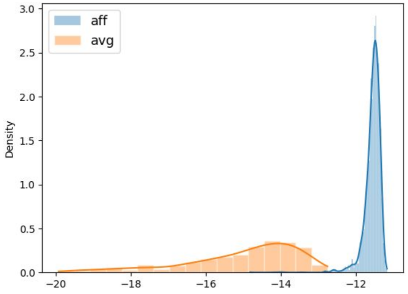
ResNet18

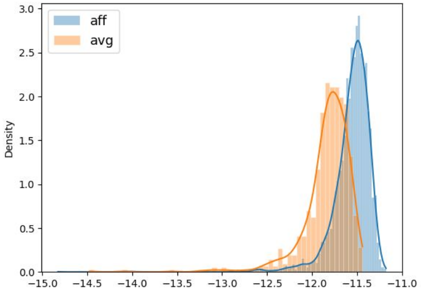




Increased Overlap

WideResnet





Semantic shifts

Covariate shifts

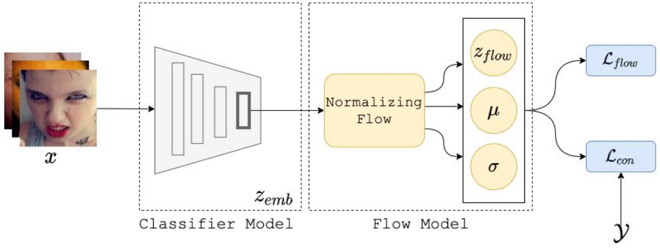
AffectNet trained on CIFAR10

Summary

$$\mathcal{L}_{con} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{S_{flow}(z_i, z_p, \mathcal{N}_i)}{\sum_{a \in A(i)} S_{flow}(z_i, z_a, \mathcal{N}_i)}$$

$$\mathcal{L} = \mathcal{L}_{con} + \lambda \mathcal{L}_{flow}$$

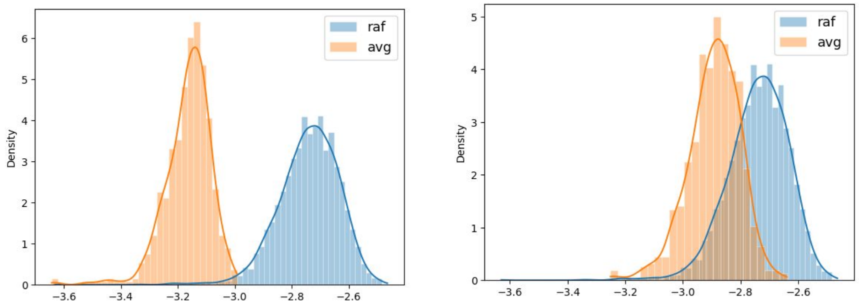
(a) Modified Contrastive Loss



(b) Class Preserving Training

D_{in} (model)	Shift	AUROC ↑	AUPR-S ↑	AUPR-E ↑	FPR-95 ↓
RAF (ResNet)	Semantic	96.35	99.57	87.50	13.05
	Covariate	87.21	97.22	50.43	64.37
RAF (WideResNet)	Semantic	97.98	99.57	91.97	10.08
	Covariate	63.51	89.89	25.22	89.75
Aff (ResNet)	Semantic	76.66	90.91	69.95	30.70
	Covariate	91.4	98.20	63.40	48.93
Aff (WideResnet)	Semantic	99.71	99.95	97.95	1.32
	Covariate	85.23	96.79	45.42	72.59

(c) Effect of Covariate Shift



(d) Correlation with Likelihood histogram