A Shift-Detection General Framework

The general framework for shift-detection can be found in the following figure, Figure 3.

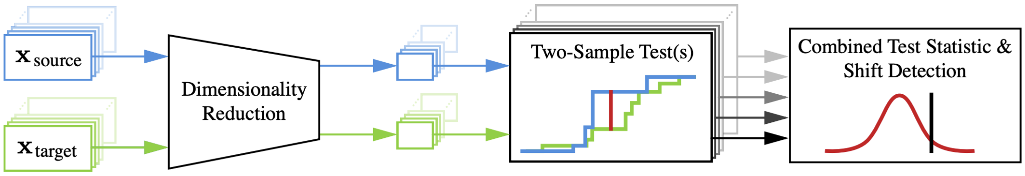

$\mathbf{x}_{\text{source}}$
$\mathbf{x}_{\text{target}}$
Dimensionality Reduction
Two-Sample Test(s)
Combined Test Statistic & Shift Detection

Figure 3: The procedure of detecting a dataset shift using dimensionality reduction and then a two-sample statistical test. The dimensionality reduction is applied to both the detection-training (source) and test (target) data, prior to being analyzed using statistical hypothesis testing. This figure is taken from [10].

B Proofs

B.1 Proof for Theorem 4.2

Proof. Define

$$\begin{aligned} \mathcal{B}_{\theta_i} &\triangleq b_i^*(m, m \cdot \hat{c}_i(\theta_i, S_m), \frac{\delta}{k}), \\ \mathcal{C}_{\theta_i} &\triangleq c(\theta_i, P). \end{aligned}$$

Consider the i$^{\text{th}}$ iteration of SGR over a detection-training set S_m, and recall that, $\theta_i = \kappa_f(x_z)$, $x_z \in S_m$ (see Algorithm 1). Therefore, θ_i is a random variable (between zero and one), since it is a function of a random variable ($x \in S_m$). Let $\mathbf{Pr}_{S_m}\{\theta_i = \theta'\}$ be the probability that $\theta_i = \theta'$.

Therefore,

$$\begin{aligned} &\mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i}\} \\ &= \int_0^1 d\theta' \mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i} | \theta_i = \theta'\} \cdot \mathbf{Pr}_{S_m}\{\theta_i = \theta'\} \\ &= \int_0^1 d\theta' \mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta'} < \mathcal{B}_{\theta'}\} \cdot \mathbf{Pr}_{S_m}\{\theta_i = \theta'\}. \end{aligned}$$

Since $\mathcal{B}_{\theta_i}$ is obtained using Lemma 4.1 (see Algorithm 1), and $\theta_i = \theta'$,

$$\mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i}\} = \mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta'} < \mathcal{B}_{\theta'}\} < \frac{\delta}{k},$$

so we get,

$$\begin{aligned} &\mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i}\} \\ &= \int_0^1 d\theta' \mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta'} < \mathcal{B}_{\theta'}\} \cdot \mathbf{Pr}_{S_m}\{\theta_i = \theta'\} \\ &< \int_0^1 d\theta' \frac{\delta}{k} \cdot \mathbf{Pr}_{S_m}\{\theta_i = \theta'\} \\ &= \frac{\delta}{k} \cdot \left(\int_0^1 d\theta' \mathbf{Pr}_{S_m}\{\theta_i = \theta'\}\right) \\ &= \frac{\delta}{k}. \end{aligned} \tag{5}$$

The following application of the union bound completes the proof,

$$\mathbf{Pr}_{S_m}\{\exists i : \mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i}\} \leq \sum_{i=1}^{k} \mathbf{Pr}_{S_m}\{\mathcal{C}_{\theta_i} < \mathcal{B}_{\theta_i}\} < \sum_{i=1}^{k} \frac{\delta}{k} = \delta.$$

□

C Exploring Model Sensitivity: Evaluating Accuracy on Shifted Datasets

In this section, we present Table 3, which displays the accuracy (when applicable) as well as the degradation from the original accuracy over the ImageNet dataset, of the considered models on each of the simulated shifts mentioned in Section 6.1.1.

Shift Dataset	**ResNet50**		**MovileNetV3**		**ViT-T**	
	Acc.	ImageNet Degradation	Acc.	ImageNet Degradation	Acc.	ImageNet Degradation
FGSM $\epsilon = 7 \cdot 10^{-5}$	76.68%	-3.7%	62.09%	-3.15%	72.51%	-2.95%
FGSM $\epsilon = 1 \cdot 10^{-4}$	75.19%	-5.19%	60.72%	-4.52%	71.49%	-3.97%
FGSM $\epsilon = 3 \cdot 10^{-4}$	66.15%	-14.23%	52.09%	-13.15%	65.06%	-10.4%
FGSM $\epsilon = 5 \cdot 10^{-4}$	59.23%	-21.15%	44.45%	-20.79%	58.9%	-16.56%
PGD $\epsilon = 1 \cdot 10^{-4}$	74.64%	-5.74%	60.63%	-4.61%	71.35%	-4.11%
GAUSSIAN $\sigma = 0.1$	79.02%	-1.36%	62.82%	-2.42%	71.79%	-3.67%
GAUSSIAN $\sigma = 0.3$	74.63%	-5.75%	55.06%	-10.18%	50.86%	-24.6%
GAUSSIAN $\sigma = 0.5$	68.56%	-11.82%	42.55%	-22.69%	22.25%	-53.21%
GAUSSIAN $\sigma = 1$	46.1%	-34.28%	13.82%	-51.42%	0.56%	-74.9%
ZOOM 50%	65.55%	-14.83%	36.96%	-28.28%	46.04%	-29.42%
ZOOM 70%	74.31%	-6.07%	53.53%	-11.71%	62.69%	-12.77%
ZOOM 90%	78.6%	-1.78%	61.28%	-3.96%	72.08%	-3.38%
ROTATION $\theta = 5°$	76.7%	-3.68%	62.42%	-2.82%	71.27%	-4.19%
ROTATION $\theta = 10°$	72.4%	-7.98%	58.22%	-7.02%	67.29%	-8.17%
ROTATION $\theta = 20°$	68.29%	-12.09%	49.96%	-15.28%	62.38%	-13.08%
ROTATION $\theta = 25°$	70.08%	-10.3%	50.95%	-14.29%	60.97%	-14.49%

Table 3: Shifted dataset accuracy and comparison with ImageNet. We displays the accuracy results for each shifted dataset and model combination, along with the accuracy degradation when compared to the original ImageNet dataset.

D Extended Empirical Results

In this section, we present a detailed analysis of our empirical findings on the ResNet50 architecture. We report the results for each window size, $|W_k| \in \{10, 20, 50, 100, 200, 500, 1000\}$, and for several shift cases discussed in Section 6.1.1. In particular, we show the detection performance of all the discussed methods, for the following shifts: FGSM (Table 4), ImageNet-O (Table 5), ImageNet-A (Table 6), and the Zoom out shift, 90% (Table 7).

Method		**Window size** AUROC ↑ / AUPR-In ↑ / AUPR-Out ↑ / DetectionError ↓ / TNR@95TPR ↓						
		10	20	50	100	200	500	1000
KS	Softmax	32/45/40/47/92	47/55/46/44/91	64/72/59/34/69	72/72/75/38/77	80/87/69/18/36	**100/100/100/2/4**	**100/100/100/0/0**
	Embeddings	54/58/55/**43/86**	32/39/48/49/100	54/64/49/39/80	41/44/48/50/99	37/48/45/47/92	60/70/59/30/60	71/77/61/33/68
MMD	Softmax	36/48/42/45/90	44/56/44/42/82	51/53/51/48/93	41/44/54/49/97	50/52/50/48/94	48/52/51/45/93	55/55/55/47/94
	Embeddings	61/56/60/48/95	57/57/59/45/93	72/73/67/36/73	63/70/56/38/71	63/69/55/37/75	67/70/61/39/74	70/79/59/28/54
Single-instance	SR	34/45/40/47/93	69/68/72/42/82	43/52/50/45/90	54/54/61/47/93	62/64/58/42/86	66/73/59/35/72	72/69/73/43/86
	Entropy	42/49/44/47/94	65/60/65/47/92	49/55/49/45/89	59/53/63/49/98	60/66/58/39/77	59/59/56/45/90	64/61/63/46/90
Ours		**71/64/75**/45/92	**77/82/75/25/51***	**88/90/85/20/39**	**84/86/84/25/49**	**99*****/99*****/99*****/3*****/5***	98/98/98/5/10	**100/100/100**/2/2

Table 4: Comparison of different evaluation metrics over **ResNet50** with the discussed baselines methods, over the FGSM shift with $\epsilon = 0.0001$. The best performing method is highlighted in **bold**; we add the superscript * to the bolded result when it is statistically significant.

Method		**Window size** AUROC ↑ / AUPR-In ↑ / AUPR-Out ↑ / DetectionError ↓ / TNR@95TPR ↓						
		10	20	50	100	200	500	1000
KS	Softmax	62/59/60/46/94	70/61/75/48/94	98/98/98/6/11	99/99/99/5/10	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
	Embeddings	85/89/76/18/35	**97/98/97/6*****/12***	99/99/99/5/9	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
MMD	Softmax	43/53/47/44/87	74/75/73/39/72	94/92/96/33/51	97/97/98/31/27	97/97/98/32/26	**100/100/100/0/0**	**100/100/100/0/0**
	Embeddings	**96/97/97*****/10/20**	95/94/97/33/39	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	94/95/97/31/46	**100/100/100/0/0**
Single-instance	SR	62/63/60/43/87	31/41/38/48/98	47/56/44/41/86	56/57/63/45/85	51/51/53/48/97	37/41/42/50/100	42/44/47/49/100
	Entropy	64/66/68/40/81	39/42/53/50/99	41/52/45/44/90	58/66/52/37/75	54/51/55/49/99	52/55/53/48/90	84/85/84/26/52
Ours		61/61/61/47/92	84/89/77/18/37	99/99/99/5/8	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**

Table 5: Comparison of different evaluation metrics over **ResNet50** with the discussed baselines methods, over the ImageNet-O shift. The best performing method is highlighted in **bold**; we add the superscript * to the bolded result when it is statistically significant.

Method		**Window size** AUROC ↑ / AUPR-In ↑ / AUPR-Out ↑ / DetectionError ↓ / TNR@95TPR ↓						
		10	20	50	100	200	500	1000
KS	Softmax	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
	Embeddings	98/98/98/7/14	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
MMD	Softmax	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
	Embeddings	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
Single-instance	SR	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
	Entropy	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**
Ours		**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**	**100/100/100/0/0**

Table 6: Comparison of different evaluation metrics over **ResNet50** with the discussed baselines methods, over the ImageNet-A shift. The best performing method is highlighted in **bold**; we add the superscript * to the bolded result when it is statistically significant.

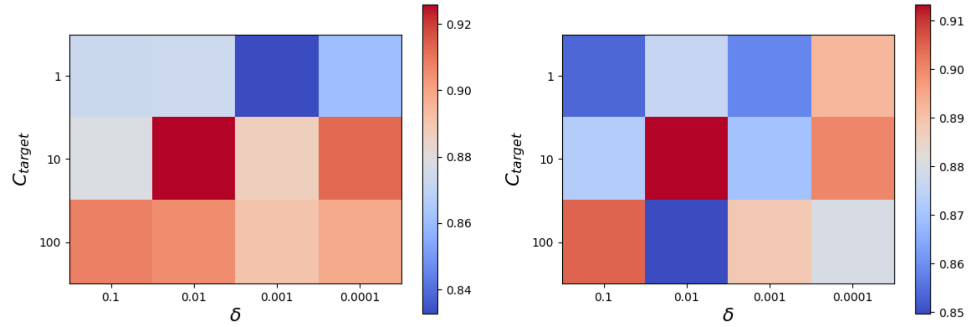
Method		Window size AUROC ↑ / AUPR-In ↑ / AUPR-Out ↑ / DetectionError ↓ / TNR@95TPR ↓						
		10	20	50	100	200	500	1000
KS	Softmax	42/49/48/47/93	53/52/54/47/97	61/69/57/37/74	71/72/69/39/77	**91**/**92**/**91***/**15**/**30**	**100***/**100***/**100***/**2***/**3***	**100**/**100**/**100**/**0**/**0**
	Embeddings	46/51/52/46/92	25/45/38/44/87	56/58/54/44/86	46/46/53/49/98	35/44/42/49/97	27/38/38/50/99	41/46/46/47/97
MMD	Softmax	48/51/50/**46**/93	46/44/52/50/100	51/57/52/45/88	50/56/52/46/88	52/57/49/46/88	51/57/47/44/90	53/53/54/49/96
	Embeddings	57/56/61/48/**91**	54/63/49/41/83	68/67/73/40/82	37/53/41/43/83	28/41/38/49/97	31/38/41/50/100	18/35/34/49/100
Single-instance	SR	28/38/38/50/100	45/45/48/51/99	44/53/53/43/87	52/59/53/42/84	65/68/59/41/85	74/73/78/38/78	84/84/83/31/62
	Entropy	29/38/39/50/100	50/47/56/51/100	51/53/58/46/91	61/63/66/41/83	71/74/61/36/72	76/73/81/39/77	87/86/87/29/58
Ours		**70**/**61**/**77**/47/95	**69**/**73**/**68**/**35**/**70**	**81**/**86**/**75**/**22**/**43**	**72**/**75**/**71**/**34**/**67**	76/82/71/23/46	94/94/94/15/31	**100**/**100**/**100**/**0**/**0**

Table 7: Comparison of different evaluation metrics over **ResNet50** with the discussed baselines methods, over the Zoom out (90%) shift. The best performing method is highlighted in **bold**; we add the superscript * to the bolded result when it is statistically significant.

E Ablation Study

In this section, we conduct multiple experiments to analyze the various components of our framework; all those experiments are conducted using a ResNet50. We explore several hyper-parameter choices, including C_{target}, δ, and κ_f. More specifically, we consider $C_{\text{target}} \in \{1, 10, 100\}$, and $\delta \in \{0.1, 0.01, 0.001, 0.0001\}$, and two different CFs κ_f, namely SR and Entropy-based.

To evaluate the performance of our detectors under varying hyper-parameters, we have selected a single metric that we believe to be the most important, namely, AUROC [46]. Additionally, since performance may vary depending on window size, we display the average AUROC across all window sizes that we have considered in our experiments. These window sizes include: $\{10, 20, 50, 100, 200, 500, 1000\}$. In Figure 4, we summarize our findings by displaying the average AUROC value as a function of the chosen hyper-parameters. These results are presented as heatmaps.



(a) $\kappa_f = (1 -$ Entropy$)$.

(b) $\kappa_f =$ SR.

Figure 4: AUROC performance of our detector under different choices of hyper-parameters.

Figure 4a, displays our AUROC detector's performance when we use Entropy-based as our CF. We observe that the optimal choice of hyper-parameters is $\delta = 0.01$ and $C_{\text{target}} = 10$, resulting in the highest performance. However, increasing the value of C_{target} leads to a more consistent and robust detector, as changes in the value of δ do not significantly affect the detector's performance. Additionally, we note that using $C_{\text{target}} = 1$ yields relatively poor performance, indicating that a single coverage choice is insufficient to capture the characteristics of the distribution represented by the sample S_m. Similar results are obtained when using SR as the CF, as shown in Figure 4b. These results suggest that selecting a high value of C_{target} and a low value of δ is the most effective approach for ensuring a robust detector. Finally, the heatmaps demonstrate that Entropy-based CF outperforms (by a low margin) SR, in terms of detection performance.