A PRELIMINARIES

A.1 MODELING REPRESENTATION SPACE OF SIGMOID-BASED DETECTOR

In this section, we describe modeling the representation space of a sigmoid-based object detector by fitting a multivariate Gaussian distribution. We denote the random variable of the input and its label of a linear classifier as $\mathbf{x} \in \mathcal{X}$ and $y = \{y_c\}_{c=1,\cdots,C} \in \mathcal{Y}, y_c = \{0, 1\}$, respectively. Then, the posterior distribution defined by the linear classifier whose output is formed by the sigmoid function can be expressed as follows:

$$P(y_c = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(w_c \mathbf{x} b_c\right)} = \frac{\exp\left(w_c \mathbf{x} + b_c\right)}{\exp\left(w_c \mathbf{x} + b_c\right) + 1}, \tag{1}$$

where w_c and b_c are the weights and bias of the linear classifier for a category c, respectively.

Gaussian Discriminant Analysis (GDA) models the posterior distribution of the classifier by assuming that the class conditional distribution ($P(\mathbf{x}|y)$) and the class prior distribution ($P(y)$) follow the multivariate Gaussian and the Bernoulli distributions, respectively, as follows:

$$\begin{aligned} &P(\mathbf{x}|y_c = 0) = \mathcal{N}(\mu_0, \Sigma_0), \qquad &&P(\mathbf{x}|y_c = 1) = \mathcal{N}(\mu_1, \Sigma_1), \\ &P(y_c = 0) = \beta_0 / \left(\beta_0 + \beta_1\right), \qquad &&P(y_c = 1) = \beta_1 / \left(\beta_0 + \beta_1\right), \end{aligned} \tag{2}$$

where $\mu_{\{0,1\}}$ and $\Sigma_{\{0,1\}}$ are the mean and covariance of the multivariate Gaussian distribution, and $\beta_{\{0,1\}}$ is the unnormalized prior for the category c and the background.

For the special case of GDA where all categories share the same covariance matrix (*i.e.*, $\Sigma_0 = \Sigma_1 = \Sigma_c$), known as Linear Discriminant Analysis (LDA), the posterior distribution ($P(y_c|\mathbf{x})$) can be expressed with $P(\mathbf{x}|y_c)$ and $P(y_c)$ as follows:

$$\begin{aligned} P(y_c = 1|\mathbf{x}) &= \frac{P(y_c = 1)P(\mathbf{x}|y_c = 1)}{P(y_c = 0)P(\mathbf{x}|y_c = 0) + P(y_c = 1)P(\mathbf{x}|y_c = 1)} \\ &= \frac{\exp\Big(\left(\mu_1 \mu_0\right)^\top \Sigma_c^1 \mathbf{x} - \frac{1}{2}\mu_1^\top \Sigma_c^1 \mu_1 + \frac{1}{2}\mu_0^\top \Sigma_c^1 \mu_0 + \ln \beta_1/\beta_0\Big)}{\exp\Big(\left(\mu_1 \mu_0\right)^\top \Sigma_c^1 \mathbf{x} - \frac{1}{2}\mu_1^\top \Sigma_c^1 \mu_1 + \frac{1}{2}\mu_0^\top \Sigma_c^1 \mu_0 + \ln \beta_1/\beta_0\Big) + 1}. \end{aligned} \tag{3}$$

Note that the quadratic term is canceled out since the shared covariance matrix is used. The posterior distribution derived by GDA in eq. 3 then becomes equivalent to the posterior distribution of the linear classifier with the sigmoid function in eq. 1 when $w_c = \left(\mu_1 \mu_0\right)^\top \Sigma_c^1$ and $b_c = -\frac{1}{2}\mu_1^\top \Sigma_c^1 \mu_1 + \frac{1}{2}\mu_0^\top \Sigma_c^1 \mu_0 + \ln \beta_1/\beta_0$. This implies that the representation space formed by $\mathbf{x}$ can be modeled by a multivariate Gaussian distribution.

Based on the above derivation, if $\mathbf{x}$ is the output of the penultimate layer of an object detector for a region proposal, and a linear classifier defined by w_c and b_c is the last layer of the object detector, it can be said that the representation space of the object detector for a category c can be modeled with a multivariate Gaussian distribution. In other words, the representation space for a category c can be represented by two parameters μ_1 (*i.e.*, μ_c) and Σ_c of the multivariate Gaussian distribution.

Discussion. The sigmoid function can be viewed as a special case of the softmax function defined for a single category as both functions take the form of an exponential term for the category-of-interest normalized by the sum of exponential terms for all considered categories. Therefore, it is straightforward to derive the modeling for the sigmoid-based detector from the previous work Lee et al. (2018), who shows that the softmax-based classifier can be modeled with a multivariate Gaussian distribution in the representation space. However, our derivation is still meaningful in that it extends the applicability of an existing modeling limited to a certain type of classifier (*i.e.*, based on softmax) to general object detectors (*i.e.*, based on sigmoid). Most object detectors, especially one-stage detectors, generally use the sigmoid function, which does not consider other categories when calculating the model output for a certain category, since more than one category can be active on a single output.

A.2 Cross-entropy with Mixture of Delta Distributions and Multivariate Gaussian Distribution

In this section, we derive the cross-entropy of two distributions that are modeled by a mixture of delta distributions and a multivariate Gaussian distribution as the normalized sum of Mahalanobis distances. Assume that the data distributions P and Q in two sets $\mathcal{D}_P$ and $\mathcal{D}_Q$ can be modeled by density functions (p and q) that take the form of a mixture of delta distributions and a multivariate Gaussian distribution, respectively, as follows:

$$p(\mathbf{x}) = \frac{1}{|\mathcal{D}_P|} \sum_{\mathbf{x}' \in \mathcal{D}_P} \delta(\mathbf{x} - \mathbf{x}'), \tag{4}$$

$$q(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^k \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{5}$$

where $\delta(\mathbf{x})$ is a Dirac delta function whose value is zero everywhere except at $\mathbf{x} = \mathbf{0}$ and whose integral over $\mathcal{X}$, which is the entire space of $\mathbf{x}$, is one. μ and Σ are two parameters of the multivariate Gaussian distribution, which can be calculated empirically over all $\mathbf{x} \in \mathcal{D}_Q$.

Then, the cross entropy, which statistically measures the difference from Q to P where Q is treated as the reference distribution, can be expressed as:

$$\begin{aligned}\mathcal{H}(P,Q) =& -\int_{\mathcal{X}} p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} \\ =& -\int_{\mathcal{X}} \frac{1}{|\mathcal{D}_P|} \sum_{\mathbf{x}' \in \mathcal{D}_P} \delta(\mathbf{x}-\mathbf{x}') \ln\left(\frac{1}{\sqrt{2\pi}^k \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)\right) d\mathbf{x}.\end{aligned} \tag{6}$$

Using two basic rules of i) $\int_{\mathcal{X}} \left(\sum_n f_n(x)\right) dx = \sum_n \left(\int_{\mathcal{X}} f_n(x) dx\right)$ if the summation is performed on a finite set, and ii) $\int_{\mathcal{X}} \delta(x-a) f(x) dx = f(a)$ if $f(x)$ is continuous on $\mathcal{X}$, the cross entropy in eq. 6 can be derived as:

$$\begin{aligned}\mathcal{H}(P,Q) &= -\frac{1}{|\mathcal{D}_p|} \sum_{\mathbf{x} \in \mathcal{D}_p} \ln\left(\frac{1}{\sqrt{2\pi}^k \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)\right) \\ &= \frac{1}{2|\mathcal{D}_p|} \sum_{\mathbf{x} \in \mathcal{D}_p} (\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) + \ln\left(\sqrt{2\pi}^k \det(\Sigma)^{1/2}\right).\end{aligned} \tag{7}$$

Note that conditions for realizing the two basic rules are satisfied in our scenario, as i) the summation is computed on a finite set $\mathcal{D}_P$, and ii) a log of the multivariate Gaussian distribution is continuous on $\mathcal{X}$.

In our scenario where the cross-entropy is used to compare the distribution gaps of different test datasets (here, $\mathcal{D}_P$s) while the reference dataset ($\mathcal{D}_Q$) is fixed, and it is computed on the representation space of the detector, the cross-entropy can be expressed as:

$$\mathcal{H}(P,Q) = \frac{1}{2|\mathcal{D}_P|} \sum_{\mathbf{x} \in \mathcal{D}_P} (f(\mathbf{x})-\mu)^\top \Sigma^{-1}(f(\mathbf{x})-\mu) + C, \tag{8}$$

where $f(\cdot)$ is the output of the detector in the representation space. The second term of eq. 7 can be regarded as a constant since k (the dimension of the representation space) and Σ (parameter of the reference dataset $\mathcal{D}_Q$) are not affected by the test dataset $\mathcal{D}_P$.

B Implementation Details

PTL. We followed the original PTL paper Shen et al. (2023) for all architectural details and training specifications of PTL except for the numbers of training epochs and iterations. The numbers of training epochs (used in sim2real transformer training) and training iterations (used in detector training) are modified to adopt a training time curtailment strategy. Specifically, in the original PTL,

Table 1: ***Wall-clock* Training time breakdown** for sim2real transformer training and detector training. Training time is shown in *mins*. The numbers in the parentheses indicate training epochs and iterations for the corresponding PTL iteration for sim2real transformer training and detector training, respectively.

(a) sim2real transformer, Vis-20

from-prev-iter	0	1	2	3	4
	28 (100)	41 (100)	56 (100)	69 (100)	83 (100)
✓	28 (100)	8 (20)	12 (20)	14 (20)	17 (20)

(b) detector, Vis-20

from-prev-iter	0	1	2	3	4	5
	40 (6.0k)	36 (6.0k)	32 (6.0k)	28 (6.0k)	25 (6.0k)	22 (6.0k)
✓	40 (6.0k)	24 (1.2k)	21 (1.2k)	19 (1.2k)	17 (1.2k)	15 (1.2k)

(c) sim2real transformer, Vis-50

from-prev-iter	0	1	2	3	4
	87 (100)	101 (100)	114 (100)	128 (100)	142 (100)
✓	87 (100)	20 (20)	23 (20)	26 (20)	29 (20)

(d) detector, Vis-50

from-prev-iter	0	1	2	3	4	5
	40 (6.0k)	38 (6.0k)	36 (6.0k)	35 (6.0k)	33 (6.0k)	31 (6.0k)
✓	40 (6.0k)	25 (1.2k)	24 (1.2k)	23 (1.2k)	22 (1.2k)	21 (1.2k)

the sim2real transformer and detector are trained for 100 epochs and 6.0k iterations, respectively, but when adopting the strategy, they are trained for 20 epochs and 1.2k iterations, respectively, after the 0th iteration.

Random selection. For random selection, we used the PTL implementation after modifying the synthetic data selection. In particular, while PTL is designed to select synthetic images by weighting images closer in domain gap to the training set, this selection is modified to randomly select synthetic data. All other parts except this selection process of the PTL training pipeline were used unchanged.

Subsets of synthetic data pool The Archangel-synthetic dataset **?** was originally created by varying the five rendering parameters as follows: 10 altitudes (from $5m$ to $50m$ at $5m$ interval), 6 radii (from $5m$ to $30m$ at $5m$ interval), 12 angles (from 0° to 330° at 30° interval) 8 human characters (Juliet, Kelly, Lucy, Mary, Romeo, Scott, Troy, and Victor), and 3 human poses (stand, prone, squat). Each subset of the synthetic data pool is built using a sparse set of each rendering parameter, as follows:

- SAlt: using sparser 5 altitudes from $10m$ to $50m$ at $10m$ interval.
- SRad: using sparser 3 radii from $10m$ to $30m$ at $10m$ interval.
- SAng: using sparser 6 angles from 0° to 300° at 60° interval.
- SCha: using sparser 4 human characters of Juliet, Kelly, Romeo, and Scott.
- SPos: using sparser 1 human pose of standing.

For each subset, all other parameters were the same as those of the original pool, except for the rendering parameter indicated to be used sparsely.

C Numerical Results

In this section, we present the numerical results of the graphs used for analysis in the main manuscript and additional results not presented in the main manuscript.

C.1 Curtailment of PTL Training Time

The reduced training time and altered accuracy by adopting the *tune-from-previous-iteration* strategy is reported in the main manuscript. Here, we also present the reduced time for two separate components of the PTL training pipeline that are affected by the strategy: detector training and sim2real transformer training (Table 1). The corresponding training times (in *mins*) with and without the strategy for every PTL iteration in Vis-20/50 settings are shown in the table. It is noteworthy that training time per PTL iteration is longer in Vis-50 than Vis-20 when using the same numbers of training iterations and epochs. This is because in our experimental setting, the real images are larger in size than the synthetic images (the image sizes of VisDrone images used as real data and the Archangel-Synthetic images used as synthetic data are 2000×1500 and 512×512, respectively) and thus require more computation time, and account for a larger portion of the training set in Vis-50.

Table 2: **Numerical results with the size of real dataset**. In each bin presenting accuracy, the mean and standard deviation of AP@.5 and AP@[.5:.95] calculated over 3 runs are reported.

(a) Same-domain and cross-domain accuracy in the Vis-N

setup	w/ synth	VisDrone	Okutama	ICG	HERIDAL	SARD
Vis-20		3.43±0.57 / 0.98±0.12	18.38±8.74 / 4.73±2.61	2.14±0.70 / 0.43±0.15	7.11±3.45 / 2.13±1.08	7.24±3.32 / 2.07±1.14
	✓	6.18±0.47 / 1.82±0.33	29.93±3.01 / 7.18±0.90	29.30±2.70 / 8.09±0.33	28.61±2.91 / 9.57±1.13	34.96±3.26 / 11.53±1.06
Vis-50		6.13±0.28 / 1.84±0.21	25.65±4.62 / 6.98±1.56	6.57±2.41 / 1.48±0.89	12.12±3.35 / 3.93±1.17	17.73±1.58 / 4.92±0.61
	✓	8.91±0.20 / 2.79±0.08	37.67±0.59 / 9.80±0.25	32.86±5.36 / 9.64±2.03	36.88±3.79 / 12.16±1.93	45.75±2.16 / 15.92±1.73
Vis-100		7.91±0.13 / 2.36±0.07	31.37±1.49 / 8.21±0.29	7.60±1.51 / 1.81±0.32	14.64±6.04 / 4.59±1.47	18.27±1.25 / 5.42±0.50
	✓	10.56±0.49 / 3.29±0.23	41.18±3.35 / 10.86±1.07	35.69±1.65 / 11.02±1.35	38.54±7.75 / 13.37±2.36	48.15±1.18 / 17.05±0.87
Vis-200		10.55±1.41 / 3.18±0.55	38.58±4.81 / 10.25±1.66	6.50±3.17 / 1.39±0.63	14.76±9.76 / 4.68±2.72	21.87±7.48 / 6.98±1.97
	✓	12.78±0.48 / 4.10±0.28	46.62±0.93 / 12.37±0.21	30.48±0.34 / 8.21±0.43	37.60±2.12 / 13.74±1.46	49.60±1.78 / 17.73±0.49

(b) Same-domain accuracy w/o synthetic data and its ratio to the Vis-N (w/ synthetic data) when using the same number of real images

	testset	# of real image			
		20	50	100	200
accuracy	Okutama	37.93±1.75 / 9.93±0.16	51.31±2.05 / 14.13±0.90	55.98±0.75 / 16.68±0.45	64.76±0.97 / 20.28±0.17
ratio to Vis-N (w/ synth)		0.79 / 0.72	0.73 / 0.69	0.74 / 0.65	0.72 / 0.61
accuracy	ICG	40.36±0.96 / 10.63±0.38	60.95±1.76 / 19.57±1.19	73.23±0.76 / 27.53±0.95	84.21±1.50 / 35.98±1.20
ratio to Vis-N (w/ synth)		0.73 / 0.76	0.54 / 0.49	0.49 / 0.40	0.36 / 0.23
accuracy	HERIDAL	41.39±2.86 / 12.75±1.82	58.97±2.86 / 19.76±0.96	65.78±0.70 / 26.27±1.25	71.53±0.49 / 31.18±1.70
ratio to Vis-N (w/ synth)		0.69 / 0.75	0.63 / 0.62	0.59 / 0.51	0.53 / 0.44
accuracy	SARD	33.44±6.36 / 8.52±2.20	51.35±1.68 / 15.28±0.78	66.81±3.15 / 23.75±1.79	75.76±1.62 / 30.17±0.98
ratio to Vis-N (w/ synth)		1.05 / 1.35	0.89 / 1.04	0.72 / 0.72	0.65 / 0.59

C.2 Scalability Behavior of Real Data

In Table 2, we present numerical results used to generate Fig. 2 of the main manuscript. Specifically, the numbers in Table 2a correspond to Fig. 2(a) and (b) of the main manuscript while the numbers in Table 2b are matched to Fig. 2(c). The results using AP@.5 follow a similar trend to those using AP@[.5:.95], which have been analyzed in the main manuscript.

C.3 Scalability Behavior of Synthetic Data

In Table 3, we present numerical results used to generate Fig. 4 of the main manuscript. The results using AP@.5 follows a similar trend to those using AP@[.5:.95], which have been analyzed in the main manuscript.

In Figure 1, we also show the scatter plots of detection scores and Mahalanobis distances for different numbers of synthetic images used in training. Among different experimental settings, we present the scatter plots for three cases: i) testing on the Okutama-Action dataset in Vis-50, ii) testing on the SARD dataset in Vis-100, and iii) testing on the HERIDAL dataset in Vis-200. For reference, the first case is the same as the scatter plots in Fig. 3(b) and Fig. 5 of the main manuscript. To better focus on the distribution of each scatter plot, scatter plots are shown separately for each size of synthetic data. In the main manuscript, these scatterplots are shown together in one figure to emphasize the differences between the plots. The observations of change in the scatter plots for the other two cases are similar to those in the first case, which has been analyzed in the main manuscript.

References

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proc. NeurIPS*, 2018.

Yi-Ting Shen, Hyungtae Lee, Heesung Kwon, and Shuvra Shikhar Bhattacharyya. Progressive transformation learning for leveraging virtual images in training. In *Proc. CVPR*, 2023.

Table 3: **Numerical results with the size of synthetic dataset.** 'Random' denotes random selection.

(a) Vis-20

method	test set	# of synthetic image: 0	100	500	1000	2000
PTL	VisDrone	3.43±0.57 / 0.98±0.12	5.88±0.58 / 1.73±0.21	6.48±0.56 / 1.89±0.13	6.28±0.86 / 1.76±0.28	6.18±0.47 / 1.82±0.33
Random			5.90±0.29 / 1.63±0.07	6.40±0.85 / 1.79±0.28	6.41±0.20 / 1.85±0.09	6.09±0.69 / 1.75±0.28
PTL	Okutama	18.38±8.74 / 4.73±2.61	28.01±2.89 / 6.73±0.80	30.54±1.06 / 7.24±0.36	29.63±1.21 / 6.90±0.41	29.93±3.01 / 7.18±0.90
Random			27.54±1.40 / 6.18±0.62	28.20±1.74 / 6.30±0.34	27.20±0.90 / 6.08±0.24	27.80±3.03 / 6.08±0.99
PTL	ICG	2.14±0.70 / 0.43±0.15	15.16±3.85 / 3.58±1.24	23.03±4.68 / 6.13±1.79	26.70±1.86 / 7.47±1.37	29.30±2.70 / 8.09±0.33
Random			8.97±0.67 / 1.97±0.20	29.62±2.03 / 8.06±0.30	26.69±6.21 / 7.39±1.14	33.70±0.85 / 9.54±0.87
PTL	HERIDAL	7.11±3.45 / 2.13±1.08	20.24±3.85 / 5.87±1.83	26.50±3.02 / 7.86±1.10	26.98±5.38 / 8.49±1.14	28.61±2.91 / 9.57±1.13
Random			14.82±2.57 / 3.94±1.00	23.37±2.19 / 6.77±0.92	25.98±2.92 / 7.41±0.73	29.22±4.55 / 9.02±1.95
PTL	SARD	7.24±3.32 / 2.07±1.14	24.51±3.39 / 7.37±1.33	34.74±1.93 /11.11±1.14	35.65±3.07 /11.90±0.78	34.96±3.26 /11.53±1.06
Random			22.15±3.13 / 6.52±1.44	34.98±5.04 /11.18±1.77	37.35±3.35 /11.78±0.67	40.01±2.07 /13.36±0.85

(b) Vis-50

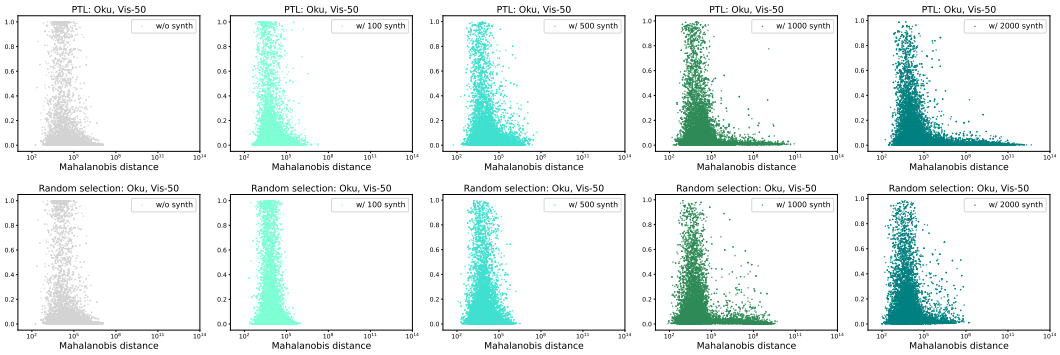
method	test set	# of synthetic image: 0	100	500	1000	2000
PTL	VisDrone	6.13±0.28 / 1.84±0.21	8.48±0.26 / 2.55±0.10	9.27±0.29 / 2.84±0.12	9.39±0.12 / 2.98±0.07	8.91±0.20 / 2.79±0.08
Random			8.17±0.28 / 2.43±0.22	9.23±0.29 / 2.77±0.15	8.97±0.08 / 2.70±0.11	9.01±0.56 / 2.69±0.04
PTL	Okutama	25.65±4.62 / 6.98±1.56	35.21±4.67 / 9.45±1.70	37.94±1.84 / 9.88±0.76	37.17±2.10 / 9.63±0.95	38.85±3.34 /10.04±1.10
Random			32.66±5.86 / 8.46±1.92	34.48±5.68 / 8.44±2.03	33.68±6.44 / 8.12±2.04	33.29±4.53 / 7.92±1.20
PTL	ICG	6.57±2.41 / 1.48±0.89	16.87±2.23 / 4.16±0.71	29.70±3.13 / 7.27±1.27	30.94±6.70 / 8.57±2.54	32.86±5.36 / 9.64±2.03
Random			14.43±2.72 / 3.24±0.90	28.78±4.90 / 7.05±1.65	31.45±1.49 / 8.72±1.44	35.51±2.12 / 9.69±0.33
PTL	HERIDAL	12.12±3.35 / 3.93±1.17	22.81±1.43 / 7.22±0.59	31.62±1.32 /10.27±0.74	33.24±1.66 /11.31±1.30	36.88±3.79 /12.16±1.93
Random			21.71±2.14 / 6.56±0.72	29.87±3.93 / 9.73±1.39	32.11±5.48 /10.41±2.96	34.06±5.59 /11.08±2.31
PTL	SARD	17.73±1.58 / 4.92±0.61	32.18±3.75 / 9.77±1.06	43.67±4.01 /14.33±1.66	43.59±2.31 /14.15±2.17	45.75±2.16 /15.92±1.73
Random			30.74±1.61 / 9.41±0.65	38.52±0.88 /12.10±0.96	44.28±2.83 /14.30±1.68	45.56±3.26 /15.21±1.64

(c) Vis-100

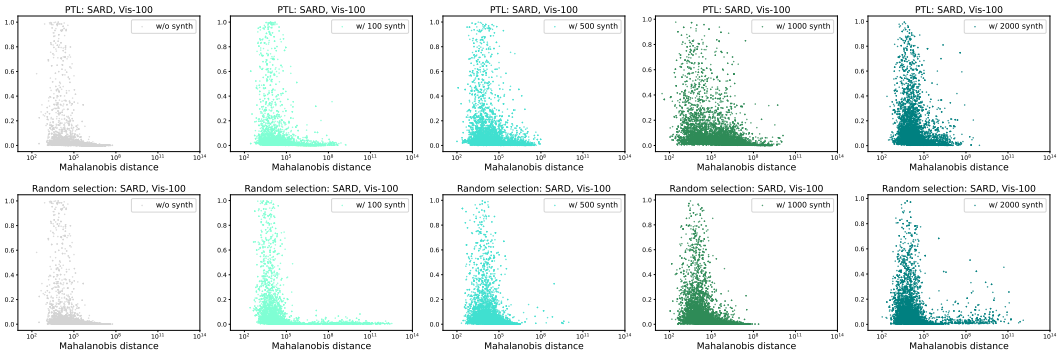
method	test set	# of synthetic image: 0	100	500	1000	2000
PTL	VisDrone	7.91±0.13 / 2.36±0.07	9.58±0.57 / 2.94±0.21	10.79±0.28 / 3.40±0.09	10.82±0.45 / 3.41±0.07	10.56±0.49 / 3.29±0.23
Random			9.13±0.60 / 2.64±0.16	10.66±0.10 / 3.23±0.09	10.67±0.43 / 3.28±0.14	10.13±0.27 / 3.11±0.03
PTL	Okutama	31.37±1.49 / 8.21±0.29	36.61±4.46 / 9.89±1.28	40.37±4.12 /10.41±0.91	40.76±4.69 /10.48±1.12	41.18±3.35 /10.86±1.07
Random			34.92±2.86 / 9.05±0.85	38.12±2.66 / 9.51±0.42	38.80±2.42 / 9.77±0.26	38.18±1.79 / 9.50±0.40
PTL	ICG	7.60±1.51 / 1.81±0.32	18.19±3.47 / 4.47±0.92	31.81±3.63 / 8.90±1.40	35.38±8.43 /10.17±2.30	35.69±1.65 /11.02±1.35
Random			16.33±1.34 / 3.66±0.57	31.67±3.23 / 7.51±1.20	32.98±2.74 / 8.90±1.23	38.75±2.49 /10.76±0.97
PTL	HERIDAL	14.64±6.04 / 4.59±1.47	25.14±9.32 / 8.23±2.54	35.00±6.30 /12.15±1.58	37.31±4.15 /13.38±0.73	38.54±7.75 /13.37±2.36
Random			23.77±7.78 / 7.43±2.13	34.28±7.17 /11.83±2.37	35.89±4.74 /12.69±1.33	40.59±5.17 /14.18±2.50
PTL	SARD	18.27±1.25 / 5.42±0.50	31.16±5.12 / 9.58±1.86	42.93±2.58 /13.98±1.20	46.04±2.30 /15.58±0.73	48.15±1.18 /17.05±0.87
Random			30.55±2.77 / 9.37±0.68	40.69±0.34 /13.13±0.12	44.10±3.84 /14.74±1.62	45.62±5.05 /15.13±2.15

(d) Vis-200

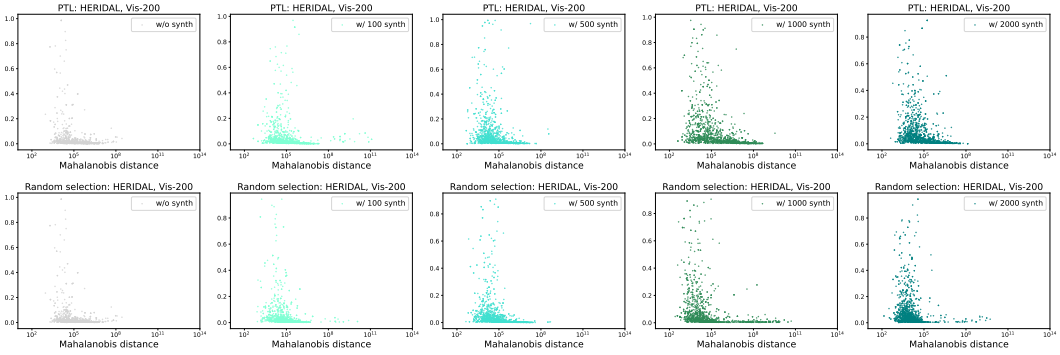
method	test set	# of synthetic image: 0	100	500	1000	2000
PTL	VisDrone	10.55±1.41 / 3.18±0.55	11.65±0.87 / 3.55±0.35	12.74±0.90 / 3.99±0.35	12.96±0.68 / 4.16±0.43	12.78±0.48 / 4.10±0.28
Random			11.35±1.23 / 3.42±0.52	12.07±1.11 / 3.72±0.46	12.63±0.59 / 3.95±0.36	12.59±1.17 / 3.97±0.49
PTL	Okutama	38.58±4.81 /10.25±1.66	40.23±2.20 /10.71±1.12	45.56±0.44 /12.44±0.51	47.79±2.47 /12.99±0.39	46.62±0.93 /12.37±0.21
Random			39.99±2.82 /10.43±0.48	39.95±2.33 /10.50±0.71	41.88±1.66 /11.02±0.74	41.20±0.92 /10.55±0.03
PTL	ICG	6.50±3.17 / 1.39±0.63	9.50±1.65 / 2.22±0.44	20.56±2.32 / 5.01±0.40	25.67±4.82 / 6.44±1.65	30.48±0.34 / 8.21±0.43
Random			9.27±2.16 / 2.49±0.16	18.29±5.19 / 4.45±1.31	23.29±5.46 / 5.98±1.31	27.29±4.52 / 7.41±1.46
PTL	HERIDAL	14.76±9.76 / 4.68±2.72	19.87±3.79 / 6.34±0.69	30.51±6.31 /10.34±1.39	34.76±8.89 /12.21±2.76	37.60±2.12 /13.74±1.46
Random			17.66±7.26 / 5.72±2.29	26.26±9.95 / 8.55±3.30	29.62±6.13 /10.09±1.44	33.74±8.06 /11.69±2.56
PTL	SARD	21.87±7.48 / 6.98±1.97	30.17±4.28 / 9.67±1.03	40.01±2.13 /13.35±0.57	46.91±5.03 /16.13±1.69	49.60±1.35 /17.73±0.49
Random			27.51±5.72 / 8.74±1.69	38.96±5.45 /12.43±2.31	44.59±1.56 /14.80±0.82	43.25±5.80 /14.85±2.10



(a) Train: Vis-50, Test: Okutama-Action



(b) Train: Vis-100, Test: SARD



(c) Train: Vis-200, Test: HERIDAL

Figure 1: **Scatter plot of detection scores and Mahalanobis distances** with various numbers of synthetic images. For each case, plots in the first row and the second row represent the scatter results for PTL and random selection, respectively. Each of the five plots in each row shows the results without using synthetic images, or with using 100, 500, 1000, or 2000 synthetic images, in order.