Improving Single Domain-Generalized Object Detection: A Focus on Diversification and Alignment

Muhammad Sohail Danish[1], Muhammad Haris Khan[1],
Muhammad Akhtar Munir[1,2], M. Saquib Sarfraz[3], Mohsen Ali[2]

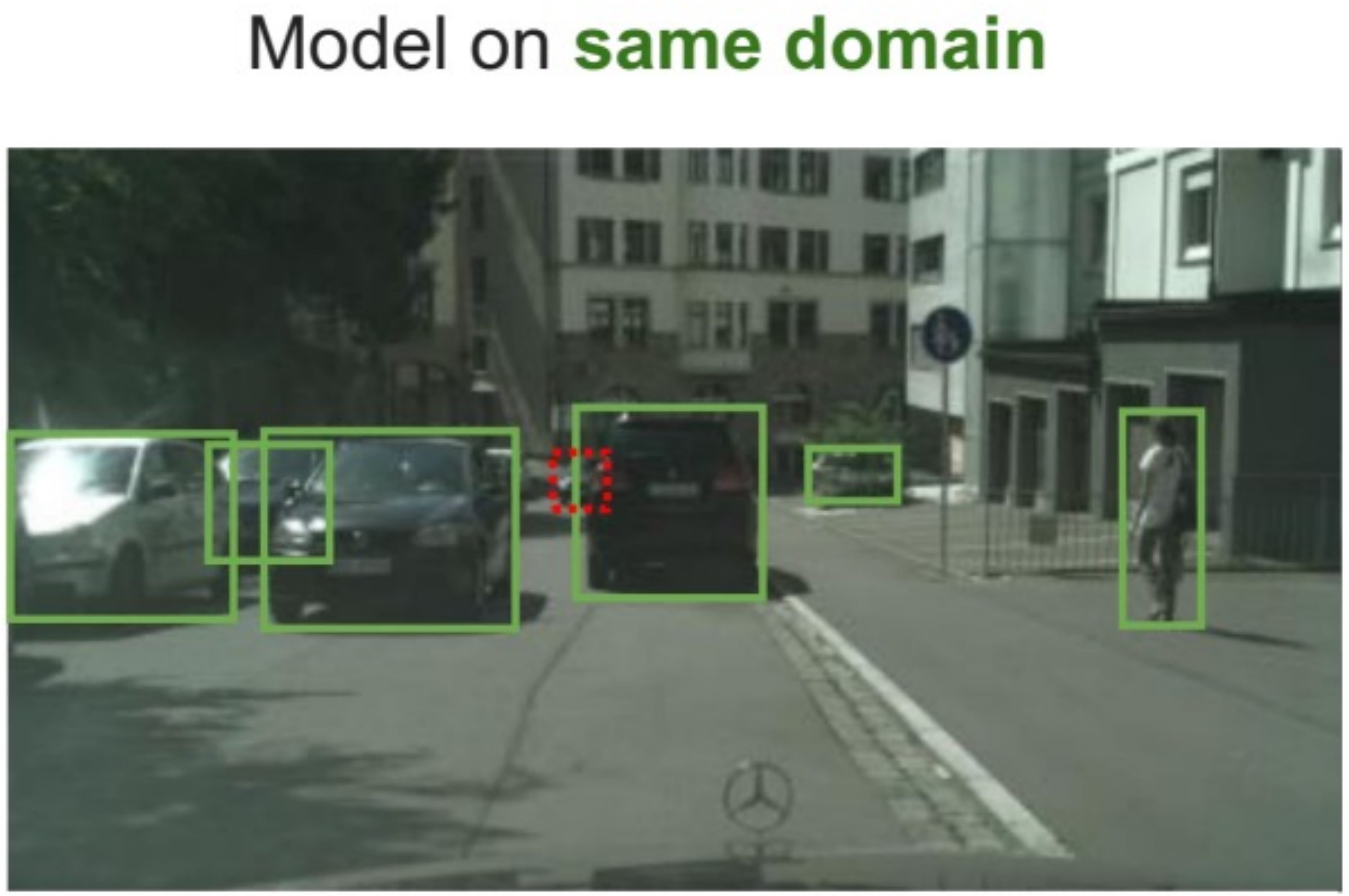
1 Mohamed Bin Zayed University of Artificial Intelligence
2 Information Technology University
3 Karlsruher Institut für Technologie

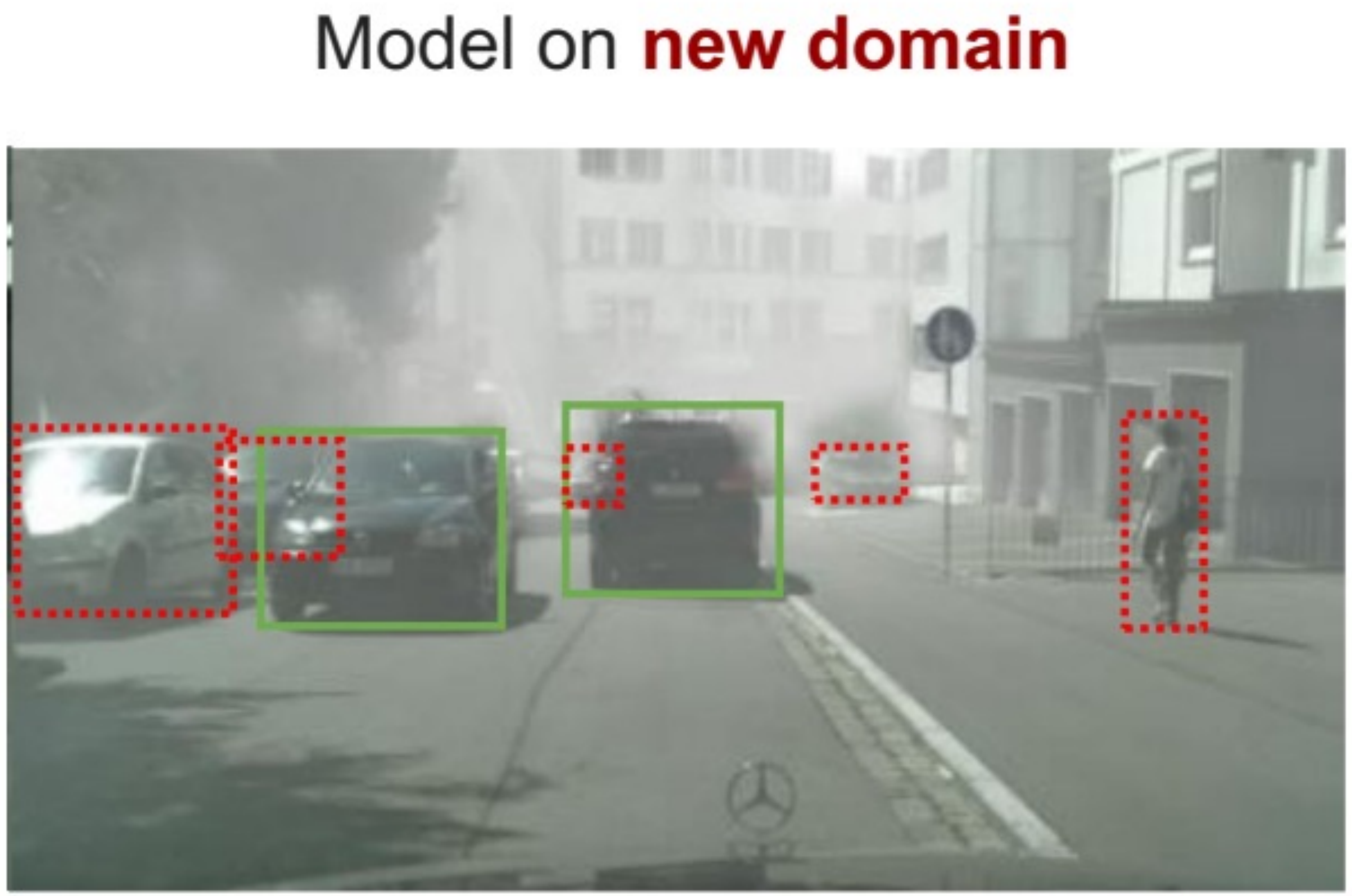
Dynamic World

World is very dynamic, very likely to encounter new domains

Oza, Poojan, et al. "Unsupervised Domain Adaptation of Object Detectors: A Survey." arXiv preprint arXiv:2105.13502 (2021).

Encountering new domain

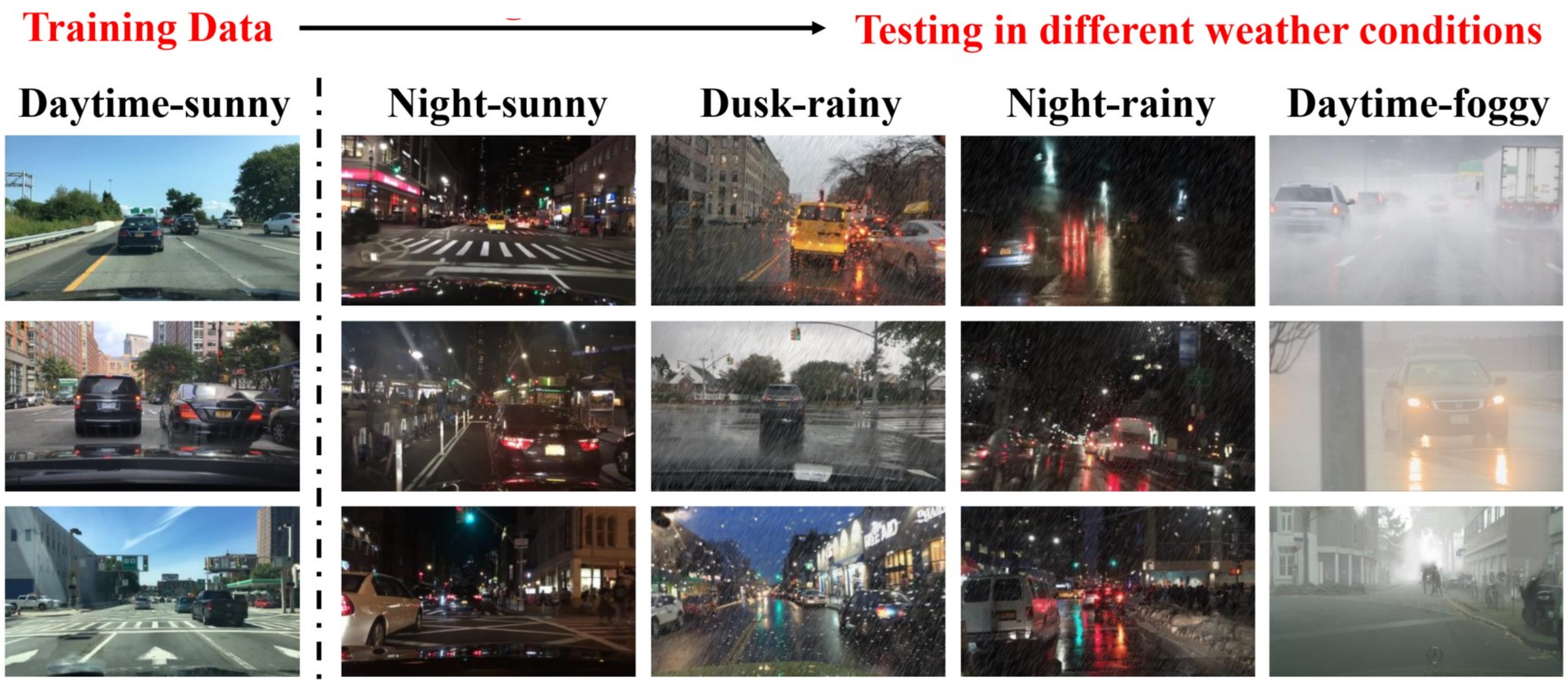




Model suffer from performance degradation upon encountring a new domain

Oza, Poojan, et al. "Unsupervised Domain Adaptation of Object Detectors: A Survey." arXiv preprint arXiv:2105.13502 (2021).

Problem Statement

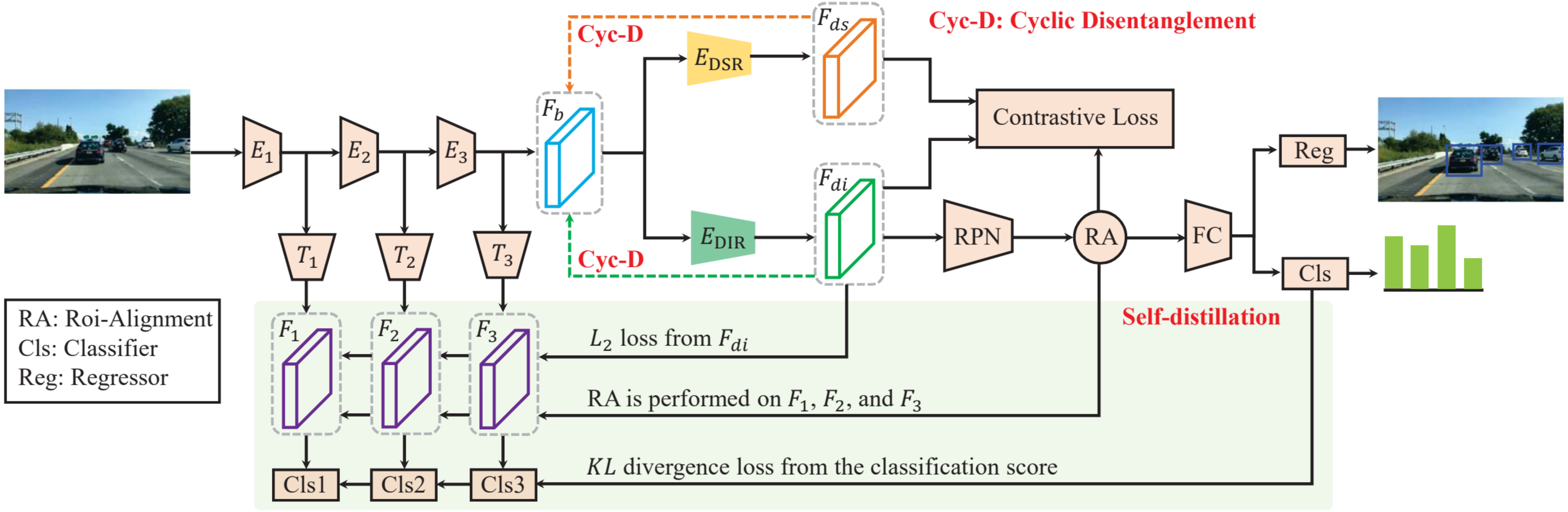


Given data sampled from single source domain, train a model that does not suffer performance degradation over other unseen target domains.

Aming Wu, et al. "Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation." CVPR. 2022.

Recent Work

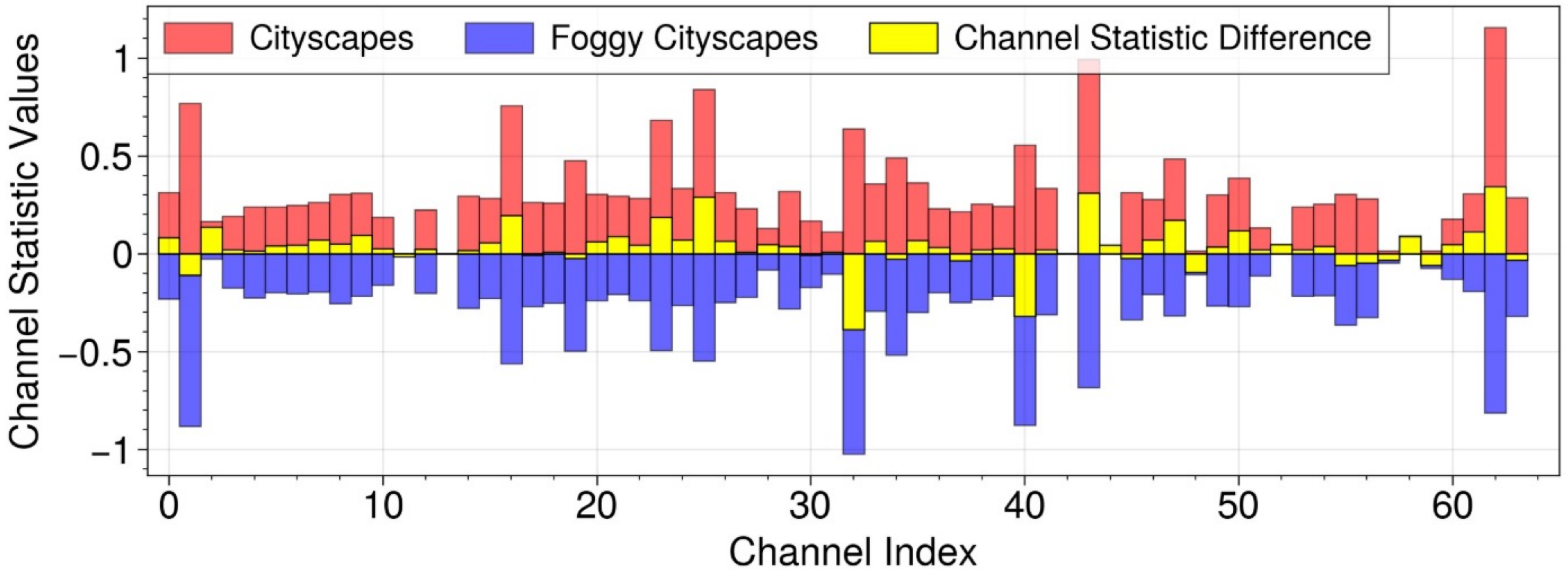
Single-Domain Generalized Object Detection in Urban Scene



- Extract Domain-invariant representations (DIR) to improve DG
- Self-distillation promotes invariant feature is shallow layers of backbone
- Boosts the source domain results at the cost of reduced generalization ability

Aming Wu, et al. "Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation." CVPR. 2022.

TOWARDS ROBUST OBJECT DETECTION INVARIANT TO REAL-WORLD DOMAIN SHIFTS – ICLR2023



$$\text{IN} \rightarrow \text{Stage 1} \xrightarrow{x_1} \alpha_1 x_1 + (\beta_1 - \alpha_1)\mu_{1,c} \rightarrow \text{Stage 2} \xrightarrow{x_2} \alpha_2 x_2 + (\beta_2 - \alpha_2)\mu_{2,c} \rightarrow \text{Stage 3-5} \rightarrow \text{OUT}$$

$x_i \in \mathcal{R}^{B\times C\times H_i\times W_i}$ $\mu_{i,c} = \frac{1}{H_i W_i}\sum_{H_i}\sum_{W_i} x_i \in \mathcal{R}^{B\times C}$ $\alpha_i, \beta_i \sim Gaussian(1, 0.75) \in \mathcal{R}^{B\times C}$

- Perturbing the feature channel statistics of source domain can synthesize new latent styles and overcome domain style overfitting

Our Approach

Proposed Solution

- We intend to make an object detector **domain invariant** by using single training domain
- Our method has two main components
  1. **Diversifying** the single domain by augmentations for segregating domain-specific features during model training
  2. **Aligning** the model prediction across different views of the same image to improve the generlization and better calibration

Preliminaries

- **Source:** $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N^S}$ is the training domain where x_i is image and y_i is label
- **Target:** $\{\mathcal{D}_t\}_{t=1}^{T}$ is set of T unseen target domains
- $\phi(\,.\,)$ is a visual corruption function which convert image from $\mathcal{D}_s$ into different domain $\mathcal{D}_\phi$ where $\phi \sim \Phi$
- We define a domain invariant object detector as

Assuming that, for an input image x, an object detection model $\mathcal{F}_{det}$ predicts class probability distribution $\widehat{p_n}$ and bounding box coordinates $\widehat{b_n} \in R^4$ for the n^{th} proposal.
Let x^s be an image from $\mathcal{D}_s$ and $x^\phi = \phi(x^s)$ be the transformation of x^s, denoted as x^ϕ, where $\phi \sim \Phi$. The model $\mathcal{F}_{det}$ is domain invariant if:

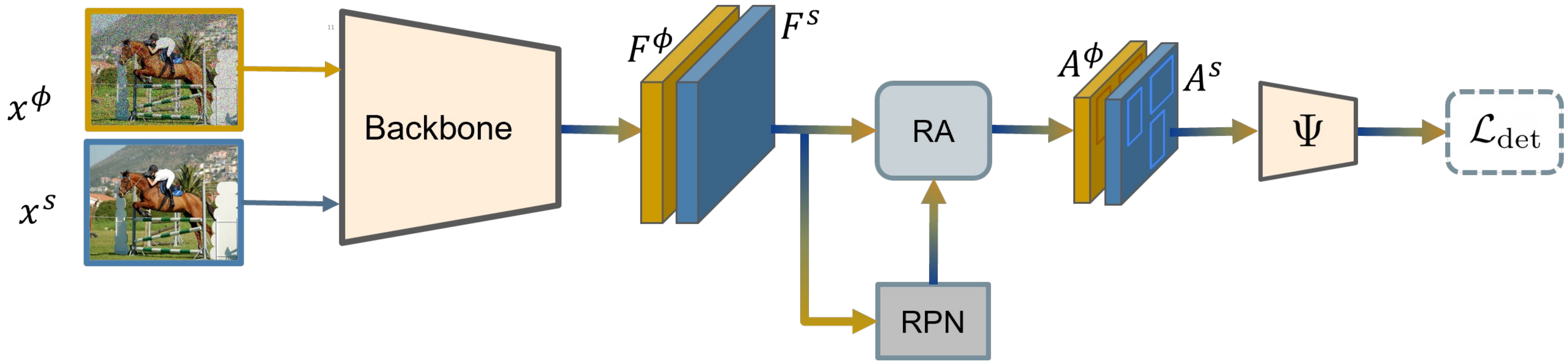
$$\widehat{p_n^s} = \widehat{p_n^\phi} \qquad (1)$$

Object Classification Constraint

$$1 - \mathrm{IoU}\left(\widehat{b_n^s}, \widehat{b_n^\phi}\right) = 0 \qquad (2)$$

Object Localization Constraint

Faster R-CNN



- $F \in R^{m\times w\times h}$ is the feature map output from the backbone
- RPN takes F as input to predict the object proposals $O \in R^{Z\times 4}$
- $A = \text{RA}(O, F) \in R^{Z\times m}$ is feature representation
- $\mathcal{L}_{det}$ is the detection Loss give as

$$\mathcal{L}_{\text{det}} = \sum_{n=1}^{Z} L_{det}(\Psi(A_n), y_n, b_n)$$

- Ψ includes the classifier and regressor, y and b are ground truth

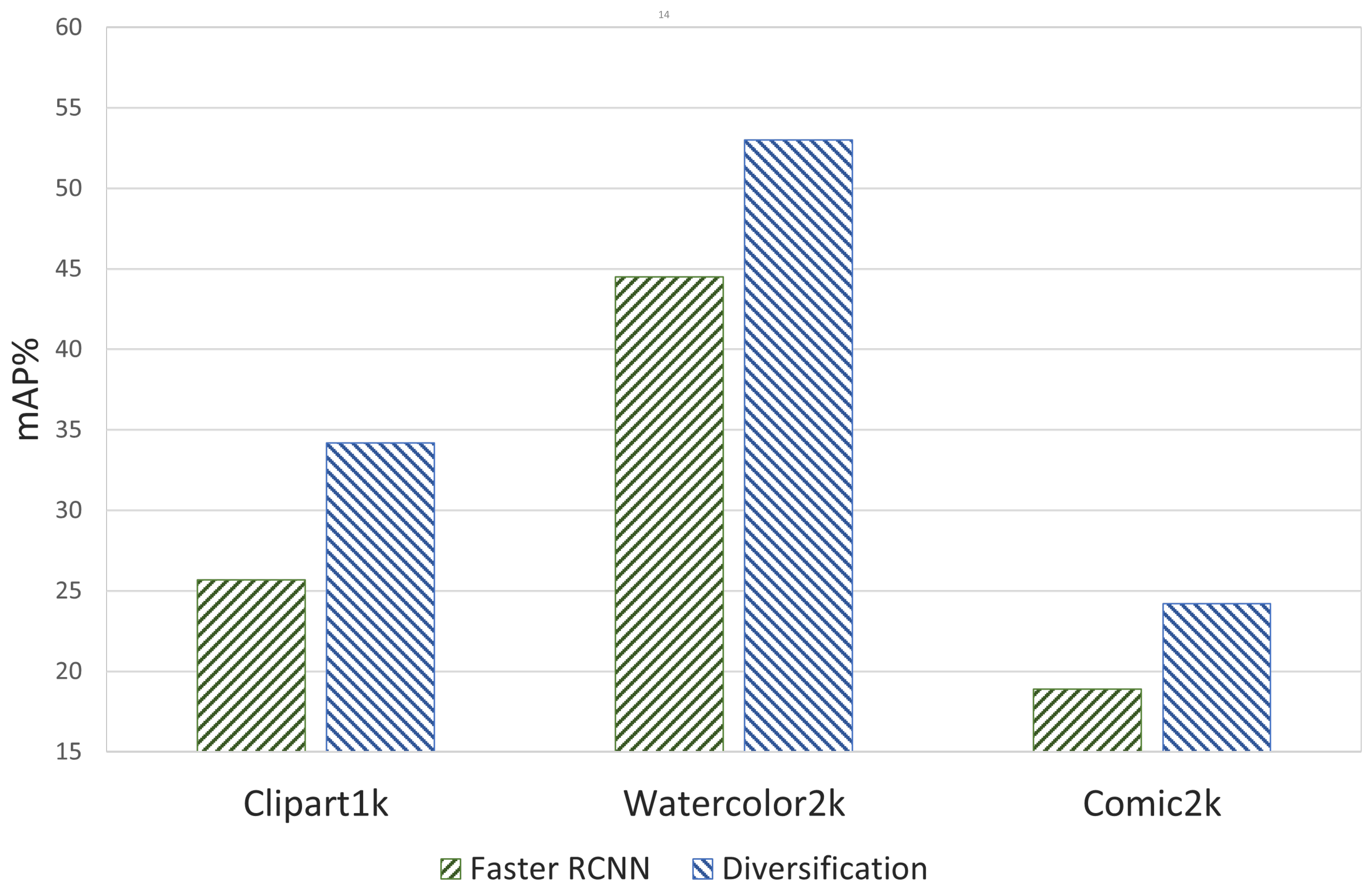
Diversifying Single Source Domain

- Diversification help to learn actual semantics instead of shortcuts
- Augment every image in the mini-batch using $\phi(\,.\,)$ where $\phi \sim \Phi$
- Φ contain ImageNet-C with Fourier transform-based corruptions grouped as
  - **Blur** smooth the pixels by apply blur functions including glass, Gaussian, motion, defocus
  - **Noise** add different kinds of noise e.g. Gaussian, shot, spackle, impulse
  - **Digital** either change the pixel intensities (brightness, saturation and contrast) or changes resolution using JPEG compression, pixelation, and elastic transformation
  - **Fourier-based** such as phase scaling, constant amplitude, and High Pass Filter

Examples of augmentations

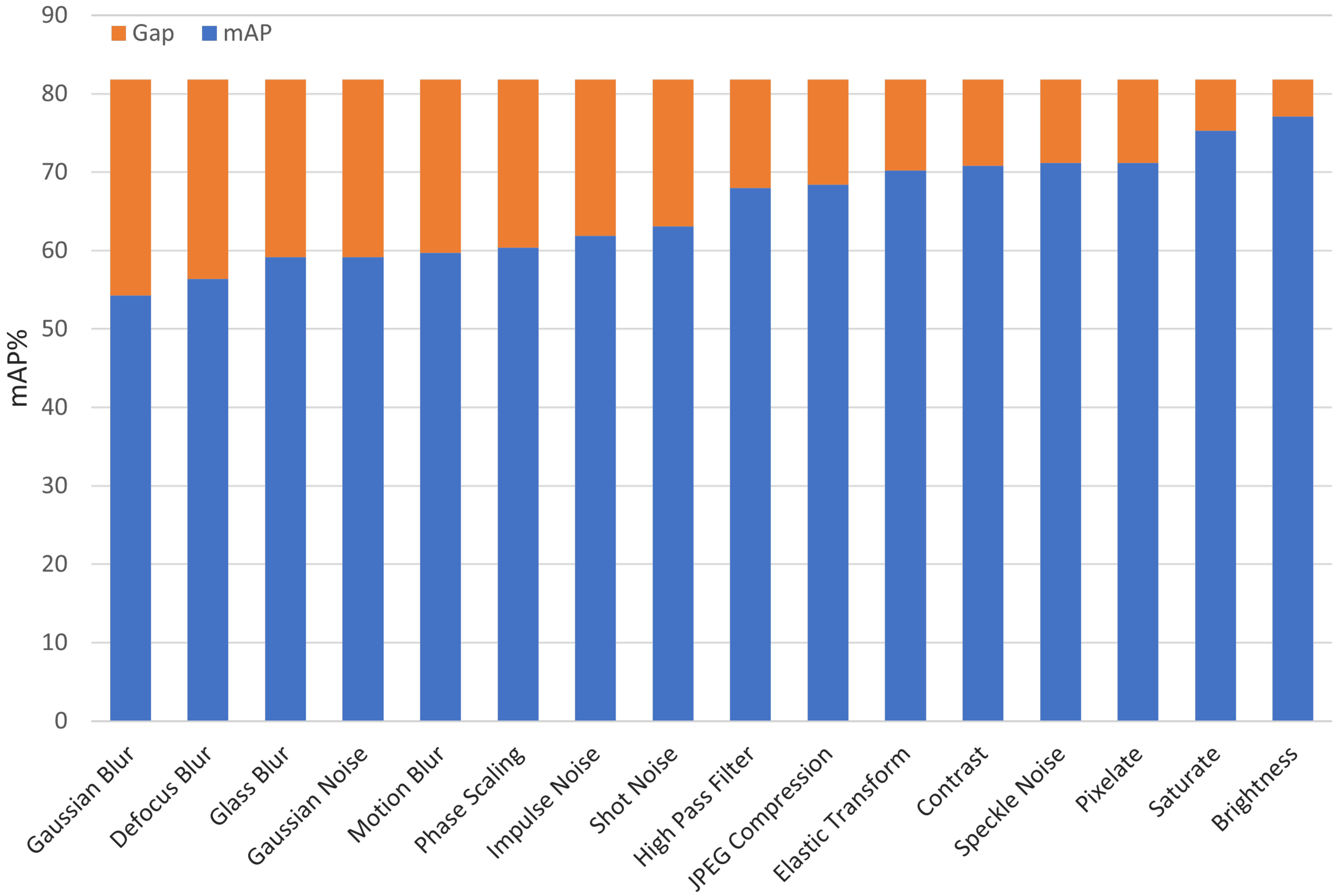
Blur | Noise | Digital | Digital + Fourier

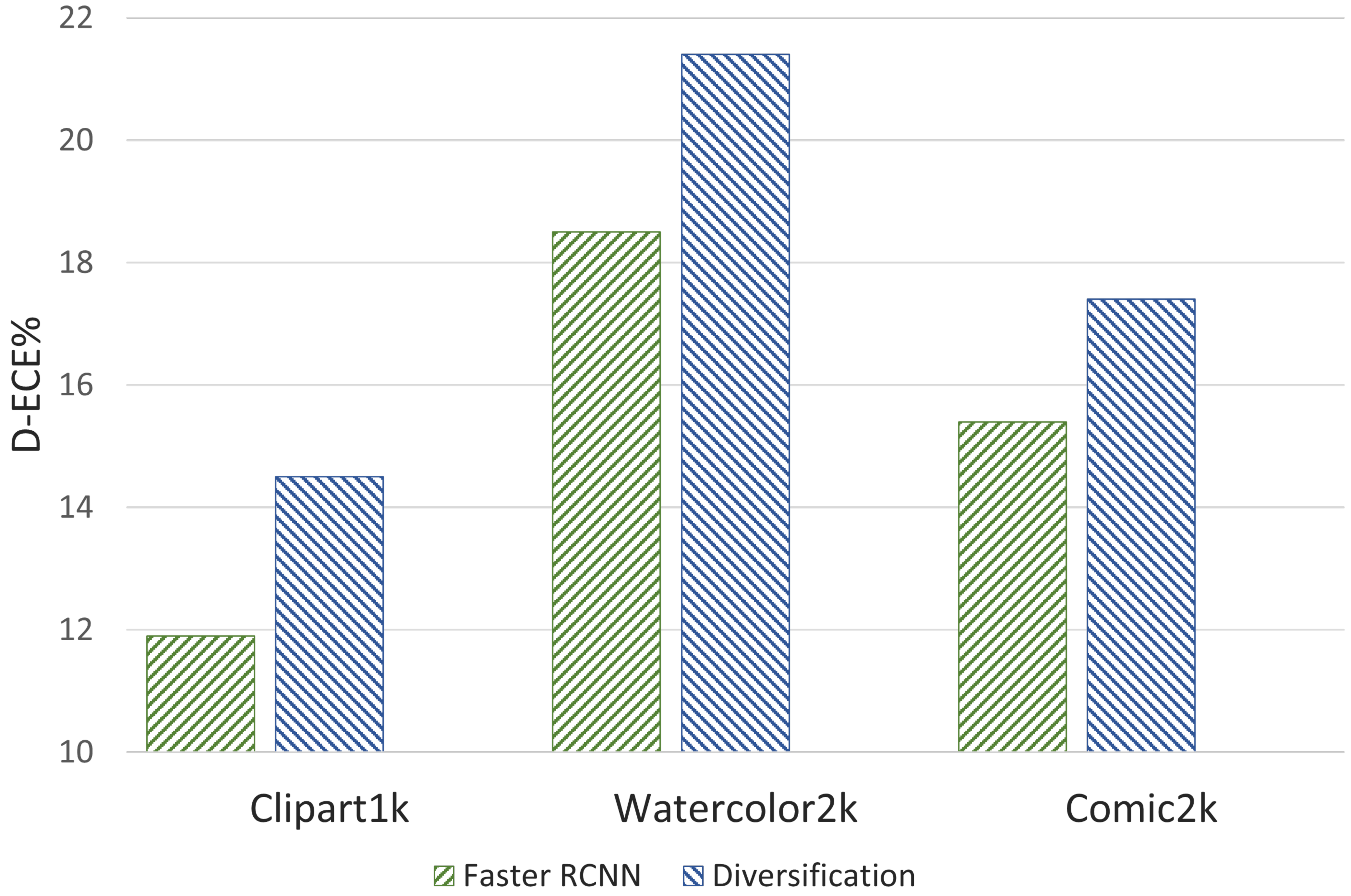
Diversifying the Single Domain



- Diversification outperforms Faster R-CNN baselines
- Model is trained on Pascal VOC (in-domain) and evaluated on Clipart1k, Watercolor2k and Comic2k (Out-of-domain)

Limitations of Diversification





- The performance misalignment on diversified and original images
- Miscalibration in out-of-domain scenarios
- **Solution:** Use proposed alignment losses

Aligning classification

- Minimize the KL divergence between the classifier output $\hat{\mathbf{p}}_n^s$ $\hat{\mathbf{p}}_n^{\phi}$
- No 1-1 correspondence between O^{φ} and O^{s}
- Obtain $\hat{\mathbf{p}}_n^{\phi'}$ by passing features from augmented and proposals from original image
- The final classification alignment loss is given by

$$\mathcal{L}_{\text{cal}} = \sum_{n=1}^{Z} \mathrm{KL}(\hat{\mathbf{p}}_n^s \| \hat{\mathbf{p}}_n^{\phi'})$$

Aligning Regression

- Obtain $\hat{\mathbf{b}}_n^{\phi'}$ similar to $\hat{\mathbf{p}}_n^{\phi'}$
- Maximize IoU between bounding box regressor output $\hat{\mathbf{b}}_n^s$ $\hat{\mathbf{b}}_n^{\phi'}$
- We achieve this minimizing the L2-squared norm

$$\mathcal{L}_{\text{ral}} = \|\hat{\mathbf{b}}_n^s - \hat{\mathbf{b}}_n^{\phi'}\|_2^2$$

Overall training objective

- Overall training loss is given as

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{cal}} + \beta\mathcal{L}_{\text{ral}}$$

- where α and β are the hyperparameters for balancing the contributions of alignment losses

Experiments and Results

Comparison on Real to artistic generalization using mAP metric (%)

Method	VOC	Clipart1k	Watercolor2k	Comic2k
Faster R-CNN	81.8	25.7	44.5	18.9
Diversification (Div.)	80.0	34.2	53.0	24.2
Div. + $\mathcal{L}_{cal}$	**82.1**	36.2	53.9	28.7
Div. + $\mathcal{L}_{ral}$	80.7	35.0	53.8	28.7
Div. + $\mathcal{L}_{cal}$ + $\mathcal{L}_{ral}$	80.1	**38.9**	**57.4**	**33.2**

After using the proposed alignment losses, we are able to boost the overall performance by 13.2%, 12.9%, and 14.3% on Clipart1k, Watercolor2k and Comic2k respectively

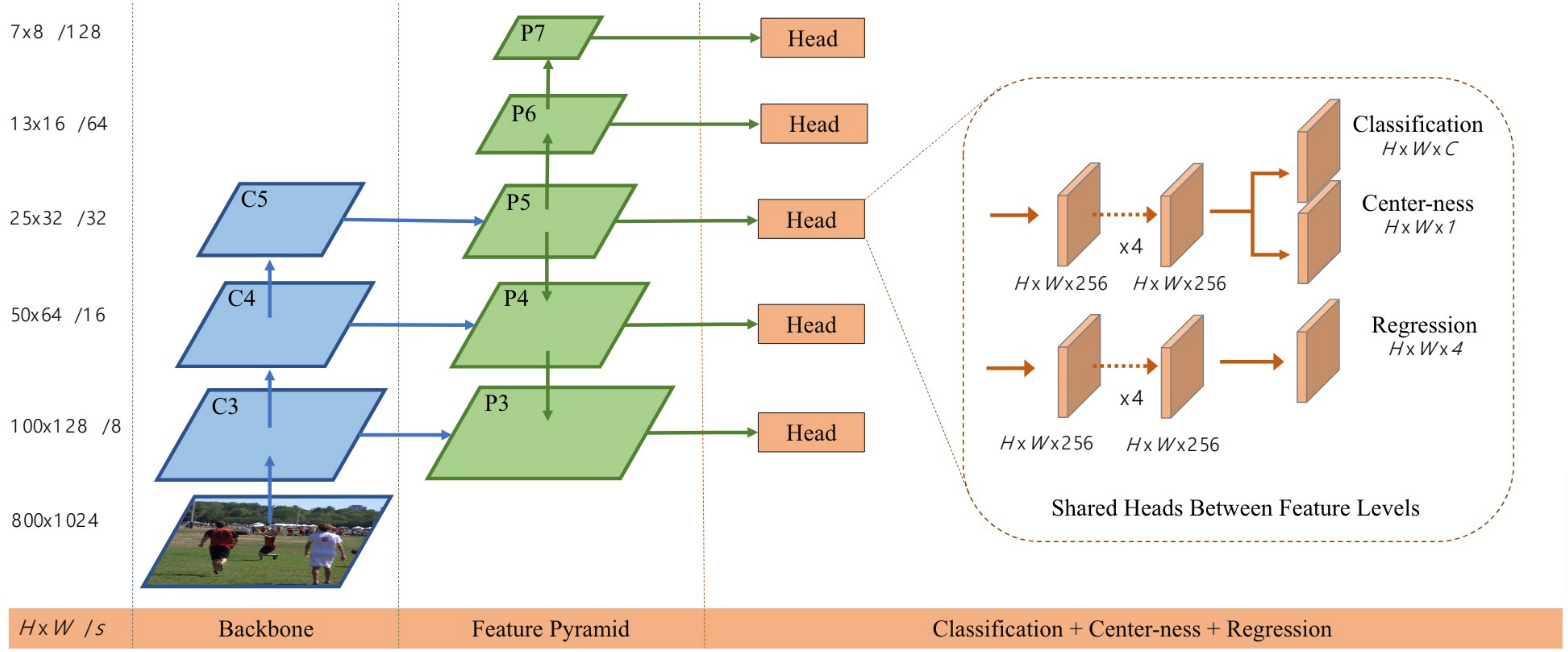
Comparison on Urban-scene detection using mAP metric (%)

Method	DS	NC	DR	NR	DF
Faster R-CNN	51.8	38.9	30.0	15.7	33.1
SW*	50.6	33.4	26.3	13.7	30.8
IBN-Net*	49.7	32.1	26.1	14.3	29.6
IterNorm*	43.9	29.6	22.8	12.6	28.4
ISW*	51.3	33.2	25.9	14.1	31.8
Wu et al.*	**56.1**	36.6	28.2	16.6	33.5
Diversification	50.6	39.4	37.0	22.0	35.6
Our Method	52.8	**42.5**	**38.1**	**24.1**	**37.2**

Our method beats all baselines and state-of-the-art method and gains 8-9% on DR and NR and 3-4 % on NC and DF.

Single Stage Detector

- We choose FCOS which is anchor-less single stage object detector to evaluate our method
- As there is no RPN involved in the FCOS, the 1-1 correspondence between detection on clean and augmented images is guaranteed



Evaluation on Single Stage Detector using mAP metric (%)

Method	VOC	Clipart	Watercolor	Comic
FCOS	78.1	24.4	44.3	15.4
Diversification	79.6	31.7	48.8	25.2
Div. + $\mathcal{L}_{cal}$	**80.1**	35.4	52.6	29.4
Div. + $\mathcal{L}_{ral}$	77.5	29.8	50.3	24.0
Div. + $\mathcal{L}_{cal}$ + $\mathcal{L}_{ral}$	77.5	**37.4**	**55.0**	**31.3**

In comparison to FCOS, our method delivers a significant gain of 13.0%, 10.7% and 15.8% on Clipart1k, Watercolor2k, and Comic2k shifts, respectively

Comparison with DA methods using mAP metric (%)

Method	Clipart	Watercolor	Comic
DA-Faster	19.8	46.0	-
SWDA	38.1	53.3	27.4
HTCN	40.3	-	-
DBGL	**41.6**	53.8	29.7
Our Method	38.9	**57.4**	**33.2**

Even though our method does not require the target domain datasets at training time, it can still achieve better results than many domain adaptation methods.

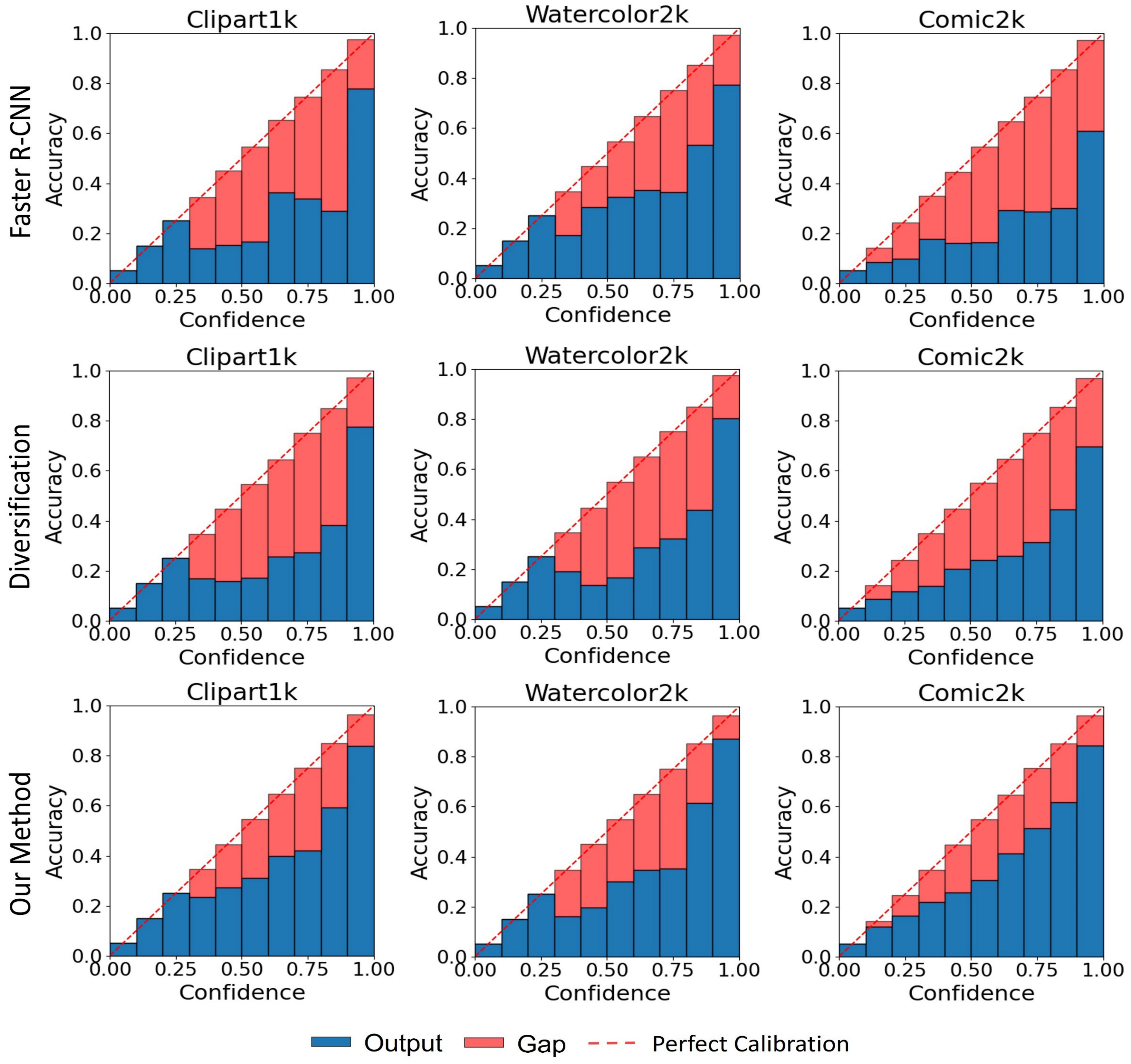
Calibration Performance using D-ECE metric (%)

Method	Clipart	Watercolor	Comic
Faster R-CNN	11.9	18.5	15.4
Diversification	14.5	21.4	17.4
Our Method	**10.7**	**14.4**	**14.3**

Method	NC	DR	NR	DF
Faster R-CNN	31.5	29.3	27.9	25.8
Diversification	33.0	30.2	28.9	25.7
Our Method	**29.3**	**24.9**	**15.8**	**20.6**

Compared to baseline, the diversification increase model calibration error, however, our method is capable of improving model calibration

Reliability Diagram



- Compared to baseline, the diversification increase model calibration error, however, our method is capable of improving model calibration

Comparison on Medical Imaging dataset

Method	HCM	LCM
Faster R-CNN	71.4	15.1
Diversification	**74.7**	25.0
Ours	70.7	**35.9**

Generalization Performance (mAP %)

Method	LCM
Faster R-CNN	8.4
Diversification	8.0
Ours	**5.5**

Calibration Performance (D-ECE %)

Our proposed method is capable of generalizing to an unseen medical imaging domain (LCM), and improving the model calibration.

Qualitative results on Real to Artistic

(a) Faster R-CNN (b) NP (c) Diversification (d) Our Method (e) Faster R-CNN (f) NP (g) Diversification (h) Our Method

By using the proposed alignment losses, our model is not only able to detect the object that were missed by baselines but also reduces the false positives.

Qualitative results on Real to Artistic

By using the proposed alignment losses, our model is not only able to detect the object that were missed by baselines but also reduces the false positives.

Thanks