Learning Intelligence
& Vision Essential
(LiVE) Group

TUE-PM-381

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Multi-view Adversarial Discriminator: Mine the Non-causal Factors for Object Detection in Unseen Domains

[Highlight Paper]

Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, Lei Zhang(✉)

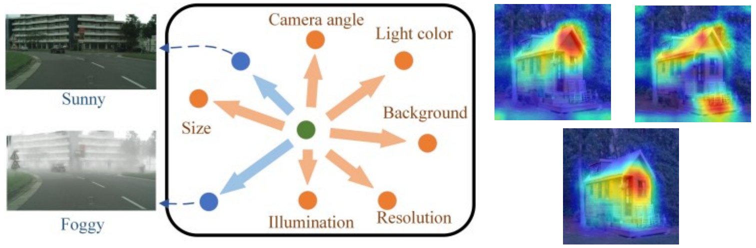
School of Microelectronics and Communication Engineering, Chongqing University, China
Hikvision Research Institute

June, 2023

Overview: Multi-view Adversarial Discriminator

Problems：

In ①limited training domains, Traditional domain adversarial learning (DAL) can't ②extract true causal features?



① ②

Results：

	Methods	Cityscapes	Foggy Cityscapes	BDD100k	Rain Cityscapes
Cityscapes	source-only		27.2	24	36.3
	MLDG		29.2	21	42.1
	FACT		25.3	26	39.9
	FSDR		31	26.2	**42.8**
	DANN+**SCG**		37.5	26.1	39.1
	MAD		**38.6**	**28**	42.3
Foggy Cityscapes	source-only	29.9		17.5	38.4
	MLDG	30.4		18	38.6
	FACT	30		20.2	38.7
	FSDR	31.3		20.4	40.8
	DANN+**SCG**	38.4		22.4	40.4
	MAD	**41.3**		**24.4**	**43.3**
BDD100k	source-only	33.6	27.2		34.3
	MLDG	24.7	17.1		20
	FACT	32.4	24.3		33.9
	FSDR	32.4	27.8		34.7
	DANN+**SCG**	35.8	29.3		33.9
	MAD	**36.4**	**30.3**		**36.1**

Solution： MAD

Spurious Correlations Generator (SCG)

In the spectrum of an image:

Extremely high and low frequency parts (non-causal features)	Mid-frequency parts (causal features)
Gaussian random	**Keep unchanged**

$$\hat{\mathbf{X}} = \mathcal{F}'(R_G(\mathbf{S}_{non}) + \mathbf{S}_{cau})$$

Faster RCNN Backbone

Multi-View Domain Classifier (MVDC)

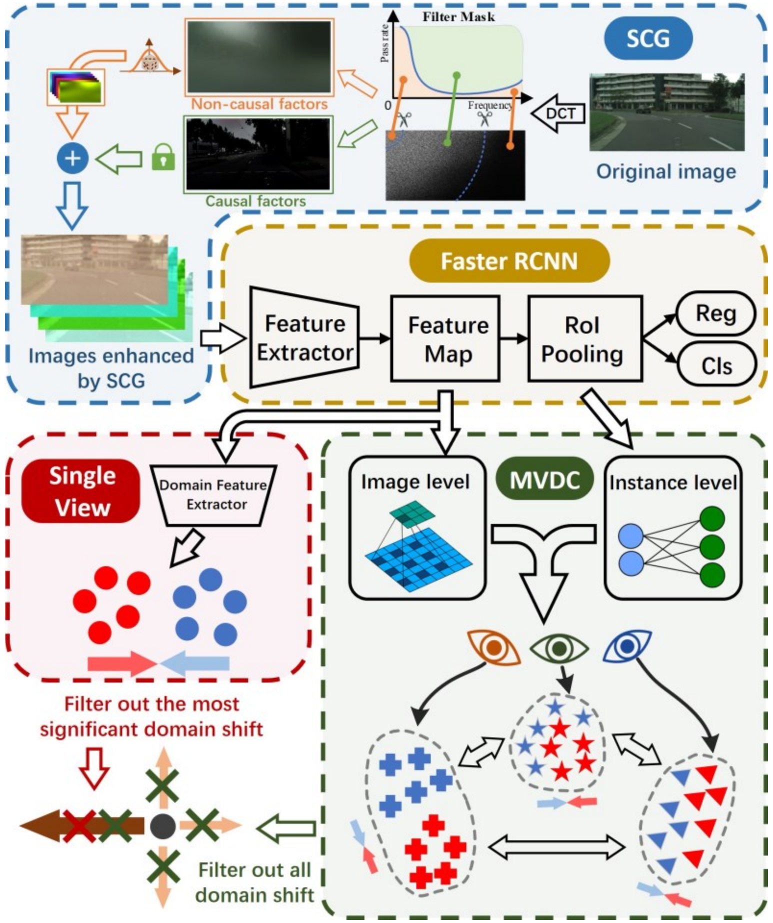
The structure of MVDC:

Feature extractor ⟺ balance ⟺ Domain classifier ~~Single-view~~ Multi-view

$$\min_{\mathcal{F}} d_{\mathcal{H}}(D_{s1}, D_{s2}) = \underbrace{\max_{\mathcal{F}} \min_{h \in \mathcal{H}} err\big(h(\mathbf{S})\big)}_{Standard\ DAL}$$

$$\Rightarrow \underbrace{\max_{\mathcal{F}} \sum_{i=1}^{M} \min_{h_i \in \mathcal{H}, e_i} err\big(h_i(e_i(\mathbf{S}))\big)}_{MVDC}$$

Each view consists of an AutoEncoder $\langle e_i | g_i \rangle$ and domain classifier h_i.



Challenges of object detection tasks in open-world

Source domain

Target domain

i.i.d.

Out-of-Distribution

Domain adaptation

[**Labeled**] **Source** samples
[**unlabeled**] **Target** samples

Extracting **common** features

Suitable for deployment in the **target** domain

Challenges faced by DA

① need to obtain the distribution of the target domain in advance

② need to retrain when encountering a new target domain

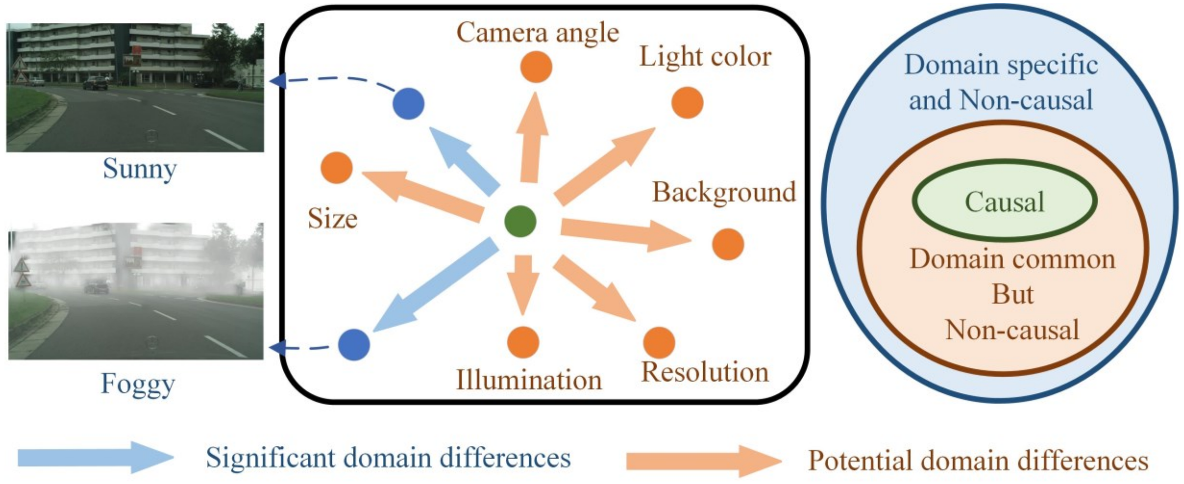
Domain Generalization Object Detection

Training one model for all scenes $\xrightarrow{\text{key}}$ Learning to extract **causal** features

Previous domain adversarial learning methods (DAL) faces two main problems :

problem ①

In limited source domains: Common features ≠ Causal features



Non-causal features

① Significant shift of limited domains

② Latent non-causal features

Nonexistent
Insignificant $\Big\}$ domain shift in training data

Cannot be eliminated by the feature extractor

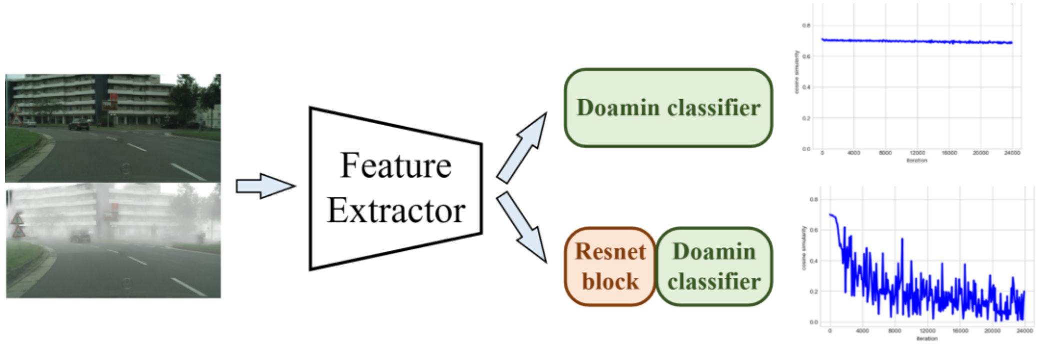
problem ②

Traditional DAL methods struggle to handle broader domain shift.

Experiment 1

Even when adversarial balancing is applied,

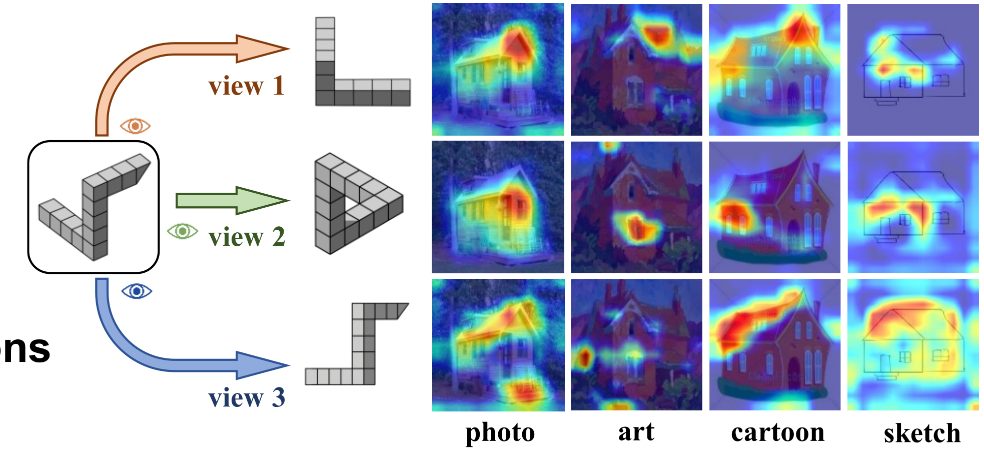
DAL still fails to eliminate all non-causal factors.



Experiment 2

We observe a Penrose triangle from different perspectives. ⟹ Different images are obtained. “L”, “Δ”, “Z”

Different domain classifiers ⟹ Focus on **different regions** of the same image



Category classifier ⟹ Improve classification accuracy

Domain classifier ⟹ Discovering more comprehensive domain differences

Multi-view Adversarial Discriminator

To Address the two mentioned problems above:

We propose a **Multi-view Adversarial Discriminator (MAD)**

- **Spurious Correlations Generator (SCG)**

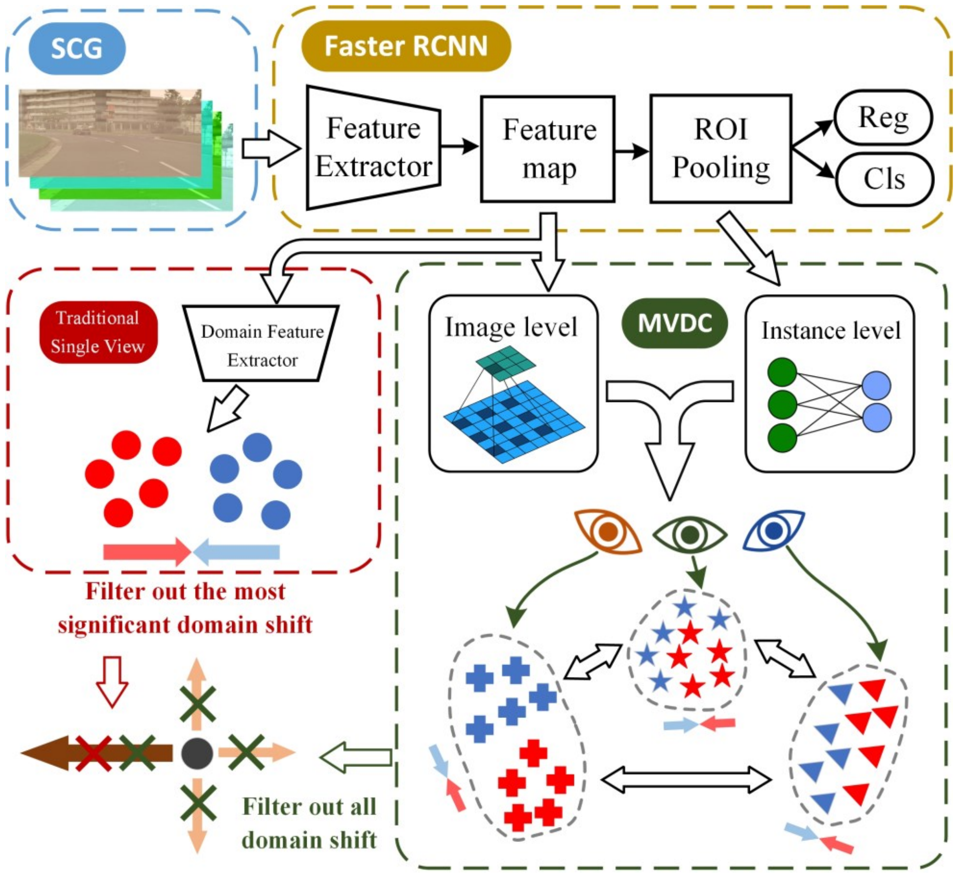
  Adding latent non-causal factors to the source domain via random transformations in the frequency domain

- **Multi-View Domain Classifier (MVDC)**

  To gain a clearer understanding of things, it is necessary to observe them from **multi-view**.

Mapping features to **multiple distinct feature spaces**

Eliminating non-causal factors exhibited in each space



The overall structure of MAD can be divided into three parts:
(1) **Yellow Part**: FasterRCNN backbone network.
(2) **Blue Part**: Spurious Correlations Generator (SCG).
(3) **Green Part**: Multi-view Domain Classifier (MVDC).
(4) **Red Part**: Traditional DAL.

Synthetic Correlation Generator (SCG)

The paper **FSDR*** verified : **Low & high frequency** components contain more **domain related information**.

In SCG:

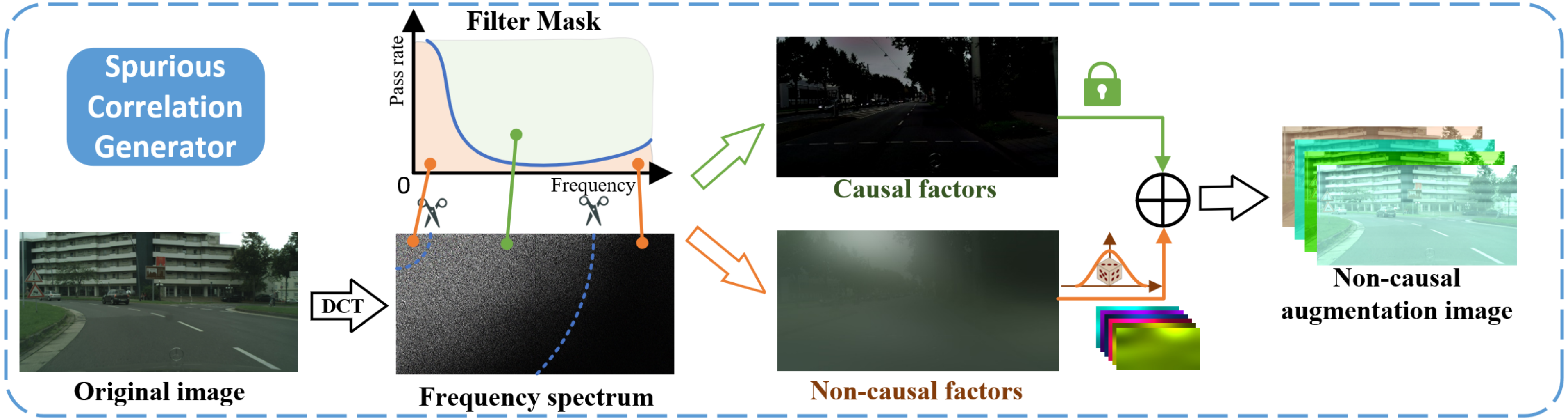
① Using DCT to obtain the spectral information of an image.

$$\mathbf{S} = \mathcal{F}(\mathbf{X})$$

② Separating the causal and non-causal components of an image using a bandpass filter.

$$\mathcal{M}(r) = e^{-\frac{u^2+v^2}{2r_H^2}} - e^{-\frac{u^2+v^2}{2r_L^2}}$$

$$\begin{cases} \mathbf{S}_{cau} = \mathcal{M}(\mathrm{r}) \cdot \mathcal{F}(\mathbf{X}) & \text{Causal} \\ \mathbf{S}_{non} = \left(1 - \mathcal{M}(\mathrm{r})\right) \cdot \mathcal{F}(\mathbf{X}) & \text{Non-causal} \end{cases}$$



* Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6891–6902, 2021.

③ Keeping the causal component unchanged, randomly modify the non-causal component.

$$R_G(\mathbf{S}_{non}) = \sum_{c=1}^{C} \mathbf{S}_{non}^{c} \cdot \left(1 + \underbrace{\mathcal{N}(0,1)}_{\text{Gaussian random}}\right)$$

④ Using IDCT $\mathcal{F}'(\cdot)$ to transform the enhanced image spectrum back to the spatial image $\widehat{\mathbf{X}}$.

$$\widehat{\mathbf{X}} = \mathcal{F}'\left(R_G\left(\left(1 - \mathcal{M}(\mathrm{r})\right) \cdot \mathcal{F}(\mathbf{X})\right) + \mathcal{M}(\mathrm{r}) \cdot \mathcal{F}(\mathbf{X})\right)$$

Differences between SCG and previous domain augmentation:

① Using a single domain without reference images.

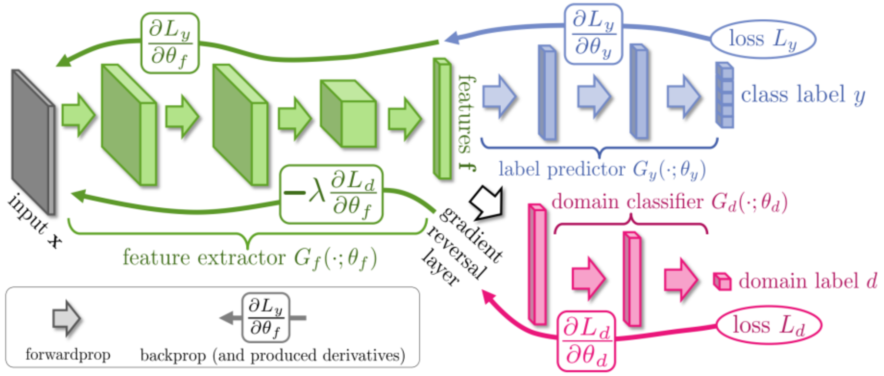
② Random augmentation can add more non-causal factors.

Images generated by SCG

Multi-View Domain Classifier (MVDC)

Single-view domain adversarial learning (DAL)

Domain differences : $d_{\mathcal{H}}(D_{s1}, D_{s2}) = 2\left(1 - 2\min_{h\in\mathcal{H}}\left(err\left(h(\mathbf{X}_{s1})\right) + err\left(h(\mathbf{X}_{s2})\right)\right)\right)$



Minimize $d_{\mathcal{H}}$: $\min_{\mathcal{F}} d_{\mathcal{H}}(D_{s1}, D_{s2})$

$$\Rightarrow \underbrace{\max_{\mathcal{F}} \min_{h\in\mathcal{H}} err\left(h(\mathbf{S})\right)}_{Standard\ DAL}$$

When maximizing and minimizing are **balanced**, the optimization is completed.

MVDC disrupts the balance

Feature extractor $\xLeftrightarrow{\text{balance}}$ ~~Single-view~~ Multi-view domain classifier

$$\min_{\mathcal{F}} d_{\mathcal{H}}(D_{s1}, D_{s2}) \Rightarrow \underbrace{\max_{\mathcal{F}} \sum_{i=1}^{M} \min_{h_i\in\mathcal{H}, e_i} err\left(h_i\left(e_i(\mathbf{S})\right)\right)}_{\text{Our MAD}}$$

~~Single domain classifier $h(\cdot)$~~

Multiple sets of AutoEncoders$\langle e_i(\cdot) | g_i(\cdot)\rangle$ & domain classifier $h_i(\cdot)$

The loss of each view

① Reconstruction loss $\mathcal{L}_{RC}$

$$\mathcal{L}_{RC}=\frac{1}{M}\sum_{m=1}^{M} MSE\left(s,g_m\left(e_m(\mathbf{S})\right)\right)$$

Ensuring the mapped features contain complete semantic information.

② Domain classifier loss $\mathcal{L}_{DC}$

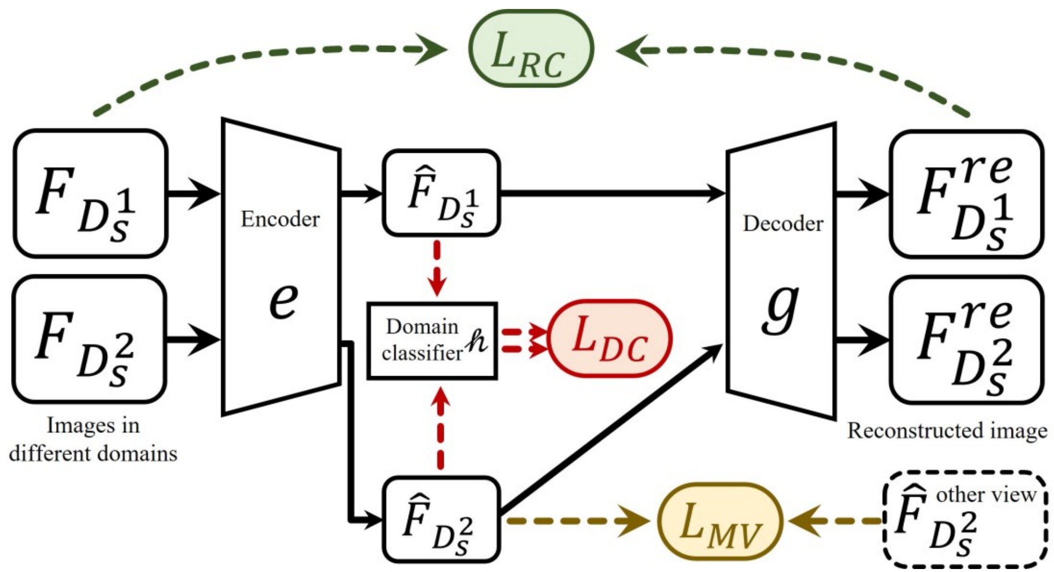
$$\mathcal{L}_{DC}=-\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} y_k\cdot log\left(p\left(D_m\left(e_m(\mathbf{S}_k)\right)\right)\right)$$

Ensuring that different features of the same view have domain discriminability.

③ View-different loss $\mathcal{L}_{MV}$

$$\mathcal{L}_{MV}=-\frac{\sum_i^M\sum_{j,i\neq j}^M\left\|e_i(\mathbf{S})-e_j(\mathbf{S})\right\|^2}{M^2-M}$$

Ensuring that each AutoEncoder maps features to different latent spaces.



Structure of one branch of MVDC

The total loss of MVDC :

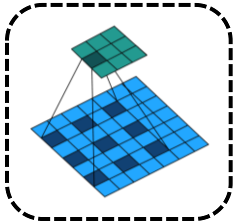
$$\mathcal{L}_{MVDC}^{(img,ins)}=\mathcal{L}_{RC}+\mathcal{L}_{DC}+\mathcal{L}_{MV}$$

Each level in object detection

Image level

Dilated convolutional layers

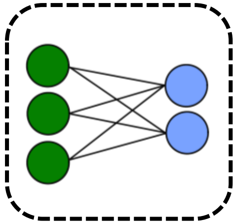
[Global Non-causal Factors of Images]

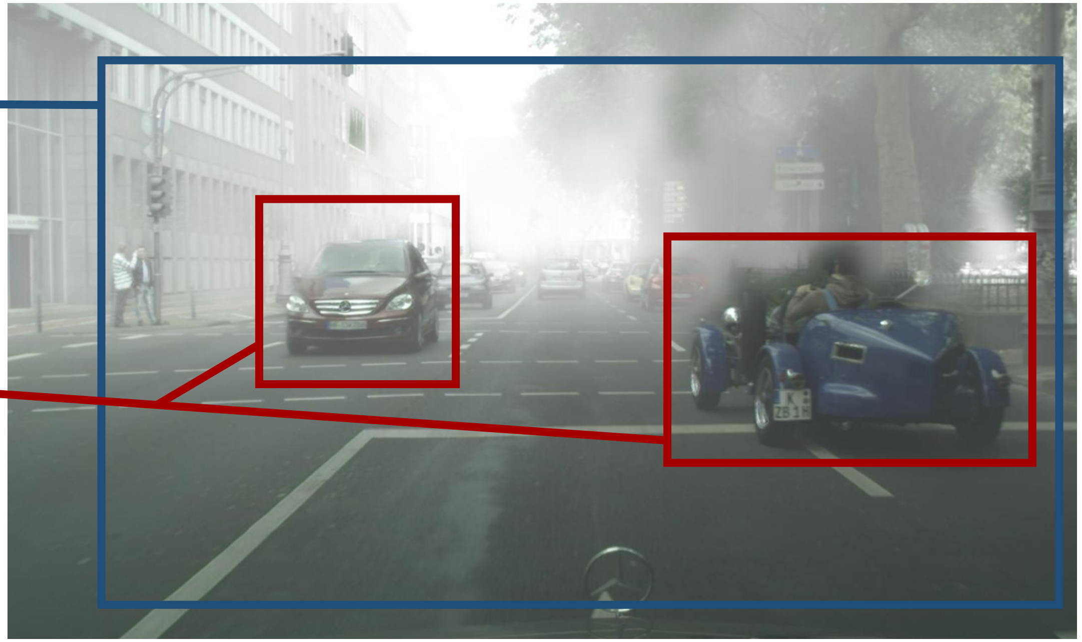


Instance level

Fully connected layers

[Non-causal factors of instances]





Consistency loss

$$\mathcal{L}_{cst} = \sum_{i,j}^{M}\sum_{n}^{N}\left\|\frac{1}{|I|}\sum_{u,v} p_i^{(u,v)} - p_{j,n}\right\|_2$$

Ensured consistency between classifiers in different level.

Overall loss of MAD

$$\mathcal{L}_{MAD} = \mathcal{L}_{det} + \lambda\left(\mathcal{L}_{MVDC}^{img} + \mathcal{L}_{MVDC}^{ins} + \mathcal{L}_{cst}\right)$$

Detection loss
(Classification & Regression)

MAD loss

Experimental setup

Datasets

7 benchmark datasets for object detection with distinctive characteristics.

Cityscapes	Foggy Cityscapes	Rain Cityscapes	BDD100k
KITTI	SIM 10k	PASCAL VOC	

Setting

Input resolution : Short edge adjusted to 600 pixels (aspect ratio unchanged)

Training epochs : 10 epochs

Learning rate : 2×10^{-3}, reduced to 2×10^{-4} after the 7th epoch

Optimization method : Stochastic gradient descent

Framework & Equipment

- PyTorch + NVIDIA TITAN XP GPU
- Mindspore + Ascend 910 computing core

Baseline Two-stage **FasterRCNN** framework, The backbone is **VGG16** pre-trained on **ImageNet**.

Validity Verification

Results on four domains (C, F, R, B) trained on single source domain

Source \ Target	Method	Cityscapes	Foggy Cityscapes	Rain Cityscapes	BDD100k
Cityscapes	Source-only	—	27.2	36.3	24.0
	MLDG	—	29.2	42.1	21.0
	FACT	—	25.3	39.9	26.0
	FSDR	—	31.0	**42.8**	26.2
	DANN+**SCG**	—	37.5	39.1	26.1
	MAD(Ours)	—	**38.6**	42.3	**28.0**
Foggy Cityscapes	Source-only	29.9	—	38.4	17.5
	MLDG	30.4	—	38.6	18.0
	FACT	30.0	—	38.7	20.2
	FSDR	31.3	—	40.8	20.4
	DANN+**SCG**	38.4	—	40.4	22.4
	MAD(Ours)	**41.3**	—	**43.3**	**24.4**
BDD100k	Source-only	33.6	27.2	34.3	—
	MLDG	24.7	17.1	20.0	—
	FACT	32.4	24.3	33.9	—
	FSDR	32.4	27.8	34.7	—
	DANN+**SCG**	35.8	29.3	33.9	—
	MAD(Ours)	**36.4**	**30.3**	**36.1**	—

MAD can achieve better results in most cross-domain scenarios

Comparison with existing methods

Methods		Dataset used	person	rider	car	truck	bus	train	motor	bike	mAP
Source-only		Single Source	27.1	39.3	36.0	14.2	31.4	9.4	26.9	33.4	27.2
DA	DAF [6]	Single Source & Target images (without labels)	31.6	43.6	42.8	23.6	41.3	21.2	28.9	32.6	33.2
	SW-DA [34]		31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
	SC-DA [52]		33.8	42.1	52.1	26.8	42.5	26.5	29.2	34.5	35.9
	MTOR [3]		30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
	ICR-CCR [43]		32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
	Coarse-to-Fine [48]		34.0	**46.9**	52.1	**30.8**	43.2	29.9	34.7	37.4	38.6
	GPA [44]		32.9	46.7	54.1	24.7	**45.7**	**41.1**	**32.4**	**38.7**	39.5
	Center-Aware [17]		**41.5**	43.6	**57.1**	29.4	44.9	39.7	29.0	36.1	**40.2**
DG	DIDN [23]	Multiple Source	31.8	38.4	**49.3**	**27.7**	35.7	26.5	24.8	33.1	33.4
	LMDG [21]	Single Source	32.2	41.7	38.9	19.2	33.0	9.1	23.5	36.3	29.2
	FACT [45]		26.2	41.2	35.9	13.6	27.7	3.0	23.3	31.3	25.3
	FSDR [19]		31.2	44.4	43.3	19.3	36.6	11.9	27.1	34.1	31.0
	MAD		**34.2**	**47.4**	45.0	25.6	**44.0**	**42.4**	**30.28**	**40.12**	**38.6**
Oracle - Train on target		Target	37.8	47.4	53.0	31.6	52.9	34.3	37.0	40.6	41.8

① MAD achieves the best performance among domain generalization object detection methods.

② MAD even surpasses some of the traditional Domain Adaptation methods.

Universal validation

The generalization ability of category "car"

C → {F, R, B, V, S, K}

Method	F	R	B	V	S	K
SourceOnly	36.0	39.0	41.3	62.0	39.2	73.4
DAF	42.8	52.9	41.4	59.2	39.0	72.1
MLDG	38.9	52.7	39.4	61.4	37.2	63.9
FACT	35.9	48.8	42.0	65.3	41.2	73.2
FSDR	43.3	52.7	**45.4**	63.4	42.2	73.8
MAD	**45.0**	**54.0**	42.4	**67.6**	**43.2**	**74.1**

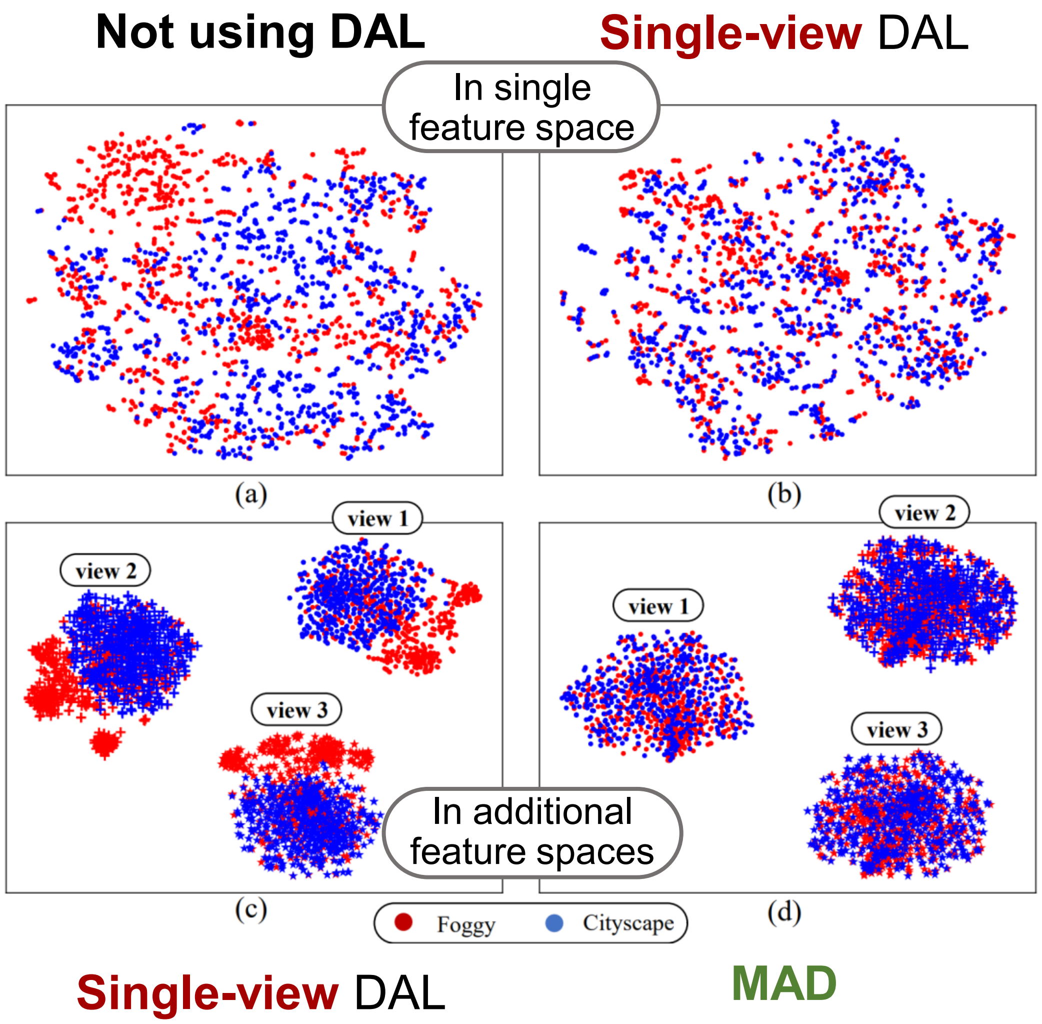
MAD exhibits generalization ability in **a wider range of domains**.

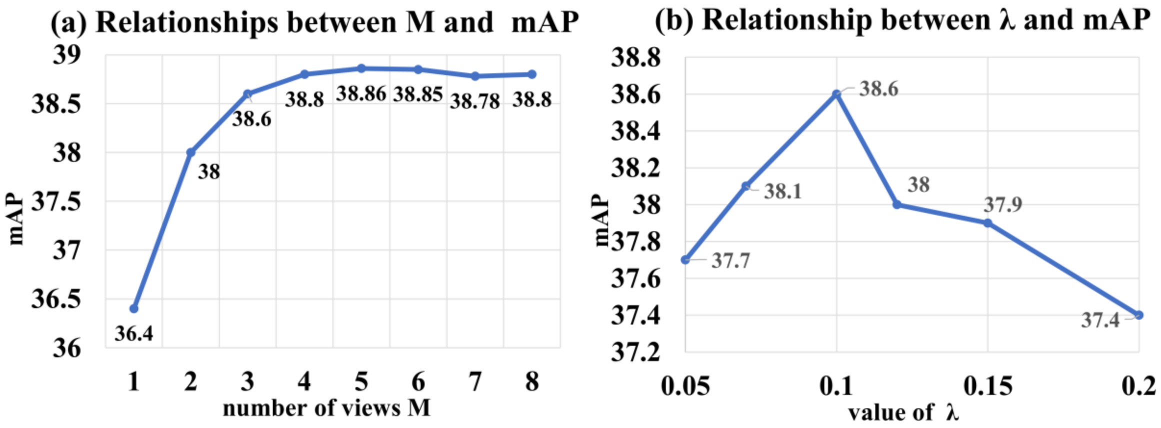
Testing MAD on Categorical Datasets

Source	Target															
	ERM				ERM+SCG				DANN+SCG				MVDC+SCG (MAD)			
PACS	P	A	C	S	P	A	C	S	P	A	C	S	P	A	C	S
P	-	61.9	26.2	31.9	-	62.8	29.3	40.1	-	63.1	35.3	43.1	-	**66.6**	**40.9**	**44.2**
A	90.6	-	67.3	57.2	90.8	-	68.7	61.7	91.4	-	70.7	64.3	**92.6**	-	**71.2**	**68.9**
C	79.5	64.1	-	65.6	78.6	64.3	-	69.0	79.2	63.6	-	69.3	**79.9**	**64.6**	-	**70.9**
S	48.0	42.8	60.5	-	49.4	51.5	62.2	-	48.7	53.8	63.4	-	**53.2**	**57.4**	**63.8**	-
VLCS	V	L	C	S	V	L	C	S	V	L	C	S	V	L	C	S
V	-	39.6	96.1	68.9	-	40.1	97.6	69.2	-	43.4	98.3	69.5	-	**47.2**	**98.5**	**71.4**
L	61.3	-	82.6	43.8	61.7	-	83.7	46.9	61.7	-	83.7	46.9	**62.2**	-	**86.7**	**51.8**
C	50.6	20.7	-	42.7	51.2	21.9	-	43.5	51.7	27.2	-	44.9	**51.8**	**29.6**	-	**46.0**
S	60.2	45.5	72.7	-	60.9	47.4	72.9	-	62.4	50.0	74.9	-	**64.0**	**51.3**	**75.4**	-

MAD is also effective in **classification tasks**.

Feature visualization & Hyperparameter analysis





① mAP increases with the number of view M. Convergence occurs when $M > 3$.

② When the loss balancing factor $\lambda = 0.1$, the network performance is optimal.

Learning Intelligence & Vision Essential (LiVE) Group

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Multi-view Adversarial Discriminator: Mine the Non-causal Factors for Object Detection in Unseen Domains

Paper

[2304.02950] Multi-view Adversarial Discriminator: Mine the Non-causal Factors for Object Detection in Unseen Domains (arxiv.org)

Code

K2OKOH/MAD (github.com)

E-mail: mingjunxu@cqu.edu.cn