Modality-Agnostic Debiasing for Single Domain Generalization

Sanqing Qu, Yingwei Pan, Guang Chen (✉), Ting Yao, Changjun Jiang, Tao Mei

Tongji University; HiDream.ai Inc

{2011444, guangchen, cjjiang}@tongji.edu.cn, {panyw.ustc, tingyao.ustc}@gmail.com, tmei@hidream.ai

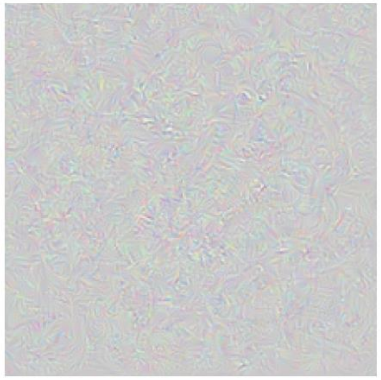
同济大学 TONGJI UNIVERSITY

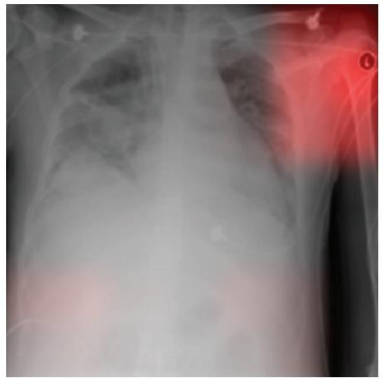
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

❑ Quick Preview

◆ Deep neural networks (DNNs) usually fail to generalize well to outside of distribution (OOD) data, due to the inherent short-cut learning phenomenon.





Article: Super Bowl 50

Paragraph: "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

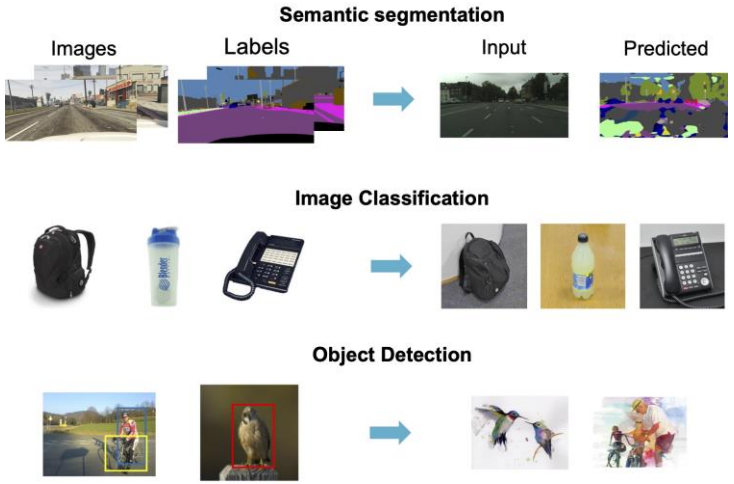
Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

**Credit to Short-cut Learning [Nature Machine Intelligence, 2020]*

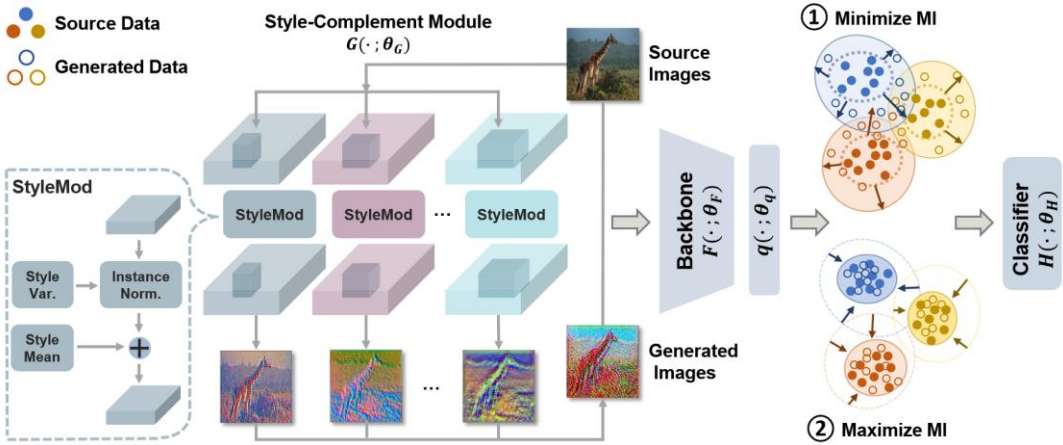
Introduction

❑ Quick Preview

- ◆ Domain generalization (DG), especially single domain generalization (single-DG) has become a promising technique to address the Out-of-distribution (OOD) issue of DNNs.
- ◆ Existing single-DG methods commonly devise various data-augmentation algorithms, typically being only applicable to one single modality (e.g., image).



*



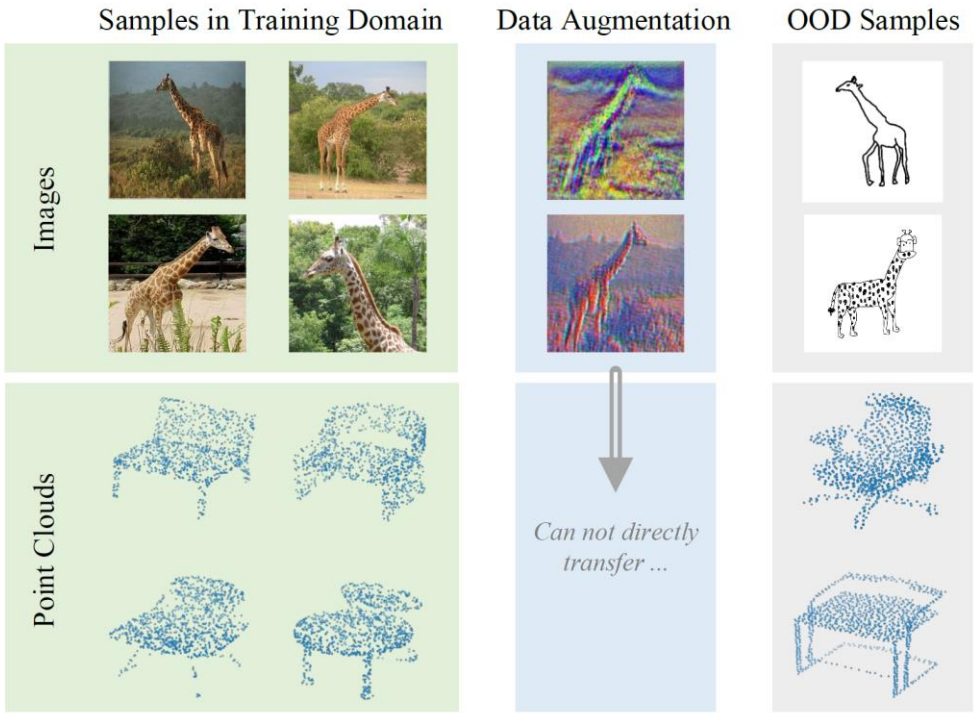
+

Credit to DANCE [NeurIPS-20]; + L2D [ICCV-21]

Introduction

❑ Quick Preview

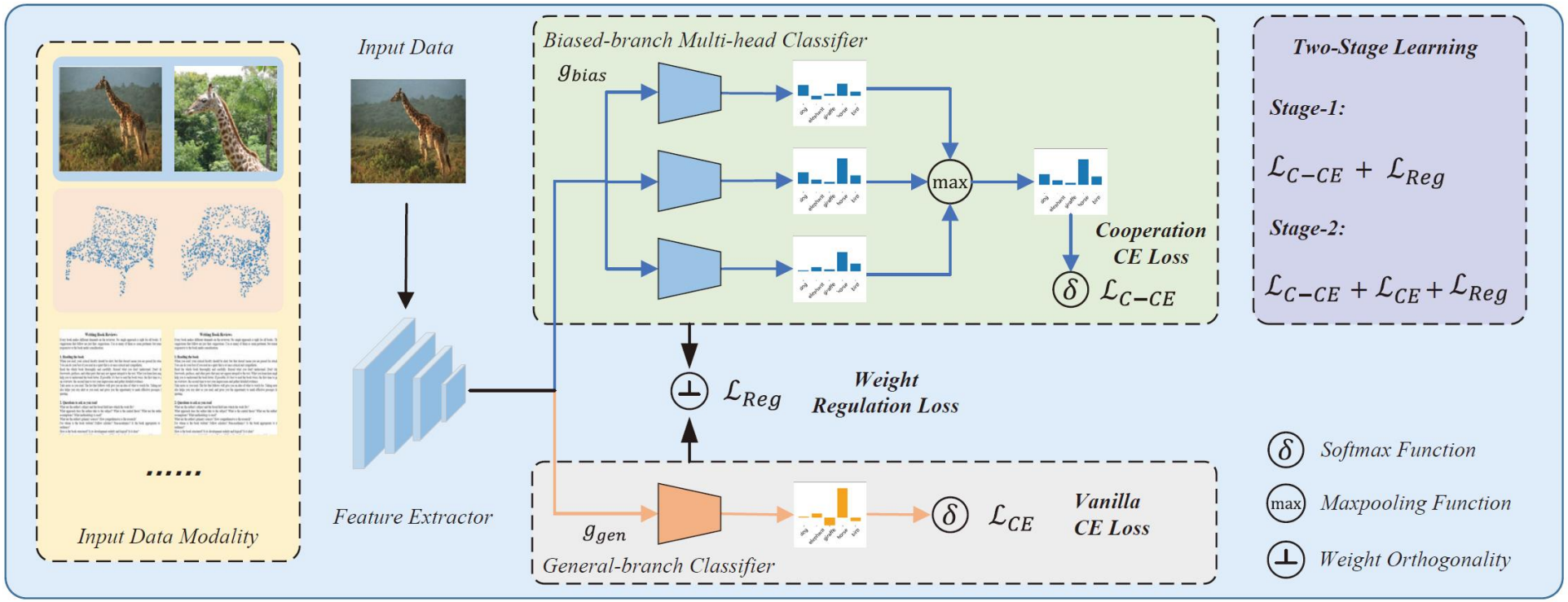
- ◆ Existing single-DG methods commonly devise various data-augmentation algorithms, typically being only applicable to one single modality (e.g., image).
- ◆ In this paper, we target a versatile Modality-Agnostic Debiasing (MAD) framework for single-DG.



Methodology

❑ Modality-Agnostic Debiasing (MAD)

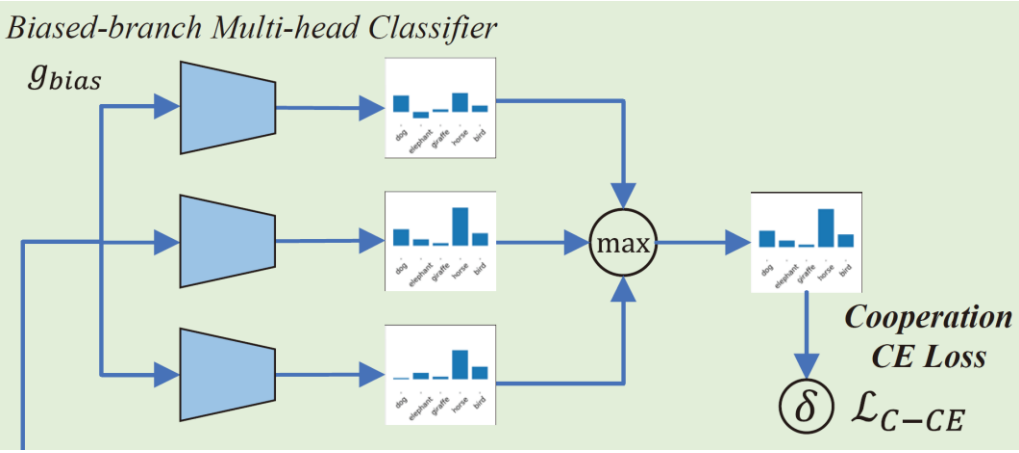
- ◆ **Two-branch framework:** Biased-branch aims to identify domain-specific features with the Cooperation CE Loss.
- ◆ **Two-stage learning mechanism:** With the information from biased-branch, the general branch is encouraged to learn domain-generalized features.



Methodology

❑ Multi-head Biased-branch

- **Multi-head design:** The weights of vanilla trained classifier can be considered as an indicator of domain-specific features, due to the inherent short-cut learning symptom. There commonly exist multiple factors that contribute to domain-specific features, e.g., background contexts, texture of objects, or other high-frequency patterns.
- **Cooperation Cross-Entropy Loss:** Domain-specific features do not represent truly domain-generalized semantics. Thereby, we do not enforce all heads of the biased-branch to correctly predict each sample. Instead, we only need one of them to accurately identify it.



$$\mathcal{L}_{C-CE} = \mathbb{E}_{x,y} \sum_{k=1}^{K} -\mathbb{1}_{[k=y]} \log \frac{\exp(\max(v_k(x)))}{\sum_{j=1}^{K} \exp(\max(v_j(x)))},$$

Methodology

❑ Single-head General-branch

- **Learning to debias:** We introduce a general-branch to capture those domain-generalized features with the guidance from the biased-branch.
- **Two-stage learning mechanism:** An intuitive solution is to directly enforce orthogonality between the biased-branch and the general-branch. However, since the biased-branch lacks knowledge of domain-specific features at the beginning. There is no guarantee that general-branch will pay more attention to those domain-generalized. We introduce a two-stage learning mechanism.

$$\mathcal{L}_{C-CE} = \mathbb{E}_{x,y} \sum_{k=1}^{K} -\mathbb{1}_{[k=y]} \log \frac{\exp(\max(v_k(x)))}{\sum_{j=1}^{K} \exp(\max(v_j(x)))},$$

$$\mathcal{L}_{CE} = \mathbb{E}_{x,y} \sum_{k=1}^{K} -\mathbb{1}_{[k=y]} \log \frac{\exp(u_k(x))}{\sum_{j=1}^{K} \exp(u_j(x))},$$

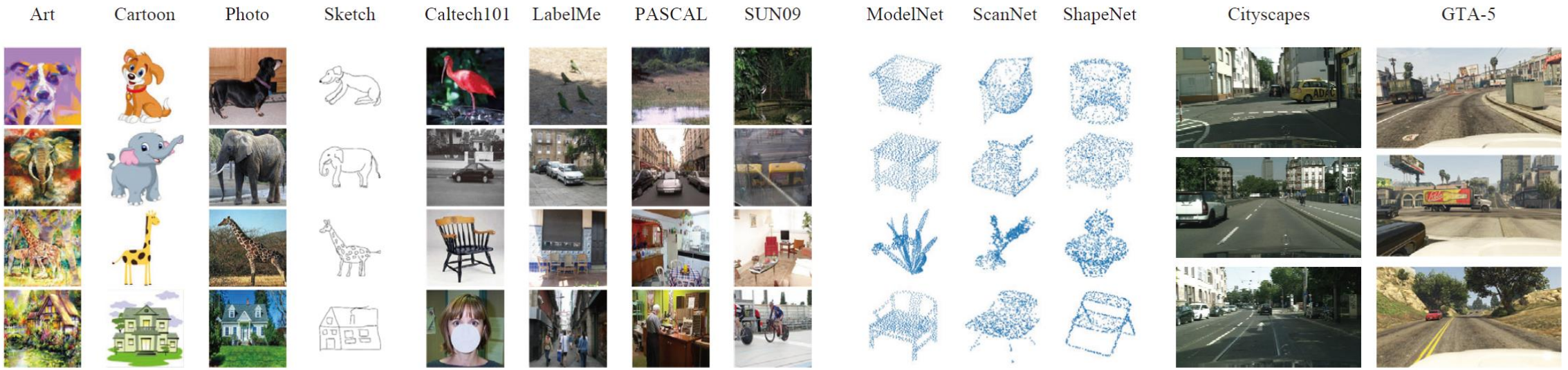
$$\mathcal{L}_{Reg} = \frac{1}{K} \sum_{k=1}^{K} \left\| W_{bias}[k,:] \, W_{gen}[k,:]^T \right\|_F^2.$$

$$\min_{f, g_{bias}, g_{gen}} \mathcal{L}_{C-CE} + \mathcal{L}_{Reg} + \mathbb{1}_{[pro \geq T]} \cdot \mathcal{L}_{CE},$$

Experiments

❑ Setup

- ◆ **Datasets:** Amazon Reviews (1D texts), PACS (2D images), VLCS (2D images), PointDA-10 (3D points); GTA-5 → Cityscapes (2D semantic segmentation).
- ◆ **Baselines:** ERM, AugMix, pAdaIN, Mixstyle, DSU, ACVC.



(a) PACS (b) VLCS (c) PointDA-10 (d) GTA-5 → Cityscapes

Experiments

❑ Results

◆ Recognition on 1D Texts, and 2D Images.

Single-DG results on 1D texts (Amazon-Review dataset).

Methods	Venue	D	E	K	B	Avg
ERM	-	74.17	73.17	73.67	71.58	73.15
ERM w/ MAD		76.08	74.33	73.33	74.67	**74.60**
Mixup [61]	ICLR 18	74.83	72.17	73.58	72.67	73.31
Mixup w/ MAD		75.33	73.58	74.33	73.75	**74.25**
Mixstyle [65]	ICLR 21	74.75	73.17	74.33	72.33	73.65
Mixstyle w/ MAD		75.17	72.75	75.00	75.25	**74.54**
DSU [30]	ICLR 22	75.00	73.45	75.25	73.08	74.20
DSU w/ MAD		76.42	74.33	76.50	75.17	**75.60**

Single-DG results on 2D images (PACS dataset).

Methods	Venue	P	A	C	S	Avg
ERM		33.65	65.38	64.20	34.15	49.34
ERM w/ MAD		32.32	66.47	69.80	34.54	**50.78**
Augmix [24]	ICLR 19	38.30	66.54	70.16	52.48	56.87
Augmixw/ MAD		36.19	68.04	73.11	54.44	**57.94**
pAdaIn [34]	CVPR 21	33.66	64.96	65.24	32.04	48.98
pAdaIn w/ MAD		34.66	65.64	70.10	42.85	**53.31**
Mixstyle [65]	ICLR 21	37.44	67.60	70.38	34.57	52.50
Mixstyle w/ MAD		41.57	69.88	71.61	41.58	**56.16**
ACVC [15]	CVPR 22	48.05	73.68	77.39	55.30	63.61
ACVC w/ MAD		52.95	75.51	77.25	57.75	**65.87**

Experiments

❑ Results

◆ Recognition on 2D Images, and 3D Point Clouds

Single-DG results on 2D images (VLCS dataset).

Methods	Venue	V	L	C	S	Avg
ERM		76.72	58.86	44.95	57.71	59.56
ERM w/ MAD		76.21	67.97	46.55	61.04	**62.95**
Augmix [24]	ICLR 19	75.25	59.52	45.90	57.43	59.53
Augmixw/ MAD		76.57	65.60	44.35	59.47	**61.50**
pAdaIn [34]	CVPR 21	76.03	65.21	43.17	57.94	60.59
pAdaIn w/ MAD		76.57	68.90	42.92	63.91	**63.08**
Mixstyle [65]	ICLR 21	75.73	61.29	44.66	56.57	59.56
Mixstyle w/ MAD		75.00	66.17	43.61	62.01	**61.70**
ACVC [15]	CVPR 22	76.15	61.23	47.43	60.18	61.25
ACVC w/ MAD		76.15	69.36	48.04	61.74	**63.82**
DSU [30]	ICLR 22	76.93	69.20	46.54	58.36	62.76
DSU w/ MAD		76.99	70.85	44.78	62.23	**63.71**

Single-DG results on 3D point cloud (PointDA-10 dataset).

Methods	Venue	SH	SC	M	Avg
ERM	-	25.69	45.09	32.94	34.57
ERM w/ MAD		31.11	48.07	34.69	**37.91**
Mixstyle [65]	ICLR 21	27.18	46.25	27.93	33.78
Mixstyle w/ MAD		29.89	51.01	33.57	**38.16**
DSU [30]	ICLR 22	25.74	43.53	31.61	33.63
DSU w/ MAD		28.92	47.69	32.72	**36.45**

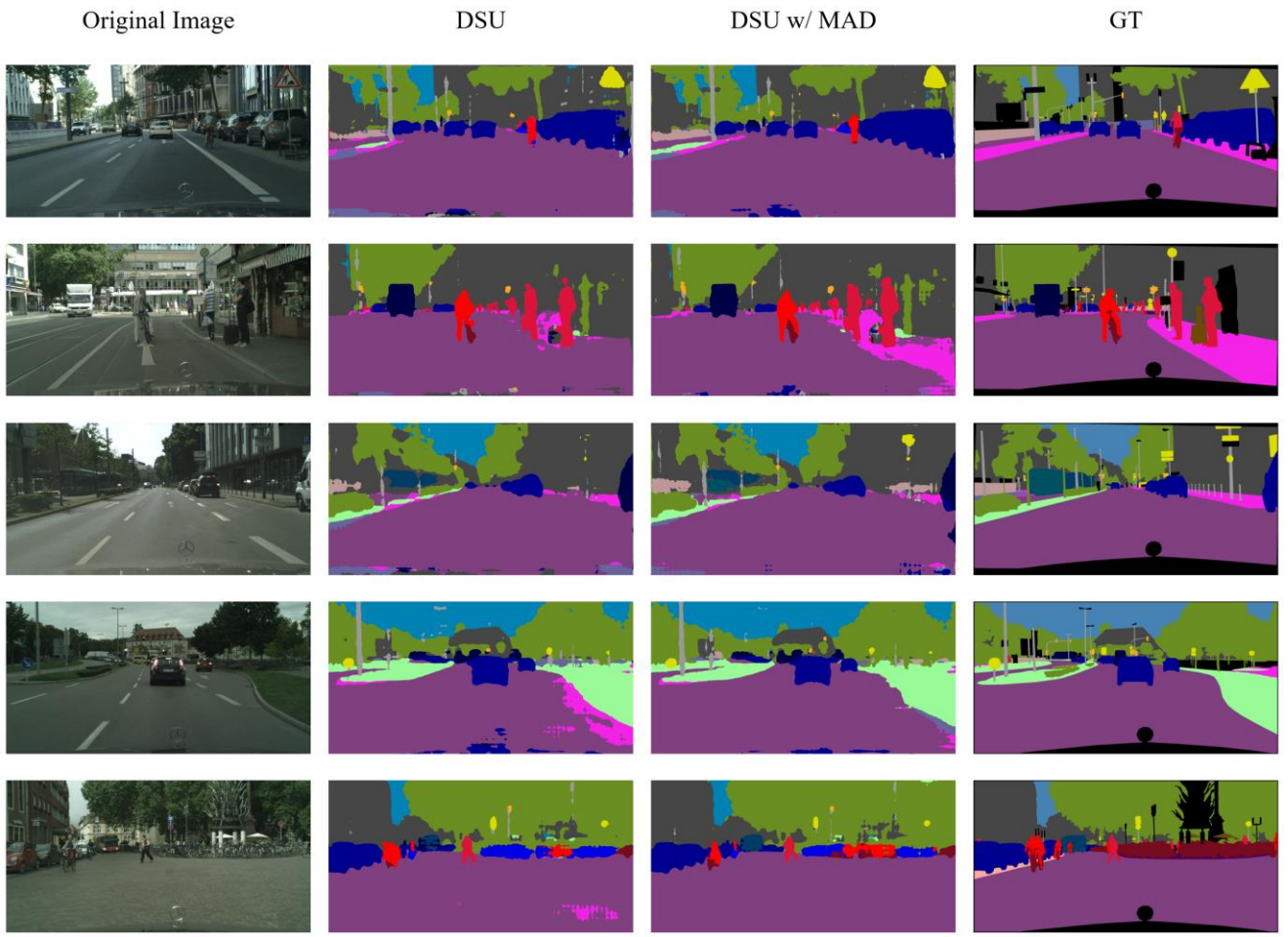
Experiments

❑ Results

◆ Semantic Segmentation on 2D Images

Table 5. Single-domain generalization on semantic segmentation (GTA-5 → Cityscapes).

Methods	Venue	mIOU(%)	mACC(%)
ERM	-	37.0	51.5
pAdaIN [34]	CVPR 21	38.3	52.1
Mixstyle [65]	ICLR 21	40.3	53.8
DSU [30]	ICLR 22	42.3	54.7
ERM w/ MAD	-	**38.9**	**52.2**
DSU w/ MAD	-	**43.8**	**57.2**



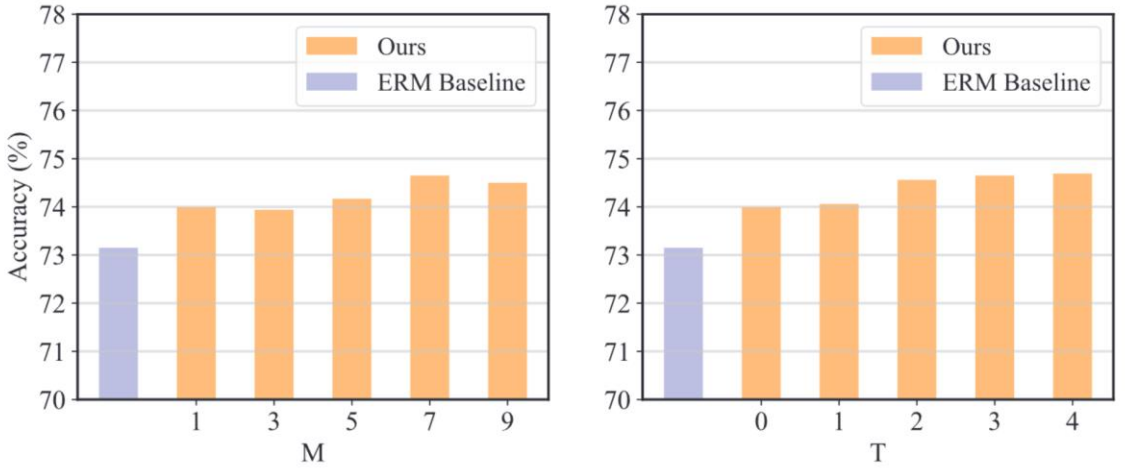
Experiments

❑ Ablation Study and Hyper-parameter Analysis

- ◆ Both the multi-head classifier design for biased-branch and two-stage learning mechanism are effective.
- ◆ MAD is not sensitive to hyper-parameter selection.

Table 6. **Ablation**. Results of the vanilla ERM, ERM w/ MAD (one-stage), ERM w/ MAD (single-head), and ERM w/ MAD. The experiments are conducted on single-domain generalization scenarios of 1D Texts (Amazon Review dataset), 2D Images (VLCS dataset), and 3D Point Clouds (PointDA-10 dataset).

Methods	1D Texts	2D Images	3D Points	Avg
ERM	73.15	59.56	34.57	55.76
MAD (one-stage)	74.00	60.41	35.67	56.69
MAD (single-head)	74.31	60.49	36.61	57.14
MAD	**74.60**	**62.95**	**37.91**	**58.49**



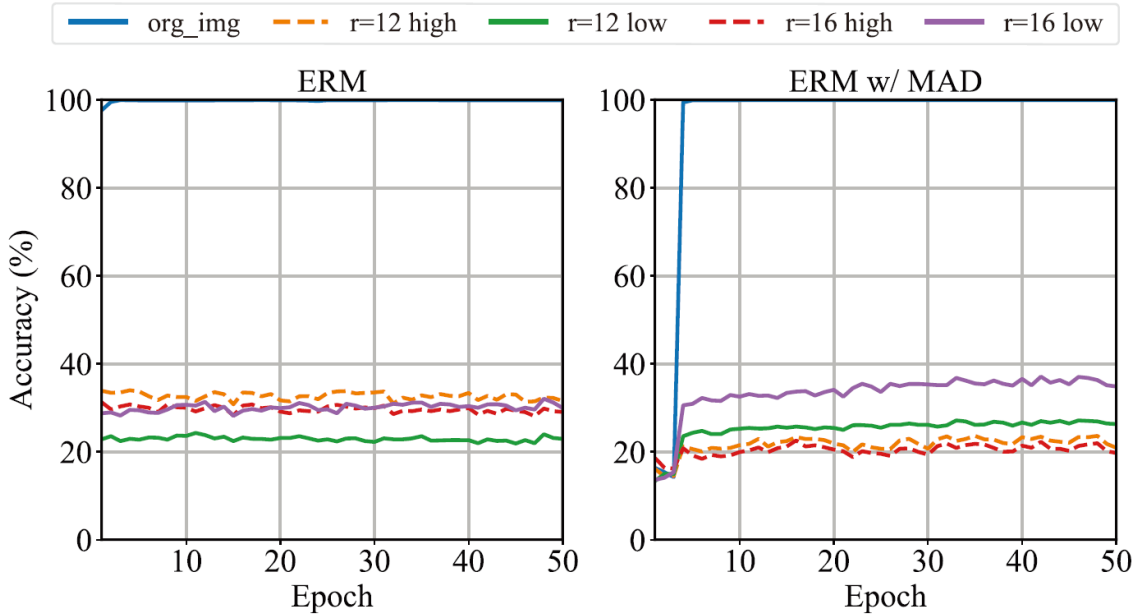
(a) sensitivity to M (T = 3) (b) sensitivity to T (M = 7)

Figure 3. Hyper-parameter sensitivity analysis on Single-DG text classification. M denotes the number of the biased-branch classifiers and T is the training threshold for the second stage.

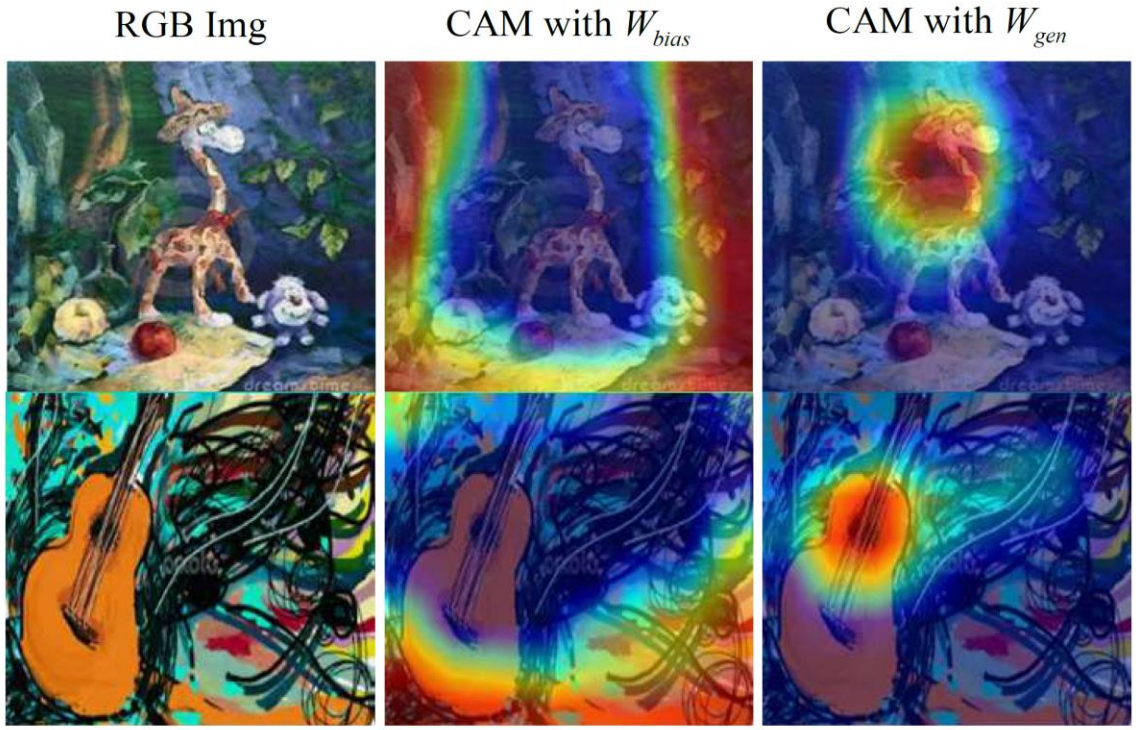
Experiments

❑ Qualitative Analysis

- ◆ MAD indeed encourages the general-branch to pay more attention to those domain-generalized features.



(a) LFC vs. HFC in "Photo" domain on PACS



Conclusion

❑ Conclusion

- Different from existing methods that introduce modality-specific data augmentation techniques, we propose a general and versatile modality-agnostic debiasing (MAD) framework for single-DG.
- We have devised a novel two-branch classifier, where a biased-branch is responsible for identifying those domain-specific (superficial) features, while the general-branch is encouraged to focus more on those domain-generalized (semantic) features.
- MAD is appealing in view that it can be seamlessly incorporated into existing methods.
- MAD does not increase the computational cost. It discards the biased-branch and only employs the feature extractor plus general-branch during inference.

Thanks for you watching!

- If you have any further question or require any further information, please feel free to contact me. Email: 2011444@tongji.edu.cn
- **Ad:** We have published a paper on universal domain adaptation titled “Upcycling Models under Domain and Category Shift” in CVPR-2023.