Accepted at WACV'22

COCOA: Context-Conditional Adaptation for Recognizing Unseen Classes in Unseen Domains

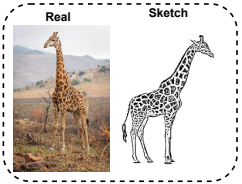
Puneet Mangla*[1], Shivam Chandhok*[2], Vineeth N Balasubramanian[1], Fahad Shahbaz Khan[2]

Indian Institute of Technology, Hyderabad, India[1]
Mohamed bin Zayed University of Artificial Intelligence, UAE[2]

Problem Setting

Zero-Shot Domain Generalization (ZSL-DG):

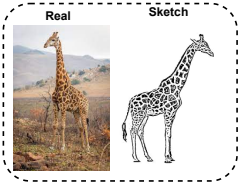
- **Training set** - $Q^{Tr} = \{(\mathbf{x}, y, \mathbf{a}_y, d) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}^s, \mathbf{a}_y \in \mathcal{A}, d \in \mathcal{D}^s\}$ where $\mathcal{X}$ is the visual space, $\mathcal{Y}^s$ set seen class labels, $\mathcal{D}^s$ represents the set of seen domains and $\mathbf{a}_y$ denotes the class-specific semantic representation.



Problem Setting

Zero-Shot Domain Generalization (ZSL-DG):

- **Test set** - $Q^{Ts} = \{(\mathbf{x}, y, \mathbf{a}_y, d) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}^u, \mathbf{a}_y \in \mathcal{A}, d \in \mathcal{D}^u\}$ where $\mathcal{Y}^u$ is the set of labels for unseen classes and $\mathcal{D}^u$ represents the set of unseen domains

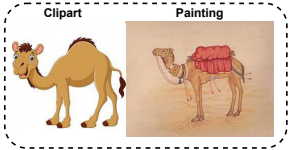


Giraffe

- Long neck
- Long legs
- Has spots

Horse

- Short neck
- Short legs
- No spots



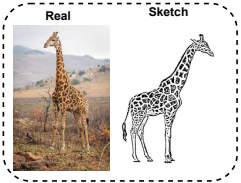
Camel

- Short neck
- Long legs
- No spots

Problem Setting

Zero-Shot Domain Generalization (ZSL-DG):

- **Goal** - Recognize unseen classes in unseen domains without having seen these novel classes and domains during training, i.e, $\mathcal{Y}^s \cap \mathcal{Y}^u \equiv \emptyset$ and $\mathcal{D}^s \cap \mathcal{D}^u \equiv \emptyset$.

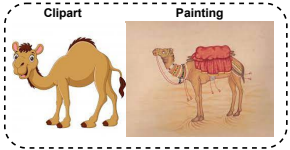


Giraffe

- Long neck
- Long legs
- Has spots

Horse

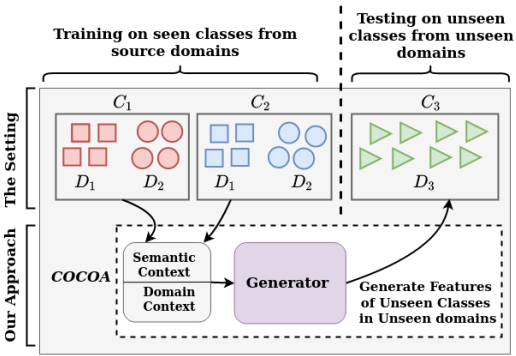
- Short neck
- Short legs
- No spots



Camel

- Short neck
- Long legs
- No spots

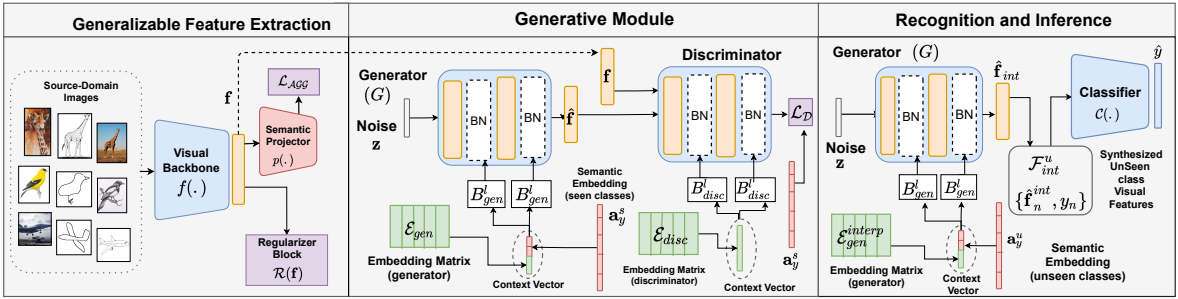
Methodology



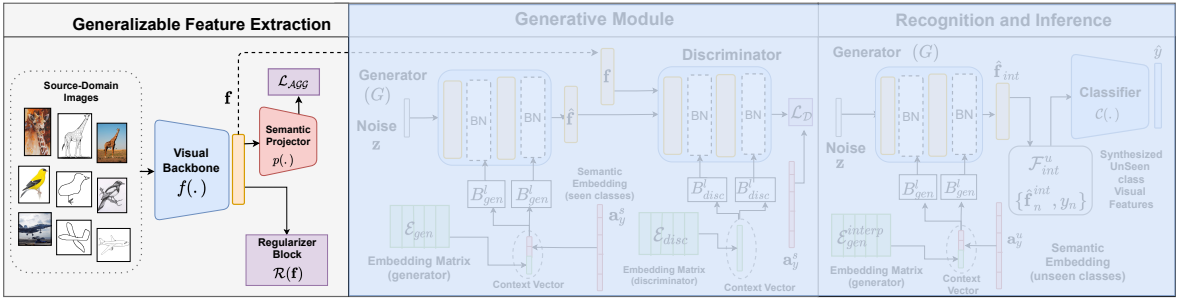
Our Approach

- Encoding semantic and domain context using COntext COnditional Adaptive (COCOA) Batch-Normalization layer helps to instill semantic and domain information into generated features.
- By jointly fusing both semantic and domain context, our method achieve best generalization to unseen classes in unseen domains at test-time.

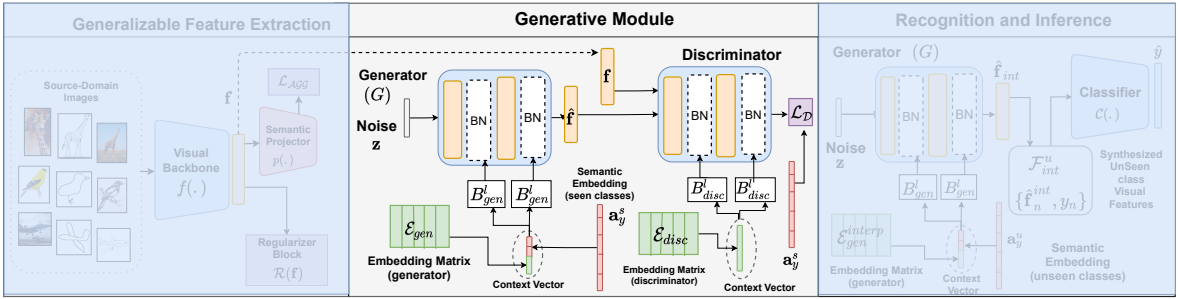
Model Architecture



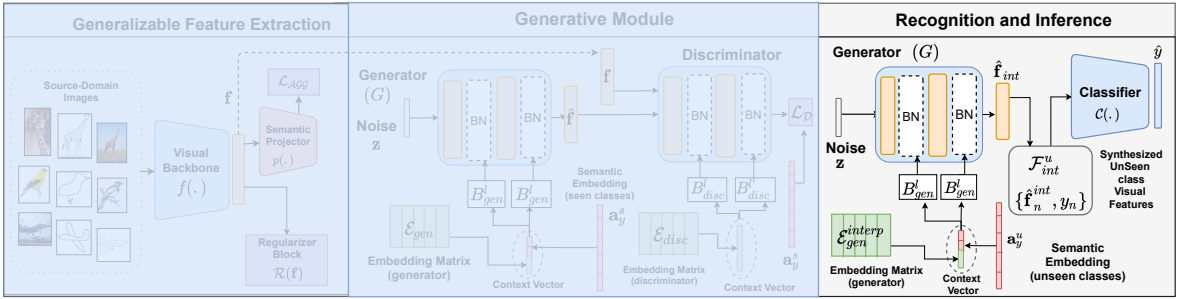
Model Architecture



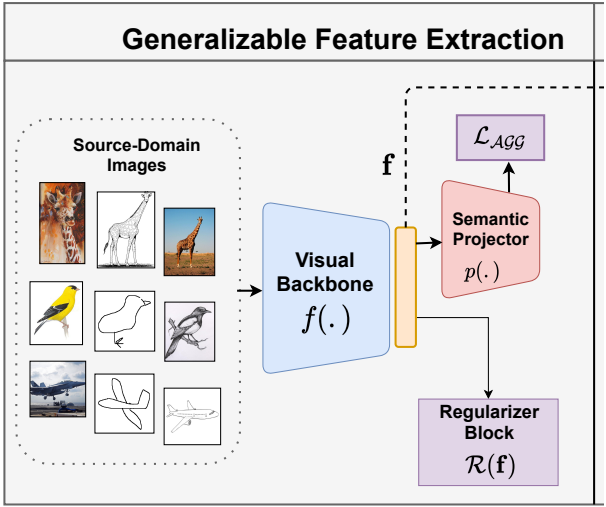
Model Architecture



Model Architecture



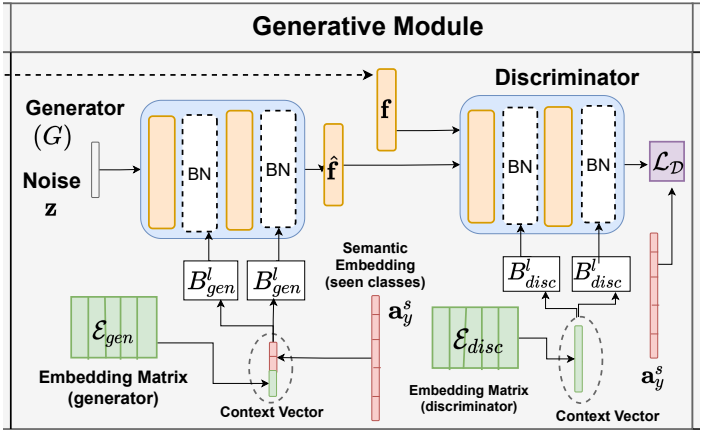
Methodology



Stage 1

- Class-level semantic (domain-invariant) and domain-specific information help improve generalization
- Visual encoder encodes discriminative information in features $\mathbf{f}$ by employing $\mathcal{L}_R = \mathcal{L}_{AGG} + \mathcal{R}(\mathbf{f})$

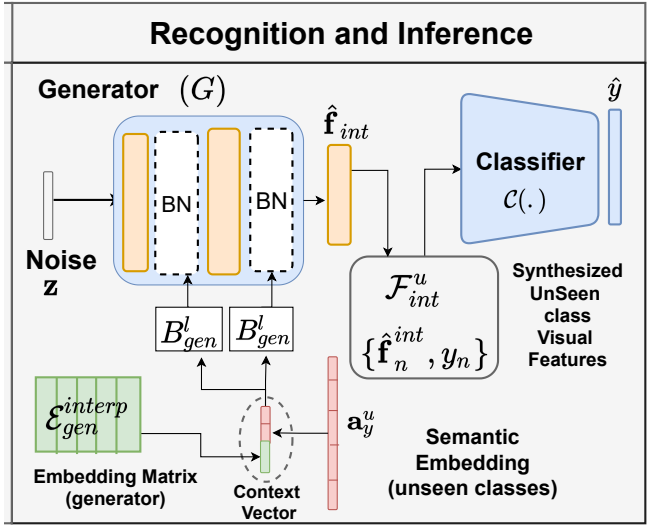
Methodology



Stage 2

- Generative model G generates features $\hat{\mathbf{f}}$ with the help of COntext COnditional Adaptation (COCOA)
- Enables model to capture both class-level and domain-specific information in generated features $\hat{\mathbf{f}}$

Methodology



Stage 3

- Trained generative model G generates features for unseen classes across domains
- Classifier $\mathcal{C}$ can handle domain shift as well as semantic shift simultaneously at test-time

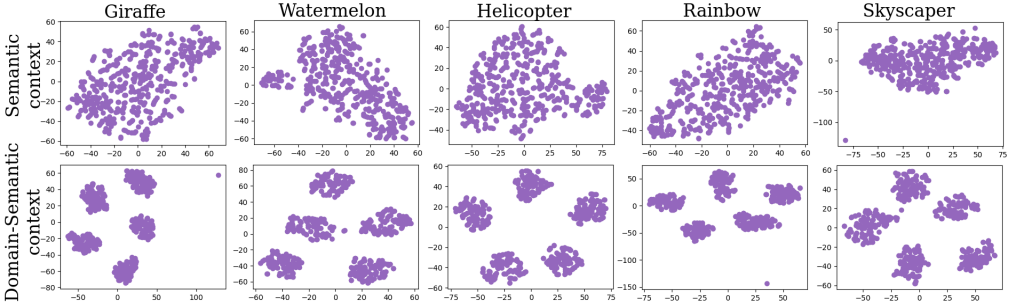
Results

Figure 1: Performance comparison with established baselines and state-of-art methods for ZSLDG problem setting scenario on benchmark DomainNet dataset

Method		Target Domain					
DG	ZSL	*clipart*	*infograph*	*painting*	*quickdraw*	*sketch*	Avg.
-	DEVISE [13]	20.1	11.7	17.6	6.1	16.7	14.4
	ALE [1]	22.7	12.7	20.2	6.8	18.5	16.2
	SPNet [51]	26.0	16.9	23.8	8.2	21.8	19.4
DANN [14]	DEVISE [13]	20.5	10.4	16.4	7.1	15.1	13.9
	ALE [1]	21.2	12.5	19.7	7.4	17.9	15.7
	SPNet [51]	25.9	15.8	24.1	8.4	21.3	19.1
EpiFCR [24]	DEVISE [13]	21.6	13.9	19.3	7.3	17.2	15.9
	ALE [1]	23.2	14.1	21.4	7.8	20.9	17.5
	SPNet [51]	26.4	16.7	24.6	9.2	23.2	20.0
Mixup-img-only		25.2	16.3	24.4	8.7	21.7	19.2
Mixup-two-level		26.6	17	25.3	8.8	21.9	19.9
CuMix[27]		27.6	17.8	25.5	9.9	22.6	20.7
f-clsWGAN [52]		20.0	13.3	20.5	6.6	14.9	15.1
AGG + f-clsWGAN		27.4	17.0	25.9	11.0	23.8	21.0
CuMix + f-clsWGAN		27.3	17.9	26.5	11.2	24.8	21.5
ROT + f-clsWGAN		27.5	17.4	26.4	11.4	24.6	21.4
COCOA$_{AGG}$		27.6	17.1	25.7	11.8	23.7	21.2
COCOA$_{ROT}$		**28.9**	**18.2**	**27.1**	**13.1**	**25.7**	**22.6**

Results

Figure 2: Individual t-SNE visualization of synthesized image features by COCOA for randomly selected unseen classes (*Giraffe*, *Watermelon*, *Helicopter*, *Rainbow*) using only semantic context (Row 1) and both domain-semantic context (Row 2)



Results

Figure 3: Performance comparison to analyse different potential ways to fuse and encode class-level semantic and domain-specific information in generated features $\hat{\mathbf{f}}$ while training the generative network G. $\mathbf{z}, \mathbf{a}_y, \mathbf{e}^{gen}$ represents *noise*, *attribute* and *domain embedding* respectively.

Fusion	Input	BN	*Clipart*	*Infograph*	*Painting*	*Quickdraw*	*Sketch*	Avg
F1	$[\mathbf{z}, \mathbf{a}_y, \mathbf{e}^{gen}]$	-	27.75	14.77	23.93	10.79	25.31	20.51
F2	$\mathbf{z} + [\mathbf{a}_y, \mathbf{e}^{gen}]$	-	28.23	14.99	24.34	10.0	23.47	20.2
F3	$\mathbf{z} + \mathbf{a}_y$	-	24.87	16.09	23.52	11.85	22.89	19.84
F4	$[\mathbf{z}, \mathbf{a}_y]$	$\mathbf{e}^{gen}$	26.56	16.9	24.89	13.0	24.9	21.25
F5	$\mathbf{z} + \mathbf{a}_y$	$\mathbf{e}^{gen}$	27.74	16.19	26.38	11.04	24.13	21.1
F6	$\mathbf{z}$	$[\mathbf{a}_y, \mathbf{e}^{gen}]$	28.9	18.2	27.1	13.1	25.7	22.6

Figure 4: Analysis on the choice of Regularizer $R(\mathbf{f})$. Note that COCOA$_{DOM(\alpha)}$ with $\alpha = 0$ is same as COCOA$_{AGG}$

Variant		*Clipart*	*Infograph*	*Painting*	*Quickdraw*	*Sketch*	Avg
COCOA$_{DOM(\alpha)}$	$\alpha = 0$	27.6	17.1	25.7	11.8	23.7	21.2
	$\alpha = 0.01$	28.14	16.1	27.19	10.55	23.8	21.1
	$\alpha = 0.05$	26.26	16.7	26.88	11.2	23.6	20.9
	$\alpha = 0.1$	26.48	15.5	26.9	11.21	23.7	20.7
COCOA$_{CuMix}$		27.7	17.5	**27.8**	12.6	25.6	22.2
COCOA$_{ROT}$		**28.9**	**18.2**	27.1	**13.1**	**25.7**	**22.6**

References

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS 2013*

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR 2013*

[51] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection networkfor zero-and few-label semantic segmentation. In *CVPR 2019*

[27] Massimiliano Mancini, Zeynep Akata, E. Ricci and Barbara Caputo.
Towards Recognizing Unseen Categories in Unseen Domains, In *ECCV 2020*