

Supplementary Material: Structured Latent Embeddings for Recognizing Unseen Classes in Unseen Domains

Shivam Chandhok¹

shivam.chandhok@mbzuai.ac.ae

Sanath Narayan²

sanath.narayan@inceptioniai.org

Hisham Cholakkal¹

hisham.cholakkal@mbzuai.ac.ae

Rao Muhammad Anwer¹³

rao.anwer@mbzuai.ac.ae

Vineeth N Balasubramanian⁴

vineethnb@iith.ac.in

Fahad Shahbaz Khan¹⁵

fahad.khan@mbzuai.ac.ae

Ling Shao²

ling.shao@ieee.org

¹ Mohamed bin Zayed University of AI, UAE

² Inception Institute of Artificial Intelligence, UAE

³ Aalto University School of Science, Espoo, Finland

⁴ Indian Institute of Technology, Hyderabad, India

⁵ Linköping University, Sweden

In this supplementary material, we additionally discuss the following ablation studies, in continuation to Sec. 4.3 of the main manuscript.

- Analysis for usefulness of structured partitioning and joint invariance
- Impact of different semantic embedding spaces
- Class-wise performance on unseen classes

1 Additional Performance Analysis

In Sec 3.3 (main manuscript), we conjecture that disentanglement of semantic and domain-specific information with the help of a structured domain agnostic latent space (as described in Sec 3.1 and 3.2) might be sufficient for standard DG setting (where the images at training and test time belong to same set of categories). However for our ZSLDG setting, this disentanglement may not hold for unseen semantic categories during testing, as previously found in [9]. In order to address this we introduced the $\mathcal{L}_{joint-inv}$ term (as described in Sec 3.3) to enable generalization to unseen classes in unseen domains. Here, we show further analysis to validate the usefulness of structured partitioning for generalization to novel (unseen) domains and usefulness of joint invariance for generalization to unseen classes.

Usefulness of structured partitioning for generalization to novel domains: In Sec 3.2,

Figure 1: Performance comparison for standard DG setting on PACS dataset using ResNet-18 backbone. Best results are highlighted in bold. We notice that learning multimodal alignment of class-specific visual cues and semantics and partitioning the latent according to semantic concepts enables us to achieve competitive performance for standard DG setting.

Model	Photo	Art	Sketch	Cartoon	Avg
AGG	94.9	76.1	69.4	73.8	78.5
DANN (JMLR'16)	94.0	81.3	74.3	73.8	80.8
MLDG (AAAI'18)	94.3	79.5	71.5	77.3	80.7
CrossGrad (ICLR'18)	94.0	78.7	65.1	73.3	77.8
MetaReg (NeurIPS'18)	94.3	79.5	72.2	75.4	80.4
D-SAM (GCLR'18)	95.30	77.33	77.83	72.43	80.72
JiGen (CVPR'19)	96.0	79.4	71.4	75.3	80.4
Epi-FCR (ICCV'19)	93.9	82.1	73.0	77.0	81.5
MASF(NeurIPS'19)	94.99	80.29	71.69	77.17	81.04
MMLD (AAAI'20)	96.1	81.28	72.29	77.16	81.83
DMG (ECCV'20)	93.35	76.90	75.21	80.38	81.46
CuMix (ECCV'20)	95.1	82.3	72.6	76.5	81.6
Ours($\mathcal{L}_{align} + \mathcal{L}_{center}$)	95.1	84.0	72.0	77.0	82.0

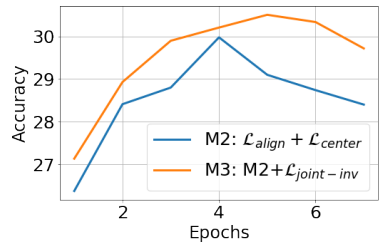


Figure 2: Impact of introducing our proposed joint invariance loss term ($\mathcal{L}_{joint-inv}$) in a standard ZSL setting. Here, the performance comparison between the models M2 and M3 is shown on the unseen classes in DomainNet by considering all the available domains in both training and testing. The introduction of $\mathcal{L}_{joint-inv}$ in M3 enables generalization to unseen classes or concepts at test time and provides consistent performance gain throughout the learning iterations.

Fig. 3 (main manuscript), we observe the impact of multimodal alignment and structured partitioning for standard DG setting. In order to validate that such an alignment and partitioning of latent space according to class-level semantic concepts enables generalization to novel domains and addresses the standard DG setting well, we show performance comparison with other standard DG methods in Fig. 1 (supplementary). We observe that learning multimodal alignment of class-specific visual cues and semantic representations; and partitioning the latent according to semantic concepts shows competitive performance for standard DG setting by enabling generalization to novel domains. Note that as discussed in Sec 3 (main manuscript), we use standard *word2vec*[1] embeddings of classnames as semantic representations, which are learned from text in an unsupervised way [2], with practically no annotation cost.

Usefulness of joint invariance term for unseen class recognition: We conduct an experiment in order to study if the loss $\mathcal{L}_{joint-inv}$ in fact enables generalization to unseen classes at test time. Here, we evaluate on the DomainNet dataset, similar to the standard ZSL setting, where we train on seen classes and and test on unseen classes; but in this case, all the domains are available together during both training and testing. Fig. 2 (supplementary) shows the performance curves for our approach without and with $\mathcal{L}_{joint-inv}$ which are represented by M2 and M3, respectively (as defined previously in Sec 4.2, main manuscript). It can be clearly seen that adding $\mathcal{L}_{joint-inv}$ loss helps enhance performance throughout the learning curve and generalize better to unseen classes than using only M2.

Impact of different Semantic spaces: In Tab. 1 (supplementary), we show the impact on performance of our proposed method, when using different embeddings as semantic representations. We consider three common and widely used semantic embedding spaces i.e *word2vec*[1], *GloVe*[3] and *fastText*[2], for this experiment. We notice a trend similar to that observed with other methods in the few-shot learning/limited-supervision literature [4].

Table 1: Impact on the performance of our proposed method with different semantic embedding spaces, on DomainNet dataset for ZSLDG setting.

Semantic Space	AVG	<i>painting</i>	<i>infograph</i>	<i>quickdraw</i>	<i>sketch</i>	<i>clipart</i>
word2vec	21.9	26.6	18.4	11.5	25.0	27.8
Glove	22.2	26.6	18.7	11.7	25.5	28.5
fastText	21.1	25.6	17.8	11.3	24.4	26.5

Specifically, we notice that *fastText* embeddings give inferior performance when compared with the other two representations. Also, *GloVE* further improves performance by 0.3% on average when compared with *word2vec* embeddings. We also note that our method is able to maintain fairly stable average performance across domains for different choice of semantic embeddings. Also, note that the performance with *word2vec* representations is same as that reported in Sec 4.1 (Tab. 1, main manuscript) since we use these representations for all our experiments, following [9].

Performance on unseen classes In Tab. 1 (main manuscript), we observed that our proposed method consistently enhances performance on all domains of the DomainNet dataset, when compared to the state-of-the-art ZSLDG method, CuMix [9]. In order to further analyse the performance, we show the class-wise distribution of accuracy across the 45 unseen classes for the different domains of DomainNet dataset, in Fig. 3 (supplementary). Note that the bars represent relative gain w.r.t current state-of-art CuMix[9]. We notice that our method improves over CuMix for majority of classes, out of the 45 unseen class set, across all domains. Specifically, we improve or match performance w.r.t CuMix for 31/45 classes on *infograph*, 29/45 classes on *sketch*, 28/45 classes on *quickdraw*, 25/45 classes on *clipart*, 25/45 classes on *painting*. Furthermore, we notice that the gains are higher and on more number of classes for harder domains like *quickdraw*, *infograph* and *sketch* which entail larger domain-shift at test-time w.r.t the source domains available during training.

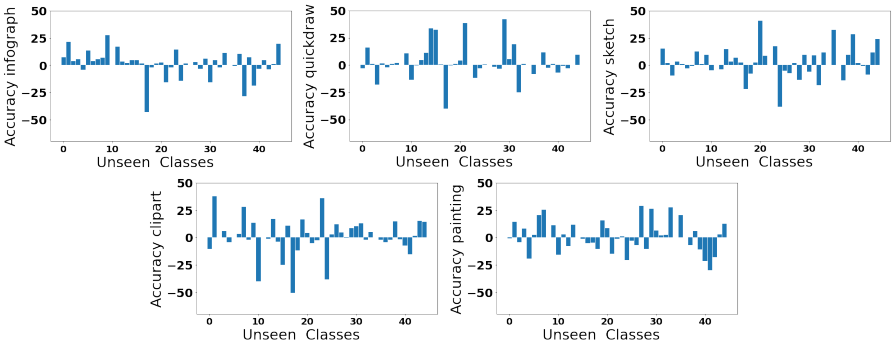


Figure 3: Class-wise accuracy gain (in %) w.r.t state-of-art CuMix [9] method for the 45 unseen classes, on different domains of DomainNet dataset, for ZSLDG setting. We notice that our model shows better performance than CuMix on majority of unseen classes especially for harder domains like *quickdraw*, *infograph* and *sketch*.

References

- [1] Mohamed Afham, S. Khan, M. H. Khan, Muzammal Naseer, and F. Khan. Rich semantics improve few-shot learning. *ArXiv*, abs/2104.12709, 2021.
- [2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [3] Massimiliano Mancini, Zeynep Akata, E. Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *ECCV*, 2020.
- [4] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.
- [5] Jeffrey Pennington, R. Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.