

Supplementary for COSMo: CLIP Talks on Open-Set Multi-Target Domain Adaptation

Munish Monga¹
munish30monga@gmail.com

Sachin Kumar Giroh¹
22m2159@iitb.ac.in

Ankit Jha^{1,2}
ankitjha16@gmail.com

Mainak Singha¹
mainaksingha.iitb@gmail.com

Biplab Banerjee¹
getbiplab@gmail.com

Jocelyn Chanussot²
jocelyn.chanussot@inria.fr

¹ Indian Institute of Technology, Bombay
Mumbai, India

² INRIA, Grenoble, France

1 Contents of the supplementary materials

We discuss the following aspects in the supplementary:

- We describe the dataset details, split ratio, and dataset statistics for the Open-Set Multi-Target Domain Adaptation (OSMT-DA) in Section 2 (Table 1).
- In Section 3, we provide the ablation study on the effect of the entropy regularization parameter and separate prompts, as shown in Figures 1 and 2, respectively.
- In Section 4, we present comprehensive results for the three datasets used in our work. Tables 2, 3, and 4 report the performance on the Office-31, Office-Home, and Mini-DomainNet datasets, respectively, using OS, OS*, and UNK as evaluation metrics. We also provide a comparison between the t-SNE visualizations from our proposed method, COSMo, and state-of-the-art methods on the Office-Home dataset, shown in Figure 3.
- Finally, in Table 5, we list the notations used in designing and training the architecture.

2 Dataset Statistics

Table 1 presents the distribution of known and unknown samples across different source domains for three datasets: Office, Office-Home, and Mini-DomainNet. It details the counts of known and unknown samples for each domain within these datasets, providing a clear overview of the data variability and composition used in our analysis.

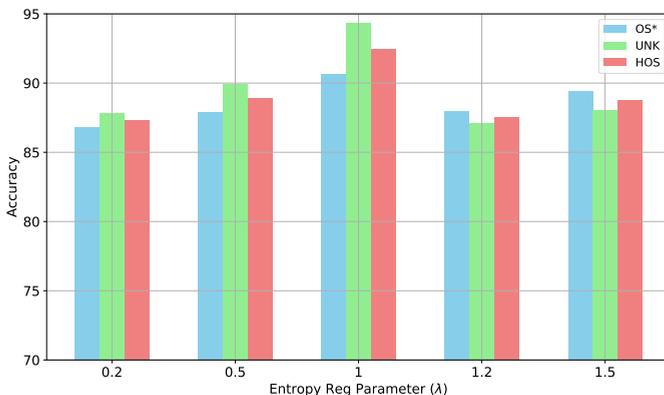
Table 1: Statistics for each dataset depicting the number of known and unknown samples for each source domain

Dataset	Source Domain	# known samples	# unknown samples
Office-31	Amazon (A)	389	904
	DSLR (D)	1059	2553
	Webcam (W)	978	2337
Office-Home	Art (A)	3396	9765
	Clipart (C)	3023	8200
	Product (P)	3062	8087
	Real (R)	2936	8295
Mini-DomainNet	Clipart (C)	55334	71108
	Painting (P)	50467	63176
	Real (R)	30524	44263
	Sketch (S)	54322	66241

To assess the efficacy of our approach in open-set Multi-Target domain adaptation, we utilize three established datasets, each offering distinct challenges and settings. The **Office-Home** dataset [9] consists of 15,500 images across four distinct domains: Art, Clip Art, Product, and Real World. It encompasses 65 categories depicting a variety of objects typically found in office and home environments. The **Office-31** dataset [8] includes 4,652 images spanning three domains: Amazon, DSLR, and Webcam, with each domain featuring 31 categories related to office supplies. Lastly, the **Mini-DomainNet**, a subset of the larger DomainNet [9] dataset, provides a broad spectrum of images across four domains—Clipart, Painting, Real, and Sketch—comprising 126 classes. Dataset split ($|C_k|/|C_u|$) for Office-31, Office-Home and Mini-DomainNet is taken as 10/21, 15/50 and 60/66 respectively.

3 Ablations

Effect of entropy regularisation parameter (λ): Figure 1 depicts the effect of varying the entropy regularization parameter (λ) on the model’s metrics: OS*, UNK, and HOS. Optimal performance is achieved at $\lambda = 1$, suggesting that a balanced entropy regularization is crucial for enhancing model accuracy.

Figure 1: Effect of varying the entropy regularization parameter λ on the Office-31 dataset.

Impact of having separate P_{kwn} and P_{unk} : Figure 2 provides a detailed analysis of the effects of using separate P_{kwn} and P_{unk} on the performance metrics: OS*, UNK, and HOS, across different source domains in the Office-31 dataset. The results demonstrate an increase in HOS scores (except on the Amazon domain) when separate prompts are implemented. A notable increase is observed in the Unknown accuracy, implying that separate prompts are able to handle the unknown classes well.

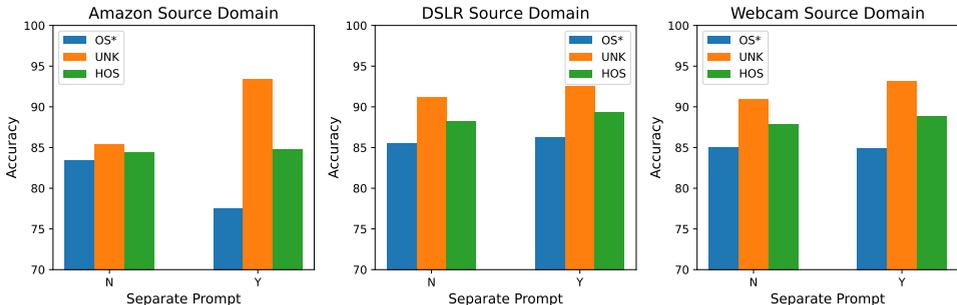


Figure 2: Impact of having separate known prompts P_{kwn} and unknown prompts P_{unk} . Here 'N' represents no separate prompt, and 'Y' represents that the separate prompts are used.

4 Comprehensive Results

t-SNE visualization: In Figure 3, we visualize and compare the t-SNE embeddings generated from the text encoder of our proposed COSMo for both known and unknown classes with other methods on the Office-Home dataset on the proposed setting. COSMo is able to segregate the known and unknown classes better. Tables 2, 3, and 4 depict the comprehensive results for the proposed setting on Office-31, Office-Home and Mini-DomainNet datasets, respectively. The results are obtained with both vision backbones: ResNet-50 [1] and ViT-B/16 [2], and all the metrics are reported (OS*, UNK and HOS).

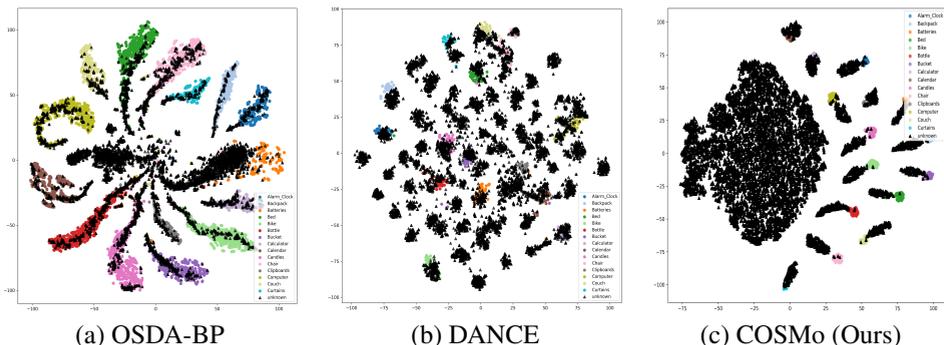


Figure 3: t-SNE visualizations on the Office-Home Dataset with Amazon as the source domain. Coloured dots represent known classes in the source domain, while black triangles denote target domain samples. For COSMo, text embeddings are used, while features from the penultimate layer are used for the other models.

Detailed results of Table 1 (Main paper): Here, we discuss the detailed results of our proposed COSMo and compare them with the referred literature.

Table 2 presents detailed results on the **Office-31** dataset. Our proposed COSMo outperforms other models across all three domains using the ViT-B/16 architecture for the OSMTDA task. With the ResNet 50 architecture, COSMo surpasses other models in two out of three domains. Additionally, COSMo achieves the highest HOS score on the Amazon domain and the lowest on the DSLR domain.

Table 2: Results on the Office-31 (10/21) dataset. Best in **bold**, second best underlined.

Method	Source Domain	RN50			ViT-B/16		
		OS*	UNK	HOS	OS*	UNK	HOS
CLIP [1]	Amazon	92.53	25.22	39.64	95.24	28.43	43.79
	DSLR	89.34	31.53	46.61	91.2	25.58	39.95
	Webcam	89.26	30.72	45.71	90.74	25.33	39.61
OSDA-BP [1]	Amazon	92.66	25.11	39.51	90.24	78.43	83.92
	DSLR	83.97	55.23	66.63	77.78	71.68	74.60
	Webcam	80.70	51.82	63.11	77.00	69.70	73.17
DANCE [1]	Amazon	96.02	64.05	<u>76.84</u>	87.82	85.18	<u>86.48</u>
	DSLR	78.12	83.00	<u>80.49</u>	78.37	96.12	<u>86.34</u>
	Webcam	76.75	82.16	<u>79.36</u>	80.84	95.98	<u>87.76</u>
AD-CLIP [1]	Amazon	93.26	25.55	40.11	100	22.68	36.97
	DSLR	92.3	22.76	36.51	79.73	35.17	48.81
	Webcam	90.46	29.65	44.66	92.07	19.6	32.32
COSMo	Amazon	74.58	81.64	77.95	90.64	94.36	92.46
	DSLR	82.93	78.69	80.76	87.05	89.82	88.41
	Webcam	84.47	76.94	80.53	87.75	90.59	89.15

Similar to the results on the Office-31 dataset, Table 3 presents a detailed comparison of our proposed COSMo model with state-of-the-art methods on the **Office-Home** dataset. COSMo consistently outperforms nearly all other models across various domains, with the exception of the Art domain when using the ViT-B/16 architecture. The Art domain poses a greater challenge compared to the other domains.

Table 3: Results on Office-Home (15/50) Dataset. Best in **bold**, second best underlined.

Method	Source Domain	RN50			ViT-B/16		
		OS*	UNK	HOS	OS*	UNK	HOS
CLIP [1]	Art	84.15	48.00	61.14	92.08	46.06	61.76
	Clipart	92.21	45.28	60.73	95.39	45.01	61.16
	Product	81.44	54.62	65.38	90.64	53.44	67.24
	Real World	79.65	51.09	62.25	90.16	49.03	63.52
OSDA-BP [1]	Art	75.36	29.88	42.8	42.44	75.7	54.39
	Clipart	71.34	49.06	58.14	35.13	31.13	33.01
	Product	67.71	41.82	51.71	54.77	57.23	55.97
	Real World	68.14	43.39	53.02	48.32	69.33	56.95
DANCE [1]	Art	67.91	81.08	<u>73.91</u>	79.63	82.83	81.2
	Clipart	65.14	84.49	<u>73.56</u>	84.38	86.5	<u>85.42</u>
	Product	58.8	85.73	<u>69.76</u>	72.57	85.68	<u>78.58</u>
	Real World	62.04	81.82	<u>70.57</u>	77.5	81.13	<u>79.28</u>
AD-CLIP [1]	Art	84.41	47.43	60.74	92.64	38.15	54.04
	Clipart	92.3	34.5	50.23	94.02	33.05	48.91
	Product	82.11	44.64	57.84	90.52	34.23	49.67
	Real World	80.38	42.36	55.48	92.16	31.26	46.69
COSMo	Art	79.31	74.74	76.96	90.25	73.4	<u>80.96</u>
	Clipart	80.59	81.99	81.28	88.92	84.79	86.8
	Product	74.8	70.42	72.54	80.78	84.13	82.42
	Real World	72.27	75.47	73.83	84.65	79.44	81.97

The **Mini-DomainNet** dataset presents a significant challenge for domain adaptation due to its large number of unknown classes and the relatively high number of known and unknown samples. Despite these difficulties, our model achieves nearly 80% HOS score, the highest among other models, across the four domains, as shown in Table 4. Regardless of the architecture used, COSMo consistently attains the best HOS score across all four domains. Notably, we observe the highest HOS score on the sketch domain and the lowest on the real domain.

Table 4: Results on Mini-Domain Net (60/66) Dataset. Best in **bold**, second best underlined.

Method	Source Domain	RN50			ViT-B/16		
		OS*	UNK	HOS	OS*	UNK	HOS
CLIP [10]	Clipart	80.47	64.59	<u>71.67</u>	89.29	65.81	<u>75.78</u>
	Painting	82.15	63.1	<u>71.38</u>	90.97	64.78	75.67
	Real	68.87	68.19	<u>68.53</u>	84.74	63.19	72.39
	Sketch	81.07	64.26	<u>71.69</u>	89.67	65.44	<u>75.66</u>
OSDA-BP [11]	Clipart	38.85	71.14	50.26	37.43	46.01	41.27
	Painting	38.58	73.25	50.54	48.3	57.28	52.41
	Real	25.73	73.54	38.12	41.65	59.08	48.86
	Sketch	38.76	70.35	49.98	40.95	35.09	37.79
DANCE [12]	Clipart	26.46	94.82	41.37	56.85	91.96	70.26
	Painting	31.7	94.29	47.45	76.07	87.29	<u>81.29</u>
	Real	14.39	97.12	25.06	61.97	89.85	<u>73.35</u>
	Sketch	28.6	96.27	44.1	60.46	94.85	73.84
AD-CLIP [13]	Clipart	83.97	41.93	55.93	91.87	37.4	53.16
	Painting	85.38	44.18	58.23	92.98	41.01	56.92
	Real	73.5	40.36	52.11	86.12	32.9	47.61
	Sketch	83.1	43.67	57.25	91.94	38.1	53.88
COSMo	Clipart	74.38	76.62	75.48	83.08	79.1	81.05
	Painting	80.37	72.69	76.34	86.58	81.85	84.15
	Real	63.36	77.06	69.54	79.33	79.03	79.18
	Sketch	72.75	80.72	76.52	81.08	84.77	82.89

Table 5: Table of Mathematical Terms and Notations

Notation	Description
$(X_s, Y_s) \in \mathcal{S}$	Source domain data and labels
X_t	Unlabeled data from all target domains combined
q	Number of target domains
C_s and C_t	Classes in the source and target domains
C_k	Known classes from the source domain, $C_k = C_s$
C_u	Unknown classes in the target domain, $C_u = C_t \setminus C_s$
D_s and D_t	Mini-batch from labeled source and unlabelled target domains
\mathcal{F}_v and \mathcal{F}_t	Pre-trained image and text encoder
$B_\theta(\cdot)$	Domain-specific bias network
β	Domain bias context token, $\beta = B_\theta(v)$
P_k^c	Known class-based prompt for class c
$P_k, bias$	Biased known class prompt
P_{kwn}	Cumulative prompt for all known classes
P_u	Unknown class-based prompt
P_{unk}	Biased unknown class prompt
τ	Text features encoded by the text encoder
λ	Hyperparameter controlling entropy regularization strength
κ_{lower}	Lower threshold for confidence in known classes
κ_{upper}	Upper threshold for confidence in unknown classes
m	Length of the context prompt

References

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [3] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation, 2019.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [5] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.
- [6] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation, 2018.
- [7] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self-supervision. 2020.
- [8] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4355–4364, 2023.
- [9] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.