Multi-Modal Proxy Learning Towards Personalized Visual Multiple Clustering – Supplementary –

Jiawei Yao[1] Qi Qian[2] Juhua Hu[1*]

[1] School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA

[2] Alibaba Group, Bellevue, WA 98004, USA

{jwyao, juhuah}@uw.edu, qi.qian@alibaba-inc.com

1. Additional Results

1.1. Clustering Analysis

We further analyze the clustering outcomes of Multi-MaP against various clusters. The results are derived by applying different embeddings to produce all clustering outputs. Then, we compare each obtained clustering result to all ground truth clusterings. The findings, as depicted in Tables 1 and 2, reveal that the top-performing outcomes exhibit a clear diagonal structure. This demonstrates that the representations generated by Multi-MaP are capable of discerning different aspects of the same data, and then produce different clusterings aligning well with different ground truth data structures.

Dataset	Clustering	C_1		C_2	
		NMI	RI	NMI	RI
ALOI [1]	Color	**1.0000**	**1.0000**	0.3164	0.6798
	Shape	0.3642	0.5491	**1.0000**	**1.0000**
Fruit [3]	Color	**0.8619**	**0.9526**	0.5379	0.6934
	Species	0.6953	0.7843	**1.0000**	**1.0000**
Fruit360 [6]	Color	**0.6239**	**0.8243**	0.4242	0.7163
	Species	0.3551	0.6333	**0.5284**	**0.7582**
Card [6]	Order	**0.3653**	**0.8587**	0.1142	0.5562
	Suits	0.1346	0.5679	**0.2734**	**0.7039**
Stanford Cars [4]	Color	**0.7360**	**0.9193**	0.4223	0.7526
	Type	0.3692	0.6245	**0.6355**	**0.8399**
Flowers [5]	Color	**0.6426**	**0.7984**	0.3277	0.7153
	Species	0.2884	0.6150	**0.6013**	**0.8103**

Table 1. Clustering analysis in six benchmark multiple clustering vision tasks.

1.2. Efficiency Analysis

To further demonstrate the efficiency of our proposed method using the frozen pre-trained model by CLIP, we compare the efficiency of different deep multiple clustering methods. The experiments are conducted on a server with a GPU GeForece RTX 2080Ti. We show the running time on Fruit dataset. The running time and color clustering performance of each method are shown in Fig. 1. Multi-MaP has significantly better performance than all the baselines in both effectiveness and efficiency. That is because our method can directly exploit the CLIP encoder to capture the image and text embeddings, without updating the encoder's parameters, so its running time is much smaller than other methods. In summary, the proposed method shows the best performance under the least running time requirement.

Clustering	Metrics	C1	C2	C3	C4
Emotion	NMI	**0.1786**	0.0362	0.0564	0.0435
	RI	**0.7105**	0.4376	0.513	0.5683
Glass	NMI	0.1104	**0.3402**	0.1163	0.1567
	RI	0.6893	**0.7068**	0.6429	0.6952
Identity	NMI	0.2342	0.3627	**0.6625**	0.325
	RI	0.5362	0.7632	**0.9496**	0.7117
Pose	NMI	0.0673	0.1258	0.1368	**0.4693**
	RI	0.4989	0.5519	0.5867	**0.6624**

Table 2. Clustering analysis on CMUface [2] datasets.

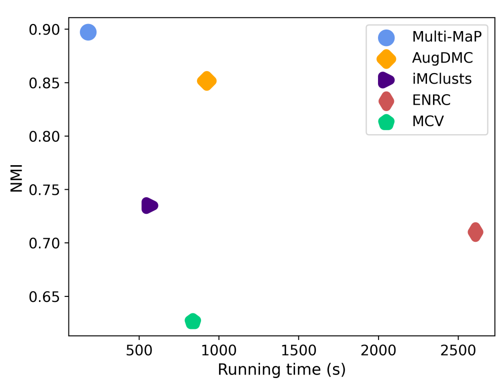

Figure 1. Performance vs. the running time on Fruit [3] dataset.

*Corresponding author

Dataset	Clustering	Multi-MaP$_{\text{woCR}}$ NMI	Multi-MaP$_{\text{woCR}}$ RI	Multi-MaP$_{\text{woC}}$ NMI	Multi-MaP$_{\text{woC}}$ RI	Multi-MaP$_{\text{woR}}$ NMI	Multi-MaP$_{\text{woR}}$ RI	Multi-MaP$_{\text{MSE}}$ NMI	Multi-MaP$_{\text{MSE}}$ RI	Multi-MaP NMI	Multi-MaP RI
ALOI [1]	Color	0.9632	0.9829	1.0000	1.0000	0.9843	0.9906	1.0000	1.0000	1.0000	1.0000
	Shape	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Fruit [3]	Color	0.7634	0.8432	0.8212	0.9274	0.8169	0.9198	0.8479	0.9296	**0.8619**	**0.9526**
	Species	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Fruit360 [6]	Color	0.5634	0.7650	0.6209	0.7825	0.6134	0.8036	0.6082	0.7943	**0.6239**	**0.8243**
	Species	0.5077	0.7368	0.5137	0.7436	0.5176	0.7363	0.5199	0.7428	**0.5284**	**0.7582**
Card [6]	Order	0.1928	0.8136	0.3560	0.8458	0.3518	0.8458	0.3605	0.8509	**0.3653**	**0.8587**
	Suits	0.2374	0.6271	0.2691	0.6632	0.2481	0.6104	0.2550	0.6596	**0.2734**	**0.7039**
CMUface [2]	Emotion	0.1692	0.6169	0.1717	0.6233	0.1709	0.6662	0.1711	0.6843	**0.1786**	**0.7105**
	Glass	0.3107	0.6902	0.3265	0.7130	0.3194	0.6908	0.3362	0.7039	**0.3402**	**0.7068**
	Identity	0.5632	0.8236	0.6236	0.8368	0.6042	0.8273	0.6396	0.8941	**0.6625**	**0.9496**
	Pose	0.4361	0.6407	0.4556	0.6492	0.4405	0.6507	0.4398	0.6479	**0.4693**	**0.6624**
Stanford Cars [4]	Color	0.5933	0.7832	0.6834	0.8665	0.6942	0.8930	0.7114	0.9105	**0.7360**	**0.9193**
	Brand	0.5562	0.7993	0.6388	0.8263	0.6207	0.7931	0.6287	0.8176	**0.6355**	**0.8399**
Flowers [5]	Color	0.5795	0.7719	0.5836	0.7838	0.6133	0.7990	0.6211	0.7936	**0.6426**	**0.7984**
	Species	0.5699	0.7604	0.5737	0.7836	0.5905	0.8012	0.5842	0.7897	**0.6013**	**0.8103**

Table 3. Components in Multi-MaP. The significantly best results with 95% confidence are in bold.

1.3. Ablation Study

To validate the effectiveness of Multi-MaP, we compare four variants of Multi-MaP that are removing the reference word constraint, removing the concept-level constraint, removing both constraints and implementing reference constraint with a high-level concept provided by a user, denoted as Multi-MaP$_{\text{woR}}$, Multi-MaP$_{\text{woC}}$, Multi-MaP$_{\text{woCR}}$ and Multi-MaP$_{\text{MSE}}$, respectively. The results are shown in Table 3. The proposed method achieved the best results, while the method that removed both reference word constraint and concept-level constraint performed the worst. This also shows that the proposed reference word constraint and concept-level constraint play an important role in the model. Moreover, Multi-MaP performs better than Multi-MaP$_{\text{MSE}}$, suggesting Multi-MaP can benefit from multiple concepts through the contrastive learning process.

References

[1] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61:103–112, 2005. 1, 2

[2] Stephan Günnemann, Ines Färber, Matthias Rüdiger, and Thomas Seidl. Smvc: semi-supervised multi-view clustering in subspace projections. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 253–262, 2014. 1, 2

[3] Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. Finding multiple stable clusterings. *Knowledge and Information Systems*, 51(3):991–1021, 2017. 1, 2

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1, 2

[5] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1, 2

[6] Jiawei Yao, Enbei Liu, Maham Rashid, and Juhua Hu. Augdmc: Data augmentation guided deep multiple clustering. In *INNS DLIA@IJCNN*, 2023. 1, 2