Towards Multi-Modal Smart Information Retrieval Systems

Muhammad Umer Anwaar

About us

Our company, vision and planned launches

The Mercateo Group and its platforms

Facts and figures

Mercateo

Procurement platform and Europe's leading B2B marketplace with > 100 million items

Unite

Neutral B2B network with pre-integrated Mercateo Shop

Made in Germany

Founded in 2000 in Munich

Growth

Revenue of €343 million in 2020

9 % year-on-year growth

Partners

> 100,000 active business customers
> 700 suppliers
> 50 system partners

International

> 600 employees in 14 subsidiaries in Europe

Unite in Europe

Availability and new countries (2021)

Unite available via Mercateo:

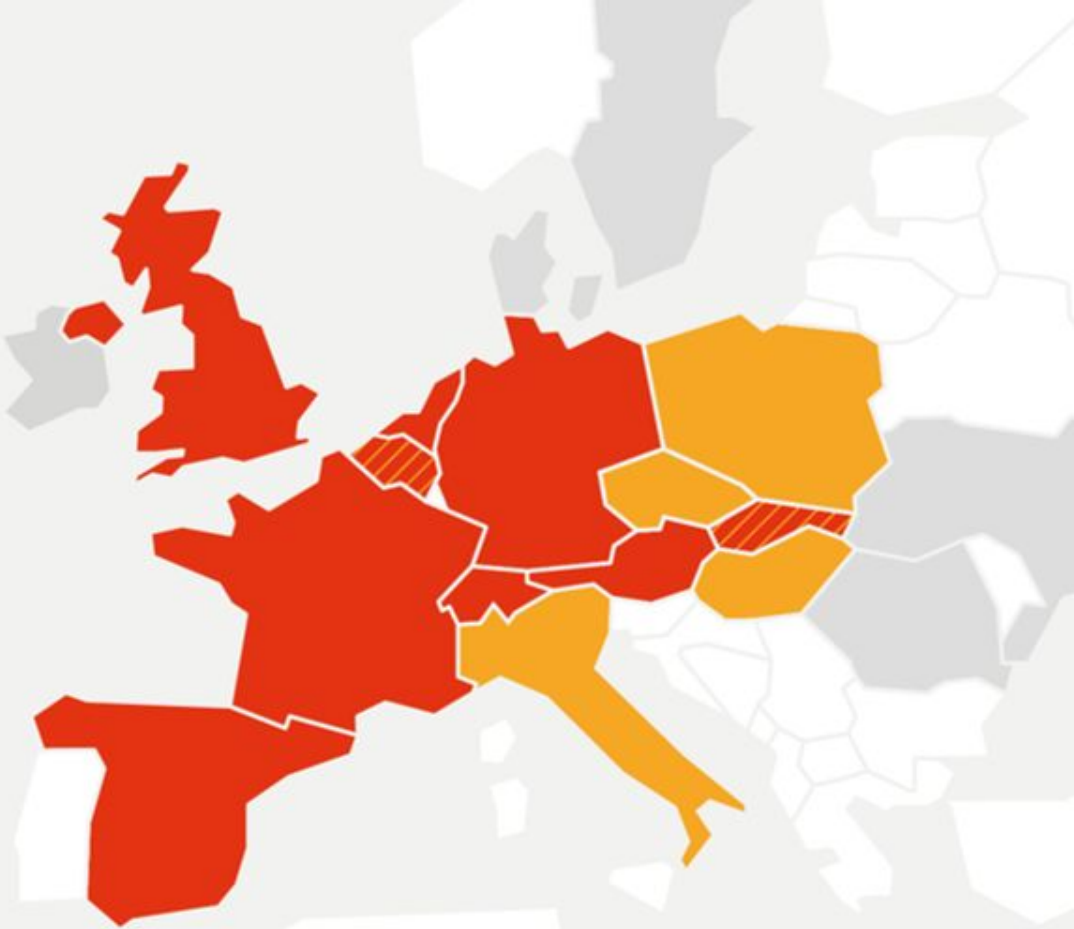
United Kingdom | The Netherlands | France | Spain | Switzerland | Germany | Austria

New countries:

Czech republic | Poland | Italy | Hungary | Belgium | Slovakia

Unite available via other e-procurement partners:

United Kingdom | The Netherlands | France | Spain | Switzerland | Germany | Austria | Belgium | Slovakia



unite mercateo

The Problem

Motivation

Motivation

Motivation

A potential scenario of a smart IR system, “enabling the customer” to express their mind better.

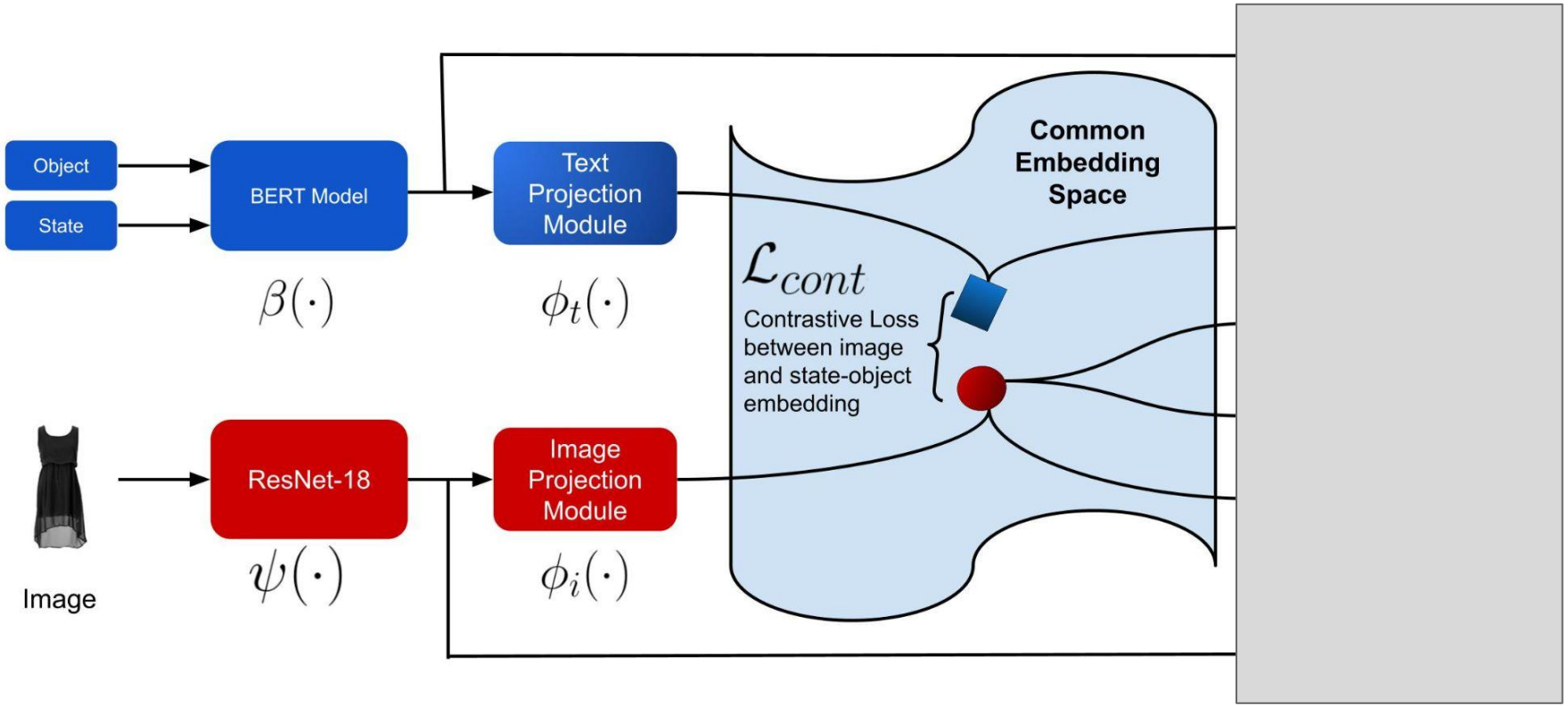
Tasks

Models which learn good state-object representations

(1) Differentiate between different states of an object and can recognise even unseen combinations of state-object. (CZSL Task)

(2) Retrieve images based on multi-modal (image-text) query, where the text describes the changes sought by the user in the query image. (Retrieval Task)

ContraNet Architecture



$$\Delta_t = \Big\|_{j=1}^{B} \phi_t(\beta(t_j)),$$

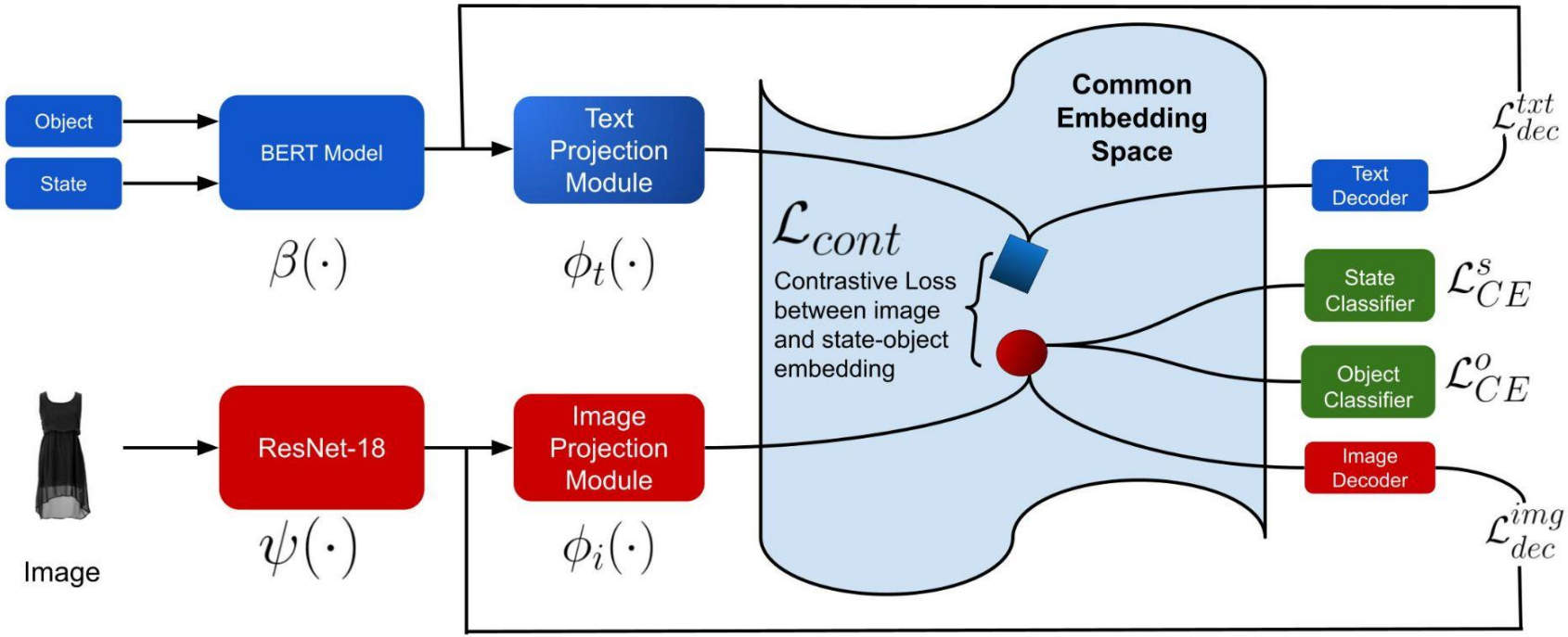
$$\Delta_i = \Big\|_{j=1}^{B} \phi_i(\psi(x_j)),$$

$$\mathcal{E} = \Delta_t * \Delta_i^T,$$

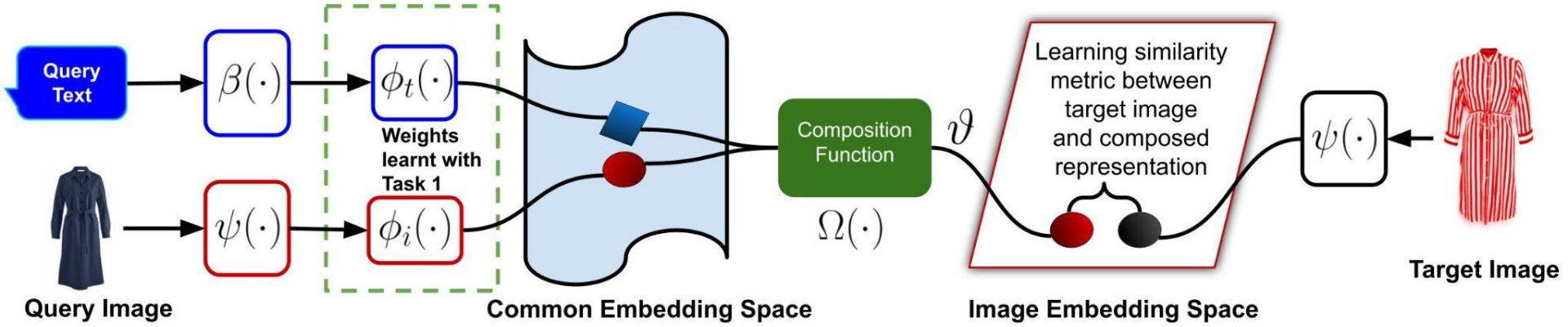
$$\mathcal{L}_{cont} = \frac{1}{2B} \sum_{j=1}^{B} -\log \left\{ \frac{\exp\{\mathcal{E}_{jj}\}}{\sum_{p=1}^{B} \exp\{\mathcal{E}_{jp}\}} \right\}$$

$$-\log \left\{ \frac{\exp\{\mathcal{E}_{jj}\}}{\sum_{p=1}^{B} \exp\{\mathcal{E}_{pj}\}} \right\},$$

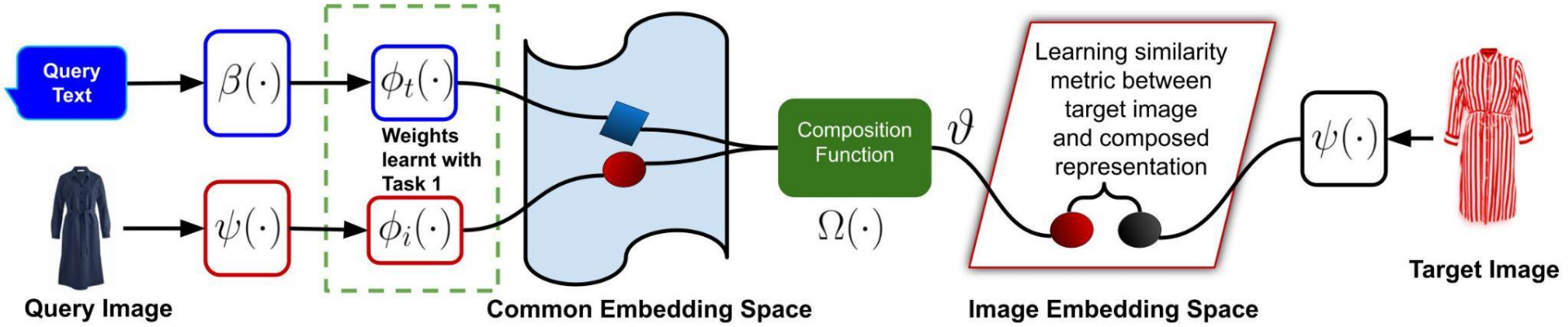
Task # 1 (CZSL Task)



Task#2 (Retrieval Task)



Task#2 (Retrieval Task)



$$\mathcal{L}_{trip} = \frac{1}{MB} \sum_{j=1}^{B} \sum_{m=1}^{M} \log \left\{ 1 + \exp\{\kappa(\vartheta_j, \psi(\tilde{y}_{j,m})) - \kappa(\vartheta_j, \psi(y_j))\} \right\}$$

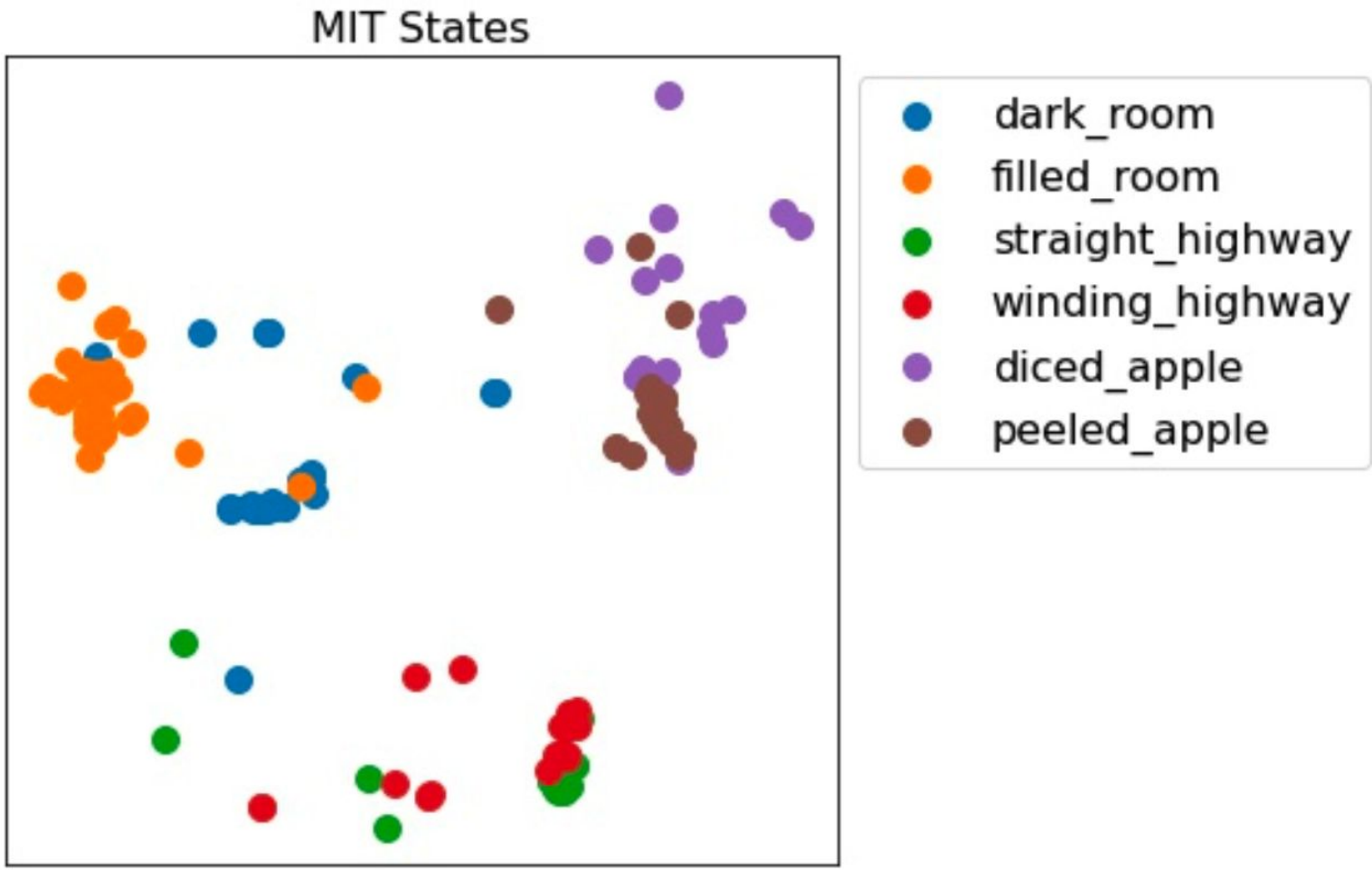
Results

Task # 1 [Traditional CZSL Split]

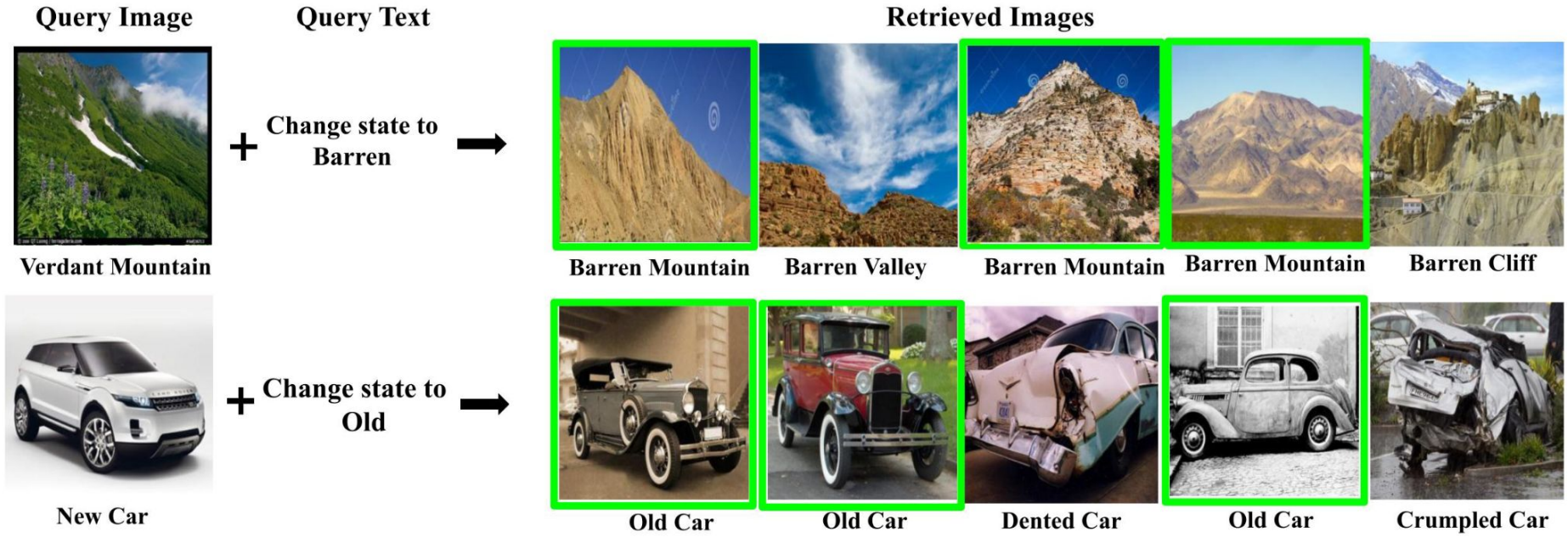
Method	MIT-States			UT-Zappos		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
AnalogousAttr	1.4	-	-	18.3	-	-
LabelEmbed	13.4	17.6	22.4	25.8	39.8	52.4
Red Wine	13.1	21.2	27.6	40.3	52.8	67.1
AoP	14.2	19.6	25.1	46.2	56.6	69.2
TAFE-Net	16.4	26.4	33.0	33.2	45.8	57.3
SymNet	19.9	28.2	33.8	52.1	67.8	76.0
ContraNet	**22.1**	**33.7**	**38.2**	**54.6**	**73.1**	**80.4**
- without $\mathcal{L}_{cont}$	14.7	18.8	24.6	39.9	50.2	62.4
- without $\mathcal{L}_{dec}$	20.4	31.2	35.9	53.3	69.7	78.6

Task # 1 [GCZSL Split]

Method	MIT-States						UT-Zappos					
	Attribute	Object	Seen	Unseen	HM	AUC	Attribute	Object	Seen	Unseen	HM	AUC
AoP	21.1	23.6	14.3	17.4	9.9	1.6	38.9	**69.9**	59.8	54.2	40.8	25.9
Red Wine	22.7	25.1	20.7	17.9	11.6	2.4	40.6	69.1	53.6	52.1	41.3	26.1
LabelEmbed	23.5	26.3	15.0	20.1	10.7	2.0	41.2	69.3	53.0	61.9	41.0	25.7
ComposeAE	23.8	26.4	21.4	22.6	14.9	2.7	41.9	68.8	57.2	58.9	44.2	29.2
TMN	23.3	26.5	20.2	20.1	13.0	2.9	40.8	69.2	58.7	60.0	45.0	29.3
SymNet	24.3	**27.3**	24.2	25.2	16.1	3.0	41.3	68.6	49.8	57.4	40.4	23.4
ContraNet	**28.9**	26.7	**28.1**	**27.4**	**17.4**	**4.7**	**52.7**	68.1	**60.7**	**62.5**	**48.9**	**34.7**
- without $\mathcal{L}_{cont}$	22.9	27.1	14.8	18.6	9.7	1.9	42.4	69.2	54.9	52.6	42.9	27.4
- without $\mathcal{L}_{dec}$	28.2	26.5	27.5	26.8	17.2	3.9	51.4	66.7	58.4	60.8	47.2	33.1


MIT States
dark_room
filled_room
straight_highway
winding_highway
diced_apple
peeled_apple

Qualitative Results: MIT-States



Qualitative Results: Fashion 200k

Task # 2: MIT-States and F200k

Method	R@1	R@5	R@10	R@1	R@10	R@50
Raw Image features only	3.3	12.8	20.9	3.5	22.7	43.7
Raw Text features only	7.4	21.5	32.7	1.0	12.3	21.8
Concatenation [Image,Text]	11.8	30.8	42.1	11.9	39.7	62.6
Show and Tell	11.9	31.0	42.0	12.3	40.2	61.8
AoP	8.8	27.3	39.1	12.2	40.0	61.7
Relationship	12.3	31.9	42.9	13.0	40.5	62.4
FiLM	10.1	27.7	38.3	12.9	39.5	61.9
SymNet	11.2	29.5	41.4	11.7	38.6	60.4
TIRG	12.2	31.9	43.1	14.1	42.5	63.8
ComposeAE	13.9	35.3	47.9	22.8	55.3	73.4
ContraNet-Ω_{res}	**14.5**	**40.7**	**51.4**	**24.0**	**58.4**	**79.2**
ContraNet-Ω_{rot}	13.9	39.1	50.8	18.5	54.8	76.3
ContraNet-Ω_{mlp}	13.7	36.9	48.8	22.9	56.7	77.5
ContraNet-Ω_{res} without Common Embedding Space	12.0	31.2	42.9	17.8	50.6	71.1

Conclusion

- Novel ContraNet Approach for both CZSL and Retrieval tasks
- Text-aware image representations via Contrastive Learning
- ContraNet outperforms the SOTA methods by a huge margin on benchmark datasets

Thanks!