NeurIPS | 2022

Thirty-sixth Conference on Neural Information Processing Systems

Coded Residual Transform for Generalizable Deep Metric Learning



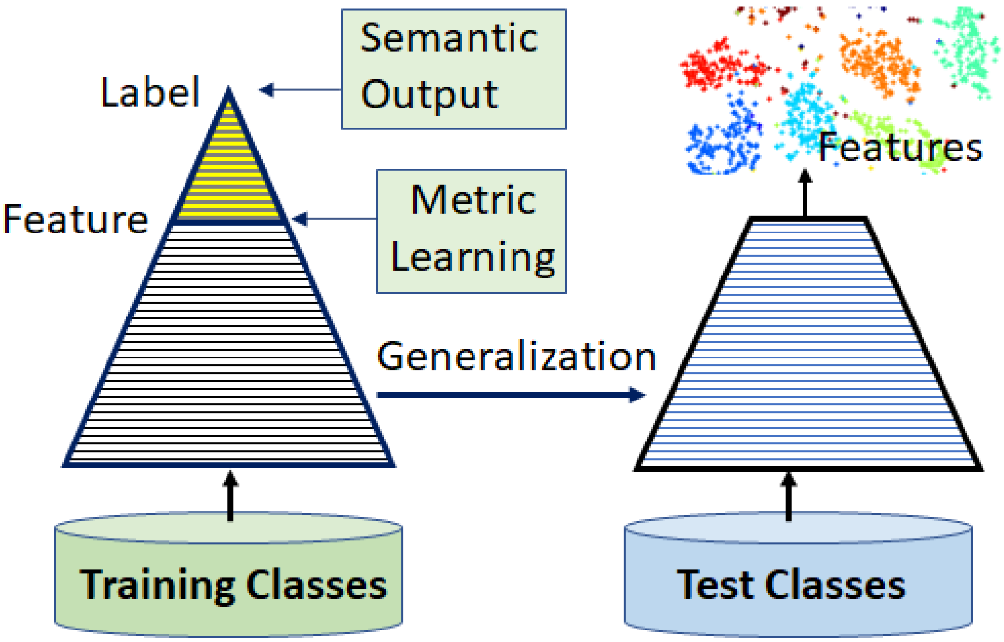



Shichao Kan[1] Yixiong Liang[1] Min Li[1] Yigang Cen[2] Jianxin Wang[1] Zhihai He[3]

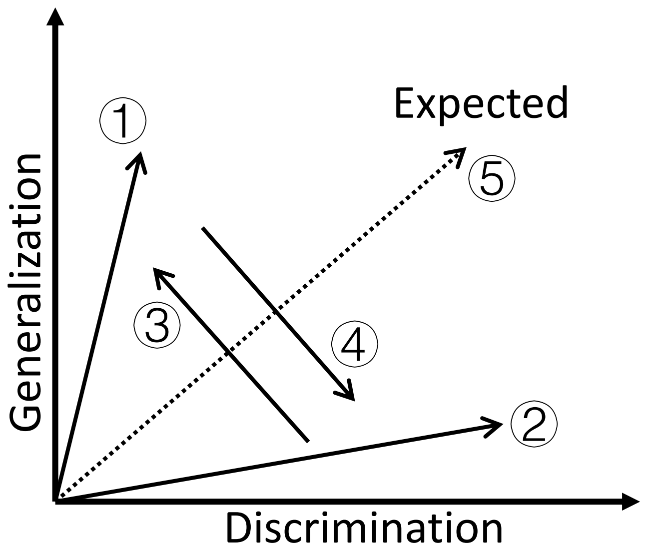
[1]Central South University & [2]Beijing Jiaotong University & [3]Southern University of Science and Technology

Problem Statement

Generalizable Deep Metric Learning



Learning discriminative features to represent images where **the test image classes are totally different from the training classes.**



Challenges

- **Discriminative**: In the embedded feature space, image features with the same semantic labels should be aggregated into compact clusters in the high-dimensional feature space while those from different classes should be well separated from each other.

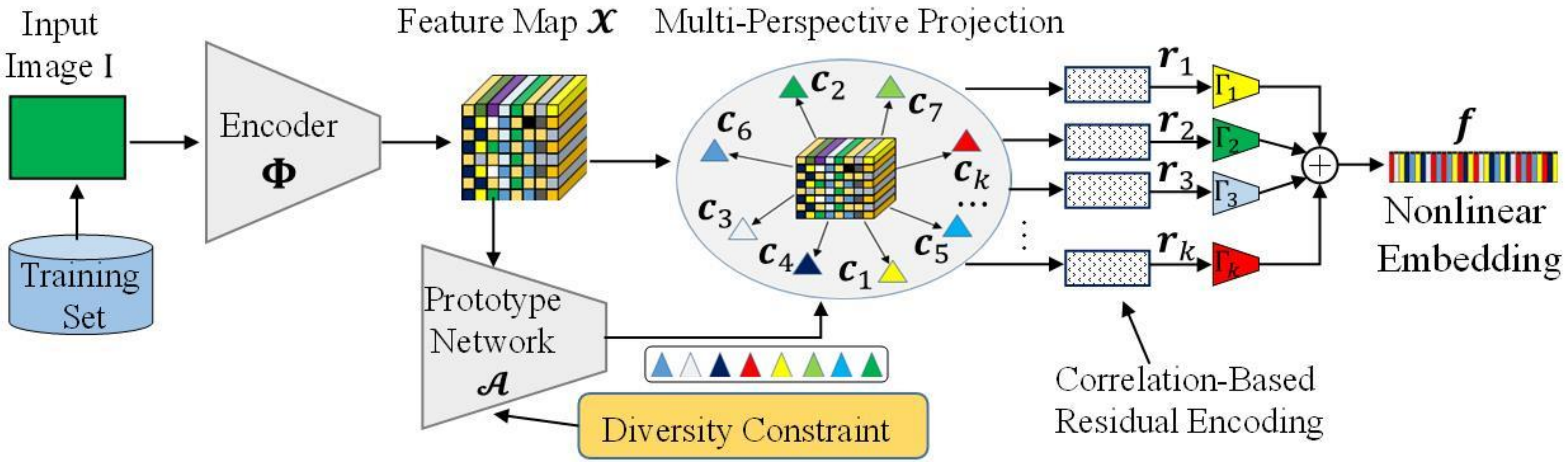
- **Generalizable**: The learned features should be able to generalize well from the training images to test images of new classes which have not been seen before.

Important Applications

- It has important applications in **image recognition**, **person re-identification, image segmentation**, **tracking**, etc.

Key Idea : Coded Residual Transform

Framework



$$\boldsymbol{r}_k = \sum_{j=1}^{H\times W} \log[1+\exp(\boldsymbol{x}_j^T \boldsymbol{c}_k)](\boldsymbol{x}_j - \boldsymbol{c}_k).$$

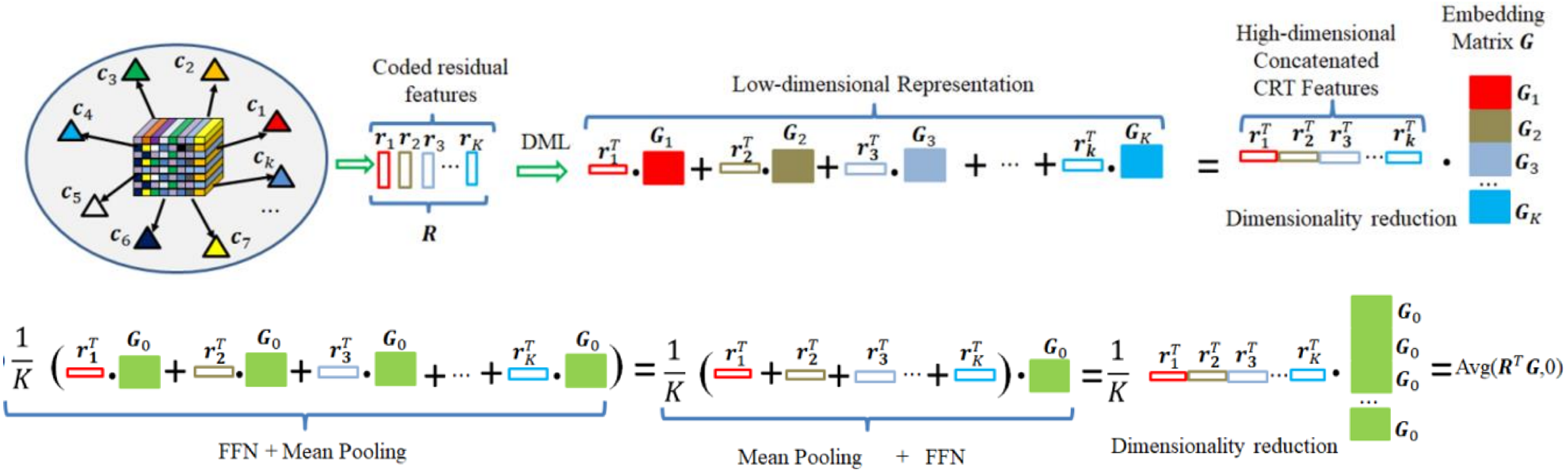
$$\boldsymbol{f} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\Gamma}_k(\mathbf{r}_k)$$

Key Idea：Coded Residual Transform

Coded Residual Feature Transform

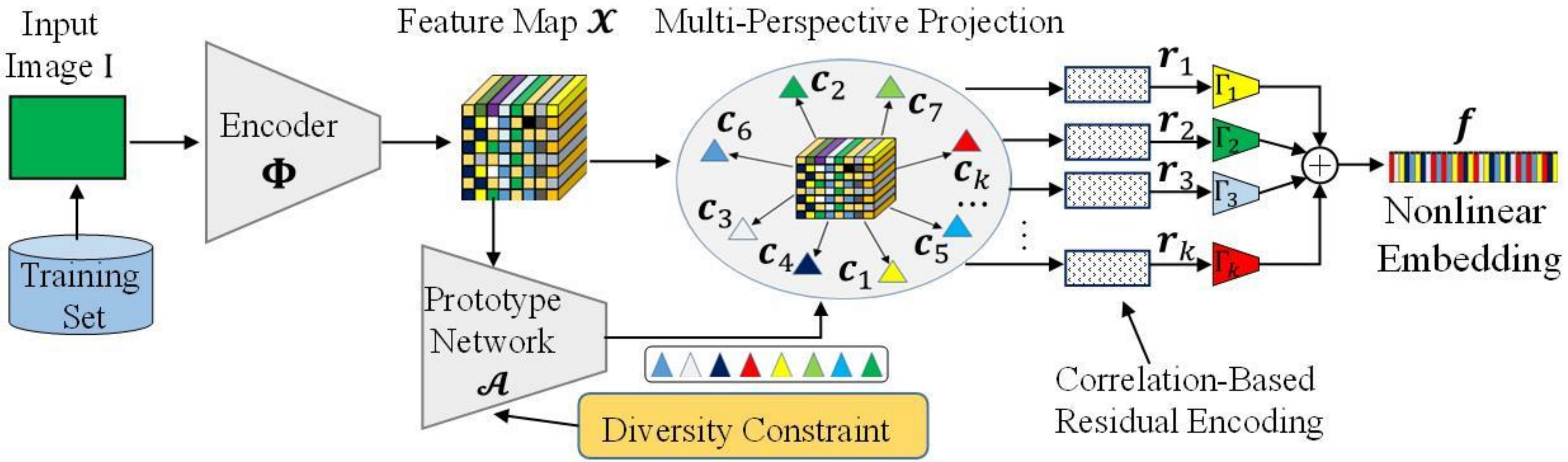
$$\boldsymbol{r}_k = \sum_{j=1}^{H\times W} \log[1+\exp(\boldsymbol{x}_j^T \boldsymbol{c}_k)](\boldsymbol{x}_j - \boldsymbol{c}_k);$$

$$\boldsymbol{f} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\Gamma}_k(\mathbf{r}_k)$$



Key Idea : Coded Residual Transform

Framework

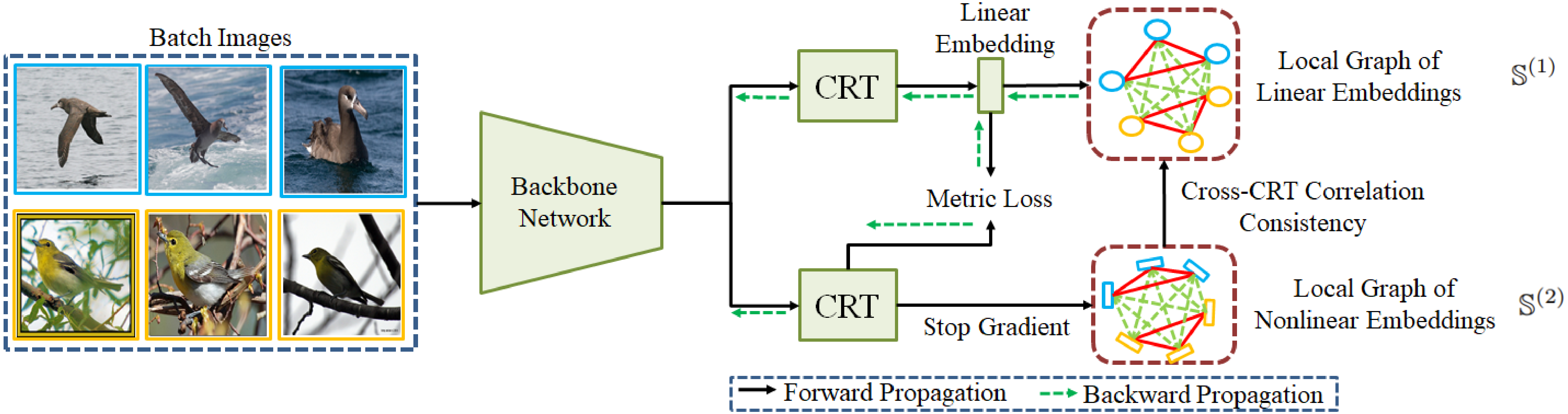


$$L_{DIV} = \frac{1}{K(K-1)} \sum_{k \neq j} \frac{|c_k \cdot c_j^T|}{\|c_k\|_2 \cdot \|c_j\|_2}$$

$$\begin{aligned} \mathrm{L}_{ms} &= \frac{1}{N} \sum_{I_i \in \mathcal{I}} \left[\frac{1}{\alpha} \log[1 + \sum_{j \in \mathcal{P}_i} \exp(-\alpha(\hat{d}_c(i,j) - \lambda))] \right] \\ &+ \sum_{I_i \in \mathcal{I}} \left[\frac{1}{\beta} \log[1 + \sum_{k \in \mathcal{N}_i} \exp(\beta(\hat{d}_c(i,j) - \lambda))] \right] \end{aligned}$$

Key Idea：Coded Residual Transform

Cross-CRT Correlation Consistency



$$\mathbb{S}^{(1)} = \left[\mathbf{s}_{ij}^{(1)}\right]_{1\leq i,j\leq N}, \quad \mathbf{s}_{ij}^{(1)} = \frac{\boldsymbol{f}_i^{(1)} \cdot [\boldsymbol{f}_j^{(1)}]^T}{\|\boldsymbol{f}_i^{(1)}\|_2 \cdot \|\boldsymbol{f}_j^{(1)}\|_2}. \qquad L_{CON} = \|\mathbb{S}^{(1)} - \mathbb{S}^{(2)}\|_1$$

$$L = L_{DIV} + \lambda_1 \mathrm{L}_{MS} + \lambda_2 \cdot L_{CON}$$

Datasets

- CUB-200-2011, Cars-196, Stanford Online Product (SOP), In-Shop Closes Retrieval (In-Shop) datasets: They are benchmark datasets of deep metric learning in image retrieval scenario.
- **CUB-200-2011** consists of 11,788 images from 200 bird categories. We use **the first 100 classes** (5,864 images) for training and **the remaining 100 classes** (5,924 images) for testing.
- **Cars-196** contains 16,185 images of 196 cars classes. We use **the first 98 classes** (8,054 images) for training and **the remaining 98 classes** (8,131 images) for testing.
- **SOP** consists of 120,053 images with 22,634 classes crawled from Ebay. we split **the first 11,318 classes** with 59,551 images for training, and **the remaining 11,316 classes** with 60,502 images for retrieval.
- **In-Shop** consists of 52,712 images with 7,986 clothing classes. We use the predefined 25,882 training images of 3,997 classes for training. The remaining 3985 classes are partitioned into a query set (14,218 images) and a gallery set (12,612 images).

Experimental Results

Performance Metric

- Recall@K, embedding space density (ESD), spectral decay (SD).

$$\text{Recall@K} = \frac{1}{|\mathcal{I}|} \sum_{I_i \in I} \begin{cases} 1, & \exists I_k \in \mathcal{G}_i^k, \text{ s.t., } y_k = y_i \\ 0, & \text{otherwise,} \end{cases}$$

The **embedding space density** metric is defined as the ratio between the average of intra-class distance and the average of inter-class distance.

$$\mathcal{D}_{\text{ESD}} = \mathcal{D}_{\text{Intra}} / \mathcal{D}_{\text{Inter}}.$$

The **spectral decay** metric is defined to be the KL-divergence between the spectrum of d singular values (obtained from Singular Value Decomposition, SVD) and a d-dimensional uniform distribution

$$\rho_{\text{SD}} = \mathcal{D}_{\text{KL}}(\mu_d, V^{\text{SV}})$$

Experimental Results

Comparison of retrieval performance on the CUB and Cars datasets.

Methods	Dim	CUB				Cars			
		R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
A-BIER [TPAMI20] [60]	512	57.5	68.7	78.3	82.6	82.0	89.0	93.2	96.1
MS [CVPR19] [11]	512	65.7	77.0	86.3	91.2	84.1	90.4	94.0	96.5
Proxy-Anchor [CVPR20] [9]	512	68.4	79.2	86.8	91.6	86.1	91.7	95.0	97.3
DRML-PA [ICCV21] [62]	512	68.7	78.6	86.3	91.6	86.9	92.1	95.2	97.4
ETLR [CVPR21] [13]	512	72.1	81.3	87.6	-	89.6	94.0	96.5	-
DCML-MDW [CVPR21] [63]	512	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0
D & C [TPAMI21] [64]	512	68.4	78.7	86.0	91.6	87.8	92.5	95.4	-
IBC [ICML21] [18]	512	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2
LSCM-GNN [TIP22] [2]	512	68.5	77.3	85.3	91.3	87.4	91.5	94.9	97.0
Group Loss++ [TPAMI22] [7]	512	72.6	80.5	86.2	91.2	90.4	93.8	96.0	97.5
IRT_R [arXiv21] [1]	384	74.7	82.9	89.3	93.3	-	-	-	
PA+DIML [ICCV21] [17]	128	66.46	-	-	-	86.13	-	-	-
Hyp-DeiT[CVPR22] [20]	128	74.7	84.5	90.1	94.1	82.1	89.1	93.4	96.3
Ours: CRT	128	**78.98**	**86.68**	**91.61**	**95.04**	**91.16**	**94.92**	**96.79**	98.03
Ours: Gain	128	**4.28**	**2.18**	**1.51**	**0.94**	**0.76**	**0.92**	**0.29**	−0.17

Experimental Results

Comparison of retrieval performance on the SOP and In-Shop datasets.

Methods	Dim	SOP				In-Shop			
		R@1	R@10	R@100	R@1000	R@1	R@10	R@20	R@30
Fusing-Net [TIP19] [8]	512	71.8	86.3	94.1	98.2	82.4	95.1	96.7	97.4
A-BIER [TPAMI20] [60]	512	74.2	86.9	94.0	97.8	83.1	95.1	96.9	97.5
MS [CVPR19] [11]	512	78.2	90.5	96.0	98.7	89.7	97.9	98.5	98.8
DRML-PA [ICCV21] [62]	512	71.5	85.2	93.0	-	-	-	-	-
ETLR [CVPR21] [13]	512	79.8	91.1	96.3	-	-	-	-	-
DCML-MDW [CVPR21] [63]	512	79.8	90.8	95.8	-	-	-	-	-
D & C [TPAMI21] [64]	512	79.8	90.4	95.2	-	90.4	97.6	-	-
IBC [ICML21] [18]	512	81.4	91.3	95.9	-	92.8	98.5	99.1	99.2
LSCM-GNN [TIP22] [2]	512	79.7	90.5	95.7	98.4	92.4	98.5	99.1	99.3
Group Loss++ [TPAMI22] [7]	512	79.2	90.1	95.8	-	90.9	97.6	98.4	98.9
IRT_R [arXiv21] [1]	384	84.0	93.6	97.2	99.1	91.5	98.1	98.7	99.0
PA+DIML [ICCV21] [17]	128	79.22	-	-	-	-	-	-	-
XBM+RTT [ICCV21] [21]	128	84.5	93.2	96.6	99.0	-	-	-	-
Hyp-DeiT[CVPR22] [20]	128	83.0	93.4	97.5	99.2	90.9	97.9	98.6	98.9
Ours: CRT	128	83.41	**93.86**	**97.66**	**99.31**	**94.48**	**99.37**	**99.68**	**99.75**
Ours: Gain	128	−1.09	**0.26**	**0.16**	**0.11**	**1.68**	**0.87**	**0.58**	**0.45**

Experimental Results

Comparisons of Recall@K (%) on the CUB, Cars, SOP and In-Shop datasets for different backbone networks.

Backbones	Methods	CUB			Cars			SOP			In-Shop		
		R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100	R@1	R@10	R@20
GoogLeNet	MS	56.53	69.13	79.54	76.49	84.76	90.38	70.10	85.73	94.41	87.33	97.73	98.56
	+CRT	**59.23**	**71.02**	**81.52**	**78.50**	**86.35**	**91.91**	**71.06**	**86.41**	**94.73**	**88.31**	**97.83**	**98.68**
	Gain	**2.7**	**1.89**	**1.98**	**2.01**	**1.59**	**1.53**	**0.96**	**0.68**	**0.32**	**0.98**	**0.1**	**0.12**
BN-Inception	MS	61.95	72.59	82.44	80.59	87.50	92.63	74.10	88.46	95.43	90.75	98.52	99.06
	+CRT	**65.78**	**76.72**	**85.31**	**81.38**	**88.38**	**93.08**	**75.65**	**89.04**	**95.64**	**91.52**	**98.61**	**99.19**
	Gain	**3.83**	**4.13**	**2.87**	**0.79**	**0.88**	**0.45**	**1.55**	**0.58**	**0.21**	**0.77**	**0.09**	**0.13**
ResNet-50	MS	62.64	73.73	83.20	79.92	87.63	92.41	76.49	89.33	95.66	90.11	97.52	98.27
	+CRT	**64.20**	**75.54**	**84.12**	**83.29**	**89.76**	**93.88**	**78.97**	**91.10**	**96.51**	**92.38**	**98.77**	**99.25**
	Gain	**1.56**	**1.89**	**0.92**	**3.37**	**2.13**	**1.47**	**2.48**	**1.77**	**0.85**	**2.27**	**1.25**	**0.98**
DeiT-S	MS	72.32	82.43	89.20	82.20	89.34	93.76	79.56	91.63	96.87	92.47	98.72	99.21
	+CRT	**74.71**	**83.83**	**89.65**	**84.26**	**90.95**	**94.90**	**81.60**	**92.65**	**97.20**	**93.31**	**98.98**	**99.35**
	Gain	**2.39**	**1.4**	**0.45**	**2.06**	**1.61**	**1.14**	**2.04**	**1.02**	**0.33**	**0.84**	**0.26**	**0.14**
MiT-B1	MS	72.25	81.85	88.54	87.36	92.26	95.23	79.67	91.55	96.66	92.20	98.64	99.21
	+CRT	**75.95**	**84.47**	**90.26**	**89.60**	**94.20**	**96.47**	**82.32**	**93.02**	**97.19**	**93.51**	**99.11**	**99.50**
	Gain	**3.7**	**2.62**	**1.72**	**2.24**	**1.94**	**1.24**	**2.65**	**1.47**	**0.53**	**1.31**	**0.47**	**0.29**

Ablation Studies

The Recall@K (%) for different component on the CUB dataset.

Model	CUB			
	R@1	R@2	R@4	R@8
The Proposed Method	**75.95**	**84.47**	90.26	**94.51**
−CRT	73.97	84.03	**90.41**	94.31
−CRT-Consistency	72.25	81.85	88.54	93.57

The Recall@K (%) for different number of projection prototypes in the first embedding branch.

K_1	CUB			
	R@1	R@2	R@4	R@8
1	75.76	84.52	90.72	**94.78**
4	75.57	84.30	90.56	94.33
16	75.44	84.25	90.24	94.46
64	**75.96**	**84.60**	**90.75**	94.31

The Recall@K (%) with (w) and without (w/o) using multi-perspective CRT feature transformation.

	CUB			
	R@1	R@2	R@4	R@8
w/o	74.34	83.59	90.23	94.14
w	**75.95**	**84.47**	**90.26**	**94.51**

Ablation Studies

The Recall@K (%) for different number of projection prototypes in the second embedding branch.

K_2	CUB			
	R@1	R@2	R@4	R@8
1	74.26	84.30	89.82	93.96
4	74.54	84.03	90.06	94.01
16	74.63	83.74	90.24	94.02
64	**75.95**	**84.47**	90.26	**94.51**
100	75.15	84.35	**90.36**	94.31

The Recall@K (%) with (w) and without (w/o) shared weights between these two embedding branches.

	CUB			
	R@1	R@2	R@4	R@8
w	75.95	84.47	90.26	**94.51**
w/o	**75.96**	**84.60**	**90.75**	94.31

Ablation Studies

Parameters count and FLOPs at resolution 227×227.

Complexity Index	GoogLeNet	BN-Inception	ResNet-50	DeiT-S	MiT-B0	MiT-B1	MiT-B2
Params(M) ↓	5.60	10.27	23.51	21.96	3.31	13.14	24.17
FLOPs(G) ↓	1.52	2.06	4.65	4.24	0.47	1.83	3.55

Experimental results on the CUB and Cars datasets for different MiT backbone networks.

Method	Backbones	CUB				Cars			
		R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
CRT	MiT-B0	69.19	79.66	86.83	92.23	85.34	91.28	94.75	97.05
	MiT-B1	75.95	84.47	90.26	94.51	89.60	94.20	96.47	97.79
	MiT-B2	78.98	86.68	91.61	95.04	91.16	94.92	96.79	98.03

Experimental results on the SOP and In-Shop datasets for different MiT backbone networks.

Method	Backbones	SOP				In-Shop			
		R@1	R@10	R@100	R@1000	R@1	R@10	R@20	R@30
CRT	MiT-B0	80.16	91.60	96.50	98.90	92.90	99.06	99.42	99.51
	MiT-B1	82.32	93.02	97.19	99.20	93.51	99.11	99.50	99.63
	MiT-B2	83.41	93.86	97.66	99.31	94.48	99.37	99.68	99.75

Ablation Studies

The Recall@K (%) for different λ_2.

λ_2	CUB			
	R@1	R@2	R@4	R@8
0	73.09	82.38	89.18	93.91
0.1	73.87	83.12	89.92	93.99
0.3	74.93	83.96	90.46	94.16
0.5	74.22	83.52	90.16	93.94
0.7	75.14	84.10	90.48	94.46
0.9	**75.95**	84.47	90.26	**94.51**
1.0	75.07	**84.54**	**90.77**	94.16

The Recall@K (%) for different batch sizes.

N	CUB			
	R@1	R@2	R@4	R@8
32	73.38	83.20	90.23	94.29
64	75.19	84.47	90.45	94.43
80	75.95	84.47	90.26	94.51
120	75.79	84.39	90.38	94.50
150	75.78	84.72	90.75	94.68
180	75.32	84.59	90.34	94.13

The Recall@K (%) for different dimensions of embedding.

The First Embedding Branch					The Second Embedding Branch				
Dim d	R@1	R@2	R@4	R@8	Dim D	R@1	R@2	R@4	R@8
64	72.57	81.70	89.10	93.62	1024	75.74	84.47	90.38	94.41
128	75.95	84.47	90.26	94.51	1024	76.52	85.33	90.94	94.65
256	76.16	84.28	90.63	94.60	1024	76.82	84.89	90.78	94.60
512	76.18	85.23	90.99	94.68	1024	75.66	85.08	90.75	94.48
1024	77.09	85.42	91.19	95.05	2048	76.40	85.43	91.46	94.99

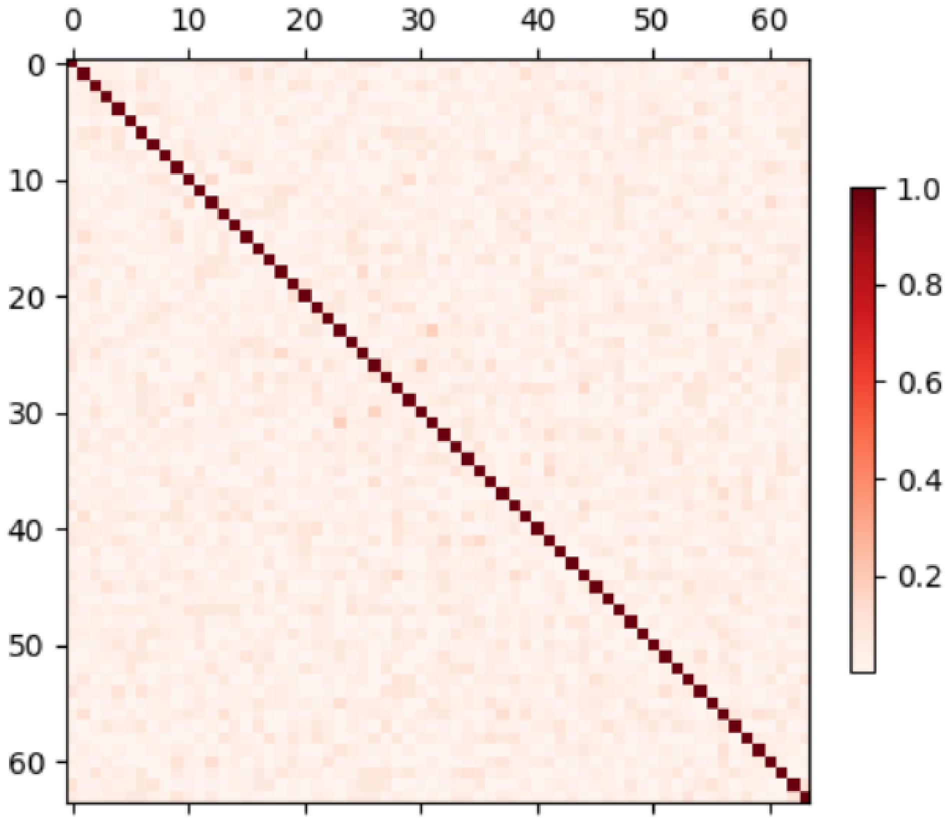
Key Idea : Coded Residual Transform

Correlation Heat Maps

Correlation heat maps between a learned set of prototypes (corresponding to 8 prototypes) and feature maps of two images. We can see that different prototypes response to different local or global views. The red dots on the top right corner indicates the background prototype.

Key Idea : Coded Residual Transform

Correlations Between Prototypes



Generalization Capability Analysis

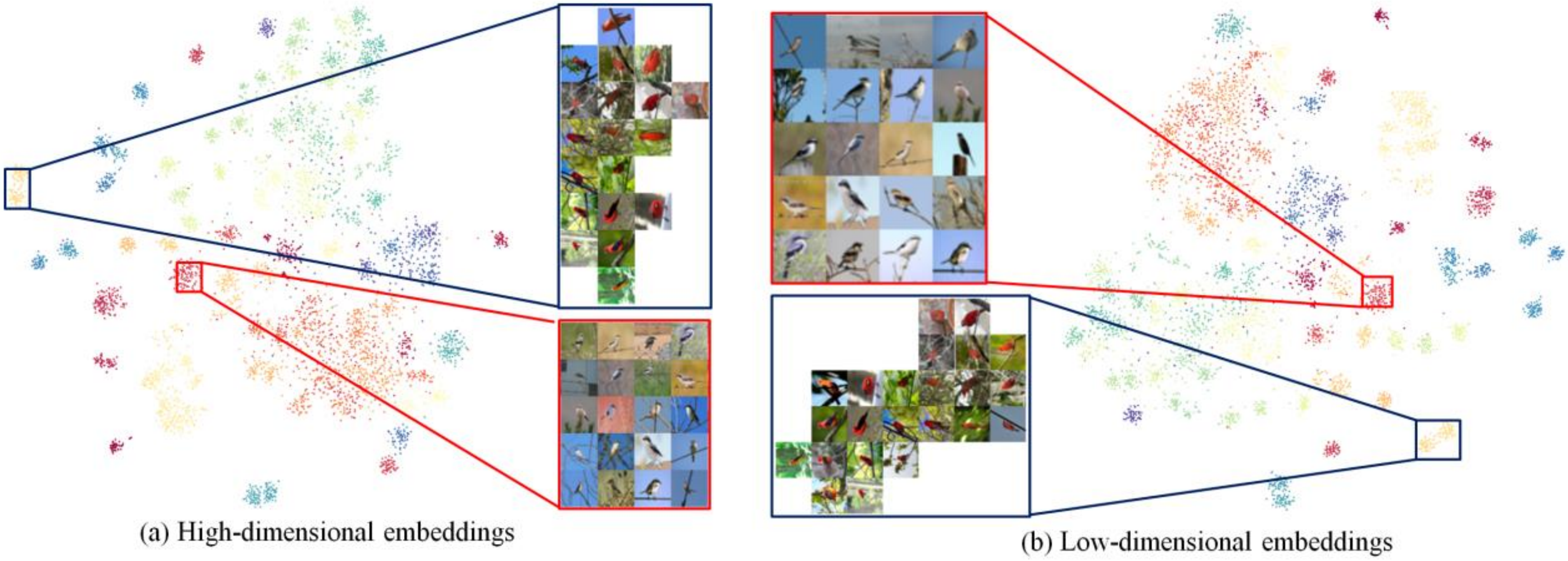
The embedding space density (↑) on the experimental datasets.

Method	CUB	Cars	SOP	In-Shop
Baseline	0.72	0.79	0.38	0.28
+CRT	**0.90**	**1.01**	**0.40**	**0.33**

The spectral decay (↓) on the experimental datasets.

Method	CUB	Cars	SOP	In-Shop
Baseline	0.27	0.24	0.31	0.23
+CRT	**0.19**	**0.15**	**0.13**	**0.10**

Visualization examples



The t-SNE visualizations of high-dimensional embeddings and low-dimensional embeddings on the CUB dataset.

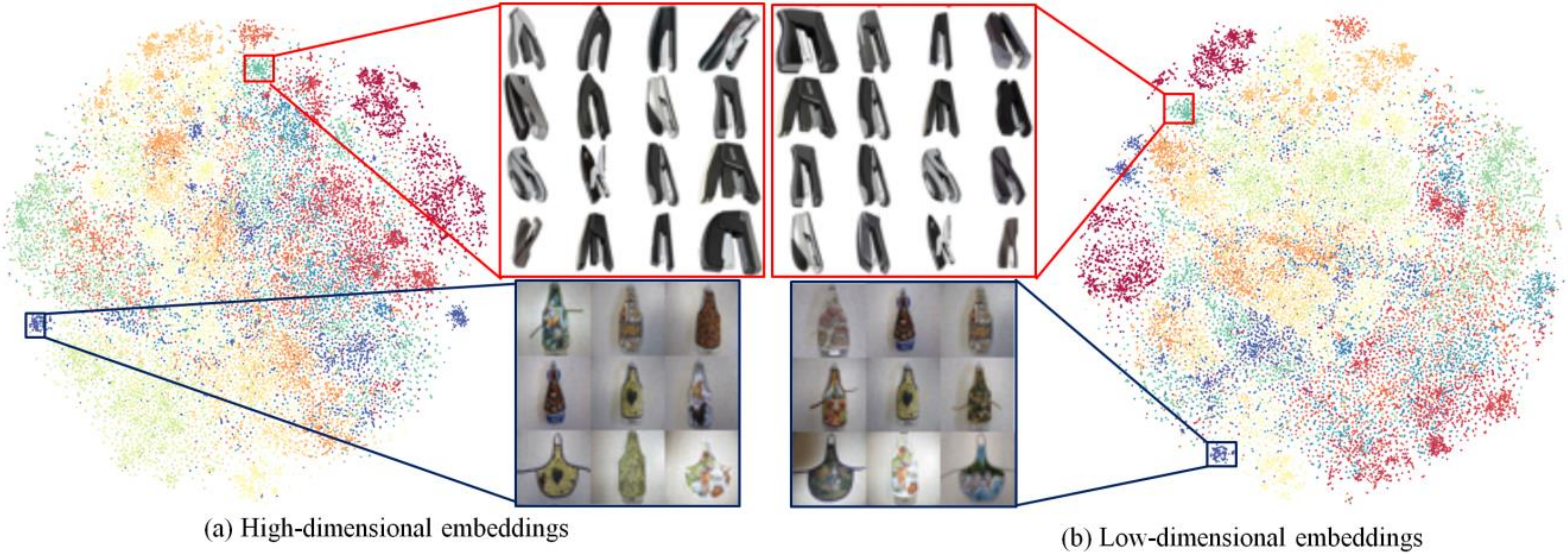
Visualization examples

(a) High-dimensional embeddings

(b) Low-dimensional embeddings

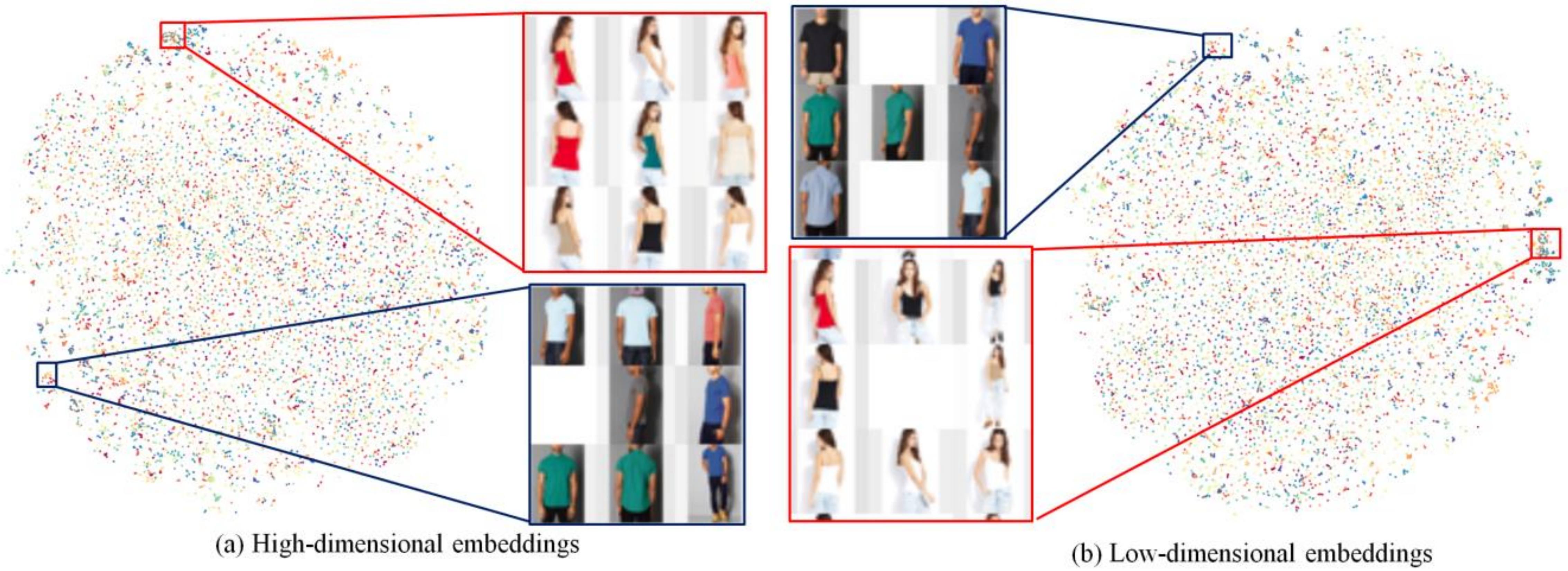
The t-SNE visualizations of high-dimensional embeddings and low-dimensional embeddings on the Cars dataset.

Visualization examples



The t-SNE visualizations of high-dimensional embeddings and low-dimensional embeddings on the SOP dataset.

Visualization examples



The t-SNE visualizations of high-dimensional embeddings and low-dimensional embeddings on the In-Shop dataset.

Conclusion

- We have developed a **coded residual transform** for generalizable deep metric learning, which consists of a **multi-perspective projection** and **coded residual transform encoder** and **a cross-CRT correlation consistency constraint**.
- It represents and encodes the feature map from a set of complimentary perspectives based on projections onto diversified prototypes.
- Unlike existing transformer-based feature representation approaches which encode the original values of features based on global correlation analysis, the proposed coded residual transform encodes the relative differences between original features and their projected prototypes.
- **One limitation** is that the memory and compute usage will be increased during training for these two embedding branches, and we shared weights between them to solve this problem in our experiments.
- **Another limitation** is that the projection prototypes were learned from the training set. It is unclear whether it is the best projection prototypes for new test classes.

THANK YOU！ Q & A