Beyond the CLS Token: Image Reranking using Pre-trained Vision Transformers

Chao Zhang Stephan Liwicki Roberto Cipolla

https://github.com/cazhang/vit-reranking

Motivations

- Visual similarity learning is an important topic
- Existing methods using global image similarity
- Conv+ViT provides better interpretability

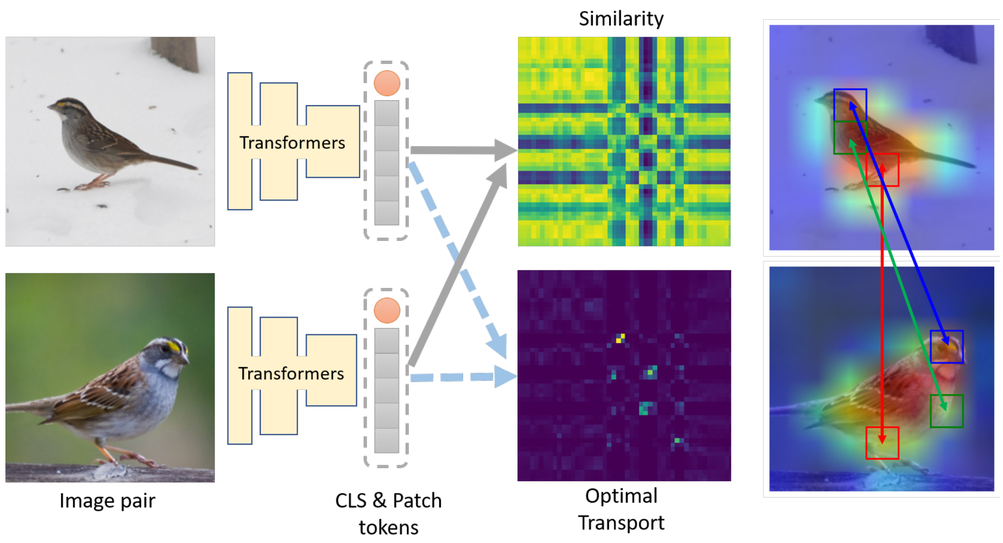
Proposed method

- Transformer-based feature extraction
- Combine global and local image similarity
- Relevancy maps and partial optimal transport

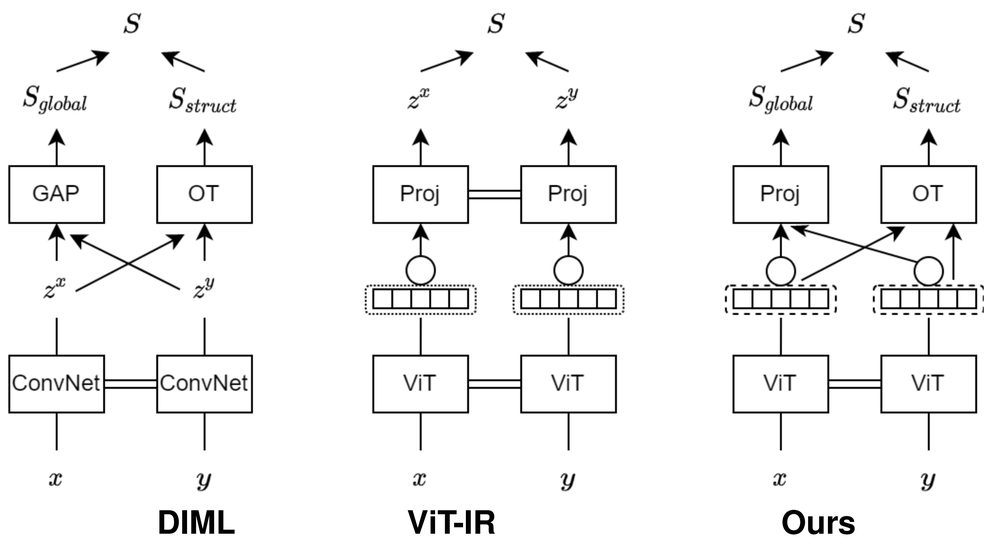
Contributions

- Improvements w/o changes to existing ViT
- Improved performance and better interpretability
- Extra training or labels not needed

Our pipeline includes: **base model training**, **structural similarity learning** and **image re-ranking**. Base model training uses global similarity only, eg. CLS token of ViT. Structural similarity learning relies on local patch features and global features. The optimal transport plan T is used to aggregate the local similarity S_{struct} and then combined with global similarity S_{global} for re-ranking.



(a) Overview of our ViT-based image re-ranking approach

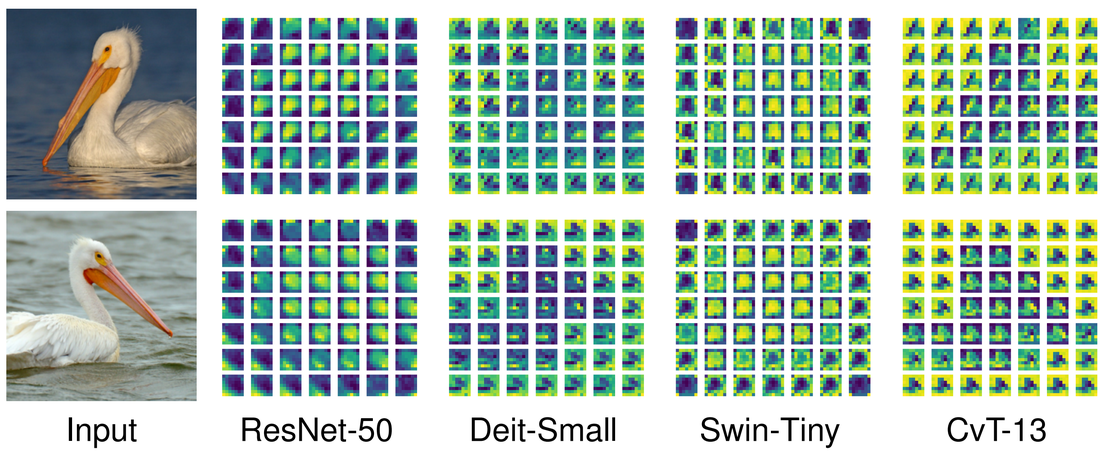


(b) High-level comparison with DIML and ViT-IR

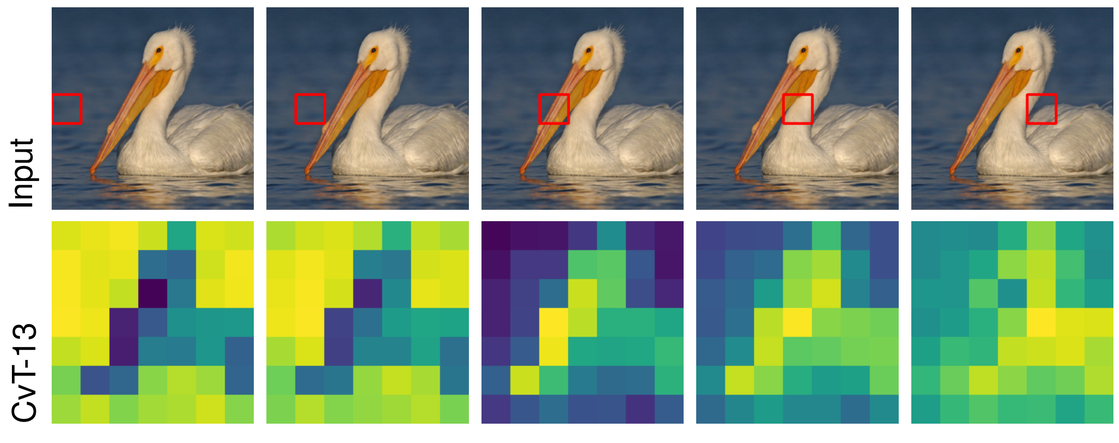
Structural Metric Learning using ViT

- Deep Metric Learning: $S_{global} = s(g^s, g^t) \in \mathcal{R}, \quad S_{struct} = s(f^s, f^t) \in \mathcal{R}^{hw \times hw}$
- Optimal Transport: $T = \text{Sinkhorn}(S_{struct}, \mu^s, \mu^t)$, where μ^s, μ^t reflect the importance of each location for the optimization.
- Re-rankng: combine S_{global} and S_{struct}, where $S_{\text{struct}}(f^s, f^t) = \sum_{1 \le i,j \le hw} s(f_i^s, f_j^t)\hat{T}_{i,j}$

Patch Similarities of Multiple Models



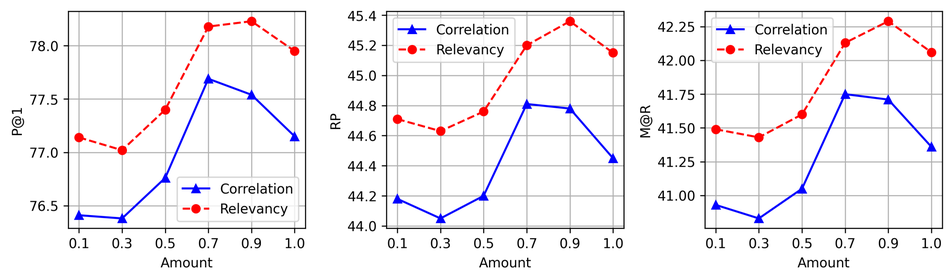
Closer Look at Convolutional Vision Transformers



Task: Image Retrieval

Method	Stage	CUB200-2011			Cars196			Online Products		
		P@1	RP	M@R	P@1	RP	M@R	P@1	RP	M@R
DIML	Base	62.12	34.50	23.40	77.43	34.25	23.57	77.41	45.09	41.74
	Struct	64.97	35.28	24.45	83.17	35.10	25.60	78.86	46.22	43.00
Deit-S	Base	70.39	39.96	29.22	76.27	32.31	21.57	78.13	45.96	42.77
	Struct	70.12	38.19	26.80	72.74	31.51	18.97	77.65	44.60	41.40
CvT-13	Base	71.75	41.94	31.19	80.55	35.30	24.80	77.15	44.74	41.53
	Struct	73.72	42.68	32.12	83.66	35.63	25.82	77.15	44.45	41.36
Swin-T	Base	74.47	**43.43**	**32.86**	83.32	37.51	27.22	79.42	47.42	44.33
	Struct	**74.98**	43.19	32.78	**85.07**	**37.79**	**27.87**	**80.02**	**47.90**	**44.86**

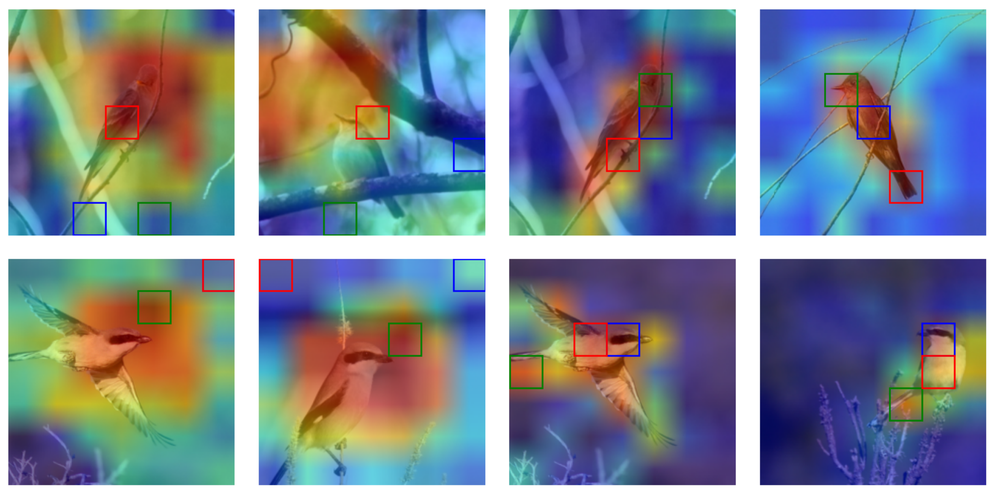
Partial Optimal Transport on SOP Dataset



Task: Place Recognition on MSLS Dataset

Method	Struct	Recall@1	Recall@5	Recall@10
NetVLAD	-	52.97	70.54	75.54
Ours (Scratch)	✗	40.95	64.05	72.30
	✓	56.62	73.24	78.38
Ours (Distill)	✗	60.68	74.59	79.46
	✓	66.35	78.24	81.76

Semantic Correspondence without Part Annotations



(a) DIML Query (b) DIML Match (c) Our Query (d) Our Match

Top-5 Retrieval Results of Car196