Beyond the CLS Token: Image Reranking using Pretrained Vision Transformers

Chao Zhang[1]
chao.zhang@crl.toshiba.co.uk
Stephan Liwicki[1]
stephan.liwicki@crl.toshiba.co.uk
Roberto Cipolla[1,2]
rc10001@cam.ac.uk

[1] Cambridge Research Lab
Toshiba Europe Ltd
Cambridge, UK

[2] Department of Engineering
University of Cambridge
Cambridge, UK

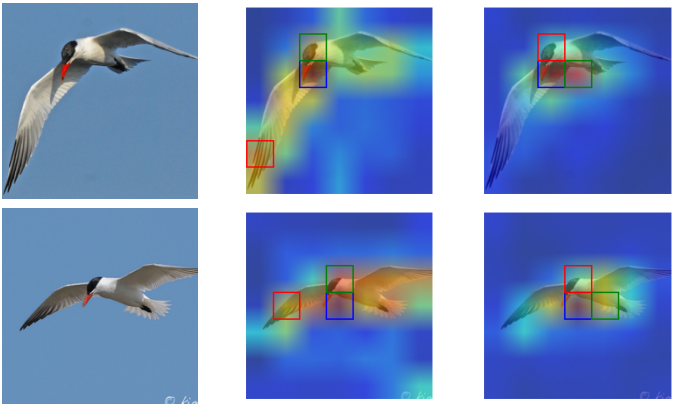
Abstract

We propose to leverage structural similarity of pretrained vision transformers for image retrieval reranking. Vision transformers have become the dominant architecture in many computer vision tasks. However, the usage of global representation (**CLS** token) makes for the lack of interpretability. Since not all patches are equally important for image similarity, our idea is to exploit a pretrained model for optimal spatial weights assigned to local patch tokens. To understand the relationship between global and local representations of vision transformers, we compare multiple transformers architectures against ResNet using similarity as an indicative measure. Our analysis suggest that the usage of convolutions within vision transformers is vital to learn suitable patch embeddings for structural similarities. We also find that local patch similarity equipped with an optimal transport solver could improve image retrieval accuracy compared to the one using global similarity only. Without re-training, our evaluations with off-the-shelf pretrained vision transformers show that the use of structural similarity not only boosts retrieval performance, but also provides visualization cues for interpretable image similarity. Evaluations on three benchmarks show that our proposed structural approach outperforms the state of the art for interpretable image retrieval.

1 Introduction

Visual similarity learning is an important topic for computer vision. It is related to a range of practical applications such as image retrieval [20, 29] and visual localization [31]. Deep metric learning (DML), leveraging state-of-the-art deep neural networks, has advanced visual similarity research recently. However, most DML methods represent images as embedding vectors and use the similarity in embedding space to encode the semantic similarity. Although often effective, DML methods are often lacking interpretability for its output.

As one of the most successful CNN architectures, ResNet [11] is widely used as backbone in DML. Hierarchical design, translation invariance and local receptive field all contribute to its success in computer vision domain. Recently, inspired by the success of transformers in natural language processing, vision transformers (ViT) [9] and other variants are



Images Cross Attention [35] Relevancy (Ours)

Figure 1: Different patches of an image contribute differently to the similarity score. A comparison of DIML [35] using cross-attention and ours using transformers' relevancy is shown. Top matches are obtained by finding the maximum values of the weighted structural similarity. The weights are obtained by solving optimal transport problem. (See Sec.3.3 for details.)

quickly adopted, and demonstrate their competitive performances both in vision task (image classification [9], semantic segmentation [18]) and in multi-modal tasks (visual question answering [33]). ViT has been applied in image retrieval task and shown better performance than CNNs [10]. However, similar to image classification, only the holistic embedding is used. Inspired by CNNs, we argue that patch embeddings from vision transformers could serve as better local features with global receptive field. In other words, patch tokens are good resources to explain structural similarity and to improve retrieval performance.

On the one hand, recent works [24, 27] exploring feature correspondence have adopted a hybrid model with modular design. Apart from the CNNs-based feature extraction module, an attention module is appended to leverage global context. We note, the architectural progress of vision transformers are borrowing ideas from CNNs. Examples include pyramid architecture [30] and local self-attention using shifted windows [18]. In this work, we attempt to compare CNNs and vision transformers from the structural similarity perspective. This will not only help understand and design better models, but also provide interpretability for the decision made by deep models.

Apart from the added interpretability, leveraging local feature maps does not require extra learning at all. In CNNs, patch level features are available before the aggregation of global average pooling and projection of fully connected layers. For ViT, local patch tokens are trained together with a special CLS token. All tokens interact each other with self-attentions. To adapt a permutation-invariant transformer to work on images, position embeddings are added to the patch embedding.

In summary, we exploit local features of pretrained vision transformers to improve image retrieval performance and to provide visual cues for similarity interpretability. In Fig.1, we highlight our method by visualizing the importance map for the image similarity as well as the top matching patches of a pair images from CUB-200. Our contributions are as follows:

1. We compare the patch representations of various vision transformers architectures with ResNet, and find that convolution operations play an important role to learn locally

smooth and globally discriminative patch embeddings.

2. We propose a training-free, transformer based framework to improve deep metric learning performance through image re-ranking.
3. We apply the attention-based relevancy maps tied to vision transformers to guide optimal transport optimization and further validate the effectiveness of partial optimal transport for dataset showing strong viewpoint and scale variations, such as SOP [20].
4. We demonstrate the effectiveness of the proposed method on three benchmarks of fine-grained image retrieval and one visual place recognition task.

2 Related work

Deep Metric Learning: Deep metric learning (DML) has recently become one of the primary frameworks for vision tasks such as image retrieval, person re-identification and face recognition. The basic idea of DML is to learn image embeddings to reflect the semantics among samples. Towards this goal, most proposed approaches focus on one of two aspects: loss functions [6, 14, 20, 25] or sampling strategies [23, 32, 36]. As the state-of-the-art methods advance performance on several benchmarks, they also become deeper and obscure, leading to over-fitting and brittle performance. Thus, there is an increasing need to interpret the decision made by the models. However, methods using embedding vectors alone often lack this interpretability. Inspired by DIML [35], we leverage the spatial structure for improved and interpretable metric learning.

Vision Transformers: Transformers have shown outstanding results in natural language understanding and computer vision. The pioneering work, Vision Transformers (ViT) [9], directly applied transformer architectures from NLP to image classification. To improve the training efficiency of ViT, DeiT [28] introduced token-based distillation with Convolutional Neural Networks (CNNs) as the teacher. Follow-up works explore the direction to combine CNNs and ViT. PVT [30] introduced the pyramid structure into ViT, which generates multi-scale feature for dense prediction tasks. CvT [33] leveraged convolutional patch embedding and convolutional attention projection to combine the best aspects of both CNNs and transformers. The Swin Transformer [18] introduced a shifted window scheme to limit self-attention within windows while allowing interaction between windows. In [10], image descriptors generated by vision transformers are used for the image retrieval task. Although improvements over CNNs are reported, it is not clear why vision transformers perform better. Unlike [10] which uses transformers' class token only, we consider both CLS token and patch tokens for image retrieval to improve interpretability and accuracy.

Optimal Transport for Feature Matching: Similar to image retrieval, inputs to feature matching are image pairs. The goal of feature matching is to establish pointwise correspondence using local features. Recently, methods combining the attention mechanism with CNNs features are the state of the art. Given keypoint descriptors, SuperGlue [24] uses a graph neural network and attention layers to solve an assignment problem. In [17], an Optimal Transport (OT) layer is adopted to obtain the semantic correspondence. Matching quality is improved by suppressing one-to-many matchings. LoFTR [27] proposes a two-stage method using coarse and fine level features with optimal transport. Given the feature

maps of two images, COTR [13] concatenate and feed feature maps to a transformer with query point as input. The output is further fed into a decoder to infer the correspondence. Among these approaches, we find two common differences with image retrieval. First, all methods require CNNs backbone for feature extraction. Second, feature matching heavily depends on datasets with dense feature correspondence for training. Examples are ScanNet [8] and MegaDepth [16]. In our work, unlike feature matching, optimal transport is exploited within a metric learning framework, in which only image level labels are available.

Interpretable Deep Vision Models: With deep learning dominating various tasks in computer vision, improving the explainability and interpretability has attracted more attention recently. For CNNs, mainstream methods either visualize feature representations [26, 37], or disentangle mixed patterns learned in each layer of CNNs [34]. Beyond classification tasks, a gradient-weighted method [26] is also adapted to embedding network in [6]. For vision transformers, a common class-agnostic method to understand its predictions is to consider the attentions as relevancy scores. Instead of taking a single attention layer, attention rollout [1] proposed to combine all attention maps in a linear way and to reassign all attention scores. A class-specific visualization method for self-attention models is proposed in [4]. It incorporates both relevancy and gradient information. Apart from visualization methods, recent work [21, 22] tried to analyze the internal representation structure of CNNs and vision transformers using classification task. We focus on improving metric learning by leveraging the representation structure of transformers. Finally we note, our method gives an indication of learning semantic feature correspondences using image labels alone.

3 Proposed Approach

We now present our approach based on the structural similarity of vision transformers for the metric learning task. First, we describe the framework called structural deep metric learning in Sec.3.1 and then review vision transformers especially the variants with convolutions in Sec.3.2. We detail our proposed method in Sec.3.3, with a novel attention-based relevancy maps and a partial extension of the optimal transport solver.

3.1 Background: Structural Deep Metric Learning

Deep Metric Learning (DML): Given a pair of images, DML uses deep neural networks to find the distance metric so that the embedding similarity reflects the semantic similarity defined by image classes. In particular, given source image x^s and target image x^t, the global similarity is given by:

$$S_{\text{global}}(f^s, f^t) = s(f^s, f^t), \tag{1}$$

where f^s and $f^t \in R^D$ are global representation of dimension D, and $s(\cdot,\cdot)$ is a similarity function (eg. Cosine similarity). For CNNs such as ResNet, f is obtained by global average pooling and fully connected layer on the feature maps of the final convolutional layer.

Structural Similarity using Optimal Transport: In order to take advantage of the spatial structures of images, a structural matching scheme called DIML was proposed in [35]. The idea is to consider both the global and structural cost for metric learning. In particular,

ResNet-50 was adopted as the backbone, and then the structural cost is computed based on the feature maps f^s and $f^t \in R^{hw \times D}$ of the final convolutional layer, where h and w are the height and width of the feature maps. Given structural cost matrix $C_{ij} \in R^{hw \times hw}$, now the aim is to minimize the overall matching cost as $\sum_{i=1}^{hw}\sum_{j=1}^{hw} C_{ij}T_{ij}$, where T is the optimal matching flow, indicating the pairwise weight towards the final similarity. In the discrete case, this problem can be formalized as an optimal transport one, in which the goal is to find the optimal transport plan. Given the two corresponding discrete distributions $\mu^s \in R^{hw}$ and $\mu^t \in R^{hw}$, the optimization problem becomes:

$$\hat{T} = \arg\min_{T}(\sum_{i=1}^{hw}\sum_{j=1}^{hw} C_{ij}T_{ij}), \text{ subject to } \quad \hat{T}\mathbf{1} = \mu^s \text{ and } \hat{T}^{\mathrm{T}}\mathbf{1} = \mu^t \tag{2}$$

To solve the above problem, Sinkhorn divergence algorithm [7] using an entropic regularizer is used to enable fast convergence. As suggested in [35], the cross-correlation between global feature and local feature maps is considered as the marginal distribution μ^s and μ^t for optimal transport. Once the optimal transport $\hat{T}$ is obtained, we define the structural similarity as follows:

$$S_{\text{struct}}(f^s, f^t) = \sum_{1 \le i,j \le hw} s(f_i^s, f_j^t)\hat{T}_{i,j} \tag{3}$$

Based on the top K candidates returned by global similarity S_{global}, the final retrieval results using structural matching can be obtained by combining S_{global} and S_{struct} together.

3.2 Vision Transformer with Convolutions

We first revisit the basics of vision transformers. Then, we describe the variant that uses convolution and show that the introduction of convolution is vital for structural metric learning.

Transformer encoders consist of alternatively stacked multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks. The layer normalization (LN) and residual connection are applied before and after each block, respectively. Specifically, let $t_l \in R^{N \times D}$ denote the output of the l^{th} transformer layer, where N is the number of tokens, and D is the feature dimension. Specifically:

$$\tilde{t}_l = \text{MSA}(\text{LN}(t_{l-1})) + t_{l-1}, \quad t_l = \text{MLP}(\text{LN}(\tilde{t}_l)) + \tilde{t}_l \tag{4}$$

where $0 < l \le L$ denotes the transformer layer and t_0 is the input.

In order to compare the local patch embedding of vision transformers to CNNs, we choose three variants of vision transformer architectures: DeiT [9], Swin Transformer [18] and CvT [33].The main difference between DeiT and CvT is the spatial awareness for transformer attention. While DeiT use pure self-attention, CvT introduces two convolution-based operations into the vision transformer, namely Convolutional Token Embedding and Convolutional Projection for attention. Both SwinT and CvT adopt a multi-stage hierarchical design similar to ResNet. Given an image represented by $H \times W$ non-overlapping patches, we visualise the cross patch similarity map in Fig.2. For all variants, $H = W = 7$. Please refer to Supp.Mat. for the zoom-in version of patch-patch similarity maps and CLS-patch similarity maps.

ResNet-50 shows smooth but blurry changes when moving to nearby patches. On the contrary, different transformer architectures demonstrate strikingly interesting patterns. For

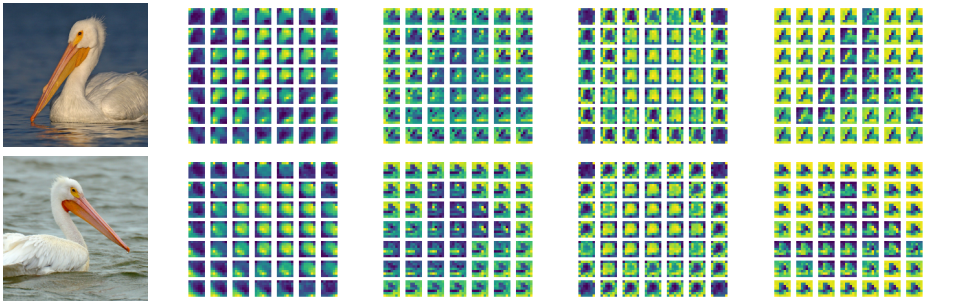


Figure 2: We show the pairwise patch similarity for a single image in each row. Consider an image I consisting of 4 patches, $P_1 \dots P_4$, we show the block similarity matrix $B_1 \dots B_4$, where B_1 is the pairwise similarity of P_1 to all patches, *e.g.* $B_1 = [s_{11}, s_{12}; s_{21}, s_{22}]$, where s_{ij} is the cosine similarity between P_i and P_j. Adjacent patches in Swin and CvT demonstrate smooth transitions and translation invariance (see also the emerging segmentation using CvT). While with DeiT, background patches show unexpected variations (see top-left of row 1)

DeiT, position embedding is added to the patch embedding to form the input tokens, leading to better discriminative representation. However, this does not show as good smoothness as ResNet. Shifted window scheme of SwinT limits attention to neighbouring patches. However, the corner patches show extreme similarity due to artefacts of the shifted window partition. CvT manages to separate background from the foreground in this example. Notice, semantic intra-class features are correlated in CvT while inter-class semantics are distinguished. For structural similarity learning, good properties of the representation should be locally smooth and semantically discriminative. Comparing to ResNet and vanilla ViT, we hypothesize that the introduction of convolution to ViT satisfies the two requirements.

3.3 Structural Metric Learning using Transformers

We now show how to perform structural metric learning using vision transformers. For two images x^s and x^t, we first obtain global and local feature maps. Global representation $\{g^s, g^t\} \in \mathcal{R}^D$ correspond to CLS token of the transformers. Spatial Feature maps $\{f^s, f^t\} \in \mathcal{R}^{hw \times D}$ are patch tokens. Next, we compute both global similarity S_{global} and structural similarity S_{struct} as follows:

$$S_{global} = s(g^s, g^t) \in \mathcal{R}, \quad S_{struct} = s(f^s, f^t) \in \mathcal{R}^{hw \times hw} \tag{5}$$

where $s(\cdot,\cdot)$ is a similarity function. With the structural similarity, we follow [35] to use an optimal transport solver to maximize the total similarity under the optimal assignment plan T as follows:

$$T = \text{Sinkhorn}(S_{struct}, \mu^s, \mu^t) \tag{6}$$

where μ^s, μ^t are corresponding discrete marginal distributions. Intuitively, μ^s, μ^t reflect the importance of each location for the optimization.

Relevancy Score as Marginal Distribution At the core of our method is how to choose the marginal distribution for vision transformers. Cross-correlation is proposed in [35], to

exploit the global and local feature maps of CNNs. It is trivial to apply this similarly to ViT, using the CLS token as global feature and the patch tokens as local features. However, one issue with cross-correlation is that the it assumes that object foreground and background is well-defined. On the contrary, vision transformers take advantage of self-attention of ViT and therefore are supposed to have learnt which parts are more important for the metric learning task. We propose to leverage the idea of relevancy scores [9] as the importance map for optimal transport distributions.

Aggregated attention is obtained by multiplying attention maps from all attention layers. It was originally used for the purpose of interpreting transformers classification [1]. In our method, the relevancy map is used to guide the optimal transport optimization for structural similarity. The relevancy map can be obtained by a forward pass of transformers, and it is theoretically applicable to almost all the transformers architectures [1] that use global attentions such as DeiT and CvT.

Given a Transformer with L layers, we compute the attention from all positions in last layer l_{L-1} to all positions in input l_0. To compute the attentions from l_i to l_j, we recursively multiply the attention weights matrices as below:

$$\tilde{A}(l_i) = \text{norm}(A(l_i) + \mathbf{I})\tilde{A}(l_{i-1}) \quad \text{if} \quad i > j \tag{7}$$

with $\tilde{A}(l_0) = A(l_0)$. We denote $\tilde{A}$ as attention rollout and A as raw attention. To account for residual connection in Transformers, identitiy matrix $\mathbf{I}$ is added to the attention map and then L2 normalized. When hierarchical structure is used in Transformers such as CvT-13, attention maps are resized to the dimension of final layer attention. To handle that the CLS token is only introduce in the last stage of CvT-13, we discard the CLS token attention from the final stage attentions by $A(l_i) = A(l_i)[1:,1:]$, where $l_i \in$ Stage2. Finally, we average the attentions of all spatial patches to obtain the approximated marginal distributions $\mu^s = \tilde{A}^s$ and $\mu^t = \tilde{A}^t$.

Partial Optimal Transport As we notice that the assumption of standard OT may not be valid for images under strong viewpoint and scale changes, such as Stanford Online Products (SOP). In other words, enforcing full matching flow for positive image pairs but without enough semantic correspondence is difficult to optimize. To alleviate this issue, OT is extended to its partial version to allow flexible amount of matching flow.

Given a cost function $C \in \mathrm{R}^{m,n}$ and the corresponding marginal distributions $u \in \mathrm{R}^m$ and $v \in \mathrm{R}^n$. Assuming u and v are unit length, so that $\|u\|_1 = \|v\|_1 = 1$. Following [3], the OT problem is converted to a partial one by adding a dummy point to both images with fraction $0 \leq s \leq 1$. The mass of the dummy point is set such that $u_{m+1} = 1 - s$ and $v_{n+1} = 1 - s$. We extend cost matrix $\bar{C}$ with $\bar{u} = [u, u_{m+1}]$, $\bar{v} = [v, v_{n+1}]$. Then we solve the extended OT to obtain $\bar{T} = \text{Sinkhorn}(\bar{C}, \bar{u}, \bar{v})$. Finally, the partial optimal transport T is obtained by discarding the last row and last column of $\bar{T}$. We validate the effectiveness of this extension on SOP dataset considering the dataset bias towards viewpoint and scale variations.

4 Experiments

To evaluate the performance of the proposed method, we conduct experiments on three datasets used widely in the image retrieval task: CUB200-2011 [29], Cars196[15], and Stanford Online Products (SOP)[20]. Details about datasets are included in the Supp. Mat.

Table 1: Comparison of multiple ViT on image retrieval. Base variants use global representation, and struct variants use structural similarity to rerank top-100 candidates.

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**

Following [35], we adopt the three evaluation metrics used in [19]: Precision@1 (P@1), R-Precision (RP), and MAP@R (M@R). We also evaluate our method on one visual place recognition benchmark MSLS[31].

It is worth noting that our method assumes a pretrained global model exists and no further training is needed for structural reranking. The global model is pretrained on ImageNet and finetuned on target datasets using global features only. In this section, we compare vision transformers of multiple variants to state-of-the-art method DIML using ResNet-50. For all the experiments, we set the truncation number $K = 100$, feature map size $h = w = 7$, embedding size $D = 128$. Margin loss is used in base model training. See Supp. Mat. for additional results.

CNNs vs. Transformers We report baseline results using global similarity only and structural similarity for reranking in Tab.1. First, Transformers shows better performance than ResNet-50 on all datasets using global representation. Second, baseline results can be improved when structural similarity is incorporated. Both CvT and SwinT benefit from structural similarity. We also see that vanilla ViT (DeiT) does not work well. This is consistent to our finding that convolution operations boost learning smooth and discriminative patch features inside ViT, as shown in Fig.2. Cross-attention weighting similar to [35] is used in this experiment.

Effects of Spatial Weighting To focus on areas of interest, spatial weighting is used to provide marginal distribution μ^s and μ^t. Different spatial weighting strategies are evaluated in Tab.2. Applying cross-attention already enhances the retrieval quality by attending to foreground patches. Furthermore, relevance map is less noisy and focus more on the object patches. In particular, it is shown to be especially beneficial for SOP, on which uniform and cross-attention do not significantly improve. In Fig.1, we visualize the spatial weighting maps with example image pair.

Partial Optimal Transport In Tab.1, applying structural similarity on the SOP dataset does not improve retrieval results significantly. We hypothesize this is due to viewpoints and scale variations. We use the partial optimal transport (OT) [3] to handle maximum a amount of the total mass, where $0 \leq a \leq 1$. We empirically find that setting $a = 0.9$ gives the best results on the SOP as shown in Fig.3.

Table 2: Effects of spatial weighting schemes used by Optimal Transport. Baseline method without weighting is a CvT-13 with global similarity. Structural methods with uniform, cross-attention and relevance map are compared.

Weighting	CUB200-2011			Cars196			Online Products		
	P@1	RP	M@R	P@1	RP	M@R	P@1	RP	M@R
None	71.75	41.94	31.19	80.55	35.30	24.80	77.15	44.74	41.53
Uniform	73.22	42.51	31.87	83.38	35.59	25.64	77.05	44.39	41.28
Cross	73.72	42.68	32.12	83.66	35.63	25.82	77.15	44.45	41.36
Relevancy	**73.85**	**43.15**	**32.68**	**83.92**	**35.70**	**25.91**	**77.95**	**45.15**	**42.06**

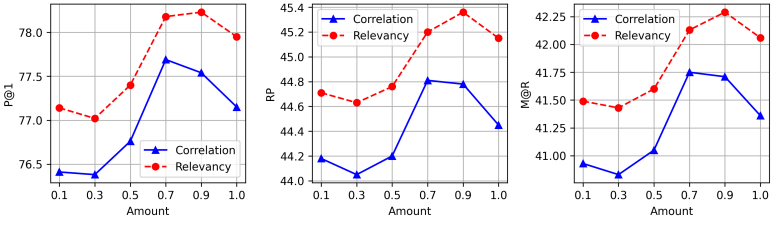


Figure 3: Effects of partial optimal transport amount a on the SOP. When $a = 1$, it is identical to standard optimal transport. We found that allowing partial OT improves the retrieval performance on the SOP.

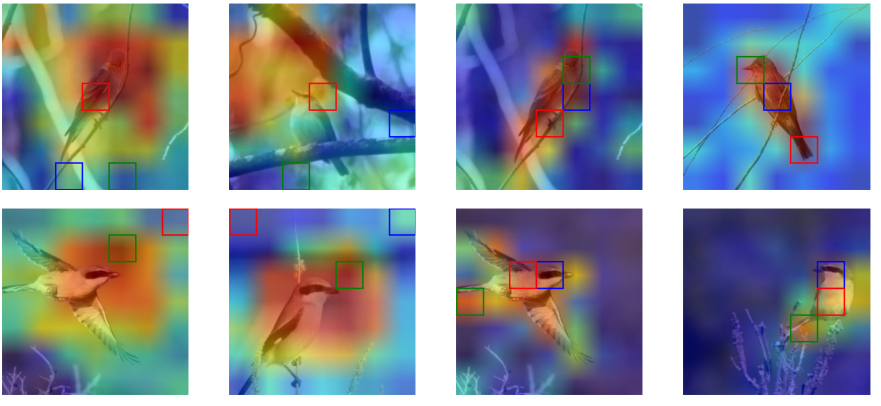
Visual Place Recognition We also evaluate the proposed method on Mapillary Streets dataset (MSLS) [31], which features variations of season, time of day, date, viewpoint and weather. Training from scratch on such a huge dataset is time-consuming and does not converge to satisfactory performance. Thus, we follow feature-based knowledge distillation approach and use ImageNet-pretrained CvT-13 as our student model. Next, we train it on MSLS similar to [12]. For the teacher model, we use MSLS-pretrained NetVLAD [2] model. Both teacher and student models use 128 dimension features. Note, inputs to teacher model are resized to 640×480, while inputs to student model are resized to 224×224. As shown in Tab.3, structural reranking massively improves Recall1 by 15.66% and 5.67%, when training scratch and with distillation. We also observe that training with distillation improves accuracy by a large margin.

Table 3: Place recognition results on MSLS. Our model is based on CvT-13. We demonstrate that structural similarity consistently improves accuracy when the model is trained from scratch or using distillation.

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

Qualitative Visualization We follow DIML and evaluate interpretability qualitatively. Specifically, we justify improved interpretability from two aspects: 1) object is more important than background; 2) similarity of semantic corresponding parts matters more. We visualize some

example images from CUB200. Cross-attention spatial weighting is used for both DIML and CvT. In Fig.4, the heatmaps indicate the spatial weighting for patches. DIML (col.1-2) produces coarse and blurry weighting, sometimes focusing on background. In contrast, our method shows only objects are considered important. The top matching patches denoted by the depicted boxes of DIML are error-prone, while our method prioritise semantic corresponding parts using the same query image. In Fig.5, we show failed semantic correspondence of CvT and emphasise that the use of OT does not resolve symmetry ambiguity.



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

Figure 4: Each row shows a comparison of DIML and our method. Col. (a-b) show a positive pair using DIML. The heatmaps indicating spatial importance for similarity are blurry and not accurately targeting objects. Our results shown in columns (c-d) suggest that importance maps are accurate and focused on birds. Please see Supp. Mat. for additional results.

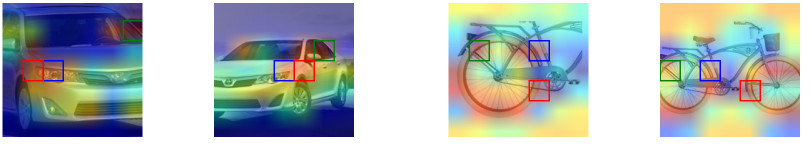


Figure 5: CvT examples of the Cars196 and SOP datasets are shown. Notice, our OT is not aware of missing features or similarities in the object.

5 Conclusions

We investigate the problem of whether pretrained vision transformers can be used for structural image reranking. We find that convolutions inside vision transformers are important to learn globally discriminative patch embeddings. We propose to use attention-based relevancy maps of vision transformers to guide optimal transport optimization. The approach demonstrates robust performance and improved interpretability on multiple benchmarks. Although vanilla ViT could not benefit from the structural similarity, it is possible to replace attention layer and MLPs with a convolutional variant. We justify improved interpretability qualitatively in this work. Some datasets such as CUB-200 provide part annotations and a quantitative analysis could further demonstrate the interpretability strength of Transformers.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[3] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. *arXiv preprint arXiv:2002.08276*, 2020.

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[5] Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting grad-cam for embedding networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2794–2803, 2020.

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] 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.

[10] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021.

[11] 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.

[12] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[13] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021.

[14] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[15] 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.

[16] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[17] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[19] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020.

[20] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[21] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[22] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Karsten Roth, Timo Milbich, and Bjorn Ommer. Pads: Policy-adapted sampling for visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6568–6577, 2020.

[24] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[27] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[30] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[31] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2626–2635, 2020.

[32] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[33] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[34] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[35] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9887–9896, 2021.

[36] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 72–81, 2019.

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[38] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.