A Light Touch Approach to Teaching Transformers Multi-view Geometry

Yash Bhalgat João F. Henriques Andrew Zisserman
Visual Geometry Group
University of Oxford
{yashsb,joao,az}@robots.ox.ac.uk

Abstract

Transformers are powerful visual learners, in large part due to their conspicuous lack of manually-specified priors. This flexibility can be problematic in tasks that involve multiple-view geometry, due to the near-infinite possible variations in 3D shapes and viewpoints (requiring flexibility), and the precise nature of projective geometry (obeying rigid laws). To resolve this conundrum, we propose a "light touch" approach, guiding visual Transformers to learn multiple-view geometry but allowing them to break free when needed. We achieve this by using epipolar lines to guide the Transformer's cross-attention maps during training, penalizing attention values outside the epipolar lines and encouraging higher attention along these lines since they contain geometrically plausible matches. Unlike previous methods, our proposal does not require any camera pose information at test-time. We focus on pose-invariant object instance retrieval, where standard Transformer networks struggle, due to the large differences in viewpoint between query and retrieved images. Experimentally, our method outperforms state-of-the-art approaches at object retrieval, without needing pose information at test-time.

1. Introduction

Recent advances in computer vision have been characterized by using increasingly generic models fitted with large amounts of data, with attention-based models (e.g. Transformers) at one extreme [12, 13, 20, 24, 34, 41]. There are many such recent examples, where shedding priors in favour of learning from more data has proven to be a successful strategy, from image classification [1, 13, 20, 29, 90], action recognition [7, 23, 27, 50, 58], to text-image matching [36, 45, 62, 71] and 3D recognition [40, 91]. One area where this strategy has proven more difficult to apply is solving tasks that involve reasoning about multiple-view geometry, such as object retrieval – i.e. finding all instances of an object in a database given a single query image. This has applications in image search [37, 39, 51, 82, 92], including identifying landmarks from images [53, 61, 85], recognizing artworks in images [80], retrieving relevant product images in e-commerce databases [14, 55] or retrieving specific objects from a scene [3, 38, 46, 60].

Figure 1. Top-4 retrieved images with (1) global retrieval (left column), (2) Reranking Transformer (RRT) [74] (middle), and (3) RRT trained with our proposed Epipolar Loss (right column). Correct retrievals are **green**, incorrect ones are **red**. The Epipolar Loss imbues RRT with an implicit geometric understanding, allowing it to match images from extremely diverse viewpoints.

The main challenges in object retrieval include overcoming variations in viewpoint and scale. The difficulty in viewpoint-invariant object retrieval can be partially explained by the fact that it requires disambiguating similar objects by small differences in their unique details, which can have a smaller impact on an image than a large varia-

tion in viewpoint. For this reason, several works have emphasized geometric priors in deep networks that deal with multiple-view geometry [22,88]. It is natural to ask whether these priors are too restrictive, and harm a network's ability to model the data when it deviates from the geometric assumptions. As a step in this direction, we explore how to "guide" attention-based networks with soft guardrails that encourage them to respect multi-view geometry, without constraining them with any rigid mechanism to do so.

In this work, we focus on post-retrieval reranking methods, wherein an initial ranking is obtained using global (image-level) representations and then local (region- or patch-level) representations are used to *rerank* the top-ranked images either with the classic Geometric Verification [57], or by directly predicting similarity scores of image pairs using a trained deep network [31, 74]. Reranking can be easily combined with any other retrieval method while significantly boosting the precision of the underlying retrieval algorithm. Recently, PatchNetVLAD [31], DELG [11], and Reranking Transformers [74] have shown that learned reranking can achieve state-of-the-art performance on object retrieval. We show that the performance of such reranking methods can be further improved by *implicitly* inducing geometric knowledge, specifically the epipolar relations between two images arising from relative pose, into the underlying image similarity computation.

This raises the question of whether multiple view relations should be incorporated into the two view architecture *explicitly* rather than *implicitly*. In the explicit case, the epipolar relations between the two images are supplied as inputs. For example, this is the approach taken in the Epipolar Transformers architecture [33] where candidate correspondences are explicitly sampled along the epipolar line, and in [88] where pixels are tagged with their epipolar planes using a Perceiver IO architecture [34]. The disadvantage of the explicit approach is that epipolar geometry must be supplied at inference time, requiring a separate process for its computation, and being problematic when images are not of the same object (as the epipolar geometry is then not defined). In contrast, in the implicit approach the epipolar geometry is only required at training time and is applied as a loss to encourage the model to learn to (implicitly) take advantage of epipolar constraints when determining a match.

We bring the following three contributions in this work: First, we propose a simple but effective *Epipolar Loss* to induce epipolar constraints into the cross-attention layer(s) of transformer-based reranking models. We only need the relative pose (or epipolar geometry) information during training to provide the epipolar constraint. Once trained, the reranking model develops an implicit understanding of the relative geometry between any given image pair and can effectively match images containing an object instance from very diverse viewpoints *without* any additional input. Second, we set up an object retrieval benchmark on top of the CO3Dv2 [63] dataset which contains ground-truth camera poses and provide a comprehensive evaluation of the proposed method, including a comparison between implicit and explicit incorporation of epipolar constraints. The benchmark configuration is detailed in Sec. 4. Third, we evaluate on the Stanford Online Products [55] dataset using both zero-shot and fine-tuning, outperforming previous methods on this standard object instance retrieval benchmark.

2. Related Work

Computing epipolar geometry. Estimating epipolar geometry given an image pair is a fairly broad problem, well-studied in multi-view geometry and computer vision [30]. Classic techniques involve predicting interest points and their descriptors [4, 44, 48, 66, 68] in the images and finding point correspondences to estimate the relative geometry [43, 52]. Several learning based methods have been proposed to provide improved interest point detection and features, e.g. R2D2 [64] SuperPoint [19], LIFT [87] and MagicPoint [18]. These features along with learning based local matching methods [69, 72, 86] and robust optimization methods [6, 9, 25] form a powerful toolbox for relative geometry estimation. We use a combination of LoFTR [72] and MAGSAC++ [6] to generate pseudo-geometry information in one of our compared methods.

Incorporating epipolar geometry in Deep Learning. Recently, many works have proposed incorporating geometric priors into deep networks to deal with problems requiring multi-view understanding, such as 3D pose estimation [33, 65, 89], 3D reconstruction [79, 88] or depth estimation [59]. Most of these approaches incorporate the epipolar geometry explicitly, e.g. Epipolar Transformers [33] compute 3D-aware features for a point by aggregating features sampled on the corresponding epipolar line, which are shown to improve multi-view 3D human-pose estimation. [88], another explicit method, proposed a few ways of featurizing multi-view geometry by encoding camera parameters or epipolar plane parameters and using them to provide geometric priors at the input-level. The epipolar plane encoding is also studied in this paper in the context of reranking transformers. Works such as [65, 78] propose implicitly incorporating geometric priors using multi-view consistency. Our work also falls in the implicit category, where we use epipolar constraints as a loss function applied to cross-attention maps to induce geometric understanding.

Image representations for retrieval. Traditionally, handcrafted descriptors such as SIFT [44], RootSIFT [4] and BoVW [57] were widely used for object retrieval. However, learned image-level (global) and region-level (local) representations [2, 5, 19, 28, 87] have shown to surpass the performance of engineered features on large-scale datasets. Local learned representations can also be simply extracted as

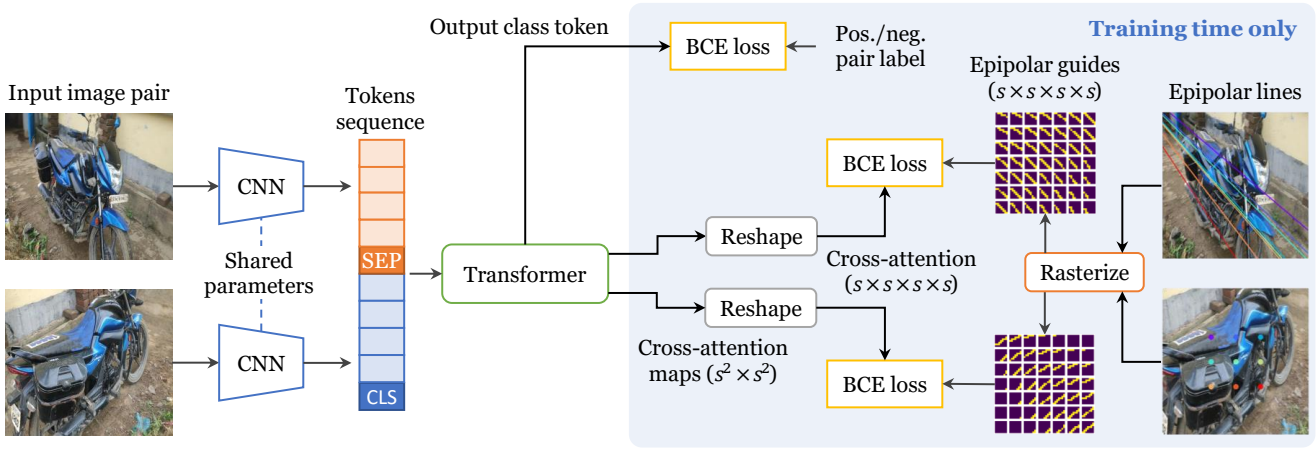

Figure 2. Overview of the proposed method. Features from two candidate images are extracted with a Convolutional Neural Network, and concatenated into a sequence of tokens for a Transformer. They are separated by a learned ⟨SEP⟩ token and end with a ⟨CLS⟩ token. The model is trained with a Binary Cross Entropy (BCE) loss to predict whether the two images match. During training, epipolar lines relating the two views (obtained with ground truth camera information) are rasterized into 4D tensors. These "epipolar guides" denote matches that are geometrically plausible given the viewpoints, and are used to train the Transformer's cross-attention maps using BCE losses.

feature volumes from convolution neural networks or transformer backbones. Global representations are obtained by a combination of (1) downsampling/pooling operations inside a deep network, (2) learned clustering-based pooling operations [2] and/or specialized pooling operations such as R-MAC [76]. Hybrid approaches that combine global and local features have also recently been proposed [11, 31].

Post-retrieval reranking. Early reranking methods, such as [35, 57], used Geometric Verification (GV) with local features to compute geometric consistency between the query and reference images. This improved the precision of the top-ranked retrievals. Query Expansion (QE) was used to improve the recall. Popular QE variants such as average-QE and α-QE compute an updated query descriptor from the global descriptors of the top retrieved images [15, 16, 75] and use it to retrieve a new set of top-ranked images. GV can be combined with many deep learning based retrieval methods used today, e.g. [11] uses RANSAC based GV on local features from its backbone model. Since RANSAC-based GV can be prohibitively slow for practical applications, [31] proposes a rapid spatial scoring technique as an efficient alternative. Recently, transformer based methods [21, 74] have been introduced for retrieval and reranking. Our work builds on top of Reranking Transformers [74].

Retrieval with 3D information. Recently, methods using 3D data [42, 81], structural cues [54] or view synthesis [73] have been proposed. Our goal in this work is to build image representations that capture 3D priors and can be used to retrieve images with large variations in pose or scale.

3. Method

We describe two variants of our method that *implicitly* or *explicitly* encourage Transformers to use geometric constraints in their predictions. Our work is built on top of Reranking Transformers (RRT) [74], a state-of-the-art approach for object retrieval with reranking. The explicit version, inspired by recent work [88], serves both as a baseline and as a contrast to our proposed implicit approach. We first provide a brief review of RRTs for the reader (Sec. 3.1) and then describe our proposed Epipolar Loss (Sec. 3.3), as well as Epipolar Positional Encodings (Sec. 3.4). The implementation details are given in Sec. 5.2.

3.1. Review of Reranking Transformers

Post-retrieval reranking is a popular technique used to boost the precision of object retrieval methods, wherein an initial ranking is obtained using global (image-level) descriptors and then local (region-level) descriptors along with the global ones are used to *rerank* the top-ranked images. In [74], each image ($\mathcal{I}$) is processed through a ResNet-50 [32] model to extract local features from the last convolution layer, with size $s \times s \times c$ ($s = 7$, $c = 2048$). Each of the s^2 local feature vectors is linearly projected from size c to a smaller size $m = 128$. Let these be denoted by $\mathbf{x}^l \in \mathcal{R}^{s^2 \times m}$ and their 2D positions in the feature volume by $\mathbf{p}_i \in \mathcal{R}^2$. The global features, computed as the mean of the local features, are used for initial ranking. Then, a light-weight transformer model (4 self-attention heads, 6 layers) is used to rerank these top predictions. With $\mathcal{I}$ as the query and $\bar{\mathcal{I}}$ as a reference image from the top predictions, as well as class ⟨CLS⟩ and separator ⟨SEP⟩ tokens (consisting of learnable embeddings), the input to the transformer model is constructed as the concatenation of tokens:

$$X(\mathcal{I}, \bar{\mathcal{I}}) = [\langle \text{CLS} \rangle, f(\mathbf{x}_1^l), \dots f(\mathbf{x}_{s^2}^l), \\ \langle \text{SEP} \rangle, \bar{f}(\bar{\mathbf{x}}_1^l), \dots \bar{f}(\bar{\mathbf{x}}_{s^2}^l)]$$

where $f(\mathbf{x}_i^l) = \mathbf{x}_i^l + \psi(\mathbf{p}_i) + \beta$, $\bar{f}(\bar{\mathbf{x}}_i^l) = \bar{\mathbf{x}}_i^l + \psi(\bar{\mathbf{p}}_i) + \bar{\beta}$, $\psi(.)$ is the frequency position encoding [83] and $\beta, \bar{\beta}$ are learnable embeddings that differentiate descriptors of $\mathcal{I}, \bar{\mathcal{I}}$. Sec. 5.2 provides training details for the reranking model.

3.2. Review of Epipolar Geometry

Epipolar geometry limits the possible image correspondences for projections of an observed 3D point from different viewpoints. A central concept is the epipolar line. Consider a 2D point $\mathbf{x}$ in one image. It may correspond to an infinity of 3D points – one for each possible depth – which lie on a 3D line that extends from the camera center and passes through $\mathbf{x}$ in the image plane. This 3D line, when projected into a *second* image captured from another viewpoint, is an epipolar line of $\mathbf{x}$. This mapping from a point in one image to its epipolar line in another image can be seen in Fig. 2 (right). Epipolar geometry can be used to effectively constrain matches across viewpoints: starting from a point in one image, it can only match points in another image that lie along its epipolar line. Epipolar geometry can be computed directly from two images, either from their relative pose or from correspondences, without requiring any information about depth or 3D geometry of the observed scene. Mathematically it is represented by a 3×3 fundamental matrix. For a more detailed exposition, please refer to [30].

3.3. Epipolar Loss

Given the feature volumes $\mathbf{x}^l, \bar{\mathbf{x}}^l \in \mathcal{R}^{s^2 \times m}$ as input tokens (in addition to $\langle$CLS$\rangle$, $\langle$SEP$\rangle$), let $\mathbf{y}_{L-1}, \bar{\mathbf{y}}_{L-1}$ denote the corresponding inputs to the last transformer layer in the RRT model. The *raw* cross-attention between these outputs can be computed as $A^{12} = Q\overline{K}^T$ and $A^{21} = \overline{Q}K^T$, where W_Q, W_K are query and key projection matrices and $Q = W_Q\mathbf{y}_{L-1}$, $K = W_K\mathbf{y}_{L-1}$,$\overline{Q} = W_Q\bar{\mathbf{y}}_{L-1}$,$\overline{K} = W_K\bar{\mathbf{y}}_{L-1}$.

Next, given the epipolar geometry between the input images, for every location $i \in \{1, \ldots, s^2\}$ in $\mathbf{x}^l$, we can find the set of locations $\bar{e}_i$ in $\bar{\mathbf{x}}^l$ that lie on the corresponding epipolar line. Similarly, for each location $i \in \{1, \ldots, s^2\}$ in $\bar{\mathbf{x}}^l$, we can find the corresponding set of locations e_i in $\mathbf{x}^l$. We want to encourage the network, for a given position in the first volume, to only *attend* to corresponding epipolar positions in the other volume. This is done by penalizing attention values that have high values outside the epipolar lines, and encouraging the attention along epipolar lines to be high. This is achieved by using a Binary Cross Entropy (BCE) loss on the *raw* cross-attention maps $\{A^{12}, A^{21}\}$:

$$L^{12}(i,j) = \text{BCE}(\sigma(A^{12}(i,j)), \mathbb{1}(i,j))$$
$$L^{21}(i,j) = \text{BCE}(\sigma(A^{21}(i,j)), \mathbb{1}(i,j))$$
$$L_{EPI} = \sum_{i=1}^{s^2}\sum_{j=1}^{s^2} L^{12}(i,j) + L^{21}(i,j) \quad (1)$$

where σ is a sigmoid function, and $\mathbb{1}(i,j)$ is a special indicator function that is 1 when location j in the other feature map lies on the epipolar line corresponding to location i in the current map. The training process is illustrated in Fig. 2.

Max-Epipolar Loss. In the Epipolar Loss proposed above, every point on the corresponding epipolar line is encouraged to have high attention even if it is not the *actual* matching point in 3D. We also propose a variant called *Max-Epipolar Loss*, wherein we select only the point on the epipolar line with the maximum predicted cross-attention value and encourage the attention for that point to be high.

$$L_{MaxEPI} = L_{zero} + L_{max} \quad (2)$$

where

$$L_{max} = \sum_i \text{BCE}\left(\max_{j \in e_i} \sigma(A(i,j)), 1\right)$$
$$L_{zero} = \sum_{\forall i,j, \mathbb{1}(i,j)=0} \text{BCE}(\sigma(A(i,j)), 0)$$

where e_i is the set of locations in the other feature map that lie on the epipolar line corresponding to location i in the current map. L_{zero}, L_{max} are applied to both A^{12} and A^{21}.

Note that the epipolar loss is applied at training time, so epipolar geometry is required only during training. The epipolar geometry can be obtained from the relative pose between the images, or from the images directly. However, as will be shown in Sec. 5.6, once trained with our proposed L_{EPI}, the attention map extracted from the trained RRT model for a previously *unseen* pair of images shows patterns corresponding to the actual epipolar lines (*without* any input epipolar geometry information). This demonstrates that the model's predictions are epipolar-geometry-aware, and at test time this leads to improved reranking performance as erroneous point matches can be avoided.

3.4. Epipolar Positional Encoding

In contrast to Epipolar Loss where we implicitly induce awareness of epipolar line correspondence into the model, epipolar constraints can be encoded *explicitly* by annotating each pixel with an encoding that uniquely identifies its epipolar plane. The family of epipolar planes "rotates" about the line joining the two camera centers, hence it can be parameterized by a scalar angle of rotation. Inspired by [88], we introduce a baseline wherein we encode the epipolar plane angle for each token and add the encoding to the input tokens of the transformer. A random epipolar plane corresponding to a randomly chosen pixel location is used as reference to calculate the plane angle. We encode the angle with the frequency positional encoding [49, 83].

The drawback of the explicit method is that it requires the epipolar geometry (or relative pose) information during

inference. In the scenario when this information is not available, we have to rely on other ways to obtain the epipolar geometry or relative pose which may not be entirely accurate and leads to loss in performance, as will be shown in Sec. 5.3. In fact, determining whether epipolar geometry can be established between two views is essentially replacing the job of the reranking transformer in determining if two images contain the same object.

4. CO3D-Retrieve benchmark

We now describe how we repurpose the CO3Dv2 [63] dataset to create a large-scale object instance retrieval benchmark with multiple views of real objects. CO3Dv2 is a dataset of multi-view images of common object categories, consisting of 36,506 videos of object instances (one video per object instance), taken from distant viewpoints spanning all 360 degrees, and covering 51 common object categories. The dataset also contains the ground-truth camera poses for the video frames and foreground segmentation masks for the object in each image.

For the *CO3D-Retrieve* dataset, we extract 5 frames per video so that each frame is separated from the next by *approximately* 72° of rotation around the object. In total, CO3D-Retrieve contains 181,857 images of 36,506 object instances. We split the dataset into two halves for training and testing: the training dataset contains 91,106 images of 18,241 object instances, and the testing datasets contains 90,751 images from 18,265 object instances. The set of object instances seen during training and testing are disjoint and so have zero overlap with each other. For benchmarking object retrieval on CO3D-Retrieve, we evaluate with each image as the query, the other images from the same object as the query are treated as *positives*, and all the images not corresponding to the query object are treated as negatives. Fig. 3 shows example object images from the benchmark.

5. Experiments

In this section, we evaluate the epipolar-geometry aware Reranking Transformer on two datasets: our CO3D-Retrieve benchmark, and the Stanford Online Products (SOP) benchmark [55]. SOP is a popular benchmark for object retrieval containing 120,053 images of 22,634 object instances from 12 object categories. We use the standard train-test split used by all the baselines we compare with, where 59,551 images are used for training and 60,502 for testing. In Sec. 5.6, we provide a discussion on the merits of the implicit approach to incorporating epipolar constraints and explore its properties.

5.1. Baselines and metrics

Pretrained descriptors. Deep networks pretrained on large scale image datasets learn powerful image representations that can be used for retrieval. Evaluating such pretrained models without fine-tuning gives us a lower bound on the performance that a model trained on our dataset should achieve. We compare with VGG16 [70] and ResNet50 (R50) [32] models pretrained on ImageNet [17], i.e. trained for classification, not retrieval. We also compare to a NetVLAD [2] model (i.e. VGG16 backbone + NetVLAD pooling) pretrained for retrieval on Pittsburgh250k [77].

Reranking Transformers (RRT). Reranking Tranformers (RRT) [74] is a state-of-the-art method that our works builds on. We compare with different versions of the RRT method:

1. *R50 (trained)*: this baseline performs global retrieval (no reranking) and does not use RRT, but works as a foundation for subsequent baselines. The model is trained using a batch-wise contrastive loss on CO3D-Retrieve or SOP [55], for the respective experiments.
2. *R50 (frozen) + RRT*: we start from a trained R50 (i.e. baseline (1)), freeze its weights and train a RRT on top of it for reranking.
3. *R50 (finetune) + RRT*: we start from a trained R50 (i.e. baseline (1)) and we finetune the R50 backbone along with the RRT.

RRT w/ Epipolar Positional Encoding. The R50 backbone along with RRT is trained with their respective "retrieval loss" functions, and the epipolar geometry is provided as input in the form of an Epipolar Positional Encoding (Sec. 3.4). We will discuss the results of this baseline in a separate Sec. 5.5.

Evaluation metrics. Given a query image and a retrieved image, they match if they contain the same object instance. We report two metrics to evaluate retrieval performance. First, **R@K** – for a given query, if a match is within the top K retrieved images, then the query is said to be retrieved. $R@K$ is the fraction of correctly retrieved queries for a given K. We report results for $K = 1$, 10 and 50. Second, we report **Mean Average Precision (mAP)** – the mean of the Average Precision [47] over all queries.

5.2. Implementation details

Extracting the epipolar geometry. For experiments with the CO3D-Retrieve benchmark, the available ground-truth pose information for each image is used to compute the epipolar geometry for a matching pair of images. During training, we use the "RandomCrop" image augmentation which shifts the principal point location. Hence, we adjust the Fundamental Matrix computation appropriately to obtain the correct epipolar lines in the cropped images. When pose information is not available as ground-truth, such as in the SOP [55] dataset, we use an off-the-shelf method to compute the epipolar geometry. Specifically, we use a lo-

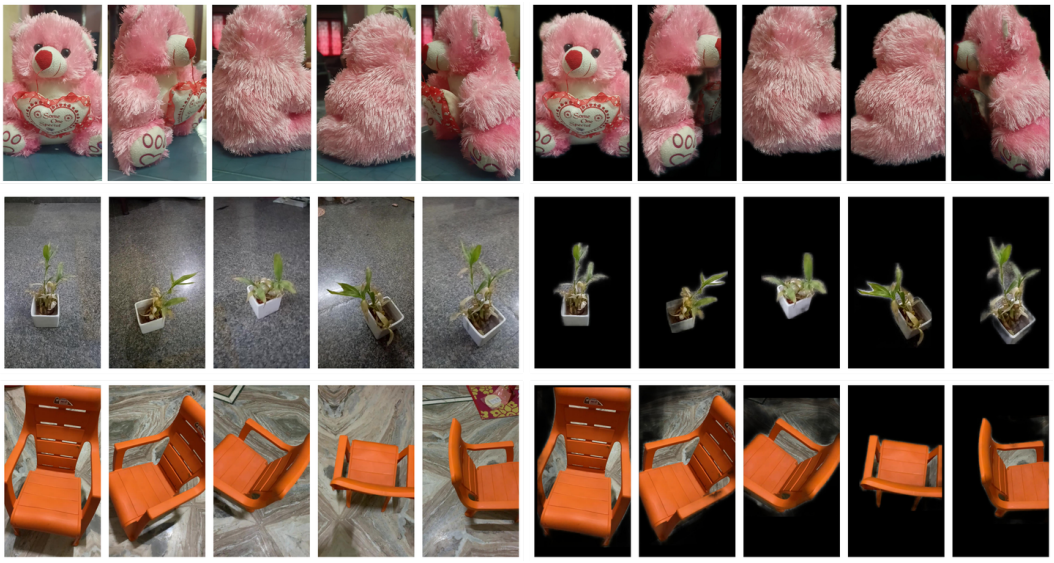

Figure 3. Example images for three object instances from the CO3D-Retrieve benchmark. The left half shows the full image, and the right half shows the masked counterparts obtained using the object masks in CO3D [63]. The number of pixels in common between views of the same object decreases from the top row to the bottom (computed using the 3D point-clouds, also available from CO3D).

cal image feature matching method, LoFTR [72], to extract high-quality semi-dense matches between the image pair. Then, we use a robust estimation method, MAGSAC++ [6] to extract the Fundamental Matrix. The epipolar geometry extracted with this method is not entirely accurate (especially for image pairs with extreme relative pose), but it provides us with a sufficient pseudo ground-truth epipolar geometry to train our models with Epipolar Loss. If the number of matches found ≤ 20 or number of inliers detected $\leq 0.2\times$ number of matches, we consider the extracted epipolar geometry unreliable and do not apply Epipolar Loss for that image pair during training. Computing epipolar geometry with this method takes $\approx$0.06 seconds per image pair on a 8-core CPU and NVIDIA P40 GPU.

Training details. We use a ResNet50 [32] for global retrieval, which is trained with a batchwise contrastive loss (batch size of 800). For an image pair$\{\mathcal{I}, \bar{\mathcal{I}}\}$ in the batch, a Binary Cross-Entropy loss is used to train the reranking model enforcing its output to be 1 if $\mathcal{I}$ and $\bar{\mathcal{I}}$ contain the same object and 0 otherwise.

In our proposed implicit method, the global retrieval model and the reranking model are trained with their respective retrieval-losses plus the Epipolar Loss. If $\mathcal{I}$ and $\bar{\mathcal{I}}$ represent the same object, then the epipolar geometry between the image pair (which is extracted as explained above) is used to compute the Epipolar Loss for training. If the image pair is not a match, then a valid epipolar geometry does not exist and we simply do not apply the Epipolar Loss for that image pair.

In the explicit method, we have to include the geometry in the input as Epipolar Positional Encodings (EPE), even when the input pair $\{\mathcal{I}, \bar{\mathcal{I}}\}$ is not a match. To handle the case when $\{\mathcal{I}, \bar{\mathcal{I}}\}$ is not a match during training and testing, we use a *random* rank-2 matrix as the Fundamental Matrix to compute the EPEs. When $\{\mathcal{I}, \bar{\mathcal{I}}\}$ is indeed a match, (a) during training, we use the ground-truth or the pseudo ground-truth (whichever is available) to compute the EPEs, (b) during testing, we do not rely on the ground-truth geometry information and always use the LoFTR/MAGSAC++ method (described above) to compute the EPEs.

The hyperparameters we use for our experiments with SOP [55] are the same as [74], except that we use 40 epochs (instead of 100) when training with the Epipolar Loss with a constant learning rate of 10^{-4}. Hyperparameters used with CO3D-Retrieve are provided in supplementary material. Our method is entirely implemented in PyTorch [56].

5.3. Results on CO3D-Retrieve

We evaluate on CO3D-Retrieve in two settings: with and without masking the background in the object images. This is because the background also provides useful visual cues for image matching and it is essential to see how the methods perform without any such extra information. Figure 3 shows examples with and without masking the background.

Table 1 shows the detailed results. We observe that pretrained models (VGG16 and R50 on ImageNet, NetVLAD on Pittsburgh250k) achieve a reasonable performance without any finetuning. However, their performance is not competitive compared with baselines specialized for CO3D-Retrieve. It's interesting to note that a simple ResNet50-based global retrieval baseline trained with batch-wise contrastive loss (i.e. the "R50 (trained)" baseline) already achieves a high $R@1$ of 86.06% on the unmasked images. Our method, which uses the Epipolar Loss to induce multi-

Table 1. Evaluation on CO3D-Retrieve benchmark. Description of all compared methods in Sec. 5.1. Baselines shown above dashed line are pretrained models and below are trained on CO3D-Retrieve. **EPE**=Epipolar Positional Encoding

Method	Full images				With masked backgrounds			
	R@1	R@10	R@50	mAP	R@1	R@10	R@50	mAP
Pretrained Models								
VGG16 [70]	66.21	85.18	91.66	22.51	63.56	81.11	89.24	16.73
R50 [32]	66.48	85.34	91.74	22.79	63.81	80.30	89.37	16.79
NetVLAD [2]	67.01	85.17	91.72	22.63	63.19	80.84	89.27	16.61
R50 (trained)	86.06	95.62	97.65	45.34	78.82	91.30	94.72	24.85
RRT [74] + R50 (frozen)	88.07	96.29	97.75	47.60	82.45	91.89	94.85	26.16
RRT + R50 (finetune) (**SOTA**)	89.20	96.85	97.89	48.81	83.28	92.13	95.05	27.33
RRT + R50 w/ EPE	88.53	96.41	97.83	47.99	82.79	91.96	94.99	26.58
RRT + R50 w/ L_{EPI} (**Ours**)	90.57	97.33	98.10	49.52	85.07	92.42	95.11	28.07
RRT + R50 w/ L_{MaxEPI} (**Ours**)	**90.69**	**97.38**	**98.10**	**49.60**	**85.17**	**92.46**	**95.14**	**28.21**

Table 2. Evaluation on Stanford Online Products [55]. Baselines shown above the dashed line are pretrained models and below the dashed line are trained on SOP. More details in Sec. 5.1. Key: *=results obtained using checkpoints from [74]; **=results reported in [74]

Method	R@1	R@10	R@50	mAP
Pretrained Models				
VGG16 [70]	55.75	70.86	79.65	11.93
R50 [32]	55.89	71.32	79.69	12.09
NetVLAD [2]	54.16	70.85	79.62	11.90
R50 (trained)*	80.74	91.87	95.54	32.90
RRT [74] + R50 (frozen)*	81.80	92.35	95.78	34.91
RRT + R50 (finetune)* (**SOTA**)	84.46	93.21	**96.04**	37.14
RRT + R50 w/ EPE	82.57	92.69	95.89	35.38
RRT + R50 w/ L_{EPI} (**Ours**)	**84.74**	**93.29**	**96.04**	**37.25**
RRT + R50 w/ L_{MaxEPI} (**Ours**)	84.53	93.27	**96.04**	37.19
*Other Metric Learning methods***				
Margin-based [67]	76.1	88.4	-	-
FastAP [10]	73.8	88.0	-	-
XBM [84]	80.6	91.6	-	-
Cross-Entropy based [8]	81.1	91.7	-	-

view geometric understanding in to the Reranking Transformer model, outperforms the state-of-the-art approach (RRT + R50 (finetune)) in both masked and unmasked settings. The margin with which the Epipolar Loss baseline outperforms "RRT + R50 (finetune)" is greater in the case of images with masked background, as this is a harder task. It can be seen that the Max-Epipolar variant of the Epipolar loss consistently gives a slight improvement.

5.4. Results on Stanford Online Products

The Stanford Online Products (SOP) dataset [55] does not contain ground-truth pose information for the object images. As detailed in Sec. 5.2, we obtain the pseudo ground-truth geometry information using LoFTR [72] for matching and MAGSAC++ [6] for robust estimation. We find that, even though these pseudo ground-truth poses are not entirely accurate, they are still useful for training with the Epipolar Loss. Table 2 shows a comprehensive comparison of our proposed methods with all the baselines. We also include deep metric learning methods [8, 10, 67, 84] with reported numbers taken from [74] into our comparisons. Our proposed implicit method outperforms all the baselines including the state-of-the-art Reranking Transformers [74].

Table 3. Zero-shot evaluation on Stanford Online Products [55] with models trained on CO3D-Retrieve.

Method	R@1	R@10	R@50	mAP
RRT + R50 (frozen)	75.53	89.43	95.01	29.27
RRT + R50 (finetune)	76.32	90.16	95.21	30.19
RRT + R50 w/ L_{EPI} (**Ours**)	**76.78**	**90.27**	**95.29**	**30.25**

We also test zero-shot retrieval, by evaluating on SOP models that were trained on CO3D-Retrieve. The results are shown in Table 3, where we can observe that the differences between all methods are reduced, but our Epipolar Loss still confers a performance advantage.

5.5. Implicit vs Explicit methods

The transformer model trained with Epipolar Loss does not require pose or epipolar geometry information at test time. The explicit method, however, requires the fundamental matrix at the input (during both training and testing) to generate the Epipolar Positional Encodings (EPE). Tables 1 and 2 show that using EPE with Reranking Transformer adversely affects the performance, compared to not using the encodings. Although reasons for this decrease are unclear, one possibility is that the encodings leak information about whether two images match or not, because when they do not match, the input epipolar encodings are arbitrary. The network may learn to rely on this signal instead of image

matching, in a case of "shortcut learning" [26]. This issue does not affect the implicit method as geometry information isn ot required at test time. When training the RRT, the Epipolar Loss is used with *only* those image pairs that contain matching images. Hence, during inference, the Transformer uses implicit geometry information only when it is *valid* (i.e. inherently for matching image pairs).

5.6. What does the implicit model learn?

After training with the Epipolar Loss (L_{EPI}), we investigate if the attention maps of the learned model show some signs of geometric-awareness. To do this, we pick two matching images $\{\mathcal{I}, \bar{\mathcal{I}}\}$ from the *test* set, i.e. these images were not seen during training, and extract the cross-attention maps from the last layer of the transformer. Since we use a $7 \times 7 \times 128$ feature volume for each image (reshaped to 49×128 for the transformer), the cross-attention maps (from $\mathcal{I}$ to $\bar{\mathcal{I}}$, and $\bar{\mathcal{I}}$ to $\mathcal{I}$) are of size 49×49 which are then reshaped back to $7 \times 7 \times 7 \times 7$. These $7 \times 7 \times 7 \times 7$ cross-attention map values indicate the attention between each *feature-pixel* of the first and second feature volume.

Fig. 4 shows predicted cross-attention maps alongside expected ground-truth maps. The latter are computed using ground-truth pose information. We observe that the attention maps obtained with L_{EPI} closely follow ground truth epipolar lines, despite this instance and associated geometry not being seen during training. Note that the attention maps obtained with L_{MaxEPI} are much sparser with peaks that lie on actual epipolar lines. In the supplementary, we show maps obtained with a pair of mismatched images.

6. Conclusion

In this work, we aimed to teach multi-view geometry to Transformer networks, and proposed a method to do so implicitly via epipolar guides. The advantages of this implicit approach over explicitly passing in geometric information to a network are two-fold: (i) ground-truth epipolar geometry (relative pose) between views is only needed at training time, not at inference; (ii) implicit losses are readily applied to existing architectures, so there is no need to design specialized architectures. We demonstrated improved performance over the state-of-the-art in object retrieval, by reranking with our method. More generally, this approach of implicitly incorporating knowledge into Transformers by a suitable loss can be employed in other scenarios. Examples include learning other geometric relations, such as a trifocal relationship over three views, as well as physical laws such as Newton's laws of motion.

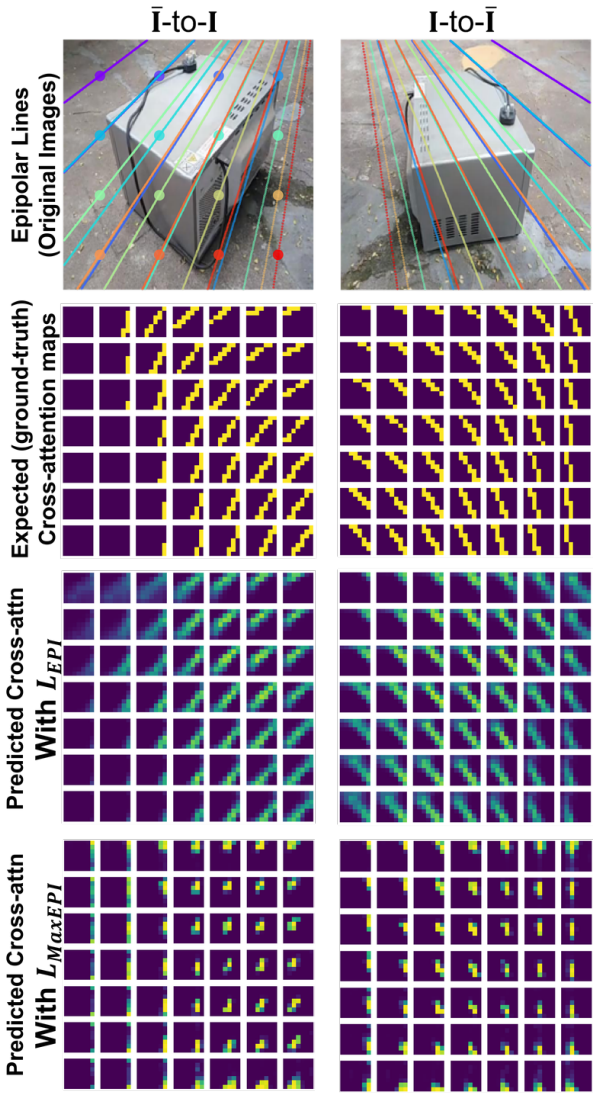

Figure 4. Visualization for a test image pair (i.e. never seen in training). **Top row**: Points shown in image $\mathcal{I}$ have correspondences on the epipolar lines of that color in image $\bar{\mathcal{I}}$. **Second row**: Expected 7×7×7×7 cross-attention maps shown as a 7×7 grid with a 7×7 patch at each grid location computed from the ground truth epipolar geometry. In the $\mathcal{I}$-to-$\bar{\mathcal{I}}$ grid (right column), a patch at grid location (i, j) shows the epipolar line in $\bar{\mathcal{I}}$ corresponding to the pixel (i, j) in the 7×7 feature space of $\mathcal{I}$. **Third row**: Predicted cross-attention maps from transformer trained with L_{EPI}. Notice how closely they match ground-truth maps, even though these are test images and do not have access to the ground truth epipolar geometry. **Bottom row**: Predicted cross-attention maps from transformer trained with Max-Epipolar Loss, L_{MaxEPI}. These are sparser and have peaks that lie close to actual epipolar lines.

Acknowledgements. We are grateful for funding from EPSRC AIMS CDT EP/S024050/1, AWS, the Royal Academy of Engineering (RF\201819\18\163), EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship RP\R1\191132. We thank the authors of [6, 72, 74] for open-sourcing their code. We also thank an anonymous reviewer for useful suggestions on the Max-Epipolar Loss.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 1

[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. 2, 3, 5, 7

[3] Relja Arandjelović and Andrew Zisserman. Smooth object retrieval using a bag of boundaries. In *Proceedings of the International Conference on Computer Vision*, 2011. 1

[4] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[5] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014. 2

[6] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 2, 6, 7, 8

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1

[8] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. Metric learning: cross-entropy vs. pairwise losses. *arXiv preprint arXiv:2003.08983*, 2020. 7

[9] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 2

[10] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1861–1870, 2019. 7

[11] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020. 2, 3

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[14] Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan, and Yinghui Xu. Weakly supervised learning with side information for noisy labeled images. In *The European Conference on Computer Vision (ECCV)*, August 2020. 1

[15] Ondřej Chum, Andrej Mikulik, Michal Perdoch, and Jiří Matas. Total recall ii: Query expansion revisited. In *CVPR 2011*, pages 889–896. IEEE, 2011. 3

[16] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007. 3

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017. 2

[19] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[20] 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. 1

[21] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021. 3

[22] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11826–11835, 2019. 2

[23] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 1

[24] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021. 1

[25] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 8

[27] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 244–253, 2019. 1

[28] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 2

[29] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 1

[30] Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2, 4

[31] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 2, 3

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*. IEEE Computer Society, 2016. 3, 5, 6, 7

[33] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 2

[34] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1, 2

[35] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008. 3

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[37] Yushi Jing and Shumeet Baluja. Pagerank for product image search. In *Proceedings of the 17th international conference on World Wide Web*, pages 307–316, 2008. 1

[38] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[39] Josip Krapac, Moray Allan, Jakob Verbeek, and Frédéric Jurie. Improving web image search results using query-relative classifiers. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1094–1101, 2010. 1

[40] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 1

[41] 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. 1

[42] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2831–2840, 2019. 3

[43] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981. 2

[44] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 2

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[46] Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, and Yaochen Li. Spatial-content image search in complex scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2503–2511, 2020. 1

[47] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. 5

[48] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005. 2

[49] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12346, pages 405–421, 2020. 4

[50] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021. 1

[51] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, and Wei-Ying Ma. Web object retrieval. In *Proceedings of the 16th international conference on World Wide Web*, pages 81–90, 2007. 1

[52] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 2

[53] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 1

[54] Amadeus Oertel, Titus Cieslewski, and Davide Scaramuzza. Augmenting visual place recognition with structural cues. *IEEE Robotics and Automation Letters*, 5(4):5534–5541, 2020. 3

[55] 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. 1, 2, 5, 6, 7

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[57] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis*, 2007. 2, 3

[58] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition*, pages 694–701. Springer, 2021. 1

[59] Vignesh Prasad, Dipanjan Das, and Brojeshwar Bhowmick. Epipolar geometry based learning of multi-view depth and ego-motion from monocular sequences. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2018. 2

[60] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 1

[61] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 1

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[63] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 2, 5, 6

[64] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 2

[65] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 750–767, 2018. 2

[66] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European conference on computer vision*, pages 430–443. Springer, 2006. 2

[67] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020. 7

[68] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[69] 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. 2

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5, 7

[71] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 1

[72] 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. 2, 6, 7, 8

[73] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 3

[74] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12105–12115, 2021. 1, 2, 3, 5, 6, 7, 8

[75] Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: refining local descriptors by feature aggregation. *Pattern recognition*, 47(10):3466–3476, 2014. 3

[76] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015. 3

[77] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013. 5

[78] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2897–2905, 2018. 2

[79] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 2

[80] Nikolai Ufer, Max Simon, Sabine Lang, and Björn Ommer. Large-scale interactive retrieval in art collections using multi-style feature aggregation. *PloS one*, 16(11):e0259718, 2021. 1

[81] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479, 2018. 3

[82] Reinier H Van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350, 2009. 1

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[84] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020. 7

[85] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020. 1

[86] Olivia Wiles, Sebastien Ehrhardt, and Andrew Zisserman. Co-attention for conditioned image matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[87] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016. 2

[88] Wang Yifan, Carl Doersch, Relja Arandjelović, João Carreira, and Andrew Zisserman. Input-level inductive biases for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6176–6186, June 2022. 2, 3, 4

[89] Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. Pcls: Geometry-aware neural reconstruction of 3d pose with perspective crop layers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9064–9073, 2021. 2

[90] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 1

[91] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 1

[92] Wengang Zhou, Yijuan Lu, Houqiang Li, Yibing Song, and Qi Tian. Spatial coding for large scale partial-duplicate web image search. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 511–520, 2010. 1