Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval

Andrew Brown, Weidi Xie, Vicky Kalogeiton, Andrew Zisserman

Visual Geometry Group, University of Oxford
{abrown,weidi,vicky,az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/smooth-ap/

Abstract. Optimising a ranking-based metric, such as Average Precision (AP), is notoriously challenging due to the fact that it is non-differentiable, and hence cannot be optimised directly using gradient-descent methods. To this end, we introduce an objective that optimises instead a *smoothed approximation* of AP, coined *Smooth-AP*. Smooth-AP is a plug-and-play objective function that allows for end-to-end training of deep networks with a simple and elegant implementation. We also present an analysis for why directly optimising the ranking based metric of AP offers benefits over other deep metric learning losses.
We apply Smooth-AP to standard retrieval benchmarks : Stanford Online products and VehicleID, and also evaluate on larger-scale datasets: INaturalist for fine-grained category retrieval, and VGGFace2 and IJB-C for face retrieval. In all cases, we improve the performance over the state-of-the-art, especially for larger-scale datasets, thus demonstrating the effectiveness and scalability of Smooth-AP to real-world scenarios.

1 Introduction

Our objective in this paper is to improve the performance of 'query by example', where the task is: given a query image, rank all the instances in a retrieval set according to their relevance to the query. For instance, imagine that you have a photo of a friend or family member, and want to search for all of the images of that person within your large smart-phone image collection; or on a photo licensing site, you want to find all photos of a particular building or object, starting from a single photo. These use cases, where high recall is premium, differ from the 'Google Lens' application of identifying an object from an image, where only one 'hit' (match) is sufficient.

The benchmark metric for retrieval quality is Average Precision (AP) (or its generalized variant, Normalized Discounted Cumulative Gain, which includes non-binary relevance judgements). With the resurgence of deep neural networks, end-to-end training has become the *de facto* choice for solving specific vision tasks with well-defined metrics. However, the core problem with AP and similar metrics is that they include a discrete ranking function that is neither differentiable nor decomposable. Consequently, their direct optimization, *e.g.* with gradient-descent methods, is notoriously difficult.

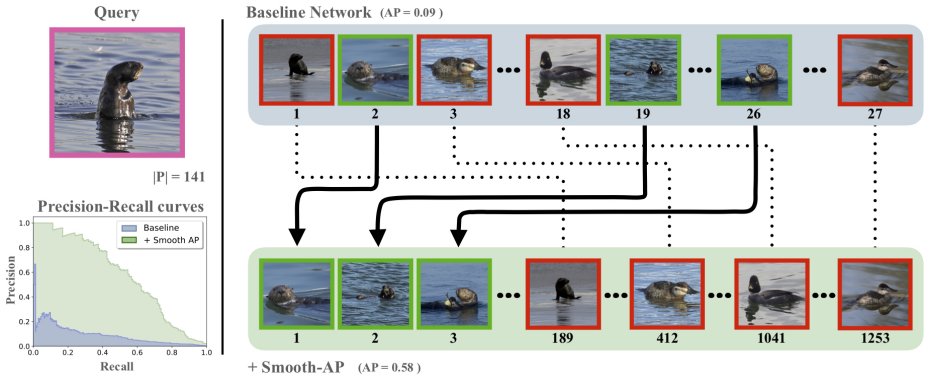


Fig. 1: **Ranked retrieval sets** before (top) and after (bottom) applying Smooth-AP on a baseline network (*i.e.* ImageNet pre-trained weights) for a given query (pink image). The precision-recall curve is shown on the left. Smooth-AP results in large boost in AP, as it moves positive instances (green) high up the ranks and negative ones (red) low down. $|\mathcal{P}|$ is the number of positive instances in the retrieval set for this query. Images are from the INaturalist dataset.

In this paper, we introduce a novel differentiable AP approximation, *Smooth-AP*, that allows end-to-end training of deep networks for ranking-based tasks. Smooth-AP is a simple, elegant, and scalable method that takes the form of a plug-and-play objective function by relaxing the Indicator function in the non-differentiable AP with a sigmoid function. To demonstrate its effectiveness, we perform experiments on two commonly used image retrieval benchmarks, Stanford Online Products and VehicleID, where Smooth-AP outperforms all recent AP approximation approaches [5,52] as well as recent deep metric learning methods. We also experiment on three further large-scale retrieval datasets (VGGFace2, IJB-C, INaturalist), which are orders of magnitude larger than the existing retrieval benchmarks. To our knowledge, this is the first work that demonstrates the possibility of training networks for AP on datasets with millions of images for the task of image retrieval. We show large performance gains over all recently proposed AP approximating approaches and, somewhat surprisingly, also outperform strong verification systems [13,35] by a significant margin, reflecting the fact that metric learning approaches are indeed inefficient for training large-scale retrieval systems that are measured by global ranking metrics.

2 Related Work

As an essential component of information retrieval [37], algorithms that optimize rank-based metrics have been the focus of extensive research over the years. In general, the previous approaches can be split into two lines of research, namely metric learning, and direct approximation of Average Precision.

Image Retrieval. This is one of the most researched topics in the vision community. Several themes have been explored in the literature, for example, one theme is on the speed of retrieval and explores methods of approximate nearest neighbors [12,27,28,30,46,58]. Another theme is on how to obtain a compact image descriptor for retrieval in order to reduce the memory footprint. Descriptors were typically constructed through an aggregation of local features, such as Fisher vectors [45] and VLAD [2,29]. More recently, neural networks have made impressive progress on learning representations for image retrieval [1,3,17,49,67], but common to all is the choice of the loss function used for training; in particular, it should ideally be a loss that will encourage good ranking.

Metric Learning. To avoid the difficulties from directly optimising rank-based metrics, such as Average Precision, there is a great body of work that focuses on metric learning [1,2,4,8,10,11,30,33,41,43,50,62,68,71]. For instance, the contrastive [11] and triplet [71] losses, which consider pairs or triplets of elements, and force all positive instances to be close in the high-dimensional embedding space, while separating negatives by a fixed distance (margin). However, due to the limited rank-positional awareness that a pair/triplet provides, a model is likely to waste capacity on improving the order of positive instances at low (poor) ranks at the expense of those at high ranks, as was pointed out by Burges *et al.* [4]. Of more relevance, the list-wise approaches [4,8,41,43,68] look at many examples from the retrieval set, and have been proven to improve training efficiency and performance. Despite being successful, one drawback of metric learning approaches is that they are mostly driven by minimizing distances, and therefore remain ignorant of the importance of shifting ranking orders – the latter is essential when evaluating with a rank-based metric.

Optimizing Average Precision (AP). The trend of directly optimising the non-differentiable AP has been recently revived in the retrieval community. Sophisticated methods [5,10,15,22,23,24,39,51,52,59,61,62,74,76] have been developed to overcome the challenge of non-decomposability and non-differentiability in optimizing AP. Methods include: creating a distribution over rankings by treating each relevance score as a Gaussian random variable [61], loss-augmented inference [39], direct loss minimization [24,59], optimizing a smooth and differentiable upper bound of AP [39,40,76], training a LSTM to approximate the discrete ranking step [15], differentiable histogram binning [5,22,23,51,62], error driven update schemes [10], and the very recent blackbox optimization [52]. Significant progress on optimizing AP was made by the information retrieval community [4,9,18,34,48,61], but the methods have largely been ignored by the vision community, possibly because they have never been demonstrated on large-scale image retrieval or due to the complexity of the proposed smooth objectives. One of the motivations of this work is to show that with the progress of deep learning research, *e.g.* auto-differentiation, better optimization techniques, large-scale datasets, and fast computation devices, it is possible and in fact very easy to directly optimize a close approximation to AP.

3 Background

In this section, we define the notations used throughout the paper.

Task definition. Given an input query, the goal of a retrieval system is to rank all instances in a retrieval set $\Omega = \{I_i, i = 0, \cdots, m\}$ based on their relevance to the query. For *each* query instance I_q, the retrieval set is split into the positive $\mathcal{P}_q$ and negative $\mathcal{N}_q$ sets, which are formed by all instances of the same class and of different classes, respectively. Note that there is a different positive and negative set for each query.

Average Precision (AP). AP is one of the standard metrics for information retrieval tasks [37]. It is a single value defined as the area under a Precision-Recall curve. For a query I_q, the predicted relevance scores of all instances in the retrieval set are measured via a chosen metric. In our case, we use the cosine similarity (though the Smooth-AP method is independent of this choice):

$$\mathcal{S}_\Omega = \left\{ s_i = \left\langle \frac{v_q}{\|v_q\|} \cdot \frac{v_i}{\|v_i\|} \right\rangle, i = 0, \cdots, n \right\}, \tag{1}$$

where $\mathcal{S}_\Omega = \mathcal{S}_P \cup \mathcal{S}_N$, and $\mathcal{S}_P = \{s_\zeta, \forall \zeta \in \mathcal{P}_q\}$, $\mathcal{S}_N = \{s_\xi, \forall \xi \in \mathcal{N}_q\}$ are the positive and negative relevance score sets, respectively, v_q refers to the query vector, and v_i to the vectorized retrieval set. The AP of a query I_q can be computed as:

$$AP_q = \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{\mathcal{R}(i, \mathcal{S}_P)}{\mathcal{R}(i, S_\Omega)}, \tag{2}$$

where $\mathcal{R}(i, \mathcal{S}_P)$ and $\mathcal{R}(i, \mathcal{S}_\Omega)$ refer to the rankings of the instance i in $\mathcal{P}$ and Ω, respectively. Note that, the rankings referred to in this paper are assumed to be *proper rankings*, meaning no two samples are ranked equally.

Ranking Function ($\mathcal{R}$). Given that AP is a ranking-based method, the key element for direct optimisation is to define the ranking $\mathcal{R}$ of one instance i. Here, we define it in the following way [48]:

$$\mathcal{R}(i, \mathcal{S}) = 1 + \sum_{j \in \mathcal{S}, j \neq i} \mathbb{1}\{(s_i - s_j) > 0\}, \tag{3}$$

where $\mathbb{1}\{\cdot\}$ acts as an Indicator function, and $\mathcal{S}$ any set, *e.g.* Ω. Conveniently, this can be implemented by computing a difference matrix $D \in \mathbb{R}^{m \times m}$:

$$D = \begin{bmatrix} s_1 & \dots & s_m \\ \vdots & \ddots & \vdots \\ s_1 & \dots & s_m \end{bmatrix} - \begin{bmatrix} s_1 & \dots & s_1 \\ \vdots & \ddots & \vdots \\ s_m & \dots & s_m \end{bmatrix} \tag{4}$$

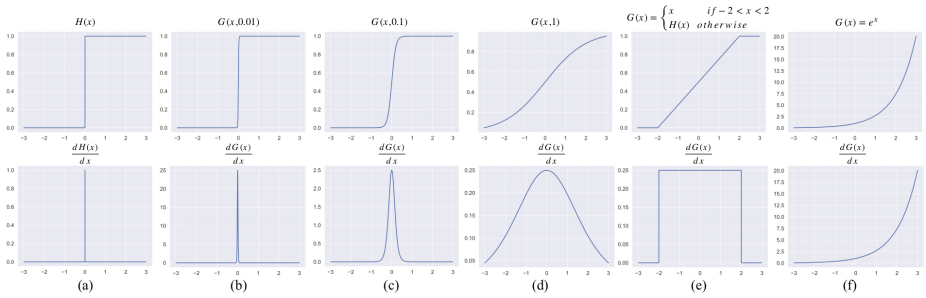


Fig. 2: The possible **different approximations** to the discrete Indicator function. First row: Indicator function (a), three sigmoids with increasing temperatures (b, c, d), linear (e), exponential (f). Second row: their derivatives.

The exact AP for a query instance I_q from Eq. 2 becomes:

$$AP_q = \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{1 + \sum_{j \in \mathcal{S}_p, j \neq i} \mathbb{1}\{D_{ij} > 0\}}{1 + \sum_{j \in \mathcal{S}_P, j \neq i} \mathbb{1}\{D_{ij} > 0\} + \sum_{j \in \mathcal{S}_N} \mathbb{1}\{D_{ij} > 0\}} \tag{5}$$

Derivatives of Indicator. The particular Indicator function used in computing AP is a Heaviside step function $\mathcal{H}(\cdot)$ [48], with its distributional derivative defined as Dirac delta function:

$$\frac{\mathrm{d}\mathcal{H}(x)}{\mathrm{d}x} = \delta(x),$$

This is either flat everywhere, with zero gradient, or discontinuous, and hence cannot be optimized with gradient based methods (Figure 2).

4 Approximating Average Precision (AP)

As explained, AP and similar metrics include a discrete ranking function that is neither differentiable nor decomposable. In this section, we first describe *Smooth-AP*, which essentially replaces the discrete indicator function with a sigmoid function, and then we provide an analysis on its relation to other ranking losses, such as triplet loss [25,71], FastAP [5] and Blackbox AP [52].

4.1 Smoothing AP

To smooth the ranking procedure, which will enable direct optimization of AP, Smooth-AP takes a simple solution which is to replace the Indicator function $\mathbb{1}\{\cdot\}$ by a sigmoid function $\mathcal{G}(\cdot;\tau)$, where the τ refers to the temperature adjusting the sharpness:

$$\mathcal{G}(x;\tau) = \frac{1}{1 + e^{\frac{-x}{\tau}}}. \tag{6}$$

Substituting $\mathcal{G}(\cdot;\tau)$ into Eq. 5, the true AP can be approximated as:

$$AP_q \approx \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{1 + \sum_{j \in \mathcal{S}_P} \mathcal{G}(D_{ij};\tau)}{1 + \sum_{j \in \mathcal{S}_P} \mathcal{G}(D_{ij};\tau) + \sum_{j \in \mathcal{S}_N} \mathcal{G}(D_{ij};\tau)}$$

with tighter approximation and convergence to the indicator function as $\tau \to 0$. The objective function during optimization is denoted as:

$$\mathcal{L}_{AP} = \frac{1}{m} \sum_{k=1}^{m} (1 - AP_k) \tag{7}$$

Smoothing parameter τ governs the temperature of the sigmoid that replaces the Indicator function $\mathbb{1}\{\cdot\}$. It defines an operating region, where terms of the difference matrix are given a gradient by the Smooth-AP loss. If the terms are mis-ranked, Smooth-AP will attempt to shift them to the correct order. Specifically, a small value of τ results in a small operating region (Figure 2 (b) – note the small region with gradient seen in the sigmoid derivative), and a tighter approximation of true AP. The strong acceleration in gradient around the zero point (Figure 2 (b)-(c) second row) is essential to replicating the desired qualities of AP, as it encourages the shifting of instances in the embedding space that result in a change of rank (and hence change in AP), rather than shifting instances by some large distance but not changing the rank. A large value of τ offers a large operating region, however, at the cost of a looser approximation to AP due to its divergence from the indicator function.

Relation to Triplet Loss. Here, we demonstrate that the triplet loss (a popular surrogate loss for ranking) is in fact optimising a distance metric rather than a ranking metric, which is sub-optimal when evaluating using a ranking metric. As shown in Eq. 5, the goal of optimizing AP is equivalent to minimizing all the $\sum_{i \in \mathcal{S}_P, j \in \mathcal{S}_N} \mathbb{1}\{D_{ij} < 0\}$, *i.e.* the violating terms. We term these as such because these terms refer to cases where a negative instance is ranked above a positive instance in terms of relevance to the query, and optimal AP is only acquired when all positive instances are ranked above all negative instances.

For example, consider one query instance with predicted relevance score and ground-truth relevance labels as:

$$\text{Instances ordered by score} : (s_0 \;\; s_4 \;\; s_1 \;\; s_2 \;\; s_5 \;\; s_6 \;\; s_7 \;\; s_3)$$
$$\text{Ground truth labels} : (1 \;\; 0 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 1)$$

the violating terms are: $\{(s_4 - s_1), (s_4 - s_2), (s_4 - s_3), (s_5 - s_3), (s_6 - s_3), (s_7 - s_3)\}$. An ideal AP loss would actually treat each of the terms unequally, *i.e.* the model would be forced to spend more capacity on shifting orders between s_4 and s_1, rather than s_3 and s_7, as that makes a larger impact on improving the AP.

Another interpretation of these violating cases can also be drawn from the triplet loss perspective. Specifically, if we treat the query instance as an "anchor", with s_j denoting the similarity between the "anchor" and negative instance, and

s_i denoting the similarity between the "anchor" and positive instance. In this example, the triplet loss tries to optimize a margin hinge loss:

$$\begin{aligned}\mathcal{L}_{\text{triplet}} = &\max(s_4 - s_1 + \alpha, 0) + \max(s_4 - s_2 + \alpha, 0) \\ &+ \max(s_4 - s_3 + \alpha, 0) + \max(s_5 - s_3 + \alpha, 0) \\ &+ \max(s_6 - s_3 + \alpha, 0) + \max(s_7 - s_3 + \alpha, 0)\end{aligned}$$

This can be viewed as a differentiable approximation to the goal of optimizing AP where the Indicator function has been replaced with the margin hinge loss, thus solving the gradient problem. Nevertheless, using a triplet loss to approximate AP may suffer from two problems: *First*, all terms are linearly combined and treated equally in $\mathcal{L}_{\text{triplet}}$. Such a surrogate loss may force the model to optimize the terms that have only a small effect on AP, *e.g.* optimizing $s_4 - s_1$ is the same as $s_7 - s_4$ in the triplet loss, however, from an AP perspective, it is important to correct the mis-ordered instances at high rank. *Second*, the linear derivative means that the optimization process is purely based on distance (not ranking orders), which makes it sub-optimal when evaluating AP. For instance, in the triplet loss case, reducing the distance $s_4 - s_1$ from 0.8 to 0.5 is the same as from 0.2 to -0.1. In practise, however, the latter case (shifting orders) will clearly have a much larger impact on the AP computation than the former.

Comparison to other AP-optimising methods. The two key differences between Smooth-AP and the recently introduced FastAP and Blackbox AP, are that Smooth-AP (i) provides a closer approximation to Average Precision, and (ii) is far simpler to implement. Firstly, due to the sigmoid function, Smooth-AP optimises a ranking metric, and so has the same objective as Average Precision. In contrast, FastAP and Blackbox AP linearly interpolate the non-differentiable (piecewise constant) function, which can potentially lead to the same issues as triplet loss, *i.e.* optimizing a distance metric, rather than rankings. Secondly, Smooth-AP simply needs to replace the indicator function in the AP objective with a sigmoid function. While FastAP uses abstractions such as Histogram Binning, and Blackbox AP uses a variant of numerical derivative. These differences are positively affirmed through the improved performance of Smooth-AP over several datasets (Section 6.5).

5 Experimental Setup

In this section, we describe the datasets used for evaluation, the test protocols, and the implementation details. The procedure followed here is to take a pre-trained network and fine-tune with Smooth-AP loss. Specifically, ImageNet pretrained networks are used for the object/animal retrieval datasets, and high-performing face-verification models for the face retrieval datasets.

5.1 Datasets

We evaluate the Smooth-AP loss on five datasets containing a wide range of domains and sizes. These include the commonly used retrieval benchmark datasets,

Table 1: **Datasets** used for training and evaluation.

dataset		# Images	# Classes	# Ims/Class
object/animal retrieval datasets	SOP train	59,551	11,318	5.3
	SOP test	60,502	11,316	5.3
	VehicleID train	110,178	13,134	8.4
	VehicleID test	40,365	4,800	8.4
	INaturalist train	325,846	5,690	57.3
	INaturalist test	136,093	2,452	55.5
face retrieval datasets	VGGFace2 train	3.31 M	8,631	363.7
	VGGFace2 test	169,396	500	338.8
	IJB-C	148,824	3,531	42.1

as well as several additional *large-scale* (>100K images) datasets. Table 1 describes their details.

Stanford Online Product (SOP) [59] was initially collected for investigating the problem of metric learning. It includes 120K images of products that were sold online. We use the same evaluation protocol and train/test split as [68].

VehicleID [65] contains $221,736$ images of $26,267$ vehicle categories, $13,134$ of which are used for training (containing $110,178$ images). By following the same test protocol as [65], three test sets of increasing size are used for evaluation (termed small, medium, large), which contain 800 classes ($7,332$ images), 1600 classes ($12,995$ images) and 2400 classes ($20,038$ images) respectively.

INaturalist [63] is a large-scale animal and plant species classification dataset, designed to replicate real-world scenarios through 461,939 images from 8,142 classes. It features many visually similar species, captured in a wide variety of environments. We construct a new image retrieval task from this dataset, by keeping 5,690 classes for training, and 2,452 unseen classes for evaluating image retrieval at test time, according to the same test protocols as existing benchmarks [68]. We will make the train/test splits publicly available.

VGGFace2 [7] is a large-scale face dataset with over 3.31 million images of $9,131$ subjects. The images have large variations in pose, age, illumination, ethnicity and profession, *e.g.* actors, athletes, politicians. For training, we use the pre-defined *training set* with $8,631$ identities, and for testing we use the *test set* with 500 identities, totalling 169K testing images.

IJB-C [38] is a challenging public benchmark for face recognition, containing images of subjects from both still frames and videos. Each video is treated as a single instance by averaging the CNN-produced vectors for each frame to a single vector. Identities with less than 5 instances (images or videos) are removed.

5.2 Test Protocol

Here, we describe the protocols for evaluating retrieval performance, mean Average Precision (mAP) and Recall@K (R@K). For all datasets, every instance of each class is used in turn as the query I_q, and the retrieval set Ω is formed out of all the remaining instances. We ensure that each class in all datasets contains several images (Table 1), such that if an instance from a class is used as the query, there are plenty of remaining positive instances in the retrieval set. For object/animal retrieval evaluation, we use the *Recall@K* metric in order to compare to existing works. For face retrieval, AP is computed from the resulting output ranking for each query, and the mAP score is computed by averaging the APs across every instance in the dataset, resulting in a single value.

5.3 Implementation Details

Object/animal retrieval (SOP, VehicleID, INaturalist). In line with previous works [5,52,54,56,70,72],we use ResNet50 [20] as the backbone architecture, which was pretrained on ImageNet [55]. We replace the final softmax layer with one linear layer (following [5,52], with dimension being set to 512). All images are resized to 256×256. At training time, we use random crops and flips as augmentations, and at test time, a single centre crop of size 224×224 is used. For all experiments we set τ to 0.01 (Section 6.4).

Face retrieval datasets (VGGFace2, IJB-C). We use two high performing face verification networks: the method from [7] using the SENet-50 architecture [26] and the state-of-the-art ArcFace [13] (using ResNet-50), both trained on the VGGFace2 training set. For SENet-50, we follow [7] and use the same face crops (extended by the recommended amount), resized to 224×224 and we L2-normalize the final 256D embedding. For ArcFace, we generate normalised face crops (112×112) by using the provided face detector [13], and align them with the predicted 5 facial key points, then L2-normalize the final 512D embedding. For both models, we set the batch size to 224 and τ to 0.01 (Section 6.4).

Mini-batch training. During training, we form each mini-batch by randomly sampling classes such that each represented class has $|\mathcal{P}|$ samples per class. For all experiments, we L2-normalize the embeddings, use cosine similarity to compute the relevance scores between the query and the retrieval set, set $|\mathcal{P}|$ to 4, and use an Adam [32] optimiser with a base learning rate of 10^{-5} with weight decay $4e^{-5}$. We employ the same hard negative mining technique as [5,52] only for the Online Products dataset. Otherwise we use no special sampling strategies.

6 Results

In this section, we first explore the effectiveness of the proposed Smooth-AP by examining the performance of various models on the five retrieval datasets.

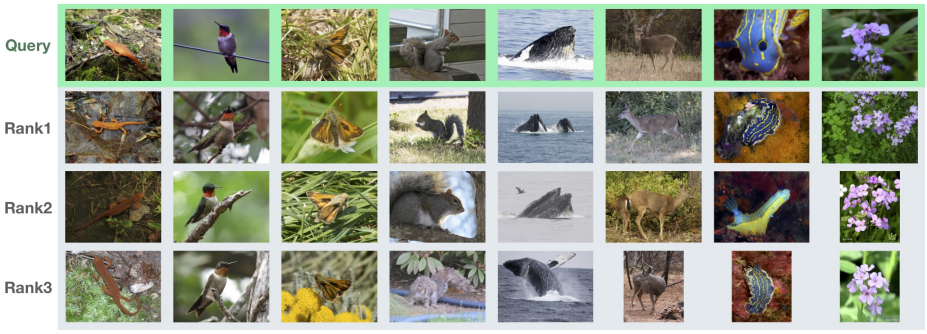


Fig. 3: **Qualitative results for the INaturalist dataset using Smooth-AP loss**. For each query image (top row), the top 3 instances from the retrieval set are shown ranked from top to bottom. Every retrieved instance shown is a true positive.

Specifically, we compare with the recent AP optimization and broader metric learning methods on the standard benchmarks SOP and VehicleID (Section 6.1), and then shift to further large-scale experiments, *e.g.* INaturalist for animal/-plant retrieval, and IJB-C and VGGFace2 for face retrieval (Sections 6.2-6.3). Then, we present an ablation study of various hyper-parameters that affect the performance of Smooth-AP: the sigmoid temperature, the size of the positive set, and the batch size (Section 6.4). Finally, we discuss various findings and analyze the performance gaps between various models (Section 6.5).

Note that, although there has been a rich literature on metric learning methods [14,16,19,31,33,36,41,42,44,47,52,59,60,62,66,68,69,70,72,73,75] using these image retrieval benchmarks, we only list the very recent state-of-the-art approaches, and try to compare with them as fairly as we can, *e.g.* no model ensemble, and using the same backbone network and image resolution. However, there still remain differences on some small experimental details, such as embedding dimensions, optimizer, and learning rates. Qualitative results for the INaturalist dataset are illustrated in Figure 3.

6.1 Evaluation on Stanford Online Products (SOP)

We compare with a wide variety of state-of-the-art image retrieval methods, *e.g.* deep metric learning methods [54,56,70,72], and AP approximation methods [5,52]. As shown in Table 2, we observe that Smooth-AP achieves state-of-the-art results on the SOP benchmark. In particular, our best model outperforms the very recent AP approximating methods (Blackbox AP and FastAP) by a 1.5% margin for Recall@1. Furthermore, Smooth-AP performs on par with the concurrent work (Cross-Batch Memory [70]). This is particularly impressive as [70] harnesses memory techniques to sample from many mini-batches simultaneously for each weight update, whereas Smooth-AP only makes use of a single

	SOP			
Recall@K	1	10	100	1000
Margin [72]	72.7	86.2	93.8	98.0
Divide [56]	75.9	88.4	94.9	98.1
FastAP [5]	76.4	89.0	95.1	98.2
MIC [54]	77.2	89.4	95.6	-
Blackbox AP [52]	78.6	90.5	96.0	98.7
Cont. w/M [70]	**80.6**	**91.6**	96.2	98.7
Smooth-AP BS=224	79.2	91.0	96.5	98.9
Smooth-AP BS=384	80.1	91.5	**96.6**	**99.0**

Table 2: **Results on Stanford Online Products.** Deep metric learning and recent AP approximating methods are compared to using the ResNet50 architecture. BS: mini-batch size.

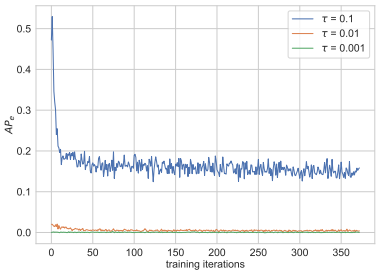


Fig. 4: **The AP approximation error**, AP_e over one training epoch for Online Products for different values of sigmoid annealing temperature, τ.

mini-batch on each training iteration.

Figure 4 provides a quantitative analysis into the effect of sigmoid temperature τ on the tightness of the AP approximation, which can be plotted via the AP approximation error:

$$AP_e = |AP_{pred} - AP| \tag{8}$$

where AP_{pred} is the predicted approximate AP when the sigmoid is used in place of the indicator function in Equation 5, and AP is the true AP. As expected, a lower value of τ leads to a tighter approximation to Average Precision, shown by the low approximation error.

6.2 Evaluation on VehicleID and INaturalist

In Table 3, we show results on the VehicleID and INaturalist dataset. We observe that Smooth-AP achieves state-of-the-art results on the challenging and large-scale VehicleID dataset. In particular, our model outperforms FastAP by a significant 3% for the Small protocol Recall@1. Furthermore, Smooth-AP exceeds the performance of [70] on 4 of the 6 recall metrics.

As we are the first to report results on INaturalist for *image retrieval*, in addition to Smooth-AP, we re-train state-of-the-art metric learning and AP approximating methods, with the respective official code, *e.g.* Triplet and ProxyNCA [53], FastAP [6], Blackbox AP [64]. As shown in Table 3. Smooth-AP outperforms all methods by $2-5\%$ on Recall@1 for the experiments when the same batch size is used (224). Increasing the batch size to 384 for Smooth-AP leads to a further boost of 1.4% to 66.6 for Recall@1. These results demonstrate that Smooth-AP is particularly suitable for *large-scale* retrieval datasets, thus revealing its scalability to *real-world* retrieval problems. We note here that these large-scale datasets (>100k images) are less influenced by hyper-parameter

Table 3: **Results on the VehicleID (left) and INaturalist (right).** All experiments are conducted using ResNet50 as backbone. All results for INaturalist are from publicly available official implementations in the PyTorch framework with a batch size of 224. † refers to the recent re-implementation [53] - we make the design choice for Proxy NCA loss to keep the number of proxies equal to the number of training classes. The VehicleID results are obtained with a batch-size of 384.

	VehicleID					
	Small		Medium		Large	
Recall@K	1	5	1	5	1	5
Divide [56]	87.7	92.9	85.7	90.4	82.9	90.2
MIC [54]	86.9	93.4	-	-	82.0	91.0
FastAP [5]	91.9	96.8	90.6	95.9	87.5	95.1
Cont. w/M [70]	94.7	96.8	**93.7**	95.8	**93.0**	95.8
Smooth-AP	**94.9**	**97.6**	93.3	**96.4**	91.9	**96.2**

	INaturalist			
Recall@K	1	4	16	32
Triplet Semi-Hard [72]	58.1	75.5	86.8	90.7
Proxy NCA† [41]	61.6	77.4	87.0	90.6
FastAP [5]	60.6	77.0	87.2	90.6
Blackbox AP [52]	62.9	79.0	88.9	92.1
Smooth-AP BS=224	65.9	80.9	89.8	92.7
Smooth-AP BS=384	**67.2**	**81.8**	**90.3**	**93.1**

Table 4: **mAP results on face retrieval datasets**. Smooth-AP *consistently* boosts the AP performance for both VGGFace2 and ArcFace, while outperforming other standard metric learning losses (Pairwise contrastive and Triplet).

VGGFace2	VF2 Test	IJB-C
Softmax	0.828	0.726
+Pairwise	0.828	0.728
+Triplet	0.845	0.740
+Smooth-AP	**0.850**	**0.754**

ArcFace	VF2 Test	IJB-C
ArcFace	0.858	0.772
+Pairwise	0.861	0.775
+Triplet	0.880	0.787
+Smooth-AP	**0.902**	**0.803**

tuning and so provide ideal test environments to demonstrate improved image retrieval techniques.

6.3 Evaluation on Face Retrieval

Due to impressive results [7,13], face retrieval is considered saturated. Nevertheless, we demonstrate here that Smooth-AP can further boost the face retrieval performance. Specifically, we append Smooth-AP on top of modern methods (VGGFace2 and ArcFace) and evaluate mAP on IJB-C and VGGFace2, *i.e.* one of the largest face recognition datasets.

As shown in Table 4, when appending the Smooth-AP loss, retrieval metrics such as mAP can be significantly improved upon the baseline model for both datasets. This is particularity impressive as both baselines have already shown very strong performance on facial verification and identification tasks, yet Smooth-AP is able to increase mAP by up to 4.4% on VGGFace2 and 3.1% on ArcFace. Moreover, Smooth-AP strongly outperforms both the pairwise [11] and triplet [57] losses, *i.e.* the two most popular surrogates to a ranking loss. As discussed in Section 4.1, these surrogates optimise a distance metric rather than a ranking metric, and the results show that the latter is optimal for AP.

Table 5: **Ablation study** over different parameters: temperature τ, size of positive set during minibatch sampling $|\mathcal{P}|$, and batch size B. Performance is benchmarked on VGGFace2-Test and IJB-C.

τ	mAP	
	VF2	IJB-C
0.1	0.824	0.726
0.01	**0.844**	**0.736**
0.001	0.839	0.733

$|\mathcal{P}| = 4$, $B = 128$

$|\mathcal{P}|$	mAP	
	VF2	IJB-C
4	**0.844**	**0.736**
8	0.833	0.734
16	0.824	0.726

$\tau = 0.01$, $B = 128$

$|\mathcal{B}|$	mAP	
	VF2	IJB-C
64	0.824	0.726
128	0.844	0.736
256.	**0.853**	**0.754**

$\tau = 0.01$, $|\mathcal{P}| = 4$

6.4 Ablation study

To investigate the effect of different hyper-parameter settings, *e.g.* the sigmoid temperature τ, the size of the positive set $|\mathcal{P}|$, and batch size B (Table 5), we use VGGFace2 and IJB-C with SE-Net50 [7], as the large-scale datasets are unlikely to lead to overfitting, and therefore provide a fair understanding about these hyper-parameters. Note that we only vary one parameter at a time.

Effect of sigmoid temperature τ. As explained in Section 4.1, τ governs the smoothing of the sigmoid that is used to approximate the indicator function in the Smooth-AP loss. The ablation shows that a value of 0.01 leads to the best mAP scores, which is the optimal trade-off between AP approximation and a large enough operating region in which to provide gradients. Surprisingly, this value (0.01) corresponds to a small operating region. We conjecture that a tight approximation to true AP is the key, and when partnered with a large enough batch size, enough elements of the difference matrix will lie within the operating region in order to induce sufficient re-ranking gradients. The sigmoid temperature can further be viewed from the margin perspective (inter-class margins are commonly used in metric learning to help generalisation [11,43,71]). Smooth-AP only stops providing gradients to push a positive instance above a negative instance once they are a distance equal to the width of the operating region apart, hence enforcing a margin that equates to roughly 0.1 for this choice of τ.

Effect of positive set $|\mathcal{P}|$. In this setting, the positive set represents the *instances* that come from the same class in the mini-batch during training. We observe that a small value (4) results in the highest mAP scores, this is because mini-batches are formed by sampling at the class level, where a low value for $|\mathcal{P}|$ means a larger number of sampled classes and a higher probability of sampling hard-negative instances that violate the correct ranking order. Increasing the number of classes in the batch results in a better batch approximation of the true class distribution, allowing each training iteration to enforce a more optimally structured embedding space.

Effect of batch size B**.** Table 5 shows that large batch sizes result in better mAP, especially for VGGFace2. This is expected, as it again increases the chance of getting hard-negative samples in the batch.

6.5 Further discussion

There are several important observations in the above results. Smooth-AP outperforms all previous AP approximation approaches, as well as the metric learning techniques (pair, triplet, and list-wise) on *three* image retrieval benchmarks, SOP, VehicleID, Inaturalist, with the performance gap being particularly apparent on the large-scale *INaturalist* dataset. Similarly, when scaled to face datasets containing millions of images, Smooth-AP is able to improve the retrieval metrics for state-of-the-art face verification networks. We hypothesis that these performance gains upon the previous AP approximating methods come from a tighter approximation to AP than other existing approaches, hence demonstrating the effectiveness and scalability of Smooth-AP. Furthermore, many of the properties that deep metric learning losses handcraft into their respective methods (distance-based weighting [59,72], inter-class margins [11,57,59,68], intra-class margins [68]), are naturally built into our AP formulation, and result in improved generalisation capabilities.

7 Conclusions

We introduce *Smooth-AP*, a novel loss that directly optimizes a smoothed approximation of AP. This is in contrast to modern contrastive, triplet, and list-wise deep metric learning losses which act as surrogates to encourage ranking. We show that Smooth-AP outperforms recent AP-optimising methods, as well as the deep metric learning methods, and with a simple and elegant, plug-and-play style method. We provide an analysis for the reasons why Smooth-AP outperforms these other losses, *i.e.* Smooth-AP preserves the goal of AP which is to optimise ranking rather than distances in the embedding space. Moreover, we also show that fine-tuning face-verification networks by appending the Smooth-AP loss can strongly improve the performance. Finally, in an effort to bridge the gap between experimental settings and real-world retrieval scenarios, we provide experiments on several large-scale datasets and show Smooth-AP loss to be considerably more scalable than previous approximations.

Acknowledgements. We are grateful to Tengda Han, Olivia Wiles, Christian Rupprecht, Sagar Vaze, Quentin Pleple and Maya Gulieva for proof-reading, and to Ernesto Coto for the initial motivation for this work. Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1. AB is funded by an EPSRC DTA Studentship.

References

1. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proc. CVPR (2016)
2. Arandjelovic, R., Zisserman, A.: All about vlad. In: Proc. CVPR (2013)
3. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Proc. ECCV (2014)
4. Burges, C.J., Ragno, R., Le, Q.V.: Learning to rank with nonsmooth cost functions. In: NeurIPS (2007)
5. Cakir, F., He, K., Xia, X., Kulis, B., Sclaroff, S.: Deep metric learning to rank. In: Proc. CVPR (2019)
6. Cakir, F., He, K., Xia, X., Kulis, B., Sclaroff, S.: Fastap: Deep metric learning to rank. `https://github.com/kunhe/Deep-Metric-Learning-Baselines` (2019)
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2018)
8. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proc. ICML (2007)
9. Chapelle, O., Le, Q., Smola, A.: Large margin optimization of ranking measures. In: NeurIPS (2007)
10. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Autocorrect: Deep inductive alignment of noisy geometric annotations. In: Proc. BMVC (2019)
11. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proc. CVPR (2005)
12. Chum, O., Mikulik, A., Perďoch, M., Matas, J.: Total recall II: Query expansion revisited. In: Proc. CVPR (2011)
13. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. CVPR (2019)
14. Duan, Y., Zheng, W., Lin, X., Lu, J., Zhou, J.: Deep adversarial metric learning. In: Proc. CVPR (2018)
15. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Sodeep: a sorting deep net to learn ranking loss surrogates. In: Proc. CVPR (2019)
16. Ge, W., Huang, W., Dong, D., Scott, M.: Deep metric learning with hierarchical triplet loss. In: Proc. ECCV (2018)
17. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: Proc. ECCV (2016)
18. Guiver, J., Snelson, E.: Learning to rank with softrank and gaussian processes. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2008)
19. Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T., et al.: Smart mining for deep metric learning. In: Proc. ICCV (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016)
22. He, K., Cakir, F., Adel Bargal, S., Sclaroff, S.: Hashing as tie-aware learning to rank. In: Proc. CVPR (2018)
23. He, K., Lu, Y., Sclaroff, S.: Local descriptors optimized for average precision. In: Proc. CVPR (2018)

24. Henderson, P., Ferrari, V.: End-to-end training of object class detectors for mean average precision. In: Proc. ACCV (2016)
25. Hermans, A., Beyer, L., Leibe, B.: In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint arXiv:1703.07737 (2017)
26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proc. CVPR (2018)
27. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Proc. ECCV (2008)
28. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE PAMI (2011)
29. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Proc. CVPR (2010)
30. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., P'erez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE PAMI (2011)
31. Kim, W., Goyal, B., Chawla, K., Lee, J., Kwon, K.: Attention-based ensemble for deep metric learning. In: Proc. ECCV (2018)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR (2014)
33. Law, M.T., Urtasun, R., Zemel, R.S.: Deep spectral clustering learning. In: Proc. ICML (2017)
34. Li, K., Huang, Z., Cheng, Y.C., Lee, C.H.: A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers. In: Proc. ICASSP (2014)
35. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proc. CVPR (2017)
36. Lu, J., Xu, C., Zhang, W., Duan, L.Y., Mei, T.: Sampling wisely: Deep image embedding by top-k precision optimization. In: Proc. ICCV (2019)
37. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
38. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB) (2018)
39. McFee, B., Lanckriet, G.R.: Metric learning to rank. In: Proc. ICML (2010)
40. Mohapatra, P., Rolinek, M., Jawahar, C., Kolmogorov, V., Pawan, K.: Efficient optimization for rank-based loss functions. In: Proc. CVPR (2018)
41. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proc. CVPR (2017)
42. Oh Song, H., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: Proc. CVPR (2017)
43. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proc. CVPR (2016)
44. Opitz, M., Waltner, G., Possegger, H., Bischof, H.: BIER: Boosting Independent Embeddings Robustly. In: Proc. ICCV (2017)
45. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: Proc. CVPR (2010)
46. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. CVPR (2007)
47. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proc. ICCV (2019)
48. Qin, T., Liu, T.Y., Li, H.: A general approximation framework for direct optimization of information retrieval measures. Information Retrieval (2010)
49. Radenovic, F., Tolias, G., Chum, O.: CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: Proc. ECCV (2016)

50. Rao, Y., Lin, D., Lu, J., Zhou, J.: Learning globally optimized object detector via policy gradient. In: Proc. CVPR (2018)
51. Revaud, J., Almazá, J., Rezende, R.S., de Souza, C.R.: Learning with average precision: Training image retrieval with a listwise loss. In: Proc. ICCV (2019)
52. Rolnek, M., Musil, V., Paulus, A., Vlastelica, M., Michaelis, C., Martius, G.: Optimizing rank-based metrics with blackbox differentiation. In: Proc. CVPR (2020)
53. Roth, K., Brattoli, B.: Deep metric learning baselines. `https://github.com/Confusezius/Deep-Metric-Learning-Baselines` (2019)
54. Roth, K., Brattoli, B., Ommer, B.: Mic: Mining interclass characteristics for improved metric learning. In: Proc. ICCV (2019)
55. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
56. Sanakoyeu, A., Tschernezki, V., Buchler, U., Ommer, B.: Divide and conquer the embedding space for metric learning. In: Proc. CVPR (2019)
57. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. CVPR (2015)
58. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. ICCV (2003)
59. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proc. CVPR (2016)
60. Suh, Y., Han, B., Kim, W., Lee, K.M.: Stochastic class-based hard example mining for deep metric learning. In: Proc. CVPR (2019)
61. Taylor, M., Guiver, J., Robertson, S., Minka, T.: Softrank: Optimising non-smooth rank metrics. In: WSDM (2008)
62. Ustinova, E., Lempitsky, V.: Learning deep embeddings with histogram loss. In: NeurIPS (2016)
63. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proc. CVPR (2018)
64. Vlastelica, M., Paulus, A., Musil, V., Martius, G., Rolnek, M.: Differentiation of blackbox combinatorial solvers. `https://github.com/martius-lab/blackbox-backprop` (2020)
65. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
66. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: Proc. ICCV (2017)
67. Wang, J., song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proc. CVPR (2014)
68. Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., Robertson, N.: Ranked list loss for deep metric learning. In: Proc. CVPR (2019)
69. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proc. CVPR (2019)
70. Wang, X., Zhang, H., Huang, W., Scott, M.R.: Cross-batch memory for embedding learning. In: Proc. CVPR (2020)
71. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NeurIPS (2006)
72. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proc. CVPR (2017)

73. Xuan, H., Souvenir, R., Pless, R.: Deep randomized ensembles for metric learning. In: Proc. ECCV (2018)
74. Yang, H.F., Lin, K., Chen, C.S.: Cross-batch reference learning for deep classification and retrieval. In: Proc. ACMMM (2016)
75. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: Proc. ICCV (2017)
76. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: SIGIR (2007)

Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval *Supplementary Material*

Andrew Brown[0000−0002−9556−2633], Weidi Xie[0000−0003−3804−2639], Vicky Kalogeiton[0000−0002−7368−6993], and Andrew Zisserman[0000−0002−8945−8573]

Visual Geometry Group, University of Oxford
{abrown,weidi,vicky,az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/smooth-ap/

Table Of Contents

1 Further qualitative results

In this Section, we provide additional qualitative results for four of the datasets used in our experiments (VGGFace2 Test set, Stanford Online Products, and INaturalist). For each dataset, we display the top retrieved instances for various queries from the baseline model, both with and without appending the Smooth-AP loss. In all cases, the query example is shown in blue. The retrieved instances belonging to the same class as the query (*i.e.* positive set) are shown in green, while the ones belonging to a different class from the query (*i.e.* negative set) are shown in red. For all retrieval examples we have shown the corresponding precision-recall curves below, in which the baseline model is represented in blue, and the Smooth-AP model for the same query instance is represented overlaid in green. For Figures S1, S2 the retrieval set is ranked from left to right starting in the top row next to the query. For Figures S3, S4, S5, S6 the retrieval set is ranked from top to bottom. In each case, the Average Precision (AP) computed over the whole retrieval set is provided either below or alongside the ranked instances.

Fig. S1: Qualitative results from the VGGFace2 Test set for the SENet-50 [7] baseline model. We show a query instance (blue) and the first 79 ranked instances in the retrieval set for the baseline model both before and after Smooth-AP was appended (ranked from left to right, starting next to the query). As shown by the precision-recall curves, Smooth-AP causes the Average Precision to jump by an impressive 52.9%, and the number of false positives (shown in red) in the top ranked retrieved instances drops considerably. The size of the positive set for each instance in the VGGFace2 test set, $|P| \approx 338$.

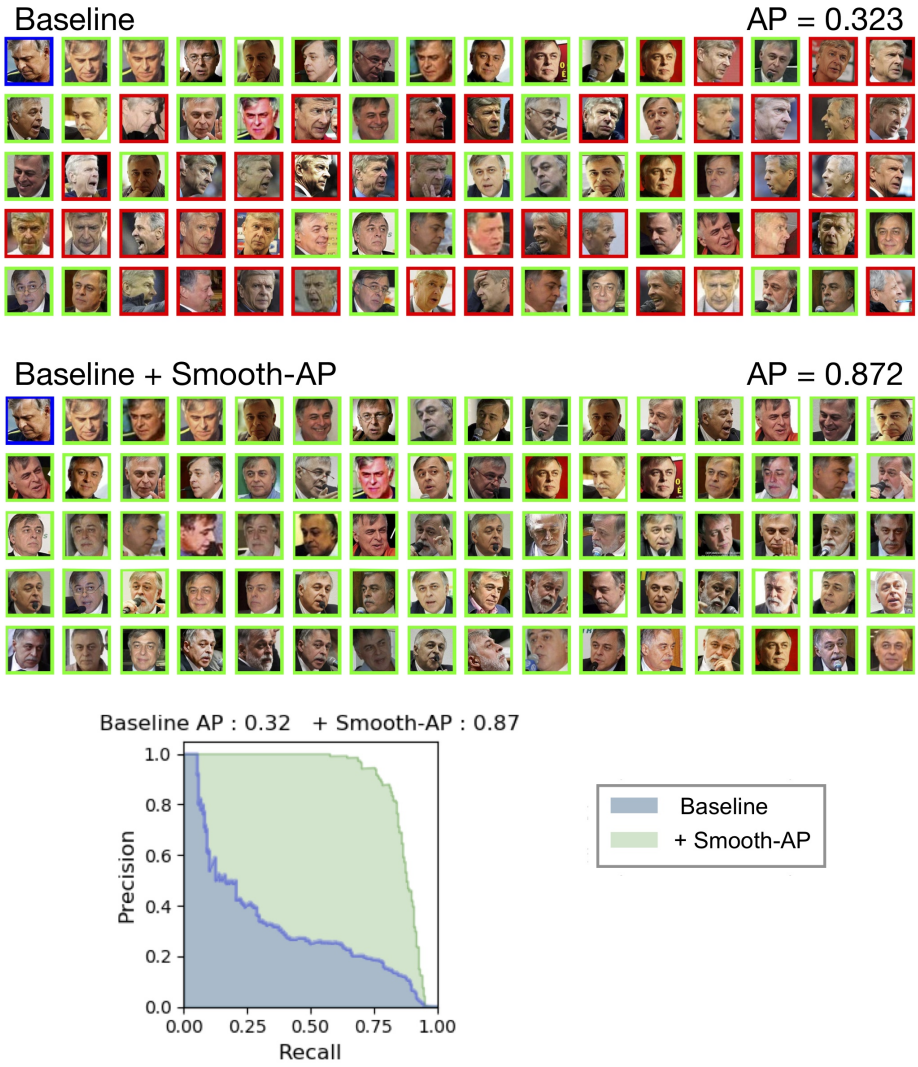


Fig. S2: Here, we show the qualitative results for a second query instance for the SENet-50 [7] baseline model on the VGGFace2 test set, both before and after appending the Smooth-AP loss. Here, we see another large increase in Average Precision of 54.9% caused by the addition of the Smooth-AP loss. We see that all false positives (shown in red) are removed from the top ranked retrieved instances after adding the Smooth-AP loss. Take the situation where each row of top-ranked instances corresponds to pages of retrieved results that a user is presented with when using a retrieval system. With the baseline model, the user comes across many false positives in the first few pages. After appending the Smooth-AP loss, the user encounters no false positives in at least the first five pages. This demonstrates the benefit to *user experience* in appending Smooth-AP to a retrieval network.

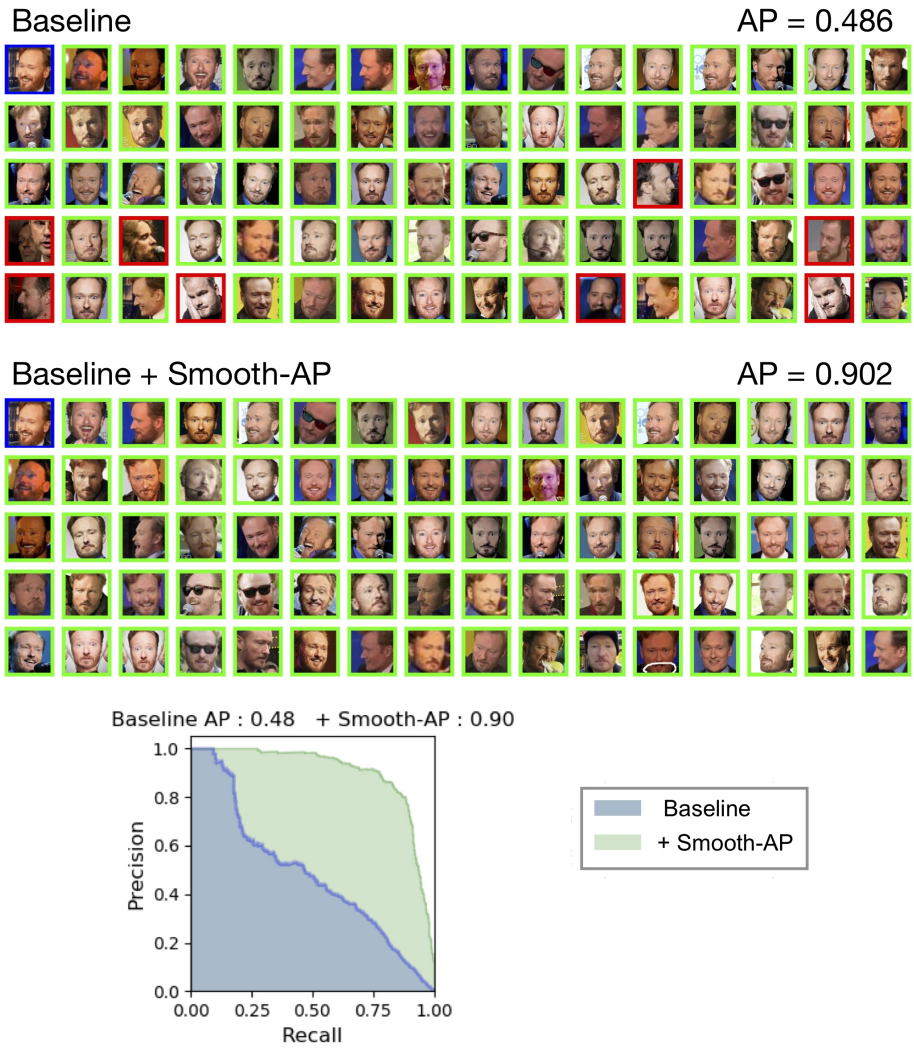


Fig. S3: Here, we show the qualitative results for a third query instance for the SENet-50 [7] baseline model on the VGGFace2 test set, both before and after appending the Smooth-AP loss. We see a large improvement in Average Precision of 41.8% after adding the Smooth-AP loss and the removal of all false positives from the top ranked retrieval results. These results confirm that Smooth-AP is indeed tailored to addressing the ranking issue.

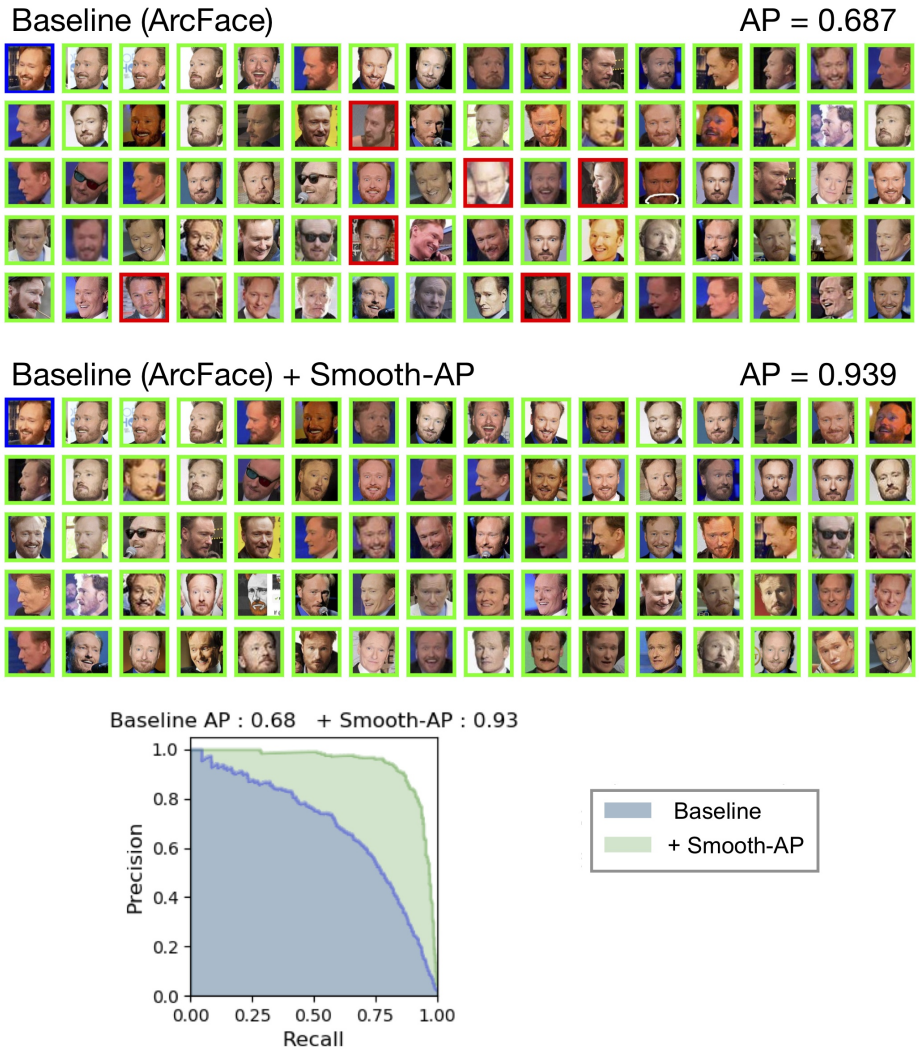


Fig. S4: Here, we show the qualitative results for a query instance from the VGGFace2 test set (same as in Figure S3) for the state-of-the-art ArcFace (ResNet-50) [13] baseline, both before and after appending the Smooth-AP loss. Appending the Smooth-AP loss to this impressive baseline leads to a large gain in Average Precision (24.8%), and again to the removal of all false positives from the top ranked retrieval results. This demonstrates that state-of-the-art face retrieval networks are far from saturated on the Average Precision metric, and through appending the simple Smooth-AP loss, this metric and the resulting user experience when using the face retrieval system can be greatly improved.

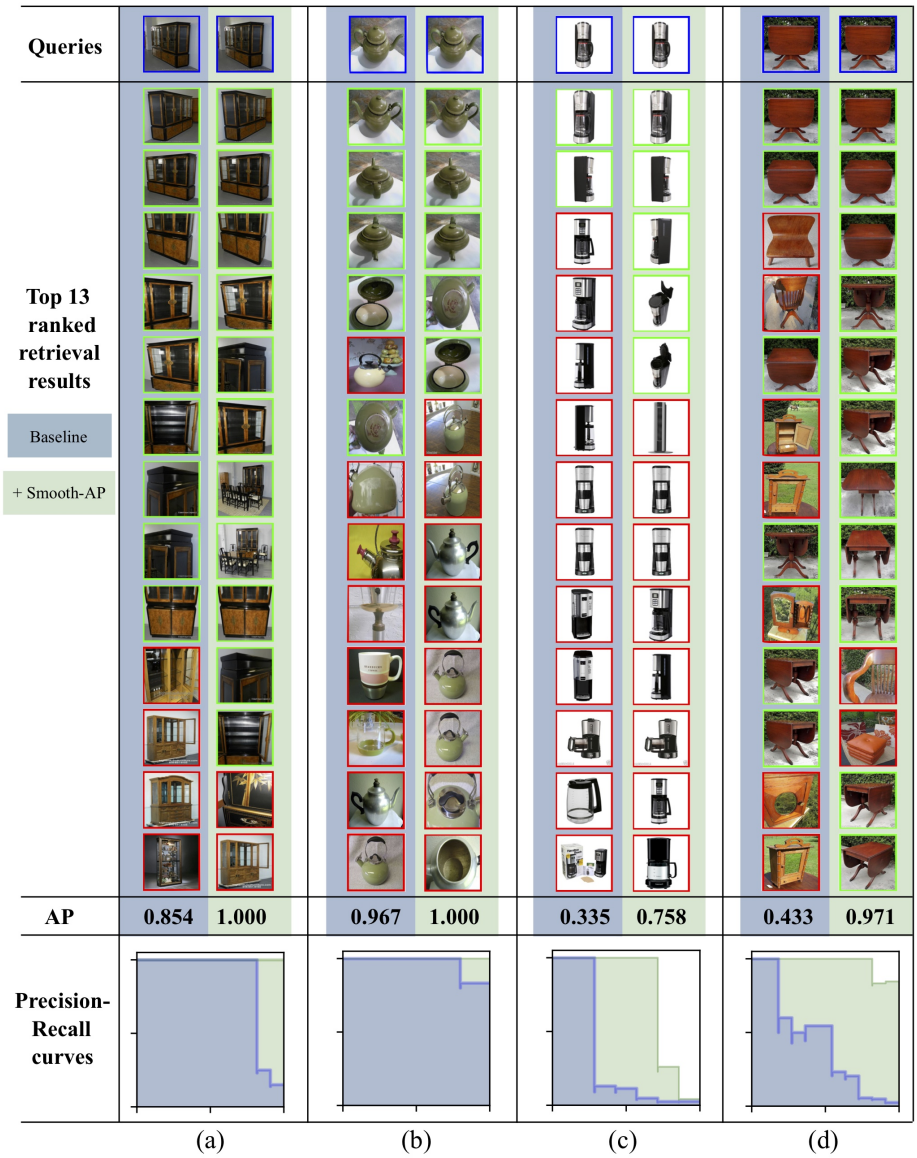


Fig. S5: This Figure shows four separate query instances from the online products dataset and the top ranked instances from the retrieval set when using the baseline model (ImageNet pre-trained weights), and after appending the Smooth-AP loss. It is noted that for this dataset, the size of the positive set ($|P|$) in the retrieval set is very small ($|P| = 11, 5, 7, 11$ for (a),(b),(c),(d) respectively), and so for cases (a) and (b) all positive instances are shown correctly retrieved above all false positives (also indicated by the AP=1.00) for the Smooth-AP model. Particularly *impressive* are the examples (b) and (c), where instances from the positive set which depict a far different pose from the query are retrieved above false positives that are very visually similar to the query.

Fig. S6: This Figure shows three separate query instances from the INaturalist dataset and the top ranked instances from the retrieval set when using the baseline model (ImageNet pre-trained weights), and after appending the Smooth-AP loss. As can be seen by the false positive retrieved instances for the baseline model, this is a highly challenging, fine-grained dataset; yet in all cases shown, appending the Smooth-AP loss leads to large gains in Average Precision. $|P| = 985$, 20, 28 for (a),(b),(c) respectively.

2 Source code

Here, we provide pseudocode for the Smooth-AP loss written in PyTorch style. The simplicity of the method is demonstrated by the short implementation.

Algorithm 1: Pseudocode for Smooth-AP in Pytorch-style.

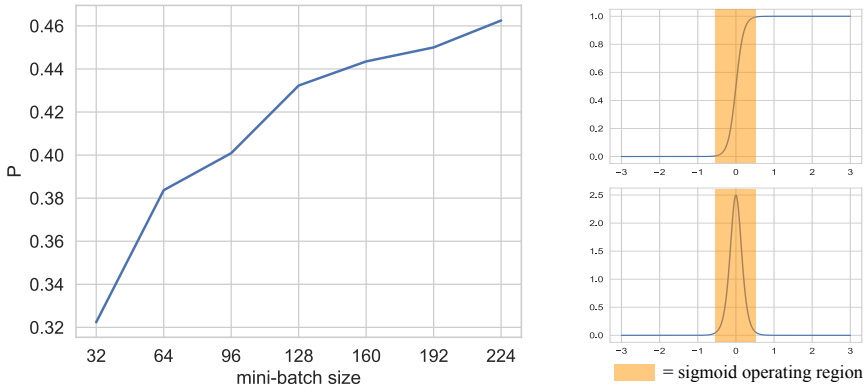
```
# scores: predicted relevance scores (1 x m)
# gt: groundtruth relevance scores (1 x m)

def t_sigmoid(tensor, tau=1.0):
    # tau is the temperature.
    exponent = -tensor / tau
    y = 1.0 / (1.0 + exp(exponent))
    return y

def smooth_ap(scores, gt):
    # repeat the number row-wise.
    s1 = scores.repeat(m, 1) # s1: m x m
    # repeat the number column-wise.
    s2 = s1.transpose # s2: m x m
    # compute difference matrix
    D = s1 - s2
    # approximating heaviside
    D_ = t_sigmoid(D, tau=0.01)
    # ranking of each instance
    R = 1 + sum(D_ * (1-eye(m)), 1)
    # compute positive ranking
    R_pos = gt.T * R
    # compute AP
    AP = (1 / sum(gt)) * sum(R_pos / R)
    return 1-AP
```

3 Details on the effects of increasing the mini-batch size on Smooth-AP

In this Section, we provide a further quantitative validation of the claims made in Section 6.4 (in the main manuscript) about the effects of mini-batch size on the Smooth-AP loss. We conjecture that a large mini-batch size increases the likelihood of relevance scores in the mini-batch being close to each other, and hence elements of the difference matrix (Equation 4 in main manuscript) falling into the narrow operating region of the sigmoid (we define the the operating region of the sigmoid as the narrow region with non-negligible gradients, see Figure S7b), meaning that non-negligible gradients are fed backwards from Smooth-AP. This conjecture can be verified by increasing the mini-batch size during training and logging the proportion of elements of the difference matrix that fall into the operating region of the sigmoid. For each mini-batch during training, a difference matrix D is constructed of size $(m * m)$ where m is the mini-batch size. The proportion of elements of D that fall into the operating region of the sigmoid

used in Smooth-AP, which we denote as P, can be computed using Equation 1 (we use a value of 0.005 to represent a non-negligible gradient). While keeping all parameters equal except mini-batch size, the average P is computed across all mini-batches in one epoch of training on the Online Products dataset for several different mini-batch sizes, with the results plotted in Figure S7a. As expected, P increases with mini-batch size due to the fact that more instances in a mini-batch means that more instances are close enough together in terms of similarity score to lie within the operating zone of the sigmoid. This in turn leads to more non-negligible gradients being fed backwards to the network weights, and hence a higher evaluation performance, as was shown in the ablation Table 5 in the main manuscript.

$$P = \frac{\sum_{i=0}^{m-1} \sum_{j=0}^{m-1} (|\frac{d\mathcal{G}(D_{ij})}{dD_{ij}}| > 0.005)}{m^2} \tag{1}$$



(a) (b)

Fig. S7: (a): P, the proportion of elements of the difference matrix that fall into the operating region of the sigmoid (shown in (b)), and hence receive non negligible gradients, for several different mini-batch sizes. This explains why Smooth-AP benefits from large mini-batch sizes.

4 Choice of hyper-parameters for the compared-to methods for the INaturalist experiments

The two AP-optimising methods that we compare to for the INaturalist experiments (Table 3) have several hyper-parameters associated with them. For FastAP [5], there is the number of histogram bins, L, and for Blackbox AP [52],

there is the value of λ and the margin. For both methods we choose the hyper-parameters that are recommended in the respective publications for the largest dataset that was experimented on, which would be closest to INaturalist in terms of number of training images. For FastAP the number of histogram bins L is set to 20, and for Blackbox AP, λ is set to 4 and the margin is set to 0.02. We note that the evaluated Recall@K scores might be increased by varying these parameters. For all experiments on the INaturalist dataset, we multiply the learning rate on the last linear layer by a factor of 2.

5 Complexity of the proposed loss

Table S1 shows the time complexities of Smooth-AP, and also the other AP-optimising loss functions that we compare to. We also measure the times of the forward and backward passes for a single training iteration when each of the different loss functions are appended onto a ResNet50 [21] backbone architecture. More specifically, we measure the forwards and backwards pass time for the backbone network, *backbone time*, and the appended loss function, *loss time*. These values for timings are averaged over all iterations in one training epoch for the Online Products dataset. The relevant notation: $\mathcal{M}$ is the number of instances in the retrieval set, which during mini-batch training is equal to the size of the mini-batch (with $p + n = \mathcal{M}$ and p, n the number of positive and negative instances, respectively). L refers to the number of bins used in the Histogram Binning technique [5,62]. Even though the complexity of the proposed Smooth-AP loss is higher, Table S1 shows that this leads to a very small overhead in computational cost on top of the ResNet50 backbone architecture ($< 3ms$ for every iteration compared to previous methods where the backbone takes 705 ms), and hence in practice has a minor impact on the *usability* of the proposed loss function. In all experiments here, all training parameters are equal ($|P| = 4$, and mini-batch size $\mathcal{M}$ of 112).

Method	*complexity*	*backbone time (ms)*	*loss time (ms)*
Blackbox AP [52]	$\mathcal{O}(n \log(n))$	705.0	3.7
FastAP [5]	$\mathcal{O}(\mathcal{M}L)$	705.0	4.2
Smooth-AP	$\mathcal{O}(\mathcal{M}^2)$	705.0	6.6

Table S1: The time complexities of different AP-optimising methods that we compare to, as well as the time taken for the forwards and backwards pass through the backbone for one iteration, *backbone time*, and the time taken for the computation of the loss, *loss time*. The slightly increased time complexity of Smooth-AP leads to a negligible increase in training time.