

Supplementary Material: Potential Field Based Deep Metric Learning

In this supplementary material, we provide additional information which we could not fit in the space available in the main paper. We do so in six sections: Sections 1 & 2 contain proofs for Proposition 1 and Corollary 1 respectively; Section 3 contains additional empirical studies that further validate the effectiveness of our method (Sec. 3.1: using a smaller embedding size, Sec. 3.2: Results when using the MLRC protocol [9]); Section 4 provides implementation details about the hyperparameters used; Section 5 empirically compares the time complexity of our method with other methods; Section 6 provides qualitative retrieval results and a t-sne visualization of the embedding space learned by PFML.

1. Proof of Proposition 1

Proposition 1: Let $Z = \{z_1 \dots z_n\}$ be a set of sample embeddings belonging to a class, then there exists a $0 < \delta < \frac{\min_{i,j} \|z_i - z_j\|_2}{2(1 + \frac{1}{n})}$ and a distance $z_{min,i} \leq \delta$ for each embedding z_i , $i \in \{1 \dots n\}$ such that the attractive potential field Ψ_{att} (Eq. 4 from paper) defined using (Z, δ) when restricted to a radial line from z_i has a minimum at each $z_{min,i}$. The field Ψ_{att}^* (by CPML, the interaction strength increasing potential defined in Eq. 2) defined by Z does not achieve such a minimum at points within δ distance from all z_i .

Proof: We first define the potential field at point \mathbf{z} caused by an individual embedding \mathbf{z}_i , $\psi_{att}(\mathbf{z}, \mathbf{z}_i)$ and $\psi_{att}^*(\mathbf{z}, \mathbf{z}_i)$ are defined as:

$$\psi_{att}(\mathbf{z}, \mathbf{z}_i) := \begin{cases} -\frac{1}{\delta^\alpha} & \text{if } \|\mathbf{z} - \mathbf{z}_i\|_2 < \delta \\ -\frac{1}{\|\mathbf{z} - \mathbf{z}_i\|_2^\alpha} & \text{otherwise.} \end{cases} \quad (1)$$

$$\psi_{att}^*(\mathbf{z}, \mathbf{z}_i) := \begin{cases} \delta^2 & \text{if } \|\mathbf{z} - \mathbf{z}_i\|_2 < \delta \\ \|\mathbf{z} - \mathbf{z}_i\|_2^2 & \text{otherwise.} \end{cases} \quad (2)$$

The potential fields created by all data points are:

$$\Psi_{att}(\mathbf{z}) = \sum_{i=1}^n \psi_{att}(\mathbf{z}, \mathbf{z}_i) \quad (3)$$

$$\Psi_{att}^*(\mathbf{z}) = \sum_{i=1}^n \psi_{att}^*(\mathbf{z}, \mathbf{z}_i) \quad (4)$$

We prove the proposition in two parts, first proving the assertion for $\Psi_{att}(\mathbf{z})$ in Part 1, and then moving on to proving the assertion for $\Psi_{att}^*(\mathbf{z})$ in Part 2.

1.1. Part 1: Proof for $\Psi_{att}(\mathbf{z})$

To prove that $\Psi_{att}(\mathbf{z})$ achieves a minimum in the radial direction at a distance $z_{min,i} \leq \delta$ distance of each embedding z_i , we observe that Ψ_{att} is continuous and bounded within the δ hyper-spheres $S_i = \{\mathbf{z} \mid \|\mathbf{z} - \mathbf{z}_i\| \leq \delta\}$. Each hypersphere is a closed bounded set.

This enables us to apply the Extreme Value Theorem (EVT) [20] to Ψ_{att} on S_i . Using EVT, we note that Ψ_{att} achieves a minimum $\Psi_{att}(\mathbf{z}^*)$ on the set S_i at some distance $z_{min,i}$.

$z_{min,i}$ may either be $< \delta$ (minimum inside the sphere) or $= \delta$ (minimum on the boundary). We analyze both cases separately, proving that the minimum for Ψ_{att} , $z_{min,i}$ on S_i is also a minimum for Ψ_{att} on the embedding space R^D (D = embedding dimension):

Case 1: If minimum \mathbf{z}^* lies inside the sphere S_i , then $z_{min,i}$ is a minimum for Ψ_{att} on R^D too because $S_i \subset R^D$ (D = embedding dimension).

Case 2: If $z_{min,i} = \delta$ (minimum \mathbf{z}^* lies on the border of sphere S_i). The proof for this intuitively relies on the fact that the derivative of the potential field of \mathbf{z}_i ($\psi_{att}(\mathbf{z}, \mathbf{z}_i)$) outside S_i , around \mathbf{z}^* is positive, and the derivative (i.e., interaction strength or force) increases as we choose a smaller δ . Consequently, the derivative of the potential field of \mathbf{z}_i dominates the force applied at \mathbf{z}^* compared to the force applied by potentials of other embeddings (which would have decayed). Hence, with a small enough delta, it is possible to ensure that the potential at points outside S_i is larger than the one on the boundary.

To formally handle this case, we first define $\Psi_{att}|_{R_i}$ as the restriction of Ψ_{att} to the line $R_i = \{\mathbf{z}_i + t \mathbf{z}^* \mid t \in \mathbb{R}\}$. Then by the definition of a local minimum

\mathbf{z}^* is a minimum of $\Psi_{att}|_{R_i}$ iff there exists an $\epsilon > 0$ such that $\Psi_{att}|_{R_i}(\mathbf{z}^*) \leq \Psi_{att}|_{R_i}(\mathbf{z}')$ for all $\mathbf{z}' \in R_i$. We note that for any $\mathbf{z}' \in S_i$, $\Psi_{att}|_{R_i}(\mathbf{z}^*) \leq \Psi_{att}|_{R_i}(\mathbf{z}')$ is trivially true by the definition of \mathbf{z}^* . For $\mathbf{z}' \notin S_i$, we use the Taylor expansion of $\Psi_{att}|_{R_i}$ to find such an ϵ and prove that such a \mathbf{z}^* is also a minimum for $\Psi_{att}|_{R_i}$ in R^D .

Specifically, the Taylor expansion for Ψ_{att} in the radial direction (co-ordinates centered at \mathbf{z}_i) is given as:

$$\Psi_{att}|_{R_i}(\mathbf{z}') = \Psi_{att}|_{R_i}(\mathbf{z}^*) + (\mathbf{z}' - \mathbf{z}^*) \cdot \left(\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} \Big|_{\mathbf{z}^*} \hat{\mathbf{r}} \right) + \text{err}(\mathbf{z}' - \mathbf{z}^*) \quad (5)$$

$$\text{here } \hat{\mathbf{r}} \text{ is the unit radial vector pointing toward } \mathbf{z}^* \text{ centered at } \mathbf{z}_i \quad (6)$$

Expanding the partial derivative, we get:

$$\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} = \sum_{i=1}^n \frac{\partial \psi_{att}(\mathbf{z}', \mathbf{z}_i)|_{R_i}}{\partial \mathbf{z}'} \quad (7)$$

Evaluating it at \mathbf{z}^* , we get:

$$\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'}(\mathbf{z}^*) = \frac{\alpha}{\|\mathbf{z}^* - \mathbf{z}_i\|_2^{\alpha+1}} + \gamma_i$$

$$\begin{aligned} \frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'}(\mathbf{z}^*) &= \frac{\alpha}{\delta^{\alpha+1}} \hat{\mathbf{r}} + \gamma_i \\ \text{where } \gamma_i &= \sum_{j=1, j \neq i}^n \frac{\partial \psi_{att}(\mathbf{z}', \mathbf{z}_j)|_{R_i}}{\partial \mathbf{z}'} \end{aligned}$$

Now $\frac{\alpha}{\delta^{\alpha+1}} > \gamma_i$ for all $\delta < \left(\frac{\alpha}{\gamma_i} \right)^{\frac{1}{\alpha+1}}$. So,

$$(\mathbf{z}' - \mathbf{z}^*) \cdot \left(\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} \Big|_{\mathbf{z}^*} \hat{\mathbf{r}} \right) = \|\mathbf{z}' - \mathbf{z}^*\| \left(\frac{\alpha}{\delta^{\alpha+1}} + \gamma_i \right) > 0 \quad \forall \delta < \left(\frac{\alpha}{\gamma_i} \right)^{\frac{1}{\alpha+1}} \quad (8)$$

Also, as $\epsilon \rightarrow 0$, $\frac{\text{err}(\mathbf{z}' - \mathbf{z}^*)}{(\mathbf{z}' - \mathbf{z}^*) \cdot \left(\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} \Big|_{\mathbf{z}^*} \hat{\mathbf{r}} \right)} \rightarrow 0$ as those are higher order terms.

Simplifying the Taylor expansion (Eq. 5)

$$\begin{aligned} \Psi_{att}|_{R_i}(\mathbf{z}') - \Psi_{att}|_{R_i}(\mathbf{z}^*) &= +(\mathbf{z}' - \mathbf{z}^*) \cdot \left(\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} \Big|_{\mathbf{z}^*} \hat{\mathbf{r}} \right) + \text{err}(\mathbf{z}' - \mathbf{z}^*) \\ &= (\mathbf{z}' - \mathbf{z}^*) \cdot \left(\frac{\partial \Psi_{att}|_{R_i}}{\partial \mathbf{z}'} \Big|_{\mathbf{z}^*} \hat{\mathbf{r}} \right) \text{ as } \epsilon \rightarrow 0 \end{aligned}$$

$$\therefore \Psi_{att}|_{R_i}(\mathbf{z}') - \Psi_{att}|_{R_i}(\mathbf{z}^*) > 0 \text{ (Using Eq. 8)} \quad (9)$$

So for $\mathbf{z}' \notin S_i$, $\Psi_{att}(\mathbf{z}_{min,i}) \leq \Psi_{att}(\mathbf{z}')$ for all $\|\mathbf{z}' - \mathbf{z}^*\| < \epsilon$.

Therefore, by definition \mathbf{z}^* is a minimum of $\Psi_{att}|_{R_i}$ on R^D in Case 2 too.

Hence, Ψ_{att} when restricted to a radial line from z_i has a minimum at a distance $z_{min,i} \leq \delta$ from $\mathbf{z}_i \forall i \in \{1 \dots n\}$ for $\delta < \operatorname{argmin}_i \left(\frac{\alpha}{\gamma_i} \right)^{\frac{1}{\alpha+1}} \cdot z_{min,i}$ lies within δ distance of the embeddings z_i by definition of the hyperspheres S_i .

1.2. Part 2: Proof for $\Psi_{att}^*(\mathbf{z})$

Outside the hyperspheres $S_i = \{\mathbf{z} - \mathbf{z}_i\| \leq \delta, \nabla \Psi_{att}^*(\mathbf{z})$ ($\mathbf{z} \notin S_i$) is given by:

$$\begin{aligned} \nabla \Psi_{att}^*(\mathbf{z}) &= \sum_{i=1}^n \nabla \psi_{att}^*(\mathbf{z} - \mathbf{z}_i) \\ \nabla \Psi_{att}^*(\mathbf{z}) &= \sum_{i=1}^n 2(\mathbf{z} - \mathbf{z}_i) = 2n\mathbf{r} - \sum_{i=1}^n 2\mathbf{z}_i = 2n \left(\mathbf{z} - \sum_{i=1}^n \frac{\mathbf{z}_i}{n} \right) \end{aligned} \quad (10)$$

It achieves a single minimum at $\mathbf{z}_{globalmin} = \frac{\sum_{i=1}^n \mathbf{z}_i}{n}$.

Now define differentiable extensions of the potentials-

$$\psi_{att}^{**}(\mathbf{z} - \mathbf{r}_0) = \|\mathbf{z} - \mathbf{r}_0\|_2^2 \quad (11)$$

To calculate $\nabla \Psi_{att}^*(\mathbf{z})$ inside the hyperspheres ($\mathbf{r} \in S_i$), we note that no two hyper spheres S_i intersect with each other as $\delta < 0.5 \min_{i,j} \|z_i - z_j\|_2$. Observing that for $\mathbf{r} \in S_i$

$$\nabla \Psi_{att}^*(\mathbf{z}) = \sum_{j=1, j \neq i}^n \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_j)$$

Using triangle inequality for $\mathbf{z} \in S_i$:

$$\left\| \sum_{j=1}^n \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_j) \right\| - \left\| \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_i) \right\| \leq \left\| \sum_{j=1, j \neq i}^n \nabla \psi_{att}^*(\mathbf{z} - \mathbf{z}_j) \right\| = \left\| \nabla \Psi_{att}^*(\mathbf{z}) \right\|$$

$$\text{Using } \left\| \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_i) \right\| \leq 2\delta \text{ (from Eq. 11)}$$

$$\left\| \nabla \Psi_{att}^*(\mathbf{z}) \right\| \geq \left\| \sum_{j=1}^n \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_j) \right\| - 2\delta$$

We know that the RHS term > 0 because $\left\| \sum_{j=1}^n \nabla \psi_{att}^{**}(\mathbf{z} - \mathbf{z}_j) \right\| > 2\delta$ for

$$\left\| \mathbf{z} - \sum_{i=1}^n \frac{\mathbf{z}_i}{n} \right\| > \frac{\delta}{n} \quad (12)$$

using Equation 10. For all such \mathbf{z} :

$$\left\| \nabla \Psi_{att}^*(\mathbf{z}) \right\| > 0 \quad (13)$$

At most one sphere S_i has an $\mathbf{r}_i \in S_i$ not satisfying Equation 12. We prove this by contradiction. Assume that another sphere S_j has $\mathbf{r}_j \in S_j$ not satisfying Equation 12. Now the distance between \mathbf{r}_i and \mathbf{r}_j satisfies:

$$\|\mathbf{r}_i - \mathbf{r}_j\| \geq d_{min} - 2\delta \quad (14)$$

here $d_{min} = \min_{i,j} \|z_i - z_j\|_2$ Using the fact that $\delta < \frac{d_{min}}{2(1 + \frac{1}{n})}$ by definition and substituting for d_{min} in equation 14 we get:

$$\begin{aligned} \|\mathbf{r}_i - \mathbf{r}_j\| &> 2\delta \left(1 + \frac{1}{n} \right) - 2\delta \\ \|\mathbf{r}_i - \mathbf{r}_j\| &> \frac{2\delta}{n} \end{aligned}$$

Both \mathbf{r}_i and \mathbf{r}_j cannot satisfy equation 12 as all points satisfying it lie within a sphere of radius $\frac{\delta}{n}$, and distance between $\mathbf{r}_i, \mathbf{r}_j$ is more than the maximum distance between points in a sphere, that is $2\frac{\delta}{n}$. Hence, no such $\mathbf{r}_j \in S_j$ can exist.

Hence for any other hypersphere $S_j, i \neq j$, for all \mathbf{z} we have $\|\nabla \Psi_{att}^*(\mathbf{z})\| > 0$, and hence no minimum exists within them. Hence, proved.

2. Proof of Corollary 1

Corollary 1: Let $Z = \{\mathbf{z}_1 \dots \mathbf{z}_n\}$ be a set of sample embeddings belonging to a class exerting an attraction field on a set of proxies $P = \{\mathbf{p}_1 \dots \mathbf{p}_m\}$. Consider the equilibrium distribution P_{eq} of proxies minimizing the potential energy. If the potential field is defined by Ψ_{att} , then the Wasserstein distance W_2 between P_{eq} and the subset of data they represent is lower than when the potential field is defined by Ψ_{att}^* .

Proof: Let the potential fields Ψ_{att} and Ψ_{att}^* be given by Equations 3 and 4 respectively.

$$\psi_{total}(\mathbf{r}) = \sum_{i=1}^n \psi(\mathbf{r} - \mathbf{z}_i) \quad (15)$$

where ψ is defined using Eq 4 from the paper(att and class subscript j omitted for clarity). The potential energies of the proxy distribution P_{eq} in these potential fields, \mathcal{U}_{proxy} and \mathcal{U}_{proxy}^* respectively are given by :

$$\begin{aligned} \mathcal{U}_{proxy} &= \sum_{\mathbf{p}_i \in P_{eq}} \Psi_{att}(\mathbf{p}_i) \\ \mathcal{U}_{proxy}^* &= \sum_{\mathbf{p}_i \in P_{eq}} \Psi_{att}^*(\mathbf{p}_i) \end{aligned}$$

At equilibrium, each proxy migrates to the nearest minimum in the field. The subset of data the proxies represent are given by the subset of m data points $Z_{subset} = \mathbf{z}_{f(k)}, k \in \{1 \dots m\}; Z_{subset} \subset Z$ which is the closest in Wasserstein distance W_2 from the proxies.

Case 1: Let the field be defined by Ψ_{att} . **Assumptions:** As the distances of proxies $\mathbf{p}_k, k \in 1 \dots m$ from embeddings \mathbf{z}_i are minimized, we assume that they migrate to a distance of $d_{min, f(k)}$ from the nearest embedding in the potential field denoted by $\mathbf{z}_{min, f(k)}$ which is located within δ distance of data point $\mathbf{z}_{f(k)}$ (using proposition 1). Assuming that the proxies are initialized using a normal distribution (commonly used) and $m \ll n$, which is typically true, we ignore the probability of more than one proxy going to the same minimum $\mathbf{z}_{min, f(k)}$ (so $f(k)$ is one-one). Therefore:

$$W_2(P_{eq}, Z_{subset}) = \inf_{\pi} \left(\frac{1}{m} \sum_{k=1}^m \|\mathbf{p}_k - \mathbf{z}_{\pi(k)}\|_2 \right)$$

here the infimum is over all permutations π of k elements

$$W_2(P_{eq}, Z_{subset}) = \left(\frac{1}{m} \sum_{k=1}^m \|\mathbf{z}_{min, f(k)} - \mathbf{z}_{f(k)}\|_2 \right)$$

Using proposition 1

$$W_2(P_{eq}, Z_{subset}) \leq \left(\frac{1}{m} \times (m\delta) \right)$$

$$W_2(P_{eq}, Z_{subset}) \leq \delta \quad (16)$$

Hence when the field is defined by Ψ_{att} we have $W_2(P_{eq}, Z_{subset}) \leq \delta$.

Case 2: Let the field be defined by Ψ_{att}^* . The proxies $\mathbf{p}_k, k \in 1 \dots m$ migrate to the nearest minimum in the potential field denoted by $\mathbf{z}_{min, f(k)}$. Using Proposition 1 proved before, we know that all minima satisfy $\min_j \|\mathbf{z}_{min, f(k)} - \mathbf{z}_j\| > \delta$ for all j except at most one $j = j'$ for which let $k = k'$.

First, we prove the corollary for the case if there exists such a $j = j'$. Let $\mathbf{z}_{g(k)}, k = \{1 \dots m\}$ represent the ordered subset of data embeddings that minimize the W_2 distance metric with the proxies $\mathbf{p}_k \in P_{eq}$. From proposition 1, we know

that:

$$\|\mathbf{z}_{min,f(k)} - \mathbf{z}_{j'}\|_2 \leq \delta \quad (17)$$

we also have:

$$\|\mathbf{z}_{g(k)} - \mathbf{z}_{j'}\|_2 \geq \min_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|_2 \geq 2\delta \left(1 + \frac{1}{m}\right) \quad (18)$$

Using the triangle inequality on the above 2 inequalities and substituting, we get:

$$\begin{aligned} \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{g(k)}\|_2 &\geq \|\mathbf{z}_{g(k)} - \mathbf{z}_{j'}\|_2 - \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{j'}\|_2 \\ \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{g(k)}\|_2 &\geq 2\delta \left(1 + \frac{1}{m}\right) - \delta \\ \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{g(k)}\|_2 &\geq \delta + \frac{2\delta}{m} \end{aligned} \quad (19)$$

The W_2^* distance is given by:

$$\begin{aligned} W_2^*(P_{eq}, Z_{subset}) &= \inf_{\pi} \left(\frac{1}{m} \sum_{k=1}^m \|\mathbf{p}_k - \mathbf{z}_{\pi(k)}\|_2 \right) \\ &\text{here the infimum is over all permutations } \pi \text{ of } k \text{ elements} \\ W_2^*(P_{eq}, Z_{subset}) &= \left(\frac{1}{m} \sum_{k=1}^m \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{g(k)}\|_2 \right) \\ W_2^*(P_{eq}, Z_{subset}) &= \left(\frac{1}{m} \sum_{k=1, k \neq k'}^m \|\mathbf{z}_{min,f(k)} - \mathbf{z}_{g(k)}\|_2 \right) + \frac{1}{m} \|\mathbf{z}_{min,k'} - \mathbf{z}_{j'}\|_2 \\ &\text{Using equation 19 and proposition 1} \\ W_2^*(P_{eq}, Z_{subset}) &\geq \frac{1}{m} \times \left((m-1)\delta + \frac{2(m-1)\delta}{m} \right) \\ W_2^*(P_{eq}, Z_{subset}) &> \delta \end{aligned} \quad (20)$$

Hence the W_2 distance between P_{eq} and Z_{subset} when the field is given by Ψ_{att} is $W_2(P_{eq}, Z_{subset}) \leq \delta$ while their distance $W_2^*(P_{eq}, Z_{subset})$ when the field is $> \delta$ as proved above. So $W_2^*(P_{eq}, Z_{subset}) > W_2(P_{eq}, Z_{subset})$. Hence, proved.

3. Additional Experimental Results

In this section of the supplement, we provide additional experiments that we could not fit in the space available in the main paper. These empirical studies further validate the effectiveness of our method.

3.1. Performance using Small Embedding size

Context: In Section 4.2, we presented results on image retrieval using an embedding space of dimension 512. However in certain settings, learning embeddings in a lower dimension space might be more useful, such as in settings where limited storage is available for storing image embeddings. While this lowers the image retrieval performance, as is to be expected, it allows for a trade-off between available memory/compute resources and the accuracy of retrieval. Hence, we compare the performance of our method in learning a lower dimensional embedding space with recent state-of-the-art baselines.

Experiment: We train a ResNet-50 network with its embedding size set to 64 (a commonly used setting) on the Cars-196 [5], CUB-200-2011 [16] and SOP [14] datasets.

Results: As seen in Tables 1 and 2, we observe that our method is able to outperform all other methods at this task; specifically, it outperforms strong Proxy-based baselines like ProxyAnchor[4] and ProxyGML[25] by more than 4.5 %, 2.2 % and 2.6% in the Recall@1 (R@1) metric on the Cars-196, CUB-200 and SOP datasets, respectively. It also outperforms the current state-of-the-art, the graph-based HIST[6] by 3% and 1.4% in terms of R@1 on the Cars-196 and CUB-200 datasets. This shows the strength of our method in learning a low-dimensional semantic representation space.

Benchmarks →	CUB-200-2011				Cars-196			
Methods ↓	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
MultiSimilarity [18]	57.4	69.8	80.0	87.8	77.3	85.3	90.5	94.2
SemiHard [12]	42.6	55.0	66.4	-	51.5	63.8	73.5	-
LiftedStruct [14]	43.6	56.6	68.6	79.6	53.0	65.7	76.0	84.3
N-Pair [13]	51.0	63.3	74.3	83.2	71.1	79.7	86.5	91.6
ProxyNCA [8]	49.2	61.9	67.9	72.4	73.2	82.4	86.4	88.7
SoftTriple [10]	60.1	71.9	81.2	88.5	78.6	86.6	91.8	95.4
Clustering [15]	48.2	61.4	71.8	81.9	58.1	70.6	80.3	87.8
ProxyAnchor [4]	61.7	73.0	81.8	88.8	78.8	87.0	92.2	95.5
ProxyGML [25]	59.4	70.1	80.4	-	78.9	87.5	91.9	-
HIST [6]	62.5	73.6	83.0	89.6	80.4	87.6	92.4	95.4
Ours	63.9 ± 0.3	74.7 ± 0.2	83.5 ± 0.1	90.1 ± 0.1	83.4 ± 0.2	89.9 ± 0.2	94.2 ± 0.1	97.1 ± 0.1

Table 1. Comparison of the Recall@ K (%) achieved by our method on the CUB-200-2011 and Cars-196 datasets with state-of-the-art baselines when using an embedding size of 64, showing that it outperforms all other methods. We compute recall for our method as an average over 5 runs as is done by other baselines which report this number.

Benchmarks →	SOP			
Methods ↓	R@1	R@10	R@100	R@1000
MultiSimilarity [18]	74.1	87.8	94.7	98.2
LiftedStruct [14]	62.5	80.8	91.9	-
N-Pair [13]	67.7	83.8	93.0	97.8
ProxyNCA [8]	73.7	-	-	-
SoftTriple [10]	76.3	89.1	95.3	-
Clustering [15]	67.0	83.7	93.2	-
ProxyAnchor [4]	76.5	89.0	95.1	98.2
ProxyGML [25]	76.2	89.4	95.4	-
HIST [6]	78.9	90.5	95.8	98.5
Ours	79.5 ± 0.2	90.8 ± 0.1	96.3 ± 0.1	98.6 ± 0.1

Table 2. Comparison of the Recall@ K (%) achieved by our method on the SOP dataset with state-of-the-art baselines when using an embedding size of 64, showing that it outperforms all other methods. We compute recall for our method as an average over 5 runs as is done by other baselines which report this number.

3.2. Evaluation under MLRC Protocol

Context: In Section 4.2, we compared the image retrieval performance of our method with other recent techniques using standard evaluation settings (backbone, embedding dimensions, image sizes) used in [4, 6, 7, 24]. Recently, some studies [9, 11] have pointed to flaws in these settings, including lack of a standardized backbone architecture, weakness of the metrics used, and lack of a standardized validation subset. Though we address some of these flaws by comparing against methods using the same experimental settings as described in Section 4.1, in this section we additionally evaluate our method under the constrained protocol proposed in [9]. The constrained protocol proposes using fixed optimization settings with no learning rate scheduling to train an Inception with BatchNorm architecture. It also introduces new, more informative metrics (the R-Precision and Mean Average Precision@R). Further details of the constrained protocol can be found in [9].

Results: We evaluate the performance of our method using models trained under the constrained protocol on the Cars-196 [5] and CUB-200-2011 [16] datasets. As seen in Tables 3 4, our method significantly outperforms all previous methods on all metrics. We significantly outperform the previous best pair-based method, the MultiSimilarity loss [18] by 7.4% and 4.8 % in terms of P@1 (128-dim embeddings) on the Cars-196 and CUB-200 datasets respectively. We also outperform the current state-of-the-art method [6], HIST by 1.9 % and 1.1 % in terms of P@1 (128-dim) on the Cars-196 and CUB-200 datasets respectively. We note that these gains are higher than the improvements made by the current state-of-the-art HIST (1.5%

Embedding type →	Concatenated (512-dim)			Separated (128-dim)		
Methods ↓	P@ 1	RP	MAP@R	P@ 1	RP	MAP@R
Contrastive [2]	81.8 ± 0.4	35.1 ± 0.5	24.9 ± 0.5	69.8 ± 0.4	27.8 ± 0.3	17.2 ± 0.4
Triplet [19]	79.1 ± 0.4	33.7 ± 0.5	23.0 ± 0.5	65.7 ± 0.6	26.7 ± 0.4	15.8 ± 0.4
N-Pair [13]	81.0 ± 0.5	35.0 ± 0.4	24.4 ± 0.4	68.2 ± 0.4	27.7 ± 0.2	16.8 ± 0.2
ProxyNCA [8]	83.6 ± 0.3	35.6 ± 0.3	25.4 ± 0.3	73.5 ± 0.2	28.9 ± 0.2	18.3 ± 0.2
Margin [21]	81.2 ± 0.5	34.8 ± 0.3	24.2 ± 0.3	68.2 ± 0.4	27.2 ± 0.2	16.4 ± 0.2
Margin/class [21]	80.0 ± 0.6	33.8 ± 0.5	23.1 ± 0.6	67.5 ± 0.6	26.7 ± 0.4	15.9 ± 0.4
N. Softmax [23]	83.2 ± 0.3	36.2 ± 0.3	26.0 ± 0.3	72.6 ± 0.2	29.3 ± 0.2	18.7 ± 0.2
CosFace [17]	85.5 ± 0.2	37.3 ± 0.3	27.6 ± 0.3	74.7 ± 0.2	29.0 ± 0.1	18.8 ± 0.1
ArcFace [3]	85.4 ± 0.3	37.0 ± 0.3	27.2 ± 0.3	72.1 ± 0.4	27.3 ± 0.2	17.1 ± 0.2
FastAP [1]	78.5 ± 0.5	33.6 ± 0.5	23.1 ± 0.6	65.1 ± 0.4	26.6 ± 0.4	15.9 ± 0.3
SNR [22]	82.0 ± 0.5	35.2 ± 0.4	25.0 ± 0.5	69.7 ± 0.5	27.5 ± 0.3	17.1 ± 0.3
MultiSimilarity [18]	85.1 ± 0.3	38.1 ± 0.2	28.1 ± 0.2	73.8 ± 0.2	29.9 ± 0.2	19.3 ± 0.2
MS+Miner [18]	83.7 ± 0.3	37.1 ± 0.3	27.0 ± 0.4	71.8 ± 0.2	29.4 ± 0.2	18.9 ± 0.2
SoftTriple [10]	84.5 ± 0.3	37.0 ± 0.2	27.1 ± 0.2	73.7 ± 0.2	29.3 ± 0.2	18.7 ± 0.1
ProxyAnchor [4]	83.3 ± 0.4	35.7 ± 0.3	25.7 ± 0.4	73.7 ± 0.4	29.4 ± 0.3	18.9 ± 0.2
HIST [6]	87.7 ± 0.2	39.9 ± 0.2	30.5 ± 0.2	79.3 ± 0.2	32.8 ± 0.2	22.3 ± 0.2
Ours	88.4 ± 0.2	40.1 ± 0.2	31.0 ± 0.3	81.2 ± 0.2	33.6 ± 0.3	22.9 ± 0.1

Table 3. Comparison of the Precision@1, R-Precision (RP) and the Mean Average Precision @ R (MAP@R) as defined in [9] achieved by our method on the Cars-196 dataset with state-of-the-art baselines under MLRC[9] settings.

Embedding type →	Concatenated (512-dim)			Separated (128-dim)		
Methods ↓	P@ 1	RP	MAP@R	P@ 1	RP	MAP@R
Contrastive [2]	68.1 ± 0.3	37.2 ± 0.3	26.5 ± 0.3	59.7 ± 0.4	32.0 ± 0.3	21.2 ± 0.3
Triplet [19]	64.2 ± 0.3	34.6 ± 0.2	23.7 ± 0.2	55.8 ± 0.3	29.6 ± 0.2	18.8 ± 0.2
N-Pair [13]	66.6 ± 0.3	36.0 ± 0.2	25.1 ± 0.2	58.1 ± 0.2	30.8 ± 0.2	19.9 ± 0.2
ProxyNCA [8]	65.7 ± 0.4	35.1 ± 0.3	24.2 ± 0.3	57.9 ± 0.3	30.2 ± 0.2	19.3 ± 0.2
Margin [21]	63.6 ± 0.5	33.9 ± 0.3	23.1 ± 0.3	54.8 ± 0.3	28.9 ± 0.2	18.1 ± 0.2
Margin/class [21]	64.4 ± 0.2	34.6 ± 0.2	23.7 ± 0.2	55.6 ± 0.2	29.3 ± 0.2	18.5 ± 0.1
N. Softmax [23]	65.6 ± 0.3	36.0 ± 0.2	25.3 ± 0.1	58.8 ± 0.2	31.8 ± 0.1	21.0 ± 0.1
CosFace [17]	67.3 ± 0.3	37.5 ± 0.2	26.7 ± 0.2	59.6 ± 0.4	32.0 ± 0.2	21.2 ± 0.2
ArcFace [3]	67.5 ± 0.3	37.3 ± 0.2	26.5 ± 0.2	60.2 ± 0.3	32.4 ± 0.2	21.5 ± 0.2
FastAP [1]	63.2 ± 0.3	34.2 ± 0.2	23.5 ± 0.2	55.6 ± 0.3	29.7 ± 0.2	19.1 ± 0.2
SNR [22]	66.4 ± 0.6	36.6 ± 0.3	25.8 ± 0.4	58.1 ± 0.4	31.2 ± 0.3	20.4 ± 0.3
MultiSimilarity [18]	65.0 ± 0.3	35.4 ± 0.1	24.7 ± 0.1	57.6 ± 0.2	30.8 ± 0.1	20.2 ± 0.1
MS+Miner [18]	67.7 ± 0.2	37.3 ± 0.2	26.5 ± 0.2	59.4 ± 0.3	31.9 ± 0.1	21.0 ± 0.1
SoftTriple [10]	67.3 ± 0.4	37.3 ± 0.2	26.5 ± 0.2	59.9 ± 0.3	32.1 ± 0.1	21.3 ± 0.1
ProxyAnchor	65.2 ± 0.2	36.0 ± 0.2	25.3 ± 0.1	56.6 ± 0.1	30.5 ± 0.1	19.8 ± 0.2
HIST [6]	69.6 ± 0.3	38.8 ± 0.1	28.2 ± 0.1	61.3 ± 0.2	33.1 ± 0.2	22.3 ± 0.1
Ours	70.1 ± 0.2	39.9 ± 0.1	29.4 ± 0.3	62.4 ± 0.1	33.8 ± 0.2	23.1 ± 0.3

Table 4. Comparison of the Precision@1, R-Precision (RP) and the Mean Average Precision @ R (MAP@R) as defined in [9] achieved by our method on the CUB-200-2011 dataset with state-of-the-art baselines under MLRC[9] settings.

and 1.1 %) over previous methods on these benchmarks. Our method also outperforms all previous methods in terms of the R-Precision and Mean Average Precision @ R metric, demonstrating the quality of the semantic metric learned by it.

Dataset →	CUB-200-2011		Cars-196		SOP	
Arch/Dim	R50/512	IBN/512	R50/512	IBN/512	R50/512	IBN/512
optimizer	Adam	Adam	Adam	Adam	Adam	Adam
learning rate	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$
weight decay	$5e^{-4}$	$5e^{-4}$	$1e^{-4}$	$1e^{-4}$	$5e^{-4}$	$5e^{-4}$
Batch Size	100	180	100	180	100	180
BN freeze	Yes	Yes	Yes	Yes	No	No
Warm-up	1	1	1	1	0	0
lr for proxies	0.01	0.01	0.01	0.01	0.01	0.01

Table 5. Hyperparameter details for Potential Field based DML for experiments described in Section 4 of the main paper.

4. Hyperparameters

For easy reproducibility, Table 5 presents further details about the hyperparameters used in our experiments on all 3 datasets described in Section 4.

5. Computational Complexity vs Number of proxies

Methods ↓	Avg. time per epoch (seconds)
ProxyNCA [8]	14.2 ±0.1
Proxy Anchor [4]	14.2 ±0.1
Potential Field (Ours) M=5	14.3±0.1
Potential Field (Ours) M=20	14.5±0.1
Potential Field (Ours) M=30	14.8±0.1

Table 6. Average time (in seconds) required to run an epoch of training on the Cars-196 dataset with a ResNet50 backbone using various methods. These results demonstrate that the time complexity of our method is similar to previous proxy-based methods. Also, note that an increase in M does not significantly alter the time complexity of our method once the number of parameters specifying proxies here (as is in most cases) is much lower than the total number of parameters (in the neural network) being trained. The standard evaluation settings of backbone, embedding dimensions, image sizes as given in Sec. 4.1 of the main paper were employed for all methods. Times were measured on a machine equipped with a single A4000 GPU over 20 epochs.

6. Visual Results

We present qualitative results for image retrieval by our method to evaluate the semantic similarity metric learned by it. Figure 1 displays 2 examples of query images from each of the 3 datasets, followed by 4 nearest images retrieved by our method, arranged in increasing order of distance. It can be seen that despite the large intra-class variation (pose, color) in the datasets, our method is able to effectively retrieve similar images.

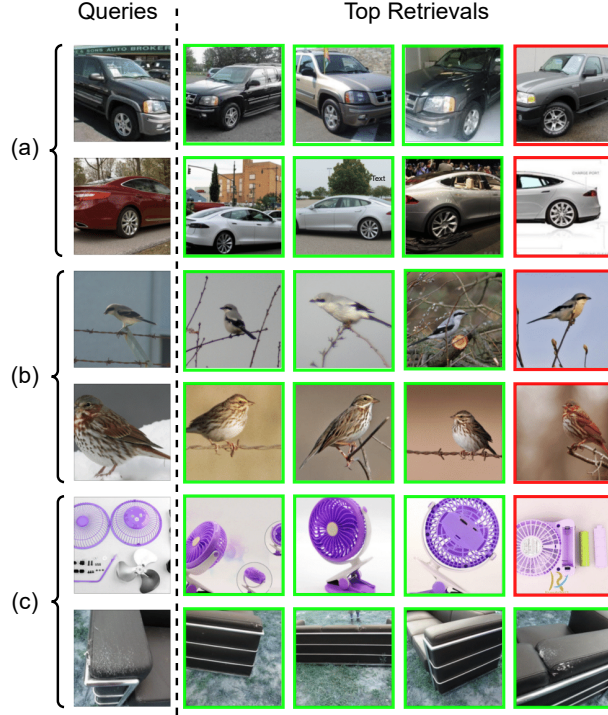


Figure 1. Example image retrieved by our method for query images from (a) Cars-196 (b) CUB-200-2011 and (c) SOP test datasets, in increasing order of distance from the query. Correct retrievals have a green border, while incorrect ones have a red one.

Figure 2 displays a t-sne visualization of the embedding space learnt by our method on the CUB-200-2011 dataset. it can be seen that images closer together share more semantic characteristics than those that are far apart.

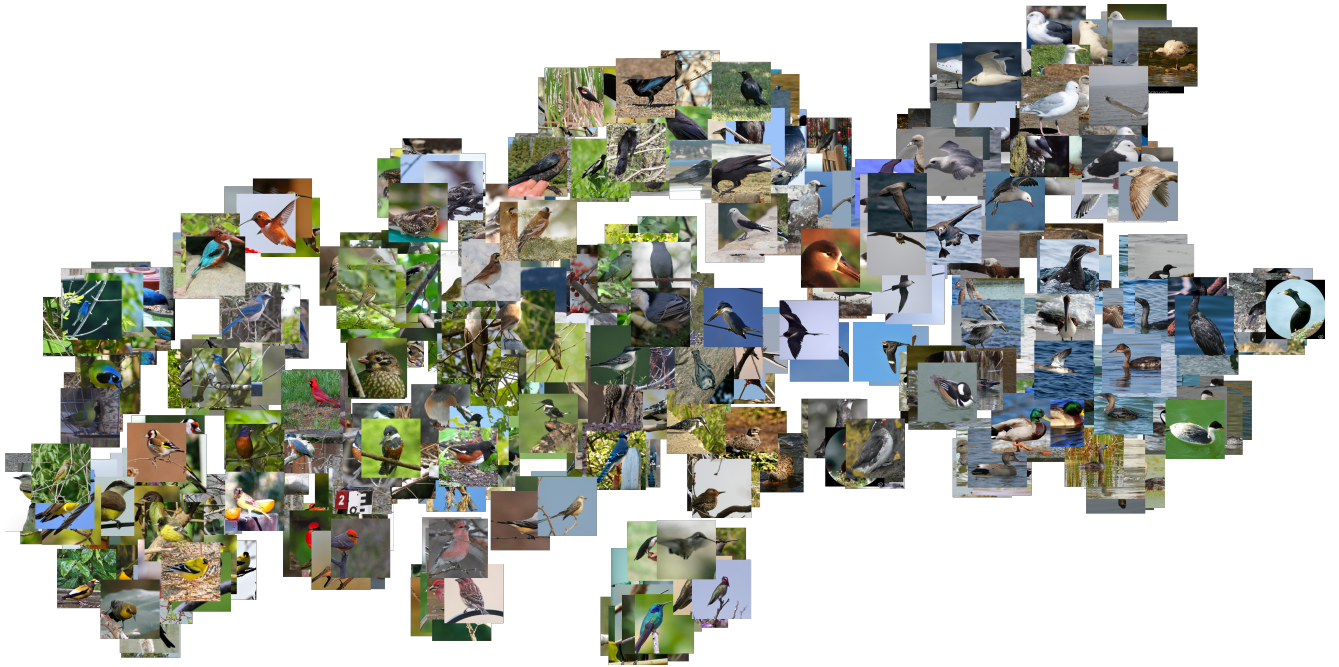


Figure 2. A t-sne visualization of a semantic representation space learnt by our method on the CUB-200 dataset

References

- [1] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1861–1870, 2019. 7
- [2] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, pages 539–546. IEEE, 2005. 7
- [3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 7
- [4] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 7, 8
- [5] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 5, 6
- [6] Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. Hypergraph-induced semantic tuple loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 212–222, 2022. 6, 7
- [7] Lizhao Liu, Shangxin Huang, Zhuangwei Zhuang, Ran Yang, Mingui Tan, and Yaowei Wang. Das: Densely-anchored sampling for deep metric learning. In *European Conference on Computer Vision*, pages 399–417. Springer, 2022. 6
- [8] Yair Movshovitz-Atias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE international conference on computer vision*, pages 360–368, 2017. 6, 7, 8
- [9] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020. 1, 6, 7
- [10] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6, 7
- [11] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 6
- [12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6
- [13] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 6, 7
- [14] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6
- [15] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2206–2214, 2017. 6
- [16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5, 6
- [17] H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 7
- [18] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7
- [19] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*. MIT Press, 2005. 7
- [20] Wikipedia contributors. Extreme value theorem, 2024. [Online; accessed October-2024]. 1
- [21] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017. 7
- [22] Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4815–4824. Computer Vision Foundation / IEEE, 2019. 7
- [23] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *British Machine Vision Conference*, 2018. 7
- [24] Wenzhao Zheng, Borui Zhang, Jiwen Lu, and Jie Zhou. Deep relational metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12074, 2021. 6
- [25] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. Fewer is more: A deep graph metric learning perspective using fewer proxies. In *Advances in Neural Information Processing Systems*, pages 17792–17803. Curran Associates, Inc., 2020. 6