Supplementary Material of NFormer: Robust Person Re-identification with Neighbor Transformer

Haochen Wang[1], Jiayi Shen[1], Yongtuo Liu[1], Yan Gao[2], Efstratios Gavves[1]
University of Amsterdam[1], Xiaohongshu Inc[2]
{h.wang3, j.shen, y.liu6}@uva.nl, wanjianyi@xiaohongshu.com, egavves@uva.nl

A. Analysis of Landmark Agent Attention

In this section, we want to prove that with enough selected samples $\mathbf{z}_l$, the approximate affinity matrix $\widetilde{\mathbf{A}}$ will approach original affinity matrix $\mathbf{A}$.

Firstly , we need to prove that the similarity between $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ is a monotonically non-decreasing function in terms of l. Input with representations $\mathbf{z} \in \mathbb{R}^{N\times d}$, where N denotes the number of the samples and d represents the dimension of each sample. The affinity matrix can be formulated as $\mathbf{A} = \mathbf{z}\mathbf{z}^\top \in \mathbb{R}^{N\times N}$. Here we replace the query $\mathbf{q}$ and key $\mathbf{k}$ with $\mathbf{z}$ to simplify the derivation because both query $\mathbf{q}$ and key $\mathbf{k}$ are obtained by $\mathbf{z}$ with arbitrary linear projection.

We define a sampling matrix $\mathbf{L} = \mathrm{diag}(l_1, ..., l_N) \in \mathbb{R}^{N\times N}$, which is a diagonal matrix. If we select the i-th sample from $\mathbf{z}$, l_i is set to be 1; if not, it is set to be 0. We note that the number of selected samples is $l = \sum_{i=1}^{N} l_i$. Then we define a selected data matrix $\mathbf{z}_l = \mathbf{L}\mathbf{z} \in \mathbb{R}^{N\times d}$. When we remove the rows which are not selected, we get $\mathbf{z}_l \in \mathbb{R}^{l\times d}$ and treat it as landmark agents in our proposed method. By multiplying the origin data matrix and the landmark agent matrix, we obtain the projected lower-dimensional data matrix $\widetilde{\mathbf{z}} = \mathbf{z}\mathbf{z}_l^\top \in \mathbb{R}^{N\times l}$. Then the approximate affinity matrix is formulated as $\widetilde{\mathbf{A}} = \widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top \in \mathbb{R}^{N\times N}$.

Proposition 1. *Consider two attention matrices: $\widetilde{\mathbf{A}}^{l_a}$ and $\widetilde{\mathbf{A}}^{l_b}$ include l_a and l_b selected samples, respectively. We define that $\widetilde{\mathbf{A}}^{l_b}$ covers all selected samples in the $\widetilde{\mathbf{A}}^{l_a}$, and $l_b > l_a$. We assume the attention matrix A is positive semi-definite matrix. Then we have*

$$\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_b})) \geq \cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_a})),$$

which denotes that the similarity $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$ is a monotonically non-decreasing function in terms of l.

Proof. Particularly, we derive the formulation of $\widetilde{\mathbf{A}}$ and obtain the relationships between $\widetilde{\mathbf{A}}$ and $\mathbf{A}$ as follows:

$$\begin{aligned} \widetilde{\mathbf{A}} &= \widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top = \mathbf{z}\mathbf{z}_l^\top(\mathbf{z}\mathbf{z}_l^\top)^\top \\ &= \mathbf{z}\mathbf{z}_l^\top\mathbf{z}_l\mathbf{z}^\top = \mathbf{z}\mathbf{z}^\top\mathbf{L}^\top\mathbf{L}\mathbf{z}\mathbf{z}^\top \\ &= \mathbf{A}\mathbf{L}\mathbf{A}. \end{aligned} \tag{1}$$

We apply $\cos(\cdot,\cdot)$ to mearsure the similarity between $\mathbf{A}$ and $\widetilde{\mathbf{A}}$. We first convert both affinity matrices into column vectors $\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\widetilde{\mathbf{A}})$. Thus the similarity can be formulated as:

$$\begin{aligned} \cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}})) &= \frac{\mathrm{vec}(\mathbf{A})^\top \mathrm{vec}(\widetilde{\mathbf{A}})}{\|\mathrm{vec}(\mathbf{A})\|\|\mathrm{vec}(\widetilde{\mathbf{A}})\|} \\ &= \frac{\mathrm{tr}(\mathbf{A}^\top\widetilde{\mathbf{A}})}{\sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})\mathrm{tr}(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}}. \end{aligned} \tag{2}$$

To simplify the similarity formulation, we decompose the affinity matrix $\mathbf{A}$. According to the finite-dimensional spectral theorem, every real symmetric matrix can be diagonalized by a real orthogonal matrix. In our case, the affinity matrix $\mathbf{A}$ is a real symmetric matrix, where $\mathbf{A}^\top = \mathbf{A}$, thus the matrix can be decomposed as:

$$\mathbf{A} = \mathbf{Q}^\top \Lambda \mathbf{Q}, \tag{3}$$

where $\mathbf{Q}$ denotes the orthogonal matrix $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$, and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_{\mathrm{N}})$ is a diagonal matrix of the eigenvalues of $\mathbf{A}$. In general, we assume that $\mathbf{A}$ is positive semidefinite matrix, $\lambda_i > 0, \forall i \in \{1, 2, ..., N\}$.

To further simplify the trace formulation, we introduce a real symmetric matrix $\mathbf{S} = \mathbf{Q}\mathbf{L}\mathbf{Q}^\top$ and $S_{ii} = \mathbf{q}_i\mathbf{L}\mathbf{q}_i^\top$, where $\mathbf{q}_i$ is the i-th eigenvector in $\mathbf{Q}$ of $\mathbf{A}$. Since $\mathbf{Q}$ is orthogonal, we obtain that $\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\mathbf{L}) = l$ and $S_{ii} \in [0, 1]$. Thus, we reformulated each trance in (2) as follows:

$$
\begin{aligned}
\mathrm{tr}(\mathbf{A}^\top\widetilde{\mathbf{A}}) &= \mathrm{tr}(\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{L}\mathbf{Q}^\top\Lambda\mathbf{Q}) \\
&= \mathrm{tr}(\Lambda^3\mathbf{S}) = \sum_i \lambda_i^3 S_{ii}, \\
\mathrm{tr}(\mathbf{A}^\top\mathbf{A}) &= \mathrm{tr}(\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{Q}^\top\Lambda\mathbf{Q}) \\
&= \mathrm{tr}(\Lambda^2) = \sum_i \lambda_i^2, \\
\mathrm{tr}(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) &= \mathrm{tr}(\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{L}\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{Q}^\top\Lambda\mathbf{Q}\mathbf{L}\mathbf{Q}^\top\Lambda\mathbf{Q}) \\
&= \mathrm{tr}(\Lambda^4\mathbf{S}) = \sum_i \lambda_i^4 S_{ii}.
\end{aligned}
\tag{4}
$$

By integrating (4) into (2), we obtain the simplified metric function as:

$$
\begin{aligned}
\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}})) &= \frac{\mathrm{tr}(\mathbf{A}^\top\widetilde{\mathbf{A}})}{\sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})\mathrm{tr}(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}} \\
&= \sqrt{\frac{(\sum_i \lambda_i^3 S_{ii})^2}{(\sum_i \lambda_i^2)(\sum_j \lambda_j^4 S_{jj})}}.
\end{aligned}
\tag{5}
$$

We set two landmark agents matrices: $\mathbf{z}_l^{l_a}$ has l_a selected samples and $\mathbf{z}_l^{l_b}$ has l_b selected samples. We note that $l_a < l_b$ and $\mathbf{z}_l^{l_b}$ includes all selected samples in the $\mathbf{z}_l^{l_a}$. Thus, we have $S_{ii}^{l_a} < S_{ii}^{l_b}$. We compare the similarity as follows:

$$
\begin{aligned}
&\frac{\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_b}))}{\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_a}))} \\
=& \sqrt{\frac{(\sum_i \lambda_i^3 S_{ii}^{l_b})^2(\sum_i \lambda_i^2)(\sum_j \lambda_j^4 S_{jj}^{l_a})}{(\sum_i \lambda_i^3 S_{ii}^{l_a})^2(\sum_i \lambda_i^2)(\sum_j \lambda_j^4 S_{jj}^{l_b})}} \\
=& \sqrt{\frac{\sum_i\sum_g\sum_j \lambda_i^3\lambda_g^3\lambda_j^4 S_{ii}^{l_b} S_{gg}^{l_b} S_{jj}^{l_a}}{\sum_i\sum_g\sum_j \lambda_i^3\lambda_g^3\lambda_j^4 S_{ii}^{l_a} S_{gg}^{l_a} S_{jj}^{l_b}}} \geq 1,
\end{aligned}
\tag{6}
$$

which demonstrates that $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_b})) \geq \cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_a}))$. This denotes that the similarity $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$ is a monotonically non-decreasing function in terms of l. □

Next, we further provide theoretical proof to show that given enough selected samples, $\widetilde{\mathbf{A}}$ will be very similar as $\mathbf{A}$.

Proposition 2. *Let $\widetilde{\mathbf{A}}$ include all training samples, e.g., $l = N$. We assume that each normalized eigenvalue λ_i/N follows the beta distribution $\mathrm{Be}(\alpha, \alpha(\mathrm{N}-1))$, where $\alpha > 0$ is a hyper-parameter depending on the datasets. When the number of training sample is large, we have*

$$
\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}})) \approx \sqrt{\frac{(\alpha+2)}{(\alpha+3)}}.
\tag{7}
$$

Thus, the upper bound of the similarity, $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$, is a monotonically increasing function in terms of α, which is larger than $\sqrt{\frac{2}{3}}$.

Proof. When all samples are selected, we have $l = N$ and $\mathbf{L} = \mathbf{I}$. Thus, we can derive the upper bound of the similarity as follows:

$$
\begin{aligned}
\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}})) &= \frac{\mathrm{tr}(\mathbf{A}^\top\widetilde{\mathbf{A}})}{\sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})\mathrm{tr}(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})}} \\
&= \sqrt{\frac{(\sum_i \lambda_i^3)^2}{(\sum_i \lambda_i^2)(\sum_j \lambda_j^4)}},
\end{aligned}
\tag{8}
$$

where $\mathrm{tr}(\Lambda) = \mathrm{tr}(\mathbf{A}) = \sum_i \lambda_i = N$, since we utilize the normalized feature vector. Here, we assume $u_i = \lambda_i/N$ is a random variable which is sampled from the beta distribution $\mathrm{Be}(\alpha, \beta)$. In our case, the average of all eigenvalues is $\hat{\lambda} = \sum_i \lambda_i/N = 1$. Thus, the average of all random samples should be $\hat{u} = 1/N$. If the number of samples is large enough, we have

$$
\hat{u} = \mathbb{E}(u) = \frac{\alpha}{\alpha+\beta} = \frac{1}{N}.
\tag{9}
$$

And we note that in the experiments, N is always very large. According to the properties of the beta distribution, we have

$$
\mathbb{E}(u^2) = \frac{\alpha+\beta+2}{\alpha+2}\mathbb{E}(u^3), \mathbb{E}(u^4) = \frac{\alpha+3}{\alpha+\beta+3}\mathbb{E}(u^3).
\tag{10}
$$

By integrating (9) and (10) into (8), we can further simplify as follows:

$$
\begin{aligned}
\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}})) &= \sqrt{\frac{(\sum_i (\frac{\lambda_i}{N})^3)^2}{(\sum_i (\frac{\lambda_i}{N})^2)(\sum_j (\frac{\lambda_j}{N})^4)}} \\
&= \sqrt{\frac{(\mathbb{E}(u^3))^2}{(\mathbb{E}(u^2))(\mathbb{E}(u^4))}} \\
&\approx \sqrt{\frac{(\alpha+2)}{(\alpha+3)}} = f(\alpha),
\end{aligned}
\tag{11}
$$

where $f(\alpha)$ is a monotonically increasing function. Since $\alpha > 0$, we obtain that the upper bound of $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$ should be larger than $\sqrt{\frac{2}{3}} \approx 0.8165$. In our experiments, we observe that the similarity achieves 0.9962 when all the samples are selected. In general, the approximate affinity matrix $\widetilde{\mathbf{A}}$ will approach the original affinity matrix $\mathbf{A}$, which shows the rationality of Landmark Agent Attention. □

B. Visualization of Ranking List

We visualize some of the ranking lists on the Market-1501 dataset in figure 1. As shown in the first rows of figure 1 (a) and (b), the persons in red boxes have similar appearance to the persons in query images while do not belong to the same identities, which show that there are some false-positive samples at the top of the ranking list obtained by the baseline model. On the contrary, those false-positive persons with similar appearances are constrained by NFormer. In the meantime, the person in the green box, which has a large lighting change and viewpoint change compared with the query image, has been kept at the top of the ranking list obtained by NFormer. This is because the NFormer could help to maintain the most discriminative feature and enable the ranking process robust to outliers by considering the neighbors information. As shown in figure 1 (c) and (d), the ranking lists of the baseline model contain many false-positive samples because of the blurred query image in (c) and similar distractors in (d). NFormer surpasses most of the false-positive persons at the top of the ranking list and brings the positive persons with low scores to the front of the lists, which shows that NFormer could bring general improvements even on very difficult samples.

C. Stability of Landmark Sampling

We further show the performance stability influenced by the landmark sampling process. Moreover, NFormer has multiple attention modules, increasing the number of sampled landmarks (5 for each attention) and improving stability. We repeated the same experiments 5 times on Market-1501 and Duke-MTMC datasets, the results are shown in Table 1. We observe only 0.02% (in terms of 95% confidence interval) on Market1501 and Duke-MTMC, which means the performance is stable even with the random sampling process. As there are lots of persons with similar appearance in the input, the affinity map is low-rank. That is, the affinity matrix can be properly approximated by a small number of landmarks.

Iterations	#1	#2	#3	#4	#5	Avg.
Market1501	91.09	91.12	91.10	91.14	91.11	91.11 ± 0.02
Duke-MTMC	83.52	83.57	83.53	83.56	83.52	83.54 ± 0.02

Table 1. mAP on Market-1501 and Duke-MTMC by 5 repeated experiments.

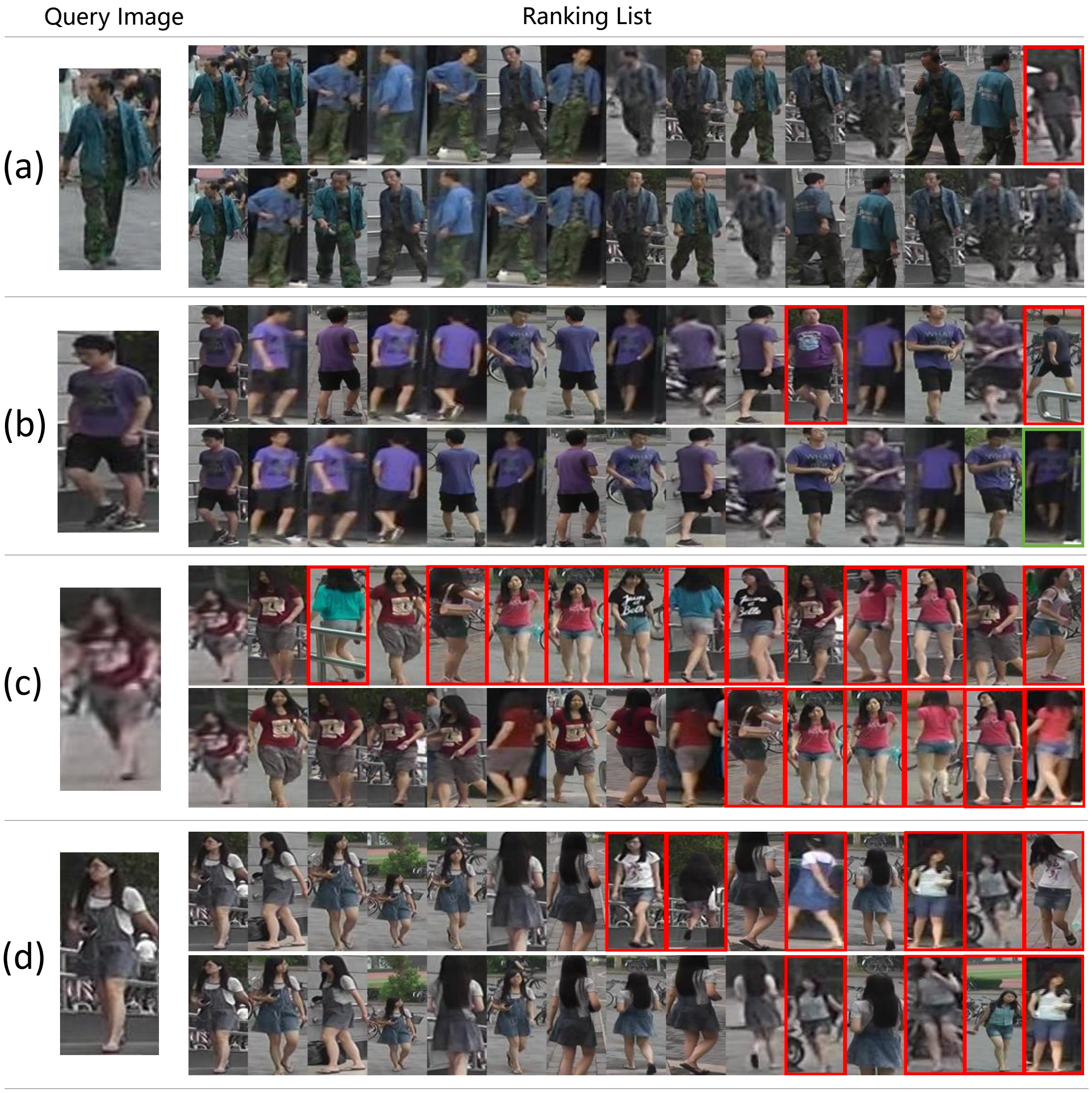

Figure 1. Visualization of ranking lists on Market-1501 datasets. The first row of each sub-figure shows the ranking list obtained by the baseline model. The second row of each sub-figure shows the ranking list obtained by NFormer. The persons who are different from query persons are marked by red boxes.