# Supplementary Material for FlashMix: Fast Map-Free LiDAR Localization via Feature Mixing and Contrastive-Constrained Accelerated Training

Raktim Gautam Goswami[1], Naman Patel[1], Prashanth Krishnamurthy[1], Farshad Khorrami[1] *

## 1. Additional Results

### 1.1. Translation and Rotation Errors

In Section 4.2 of the manuscript, we evaluated FlashMix against the leading LiDAR pose regression methods of HypLiLoc [3], NIDALoc [6], and PosePN++ [7]. Now we show results from additional methods like PosePN, PoseSOE, PoseMinkLoc [7], and PointLoc [4] for the Oxford-Radar and vReLoc datasets in Tables A and B, respectively. Results from retrieval-based methods such as PointNetVLAD [2] and DCP [5] are also presented for the Oxford-Radar dataset to provide a broader performance context. FlashMix demonstrates the lowest translation errors on the Oxford-Radar dataset and exhibits competitive performance on the vReLoc dataset, all while requiring significantly less training time.

### 1.2. Contrastive Regularization

In the manuscript, we demonstrated how integrating contrastive regularization enhances FlashMix's efficacy. Specifically, we assessed FlashMix's performance with the inclusion of the contrastive regularization losses of SigLIP [9], NTXent [1], and Barlow Twins [8], alongside the metric-based Triplet Loss. Here, we provide more details into these losses.

SigLIP is a contrastive loss defined as:

$$\mathcal{L}_{SigLIP} = -\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{|B|}\log\frac{1}{1+e^{z_{ij}(-t l_i^q . l_j^p + b)}} \tag{1}$$

where $l_i^q$ is the query instance at index $i$ in the batch, $l_j^p$ is the positive to the query instance at index $j$, respectively, $|B|$ represents the batch size, $z_{ij} = 1$ when $i = j$ and $z_{ij} = -1$ when $i \neq j$. The parameters $t$ (temperature) and $b$ (bias) govern the loss scaling and offset, respectively. Following common practice [9], the temperature $t$ is parameterized as $\exp(\bar{t})$, with $\bar{t}$ being a trainable parameter initially set to $\log\frac{1}{0.07}$, and the trainable bias $b$ starting at 0.

For query $l_i^q$ and its positive $l_i^p$, the NTXent (Normalized Temperature-Scaled Cross-Entropy) Loss [1] is defined as

$$\mathcal{L}_{i,j} = -\log\frac{\exp\left(\text{sim}\left(l_i^q, l_j^p\right)/\tau\right)}{\sum_k 1_{[k\neq i]}\exp\left(\text{sim}\left(l_i^q, l_k^p\right)/\tau\right)} \tag{2}$$

$$\mathcal{L}_{NTXent} = \sum_{i,j}\mathcal{L}_{i,j} \tag{3}$$

where $1_{[k\neq i]}$ is the indicator function, which is 1 if $k \neq i$, and 0 otherwise. The function $\text{sim}(l_i^q, l_i^p)$ calculates the cosine similarity between vectors $l_i^q$ and $l_i^p$, and $\tau$ is a temperature parameter set to 0.07.

The Barlow Twins contrastive loss, with hyperparameter $\mu$ (0.005), is formulated as:

$$\mathcal{L}_{BarlowTwins} = \sum_i (1 - C_{ii})^2 + \mu\sum_i\sum_{j\neq i} C_{ij}^2 \tag{4}$$

where $C$ is the cross-correlation matrix between the descriptors of queries and positives in a batch, and given by

$$C_{i,j} = \frac{\sum_a l_{a,i}^q l_{a,j}^p}{\sqrt{\sum_a (l_{a,j}^q)^2}\sqrt{\sum_a (l_{a,j}^p)^2}} \tag{5}$$

where $l_{a,i}^q$ and $l_{a,i}^p$ are the values at index $a$ of the projected embeddings of the query ($l_i^q$) and its positive counterpart ($l_i^p$), respectively.

For each set of query $l_i^q$, positive $l_i^p$, and negative $l_i^n$, the triplet margin loss is defined as:

$$\mathcal{L}_{TripletLoss} = \max\left\{\|l_i^q - l_i^p\|_2^2 - \|l_i^q - l_i^n\|_2^2 + m, 0\right\} \tag{6}$$

where the margin $m$ is set to 0.05.

The relocalization success rate comparison while using contrastive and metric loss regularization is shown in Table C (Table 4 of the manuscript). Not using any regularization loss resulted in the poorest performance. Among

*[1]Control/Robotics Research Laboratory (CRRL), Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, Brooklyn, NY, 11201. E-mails: {rgg9769,nkp269,prashanth.krishnamurthy,khorrami}@nyu.edu.
This paper is supported in part by the Army Research Office under grant number W911NF-21-1-0155 and by the New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

| Method | Training Time | Full6 | Full7 | Full8 | Full9 | Average |
|---|---|---|---|---|---|---|
| **PNVLAD** | - | 18.14, 3.28 | 24.57, 3.08 | 19.93, 3.13 | 15.59, 2.63 | 19.56, 3.03 |
| **DCP** | - | 16.04, 4.54 | 16.22, 3.56 | 14.87, 3.45 | 12.97, 3.99 | 15.03, 3.89 |
| **PosePN** | - | 14.32, 3.06 | 16.97, 2.49 | 13.48, 2.60 | 9.14, 1.78 | 13.48, 2.48 |
| **PoseSOE** | - | 7.59, 1.94 | 10.39, 2.08 | 9.21, 2.12 | 7.27, 1.87 | 8.62, 2.00 |
| **PoseMinkLoc** | - | 11.20, 2.62 | 14.24, 2.42 | 12.35, 2.46 | 10.06, 2.15 | 11.96, 2.41 |
| **PointLoc** | - | 12.42, 2.26 | 13.14, 2.50 | 12.91, 1.92 | 11.31, 1.98 | 12.45, 2.17 |
| **PosePN++** | 590 minutes | 9.59, 1.92 | 10.66, 1.92 | 9.01, 1.51 | 8.44, 1.71 | 9.43, 1.77 |
| **NIDALoc** | 1200 minutes | 6.71, 1.33 | 5.45, 1.40 | 6.68, 1.26 | 4.80, 1.18 | 5.91, 1.29 |
| **HypLiLoc** | 1020 minutes | 6.00, **1.31** | 6.88, **1.09** | 5.82, **0.97** | 3.45, **0.84** | 5.54, **1.05** |
| **Flash-Mix (M.L. Reg.)** | **80 minutes** | 3.153, 2.002 | **4.066**, 1.882 | **4.611**, 2.536 | 3.68, 1.791 | 3.878, 2.053 |
| **Flash-Mix (C.L. Reg.)** | **80 minutes** | **3.048**, 1.959 | 4.551, 2.049 | 4.674, 2.052 | **2.943**, 1.791 | **3.804**, 1.963 |

Table A. Mean position (m) and orientation errors (°) on Oxford-Radar Dataset. Best performance is highlighted in **bold**, lower is better.

| Methods | Training Time | Average |
|---|---|---|
| **PosePN** | 40 minutes | 0.12, 3.69 |
| **PoseSOE** | - | 0.13, 3,08 |
| **PoseMinkLoc** | - | 0.15, 4.57 |
| **PointLoc** | - | 0.12, 3.07 |
| **PosePN++** | 22 minutes | 0.13, 3.04 |
| **NIDALoc** | 38 minutes | 0.18, 3.74 |
| **HypLiLoc** | 13 minutes | **0.10, 2.50** |
| **Flash-Mix (ML Reg.)** | **5 minutes** | 0.14, 3.34 |
| **Flash-Mix (CL Reg.)** | **5 minutes** | 0.14, 3.42 |

Table B. Average of the Median position (m) and orientation errors (°) on vReLoc sequences. Best performance is highlighted in **bold**, lower is better.

| | F6 | F7 | F8 | F9 | Avg. |
|---|---|---|---|---|---|
| **No Reg. Loss** | 88.92 | 78.15 | 76.32 | 89.72 | 82.77 |
| **SigLIP** | 88.14 | 81.01 | 79.43 | 90.87 | 84.48 |
| **NTXent** | 88.63 | **82.29** | **81.58** | **92.56** | **85.92** |
| **Triplet** | **91.95** | 82.13 | 78.80 | 91.80 | 85.69 |
| **Barlow Twins** | 91.82 | 81.56 | 80.42 | 90.92 | 85.74 |

Table C. Ablation Study: Impact of Contrastive and Metric Loss regularization. The best and second best performances are highlighted in **bold** and underline, respectively.

the contrastive loss methods, NTXent achieved the highest average relocalization rate at 85.92%, closely followed by Barlow Twins with a rate of 85.74%. Meanwhile, the metric-learning-based Triplet Loss posted a rate of 85.69%.

While NTXent demonstrates higher performance, its computational cost scales quadratically with the batch size, posing significant efficiency challenges. In contrast, the computational cost for Barlow Twins scales linearly, which substantially reduces training times. Consequently, to optimize the balance between performance and computational efficiency, we integrated Barlow Twins Contrastive regularization into FlashMix. Additionally, we developed a variant of FlashMix utilizing Triplet Loss regularization, thereby offering two distinct configurations tailored to different operational needs.

### 1.3. Descriptor Aggregator

Section 4.4 of the manuscript explores various descriptor aggregation techniques, including MLP+Global Average Pooling (GAP), Multi-headed Attention (MHA)+GAP, Mixer+SALAD, and Mixer+GAP. Below, we detail each method used in our ablation studies:

**MLP+GAP**: This approach utilizes a Multilayer Perceptron (MLP) that features a linear layer followed by a ReLU nonlinearity. The point descriptors are projected to the global descriptor dimension and subsequently processed via Global Average Pooling to yield a singular global descriptor for each point cloud.

**MHA+GAP**: This method employs a transformer architecture with multi-headed attention, followed by GAP, for descriptor aggregation. The transformer configuration includes four attention heads, facilitating intricate interactions among point descriptors within each point cloud.

**Mixer+SALAD**: The Sinkhorn Algorithm for Locally Aggregated Descriptors (SALAD) technique refines the NetVLAD framework for feature-to-cluster assignment using an optimal transport mechanism. SALAD processes point features through the optimal transport block and integrates the output with a global token to construct robust global descriptors. Although this configuration demonstrated higher performance with Barlow Twins loss in Table 6 of our manuscript, its computational intensity restricted batch sizes to smaller numbers, consequently extending training times.

**Mixer+GAP**: This setup, which is the standard across all our experiments as discussed in Section 3.3.1, combines a Mixer with GAP to form the descriptor aggregator.

## References


[1] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Proceedings of the Advances in Neural Information Processing Systems, volume 29, Barcelona, Spain, December 2016. 1

[2] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4470–4479, Salt Lake City, UT, June 2018. 1

[3] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5176–5185, Vancouver, Canada, June 2023. 1

[4] Wei Wang, Bing Wang, Peijun Zhao, Changhao Chen, Ronald Clark, Bo Yang, Andrew Markham, and Niki Trigoni. Pointloc: Deep pose regressor for lidar point cloud localization. IEEE Sensors Journal, 22(1):959–968, 2021. 1

[5] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3523–3532, Long Beach, CA, June 2019. 1

[6] Shangshu Yu, Xiaotian Sun, Wen Li, Chenglu Wen, Yunuo Yang, Bailu Si, Guosheng Hu, and Cheng Wang. Nidaloc: Neurobiologically inspired deep lidar localization. IEEE Transactions on Intelligent Transportation Systems, 2023. 1

[7] Shangshu Yu, Cheng Wang, Chenglu Wen, Ming Cheng, Minghao Liu, Zhihong Zhang, and Xin Li. Lidar-based localization using universal encoding and memory-aware regression. Pattern Recognition, 128:108685, 2022. 1

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, pages 12310–12320, Vienna, Austria, July 2021. 1

[9] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, Paris, France, October 2023. 1