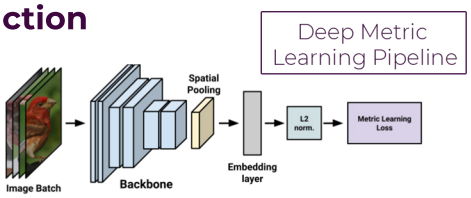
# Generalized Local Attention Pooling for Deep Metric Learning

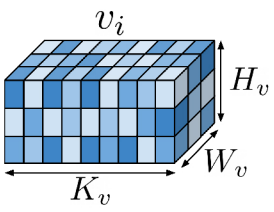
**Carlos Roig, David Varas, Issey Masuda, Juan Carlos Riveiro, Elisenda Bou-Balust**

## Introduction




The **main contribution** of this work is the **Generalized Local Attention Pooling** method which is a pooling operation that takes into account the **local relevance of the features**. Our method is capable of **outperforming other pooling operations** in the field.

## Spatial Pooling Operations




$$g_{GAP}(v_i,k)=\frac{1}{W_vH_v}\sum_{w=1}^{W_v}\sum_{h=1}^{H_v}v_i(w,h,k)$$

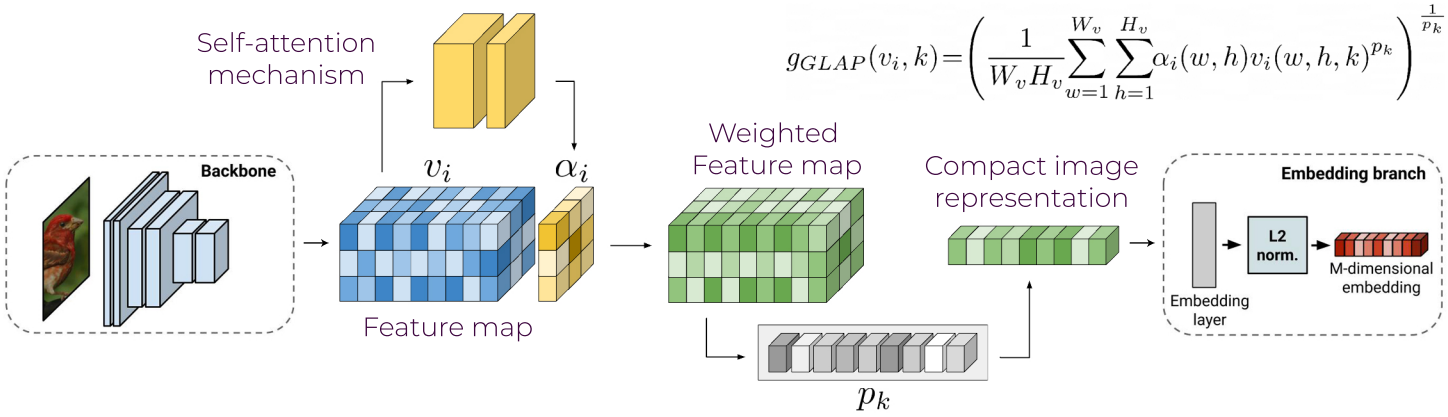
$$g_{GMP}(v_i,k)=\max_{\substack{w=1,\dots,W_v\\h=1,\dots,H_v}}v_i(w,h,k)$$

$$g_{GA+MP}(v_i,k)=g_{GAP}(v_i,k)+g_{GMP}(v_i,k)$$

$$g_{GeM}(v_i,k)=\left(\frac{1}{W_vH_v}\sum_{w=1}^{W_v}\sum_{h=1}^{H_v}v_i(w,h,k)^{p_k}\right)^{\frac{1}{p_k}}$$

## Generalized Local Attention Pooling (GLAP) - Our method

$$g_{GLAP}(v_i,k)=\left(\frac{1}{W_vH_v}\sum_{w=1}^{W_v}\sum_{h=1}^{H_v}\alpha_i(w,h)v_i(w,h,k)^{p_k}\right)^{\frac{1}{p_k}}$$




## Results

### Recall@K for different metric learning losses with the same parameters as the original method

| Recall@K | | CUB 1 | CUB 2 | CUB 4 | CUB 8 | Cars 1 | Cars 2 | Cars 4 | Cars 8 |
|---|---|---|---|---|---|---|---|---|---|
| Semihard$^{64}$ [1] | BN | 42.59 | 55.03 | 66.44 | 77.23 | 51.54 | 63.78 | 73.52 | 81.41 |
| Semihard+GLAP$^{64}$ | BN | 51.81 | 64.18 | 74.88 | 83.62 | 61.09 | 71.76 | 80.41 | 87.48 |
| ProxyNCA$^{64}$ [5] | BN | 49.21 | 61.90 | 67.90 | 72.40 | 73.22 | 82.42 | 86.36 | 64.90 |
| ProxyNCA+GLAP$^{64}$ | BN | 55.62 | 67.44 | 77.80 | 86.33 | 75.21 | 83.61 | 89.61 | 93.52 |
| Margin$^{128}$ [14] | R50 | 63.6 | 74.4 | 83.1 | 90.0 | 79.6 | 86.5 | 91.9 | 95.1 |
| Margin+GLAP$^{128}$ | R50 | 65.58 | 76.72 | 84.96 | 91.24 | 81.81 | 88.81 | 93.24 | 95.99 |
| SoftTriple$^{512}$ [16] | BN | 65.4 | 76.4 | 84.5 | 90.4 | 84.5 | 90.7 | 94.5 | 96.9 |
| SoftTriple+GLAP$^{512}$ | BN | 67.45 | 78.44 | 86.60 | 92.20 | 86.53 | 92.19 | 95.38 | 97.39 |
| MS$^{512}$ [13] | BN | 65.7 | 69.8 | 80.0 | 91.2 | 84.1 | 90.4 | 94.0 | 96.5 |
| MS+GLAP$^{512}$ | BN | 68.23 | 78.75 | 87.9 | 92.20 | 85.57 | 91.23 | 94.88 | 97.18 |
| Proxy-Anchor$^{512}$ [6] | BN | 68.4 | 79.2 | 86.8 | 91.6 | 86.1 | 91.7 | 95.0 | 97.3 |
| Proxy-Anchor+GLAP$^{512}$ | BN | 71.00 | 81.03 | 88.05 | 93.03 | 90.74 | 94.69 | 96.91 | 98.32 |
| Proxy-Anchor$^{512}$ [6] | R50 | 69.7 | 80.0 | 87.0 | 92.4 | 87.7 | 92.9 | 95.8 | 97.9 |
| Proxy-Anchor+GLAP$^{512}$ | R50 | 71.29 | 81.06 | 88.25 | 92.81 | 92.94 | 95.95 | 97.64 | 98.65 |

CUB-200-2011 [7] Cars-196 [8]

| Recall@K | | 1 | 10 | 100 | 1000 |
|---|---|---|---|---|---|
| ProxyNCA$^{64}$ [5] | BN | 73.73 | - | - | - |
| ProxyNCA+GLAP$^{64}$ | BN | 74.41 | - | - | - |
| Margin$^{128}$ [14] | R50 | 72.7 | 86.2 | 93.8 | 98.0 |
| Margin+GLAP$^{128}$ | R50 | 77.81 | 89.43 | 95.26 | 98.46 |
| MS$^{512}$ [13] | BN | 78.2 | 90.5 | 96.0 | 98.7 |
| MS+GLAP$^{512}$ | BN | 78.80 | 90.41 | 95.94 | 98.74 |
| Proxy-Anchor$^{512}$ [6] | BN | 79.1 | 90.8 | 96.2 | 98.7 |
| Proxy-Anchor+GLAP$^{512}$ | BN | 79.77 | 90.76 | 95.85 | 98.47 |

Stanford Online Products [9]

| Recall@K | | 1 | 10 | 20 | 40 |
|---|---|---|---|---|---|
| MS$^{512}$ [13] | BN | 89.7 | 97.9 | 98.5 | 99.1 |
| †MS+GLAP$^{512}$ | BN | 87.49 | 97.57 | 98.49 | 99.02 |
| Proxy-Anchor$^{512}$ [6] | BN | 91.5 | 98.1 | 98.8 | 99.1 |
| Proxy-Anchor+GLAP$^{512}$ | BN | 92.22 | 98.09 | 98.71 | 99.12 |

Inshop Clothes Retrieval [10]

## Conclusions

- The proposed **Generalized Local Attention Pooling (GLAP)** method generates a compact image representation that improves the performance of multiple metric learning approaches.
- Our GLAP method uses **higher spatial resolution**, not only increasing the performance, but also **reducing the number of parameters** required for the network.

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
[2] Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies", in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017.
[3] C. Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning", in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2840–2848.
[4] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling", in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October, 2019.
[5] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multisimilarity loss with general pair weighting for deep metric learning", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.
[6] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning", in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020.
[7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset", California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
[8] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization", in 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.
[9] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding", in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.
[10] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations", In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.