# CLOC: Contrastive Learning for Ordinal Classification with Multi-Margin N-pair Loss

Supplementary Material

## A. Datasets

### A.1. Adience dataset

This diverse dataset includes people from various cultures, ethnicities, backgrounds, and attire, along with image variations like additional people in the background. The five folds has 9,268, 6,872, 5,913, 7,104 and 7,706 images respectively. On average each fold includes 363 images in the 0-2 age group, 317 in 4-6, 317 in 8-13, 224 in 15-20, 670 in 25-32, 313 in 38-43, 114 in 48-53 and 116 in 60+ age group. Figure 1 presents sample images from the dataset.

### A.2. Historical Colour Image Dating (HID) dataset

This dataset features vehicles, scenery, roads, landscapes covering various seasons and contexts, occasionally including people. Figure 6 provides sample images from the dataset.

### A.3. Knee Osteoarthritis (KOA) dataset

The dataset includes 2,286 grade 0, 1,046 grade 1, 1,516 grade 2, 757 grade 3, and 173 grade 4 knee joints in the training set. The validation set contains 328 grade 0, 153 grade 1,212 grade 2, 106 grade 3, and 27 grade 4 samples. We combine these into a single training set. We evaluate using the provided testing split having 639 knee joints of grade 0, 296 of grade 1, 447 of grade 2, 223 of grade 3, and 51 of grade 4 and train using the rest of the dataset. Figure 7 shows sample images from the dataset.

### A.4. Indian Diabetic Retinopathy Image Dataset (IDRID) dataset

The training set features 134 fundus images from grade 0, 20 from 1, 136 from 2, 74 from 3 and 49 from 4. Following a similar distribution, the testing set includes 34 fundus images from grade 0, 5 from grade 1, 32 from grade 2, 19 from grade 3 and 13 grade 4. The figure 1 shows samples from the dataset.

### A.5. BReAst Carcinoma Subtyping dataset (BRACS) dataset

The training set has 403 images from grade 0, 757 from grade 1, 435 from grade 2, 673 from grade 3, 428 from grade 4, 705 from grade 5 and 568 from grade 6. The esting set includes 81 images from grade 0, 79 from grade 1, 82 from grade2, 83 from grade 3, 79 from grade 4, 85 from grade 5 and 81 from grade 6. The figure 8 shows samples from the dataset.

## B. Phase One Training Remarks

This section outlines some remarks from the training phase one that could lead to trivial solutions, and hence should be avoided.

**Remark 1** (Margin-collapsed solution). Assuming that the set of all-zero margins is represented by $m_{\mathcal{H}} = 0$. Although, the training can converge with all margins being $0$, a meaningful rank representation will not be learned.

*Reason:* For every configuration $(\phi^*, \gamma^*, m^*_{\mathcal{H}})$, the objective function 1 relies on the learned margins. However, when all margins are zero, the contrastive force applies to all negative pairs becomes uniform, regardless of rank differences. Despite this, the model still converges due to the cross-entropy loss, which optimizes for standard classification rather than ordinal classification. The optimal margins remain trivial at $m^*_{\mathcal{H}} = 0$, because equation 1 is minimized. Thus, $m^*_{\mathcal{H}} = 0$ is a technically optimal since the objective function is minimized, but a degenerate solution that fails to capture ordinal relationships. □

To prevent this, we initialize the margins randomly from a uniform distribution in the range $[0.5, 1.0)$, instead of initializing them close to $0$ when training. In addition, we take precautions in remark 2 to delay converging to zero.

**Remark 2** (Non-smooth activation functions). Jointly optimising features and margins could lead to margin collapse if the activation function used for margin parameters is non smooth around its lower bound, that is $\leq 0$.

*Reason:* Suppose the activation function for margin parameters $\varphi(\mathcal{H})$, is non-smooth (i.e. non-differentiable) at lower bound and lower bound is at or below zero (like ReLU). In the first training phase, when training accuracy improves, the $\ell_{CE}$ in the equation 1 is effectively minimised since it directly contributes to classification. As training further continues, to reduce the overall loss, the optimization objective then shifts to minimize MMNP, which depends on the margins. Further, minimizing MMNP drives $\varphi(\mathcal{H})$ towards its lower bound, hence collapsing margins to zero. □

Remark 2 simply says that if the model is trained for a large number of epochs in phase one and used a non-smooth activation function that is bounded at or below 0 for margin learning parametrs, eventually the model can approach to a all-zero margin solution (remark 1) and mimic a standard classification during training phase one.

To prevent this, we employ two strategies. First, we avoid using activation functions with sharp, non-smooth (i.e. non-differentiable) transitions at lower bound, such


Figure 6. Examples from Historical Colour Image Dating (HID) dataset

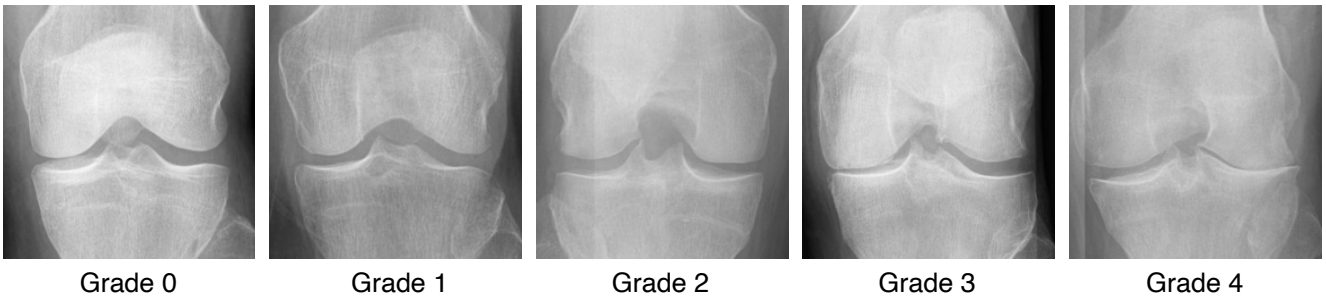


Figure 7. Examples from Knee Osteoarthritis (KOA) dataset

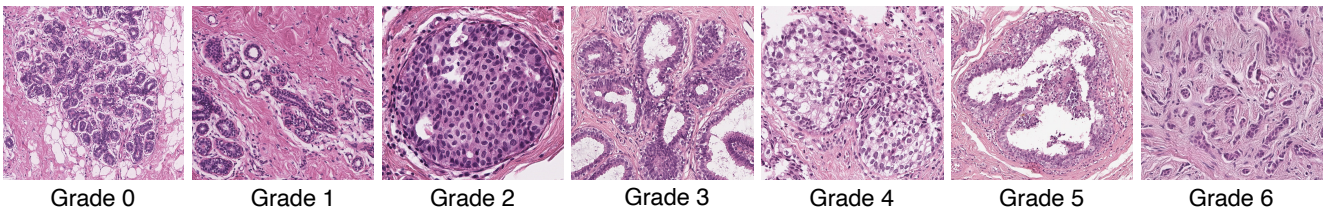


Figure 8. Examples BRACS dataset

as ReLU, on margin learning parameters. Instead, we use smoother activation functions like Softplus, which provide smoother gradients and approach 0 gradually. This helps avoid dead neurons and, consequently, margin collapse. The second strategy involves training the model in two phases. In the first phase, we optimize the model for both feature learning and margin learning objectives, with early stopping. In the second stage, we re-train the model for the feature learning objective, keeping the margins frozen at the values learned during the first stage.

| # Classes (Bin size) | Test set MAE ↓ |
|---|---|
| 79(1) | 6.094 |
| 42(2) | 3.134 |
| 28(3) | 2.018 |
| 22(4) | 1.471 |
| 18(5) | 1.121 |

Table 5. Evaluation of CLOC on CLAP2015 with different class numbers.

| Method | Test set MAE ↓ |
|---|---|
| POE [34] | 2.41 |
| GOL [31] | 2.33 |
| MWR [49] | 2.25 |
| CLOC (Ours) | 2.02 |

Table 6. Comparison with related methods using 28 classes in CLAP2015 (bin size=3).

## C. Additional experiments

### C.1. With A Large Number Of Classes

We evaluate CLOC with a VGG16 [50] backbone on the CLAP2015 dataset [17] in Tab. 5. CLAP2015 has 79 classes where unique ages are treated as classes. We first align the faces using MTCNN [63] following the method in MWR [49]. When treating each age as a separate class (79 classes), CLOC underperformed MWR (6.09 to 2.77 testset MAE). However, binning consecutive ages (sizes 2, 3, 4, 5) yielded competitive results for bin sizes of 3 and above. For fairness, we re-ran related methods using the same labels for bin size of 3, with results in Tab. 6, where we can see better performance compared to the related methods.

### C.2. Training Time Comparison

In Tab. 7, we compare runtime (in hours) for 100 epochs on the IDRID dataset using ResNet50 backbone model on a

| Method | Time (hours h) |
|---|---|
| SimCLR | 1.21 h |
| SupCon | 0.89 h |
| DINO | 1.01 h |
| ORCNN | 3.28 h |
| POE | 0.97 h |
| GOL | 1.25 h |
| MWR | 2.22 h |
| RnC | 0.51 h |
| CLOC | 1.19 h |

Table 7. Training time comparison in hours.

RTX 4090. CLOC phase one and two take 0.65h and 0.54h, respectively, totaling 1.19h.

### C.3. With Different Backbone Models

The table 8 compares CLOC's performance with different backbone models on IDRID dataset, where we can see a steady increase in accuracy as the size of the model (measured by number of parameters) increases.

| Backbone | Accuracy ↑ | MAE ↓ |
|---|---|---|
| DenseNet121 | 0.6990 | 0.4854 |
| ResNet50 | 0.7379 | 0.4078 |
| VGG16 | 0.7476 | 0.4351 |

Table 8. Ablation study with different backbone models, arranged in the increasing number of parameters DenseNet121 $<$ ResNet50 $<$ VGG16.

### C.4. More Visualizations

The Figure 9 visualizes learned representation by GOL and POE by extending the Figure 5 in the main paper.

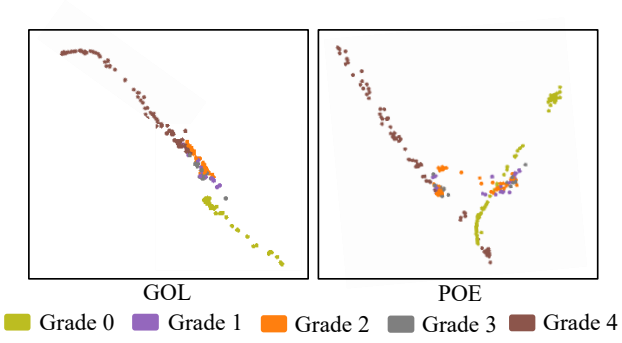


Figure 9. UMAP visualizations of learned representations by GOL [31] and POE [34] for the IDRID dataset that focuses on cancer grade classification.