Table 3: Overview of retrieval performance expressed with the ranking metric (NDCG) as explained in Section 4.1. The bottom section shows the strong results for the pre-trained PointBERT ULIP-2 model. The PointNet++ section provides the baseline results for the prior art. The PointBERT section in the center of the table serves as a comparison with the bottom section to see the effect of the ULIP-2 multi-modal pre-training. Each cell of the table shows the test result of a separate training using the training method + model on the left of the row and the target dataset on the top of the column. There are three training methods considered in this table: supervised classification, self-supervised VICReg[7] and multi-modal contrastive learning (MMCL)[8], we refer to Section 3 for more detail. The **bold** numbers represent the highest (best) score for each dataset.

Training method \Dataset	MN40	SN Norm	SN Per	MCB	Prop	Obja Easy	ScanObjectNN
PointNet++							
random init	62.0	72.9	39.2	66.8	45.5	38.5	50.5
classification	86.0	81.0	68.1	87.6	64.4	49.7	71.1
VICReg	74.5	78.8	68.5	84.9	63.1	48.8	63.8
PointBERT							
random weights	62.9	73.2	40.3	68.1	45.9	38.7	50.7
classification	81.3	82.3	74.1	90.2	67.8	51.0	73.1
VICReg	71.8	76.7	63.1	84.6	59.4	44.4	51.0
MMCL	77.1	78.8	73.1	86.5	62.0	52.7	-
PointBERT ULIP-2							
pre-trained	78.1	84.6	75.4	87.4	68.8	66.1	63.4
FT w/ classification	**93.6**	**90.4**	**85.6**	**97.2**	**85.0**	**79.9**	**93.0**
FT w/ VICReg	79.9	84.4	79.2	87.8	71.1	66.3	66.6
FT w/ MMCL	82.5	84.3	77.2	89.9	68.9	68.3	-

Table 4: Comparison between state-of-the-art view-based methods View-GCN [27], MVTN [28] and our best performing point cloud-based approach expressed with the ranking metric (NDCG) as explained in Section 4.1. The three methods are competitive, with a slight advantage for the classification fine-tuned ULIP-2 PointBERT in 4 out of 7 cases. Each cell represents a separate training of the fine-tuning method on the left and the dataset on the top. The **bold** numbers represent the highest (best) score for each dataset.

Training method \Dataset	MN40	SN Norm	SN Per	MCB	Prop	Obja Easy	ScanObjectNN
View-GCN[27]							
FT w/ classification	**93.8**	89.2	86.1	97.1	84.3	74.6	-
MVTN[28]							
FT w/ classification	93.3	90.0	**86.3**	96.8	83.5	74.1	90.6
PointBERT ULIP-2							
FT w/ classification (ours)	93.6	**90.4**	85.6	**97.2**	**85.0**	**79.9**	**93.0**

Table 5: Overview of retrieval performance expressed with the nearest neighbor classification test F1 (NN F1) metric as explained in Section 4.1. The bottom section shows the strong results for the pre-trained PointBERT ULIP-2 model. The PointNet++ section provides the baseline results for the prior art. The PointBERT section in the center of the table serves as a comparison with the bottom section to see the effect of the ULIP-2 multi-modal pre-training. Each cell of the table shows the test result of a separate training using the training method + model on the left of the row and the target dataset on the top of the column. There are three training methods considered in this table: supervised classification, self-supervised VICReg[7] and multi-modal contrastive learning (MMCL)[8], we refer to Section 3 for more detail. The **bold** numbers represent the highest (best) score for each dataset.

Training method \Dataset	M40	SN Norm	SN Per	MCB	Prop	Obja Easy	ScanObjectNN
PointNet++							
random weights	69.2	55.2	16.6	72.6	49.3	53.6	36.3
classification	86.6	75.2	53.6	89.0	64.2	71.1	77.3
VICReg	84.7	68.1	55.2	89.0	65.1	72.3	71.7
PointBERT							
random weights	71.5	56.6	17.8	77.9	50.1	54.9	37.7
classification	85.8	71.9	58.3	91.0	65.0	74.2	79.8
VICReg	80.1	64.5	46.5	88.8	58.1	67.7	50.6
MMCL	81.8	68.2	56.5	89.7	61.0	75.1	-
PointBERT ULIP-2							
pre-trained	85.6	75.6	62.1	92.3	70.3	84.3	73.8
FT w/ classification	**91.0**	**79.0**	**70.8**	**94.7**	**75.2**	**87.2**	**92.7**
FT w/ VICReg	86.2	75.6	66.0	92.1	72.4	83.3	75.9
FT w/ MMCL	88.8	75.7	64.9	93.8	70.3	86.5	-

Table 6: Comparison between state-of-the-art view-based methods View-GCN [27], MVTN [28] and our best performing point cloud-based approach expressed with the nearest neighbour classification F1 metric as explained in Section 4.1. The three methods are competitive, with a slight advantage for the classification fine-tuned ULIP-2 PointBERT in 4 out of 7 cases. Each cell represents a separate training of the fine-tuning method on the left and the dataset on the top. The **bold** numbers represent the highest (best) score for each dataset.

Training method \Dataset	MN40	SN Norm	SN Per	MCB	Prop	Obja Easy	ScanObjectNN
View-GCN[27]							
FT w/ classification	**91.4**	74.8	65.9	93.3	73.4	**87.8**	-
MVTN[28]							
FT w/ classification	90.5	78.5	**71.4**	93.4	72.7	87.0	90.2
PointBERT ULIP-2							
FT w/ classification	91.0	**79.0**	70.8	**94.7**	**75.2**	87.2	**92.7**