
Improving Dense Contrastive Learning with Dense Negative Pairs

Berk Iskender^{†*} **Zhenlin Xu**[†] **Simon Kornblith** **En-Hung Chu** **Maryam Khademi**
UIUC Amazon Google Research Google Google

Abstract

Many contrastive representation learning methods learn a single global representation of an entire image. However, dense contrastive representation learning methods such as DenseCL (Wang et al., 2021) can learn better representations for tasks requiring stronger spatial localization of features, such as multi-label classification, detection, and segmentation. In this work, we study how to improve the quality of the representations learned by DenseCL by modifying the training scheme and objective function, and propose DenseCL++. We also conduct several ablation studies to better understand the effects of: (i) various techniques to form dense negative pairs among augmentations of different images, (ii) cross-view dense negative and positive pairs, and (iii) an auxiliary reconstruction task. Our results show 3.5% and 4% mAP improvement over SimCLR (Chen et al., 2020a) and DenseCL in COCO multi-label classification. In COCO and VOC segmentation tasks, we achieve 1.8% and 0.7% mIoU improvements over SimCLR, respectively.

1 Introduction

Self-supervised learning aims to learn representations from unlabeled data via pre-text task training. Contrastive learning, as a self-supervised learning technique, performs the pre-text task of instance discrimination (Dosovitskiy et al., 2014; Wu et al., 2018; Chen et al., 2020a). Instance discrimination in contrastive learning usually trains a single global representation, and these representations are principally evaluated in terms of downstream performance on a single-label classification task. However, methods that perform well in this setting may perform suboptimally on multi-label classification tasks, where each label is associated with a distinct object in an image, but different image regions contain different semantic content. Motivated by the important application of multi-label classification in industry, we target representation learning for this task. We demonstrate that our approach also improves accuracy on dense downstream tasks such as segmentation.

Our work is inspired by DenseCL (Wang et al., 2021) which proposes to use dense features rather than global ones in contrastive learning to improve the performance in dense prediction tasks. We focus on further boosting the performance of dense contrastive learning by modifying the training scheme and the objective function. Unlike DenseCL, our proposed approach formulates negative pairs between the dense features of augmented views of different images and uses their similarities in the proposed dense contrastive loss scheme. We show that the proposed method outperforms DenseCL in various settings. We also conduct several ablation studies to better understand the effects of: (i) various methods to form dense negative pairs among augmentations of different images, (ii) cross-view dense negative and positive pairs, and (iii) an auxiliary reconstruction task.

*Correspondence to berk2@illinois.edu and maryamkhademi@google.com

[†]Work done during an internship at Google

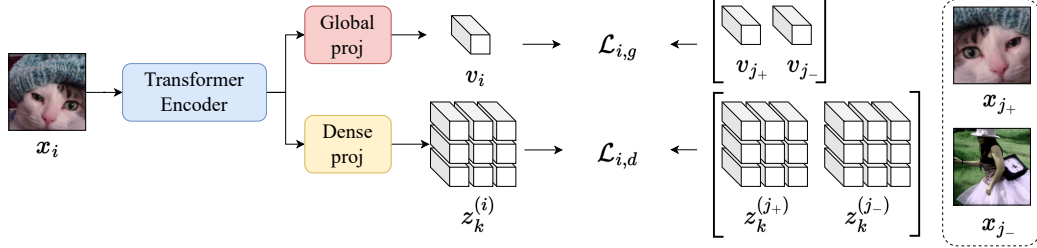


Figure 1: DenseCL++ training scheme. Global and dense positive/negative correspondences are used in the global (top row) and dense (bottom row) loss functions, respectively.

2 Related Work

SimCLR (Chen et al., 2020a) proposes a simple contrastive learning framework such that the *projected* representations of randomly augmented views of the same image sample are attracted to each other using a contrastive loss. DenseCL (Wang et al., 2021) proposes an extension of this framework better suited to dense prediction tasks. In DenseCL, the contrastive loss is applied in a dense pairwise manner which improves the performance compared to the global representation learning counterparts (Chen et al., 2020b).

On the other hand, following widespread success of the transformer architecture (Vaswani et al., 2017) in NLP tasks, Vision Transformer (ViT) (Dosovitskiy et al., 2020) adapts the architecture for visual tasks and achieves impressive results when pretrained on sufficient amount of data. Inspired by (Devlin et al., 2018), Li et al. (2021) explored the idea of introducing a reconstruction task to the contrastive learning framework using ViT as an encoder. Wang et al. (2022) further study the use of reconstruction as a pretext task, incorporating a decoder module in various self-supervised contrastive settings. These two methods use shallow convolutional networks for their decoders to preferably learn additional useful local features in the latent space. However, the authors suggest that the use of sophisticated reconstruction models may be harmful to transfer tasks, as they could lead to excessively local representations.

3 Method

3.1 Dense Contrastive Learning

Contrastive learning learns latent representations of signals for which the positive correspondences are attracted to each other and negative ones are repelled from one another. Dense contrastive learning (Wang et al., 2021) further adapts this framework for dense prediction tasks by replacing the global representations with their dense counterparts. For each image, instead of the global representation $v \in \mathbb{R}^D$, $S \times S$ many dense feature vectors $z \in \mathbb{R}^L$ are extracted.

Then, dense positive pairs are formed between dense features of the anchor view x_i and its corresponding augmented view x_{j+} by finding the most similar correspondence for each dense vector of x_i in x_{j+} as $k_+ = \arg \max_l \text{sim}(z_k^{(i)}, z_l^{(j+)})$, where $z_k^{(i)}$ is the k -th dense feature of the anchor view x_i , $z_l^{(j+)}$ is the l -th dense feature of x_{j+} , and $\text{sim}(a, b)$ calculates the cosine similarity between two feature vectors. Dense negative pairs are formed between the dense feature vectors of the anchor view and the global representations of views from other images. The dense contrastive loss is computed as

$$\mathcal{L}_{i,d} = \sum_k -\log \frac{\exp(z_k^{(i)} \cdot z_{k_+}^{(j+)})/\tau}{\exp(z_k^{(i)} \cdot z_{k_+}^{(j+)}) + \sum_{j_-} \exp(z_k^{(i)} \cdot v_{j_-})/\tau} \quad (1)$$

where $z_{k_+}^{(j+)}$ is the positive dense correspondence for the dense feature vector $z_k^{(i)}$ in the view x_{j+} , v_{j_-} is the global feature for the image x_{j_-} , and τ is the temperature parameter.

The overall loss is a linear combination of the global InfoNCE loss term $\mathcal{L}_{i,g}$ (Oord et al., 2018) and dense loss, $\mathcal{L}_i = (1 - \lambda)\mathcal{L}_{i,g} + \lambda\mathcal{L}_{i,d}$, where $\lambda \in [0, 1]$ is a weight constant.

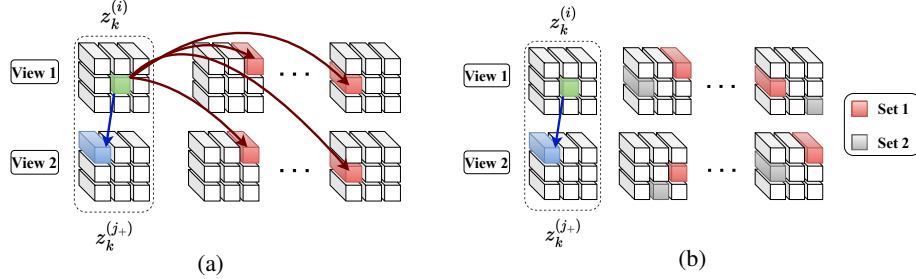


Figure 2: Dense positive and negative pair formulations for DenseCL++. Pair forming diagrams for (a) randomly sampling dense negatives, and (b) using $M = 2$ candidate negative feature sets for guided sampling from other image augmentations.

3.2 DenseCL++: DenseCL with Dense Negative Pairs

In our proposed method, instead of computing dense-global negative correspondences in dense contrastive loss $\mathcal{L}_{i,d}$, we form dense negative pairs. This leads to the following dense contrastive loss

$$\mathcal{L}_{i,d} = \sum_k -\log \frac{\exp(z_k^{(i)} \cdot z_{k_+}^{(j+)})/\tau}{\exp(z_k^{(i)} \cdot z_{k_+}^{(j+)}) + \sum_{j_-,m} \exp(z_k^{(i)} \cdot z_m^{(j-)})/\tau}, \quad (2)$$

where the global representation v_{j_-} of the view x_{j_-} is replaced by its dense features $z_m^{(j-)}$. This is illustrated in Fig. 1.

Forming dense negative pairs Multiple possibilities exist for forming dense negative pairs. The first approach is to randomly sample a single dense feature from each augmented view in the batch and use them for the computation of negative pairs with the dense features of the anchor view. We use this option as the baseline of our proposed method. A diagram is provided in Fig. 2 (a).

Another possibility is to sample M random candidate sets of dense negative features where each set contains a single feature from each view in the batch of other images. Then, the similarities between each set and the dense features of the anchor image x_i can be sorted using a specific criterion. Finally, the set with the top rank can be used to form dense negative pairs. The criterion can be chosen as the average similarity of the dense features in the set to all dense anchor features, with the hope that such sets include harder negatives. We call this alternative the guided dense negative formulation.

To more easily distinguish sets with hard dense negatives, thresholding can be applied to the similarities $q = \text{sim}(a, b)$ before computing the average

$$\bar{q} = \begin{cases} -1, & q \leq \beta \\ q, & \text{otherwise,} \end{cases} \quad (3)$$

where $\beta \in [-1, 1]$ is the threshold constant.

Since the corresponding views of an image are only used for forming dense positive pairs, it can be possible to observe high mean cosine similarities between dense features of these views. To improve discriminability and obtain a wider distribution of cross view dense features, N additional pairs can be formed from the corresponding augmented view x_{j_+} by selecting the least similar features.

4 Experiments

Framework We use ViT-S/16 (Dosovitskiy et al., 2020) as the backbone encoder in our experiments. Since we aim to compare the pretraining schemes and objectives, we do not explicitly try to optimize the backbone configuration for the specific downstream task. The projection head is chosen as a 3 layer MLP with hidden layer dimension of 4096 and output dimension of 128. For data augmentation, random cropping, flipping, color jittering, and Gaussian blur are applied as in (Chen et al., 2020a). For each experiment, the network is pretrained for 1000 epochs on the COCO/2017 (Lin et al., 2014) training dataset. The evaluation is performed on the validation subset of COCO/2017 by fixing the pretrained backbone parameters and training a linear classifier on top of the learned representations for multi-label classification. The AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 4×10^{-3} , cosine decay schedule, and a weight decay of 5×10^{-2} are used to train the model.

To evaluate the performance in multi-label classification evaluations, we report the mAP metric as described in Section 4.2 of (Veit et al., 2017), and F1-score.

Baseline experiments To obtain baseline metrics, we pretrained the same backbone with SimCLR (Chen et al., 2020a) and DenseCL (Wang et al., 2021) methods and evaluated the performance on multi-label classification with the same parameter setting described in Section 4. The dense contrastive loss weight for DenseCL is chosen as $\lambda = 0.3$, which performed the best in a linear sweep.

For SimCLR and DenseCL, we test two different methods to obtain global features. The former uses the CLS token of ViT, and the latter uses the average of the dense local features (GAP) as global features.

For DenseCL, both global and dense projection heads have three linear layers with 4096 hidden dimensionality and global and dense feature vectors are $D = L = 128$ dimensional. The positive dense pairs are formed based on their cosine similarities using the backbone encoder or global projection head output representations. The results are shown in Table 1.

DenseCL++ experiments As mentioned in Section 3.2, we replace dense-global negative comparisons in DenseCL with dense-dense counterparts. A dense feature is sampled uniformly random from each view that belongs to augmentations of different images in the batch to form negative pairs with the dense feature of the anchor view. This results in a substantial improvement for evaluation on multi-label classification. The positive pairs are again formed using their encoder output representations. The comparison for different dense loss weights and global feature aggregations techniques as explained in Section 4 is reported in Figure 3. The best performing configuration on multi-label classification evaluation is used as the baseline for DenseCL++. We report multi-label classification results for the best performing M , β , and N in Table 2. In Table 3 we show that DenseCL++ can also lead to improvements when fine-tuning models for semantic segmentation on PASCAL VOC and MS COCO. Experimental details for segmentation evaluation can be found in Appendix A. We also provide further ablation studies with various dense negative formulation techniques and reconstruction modules in Appendix C.

The results in Figure 3 show that the performance of DenseCL++ consistently improves until $\lambda = 0.9$ and degrades for $\lambda = 1$, where the drop is more dramatic for the CLS aggregation. Both CLS and GAP cases work almost equally well for the optimal weight of $\lambda = 0.9$ but due to the significant drop at $\lambda = 1$ for CLS aggregation, we report GAP aggregation results for DenseCL++ in Table 2.

Table 2: Top performances for multi-label classification on COCO for different methods. Baseline DenseCL++ forms dense negative pairs as in Fig. 2(a) whereas DenseCL++* uses guided dense negatives with $M=256$, $\beta=0.5$ and $N=64$.

Method	Agg.	Pair feature	mAP	F1
SimCLR	CLS	-	59.6	37.8
DenseCL	CLS	backbone	59.9	38.1
DenseCL++	GAP	backbone	63.4	39.0
DenseCL++*	GAP	backbone	64.1	39.1

Table 1: SimCLR and DenseCL multi-label classification results on COCO for different global feature aggregation and dense matching types.

Method	Agg.	Pair feature	mAP	F1
DenseCL	CLS	backbone	59.9	38.1
		proj head	59.8	38.1
	GAP	backbone	58.1	37.3
		proj head	57.8	37.5
SimCLR	CLS	-	59.6	37.9
	GAP	-	58.4	37.7

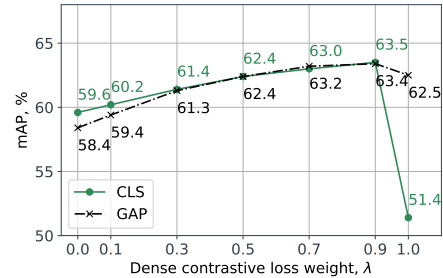


Figure 3: mAP vs. dense contrastive loss weight λ plot for DenseCL++ for different global feature aggregation settings.

Table 3: Semantic segmentation on PASCAL VOC and MS COCO. All methods use GAP aggregation type. Experimental details are provided in Appendix A.

Method	Pair feature	VOC mIoU	COCO mIoU
SimCLR	-	69.3	61.5
DenseCL++	backbone	70.0	63.3

Acknowledgement We thank our colleagues from Google Research and Brain, Dilip Krishnan, Yin Cui, Aaron Sarna, Ting Chen, and Golnaz Ghiasi who provided insight and expertise that greatly assisted this research. Also, we are grateful to Yeqing Li who provided the UViT implementation.

References

- L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.
- W. Chen, X. Du, F. Yang, L. Beyer, X. Zhai, T.-Y. Lin, H. Chen, J. Li, X. Song, Z. Wang, et al. A simple single-scale vision transformer for object localization and instance segmentation. *arXiv preprint arXiv:2112.09747*, 2021.
- X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.
- A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.
- Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.
- L. Wang, F. Liang, Y. Li, W. Ouyang, H. Zhang, and J. Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. *arXiv preprint arXiv:2201.06857*, 2022.
- X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.
- Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Appendix

A PASCAL VOC and COCO Semantic Segmentation

To further evaluate the downstream performance of the proposed pretraining scheme, we finetune our models with a UViT (Chen et al., 2021) backbone on PASCAL VOC and COCO/2014 semantic segmentation tasks. As in the multi-label classification case, we do not explicitly try to optimize the backbone configuration. Thus, the UViT architecture configuration is the same across different methods and has depth of 18 layers, hidden size of 342, number of attention heads per layer as 6, and patch size of 8. We also do not leverage the specific attention window strategy of UViT and use global attentions. DenseCL++ forms dense negative pairs using the random sampling method described in Fig. 2 (a) and has a dense contrastive loss weight $\lambda = 0.7$.

All segmentation models are finetuned using a COCO pretrained initialization as in multi-label classification experiments. PASCAL VOC segmentation evaluations use ASPP (Chen et al., 2017) decoder, DeepLabv3+ (Chen et al., 2018) head, AdamW (Loshchilov and Hutter, 2017) optimizer with a stepwise decay schedule with initial learning rate of 3×10^{-3} , weight decay as 10^{-4} , batch size of 256, and 20k training steps. COCO segmentation evaluations use FPN (Lin et al., 2017) decoder, AdamW optimizer with initial learning rate of 5×10^{-4} with cosine decay schedule, batch size of 256, and 64k training steps. The results are shown in Table 3.

B Reconstruction Decoder

Inspired by several recent studies that incorporate an auxiliary lightweight reconstruction module in the contrastive learning framework (Li et al., 2021; Wang et al., 2022), we test several different alternatives in the context of our proposed dense contrastive loss. These include convolutional and transformer-based decoders that reconstruct the augmented input image from its dense hidden representations. The reconstruction architecture is shown in Fig. 4.

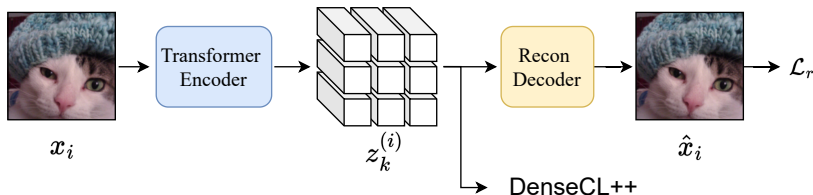


Figure 4: Reconstruction module for DenseCL++ framework.

B.1 Reconstruction Objective

The encoder and the reconstruction network parameters are updated end-to-end during pre-training to minimize the mean absolute error,

$$\mathcal{L}_r = ||x_i - \hat{x}_i||, \quad (4)$$

where \hat{x}_i is the estimated reconstruction of the input image. The reconstruction loss term is weighted by a constant $\gamma \in \mathbb{R}$. The overall loss is expressed as

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_g + \lambda\mathcal{L}_d + \gamma\mathcal{L}_r. \quad (5)$$

The decoder is discarded after pre-training of the encoder.

B.2 Convolutional Decoder

In this setting a simple convolutional neural network architecture with consecutive convolutional layers is applied on the global or dense features of the input image. Global features are reshaped to 2D before they are fed to the decoder. On the other hand, dense representations $z_k^{(i)} \in \mathbb{R}^{S \times S \times L}$ are treated as L channel inputs. In both cases, either transposed convolutional layers or convolutional

layers with fixed number of output channels and a fixed rate of bicubic upsampling are applied consecutively to the inputs at each layer until the estimated reconstruction $\hat{x} \in \mathbb{R}^{224 \times 224 \times 3}$ is obtained.

B.3 Transformer-based Decoder

It is also possible to use the transformer architecture (Dosovitskiy et al., 2020) to reconstruct images. To do so, a linear layer is applied to the dense features to project them to a lower dimensional subspace. Then, the projected representations are fed to the decoder and the image is reconstructed patch-wise as in (He et al., 2022).

C Ablation studies

C.1 Dense negative pair forming strategies

Guided dense negative sampling from other pairs. Encouraged by the performance improvement introduced by random dense negative sampling, we also experiment with a guided dense negative sampling method. We repeat the random sampling of dense features from other pairs M times and pick the set of dense features that have the largest average similarity to the anchor view features. As described before, we aim to form hard dense negatives as a result of this modification. Averaged mAP values over β and N for different M are shown in Fig. 5 (a).

Furthermore, as described in Section 3.2, hard thresholding can be applied to the similarities while computing the most similar set of dense negative features on average to the anchor image features. By doing so, it can be prevented that the average similarity being mostly decided by moderately similar dense features of the anchor image which ultimately assigns a more important role to hard dense negatives. Fig. 5 (b) shows the mAP performance for various threshold levels β .

Sampling dense negatives from the corresponding view. After observing high and condensed similarity distributions for the augmented views of the same image during training and evaluation, we also compute pairwise dense negatives by finding the least similar k dense features in the corresponding view of the anchor dense feature map. As shown in Fig. 6, incorporating additional pairwise dense negatives results in widened cross similarity distributions. Evaluation performances are provided in Fig. 5 (c).

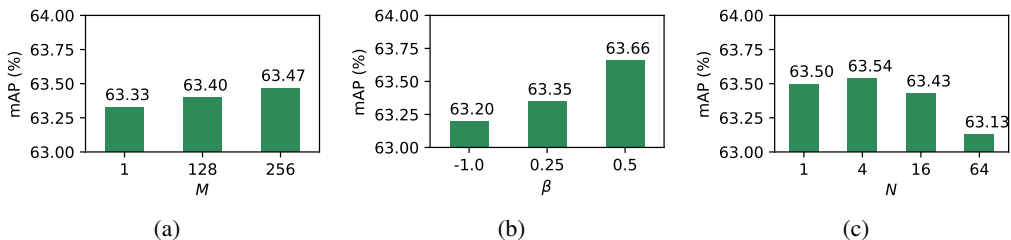


Figure 5: Averaged mAP for each fixed parameter. We conduct 36 experiments with all different configurations of M , β , N and report the averaged metrics for experiments that include the specific parameter value. Results for different (a) number of random negative index sampling sets, (b) similarity thresholds, and (c) cross-view negative pairs.

C.2 Dense positive pair forming strategies

To explore the potential effect of introducing multiple positive correspondences in our framework, we experiment with multiple positive pair sampling from the corresponding view by selecting the top- k similarity pairs as positive correspondences using the best performing DenseCL++ configuration with random dense negative sampling from other view pairs. We adapt our loss function provided in (2) such that it incorporates multiple positive dense features as proposed in (Khosla et al., 2020). The performances for increasing number of positives are provided in Fig. 7.

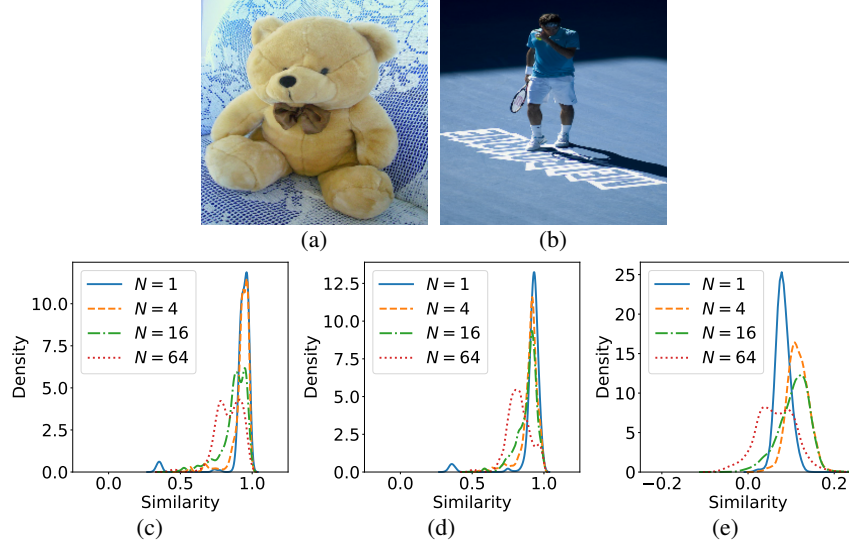


Figure 6: Intra-image and inter-image dense feature cosine similarity distributions of images (a) and (b) for increasing number of cross-view dense negatives N during training. Graphs (c) and (d) show the intra-image distributions for (a) and (b), respectively. Inter-image distributions are shown in (e).

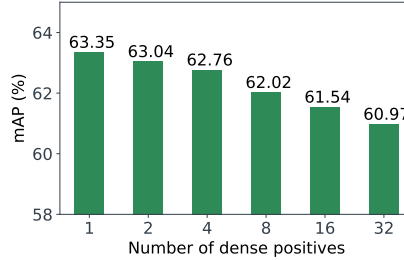


Figure 7: mAP vs. number of cross view positive pairs. The rest of the hyperparameters pre-training are selected identically to the baseline DenseCL++ configuration described in Table 2. Larger number of cross view dense positives leads to decreasing evaluation performances consistently.

C.3 Reconstruction module strategies

Convolutional decoder. Using reshaped global features as input to the convolutional reconstruction decoder with various projected representation dimensions do not provide interpretable reconstructions and hence do not have a considerable effect on the evaluation performance of the method. Thus, we report the experimental results belonging to dense feature input setting. Although incorporation of the dense convolutional decoder do not improve multi-label classification evaluation performance considerably over our DenseCL++, we observe that convolutional layers followed by a fixed interpolation scheme, e.g. bicubic, works slightly better compared to using transposed convolutional layers. Also, increasing the loss weight above a certain level and hence achieving better reconstruction estimates degrade the evaluation performance significantly.

The best performing configuration in our experiments provided 63.6% mAP and 38.9% F1-score, approximately +0.2% mAP improvement over the baseline DenseCL++ performance. In this setting, we used $\gamma = 1 \times 10^{-6}$, number of output layers per convolutional layer as 16, and 4x bicubic upsampling after each convolutional layer.

Transformer-based decoder. Similar to the convolutional decoder setting we did not observe considerable improvement for multi-label classification evaluation performance when transformer-based decoder is used. Again, prioritizing reconstruction performance causes a performance drop in the evaluation stage. Both CLS token and GAP aggregation settings for obtaining global hidden representations perform similarly. Reasonable reconstructions can be achieved for projected decoder hidden representations with dimension 32 or larger. Lower dimensional vectors cause significant block-like artifacts. In accordance with the suggestions of using a lightweight decoder in (Li et al.,

2021) and (Wang et al., 2022), we also experiment with reduced number of transformer blocks and attention heads for each block compared to (He et al., 2022) and provide results in Figure 8.

Using transformer-based decoder provides a marginal improvement (+0.3% mAP) over the baseline DenseCL++ result in the best case when number of layers and number of heads are both 4 and the loss weight is $\gamma = 5 \times 10^{-7}$ with decoder latent dimension of 64.

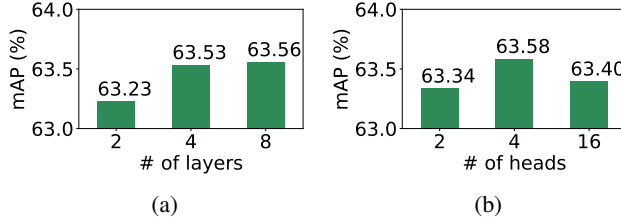


Figure 8: Averaged mAP for fixed levels of number of (a) layers and (b) heads in the transformer-based decoder for decoder hidden dimension of 64. In total, 9 different experiments were conducted with all possible configurations for number of heads and layers in the decoder and the averaged mAP for experiments with the specific parameter value are reported. Remaining pre-training hyperparameters were kept the same as in baseline DenseCL++ in Table 2.

For both decoder types, marginal gains are only possible if the reconstruction loss weight γ is chosen appropriately such that the reconstructions are imperfect. Optimizing for perfect accuracy by using larger weights significantly degrades the multi-label classification evaluation performance. The multi-label classification mAP values for various γ in Fig. 9 highlight this fact. Reconstruction samples from the validation set in Fig. 10 show how accuracy changes with respect to the loss weight.

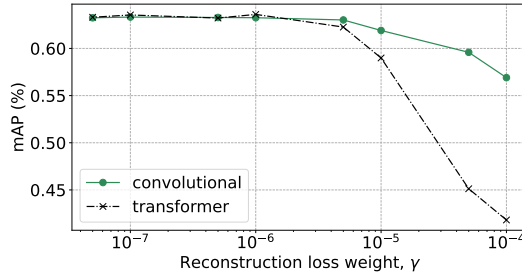


Figure 9: The multi-label classification mAP vs. reconstruction loss weight γ for convolutional and transformer-based decoders. Convolutional decoder has the configuration described in Section C.3 and the transformer-based uses 16 layers and 8 attention heads at each layer.

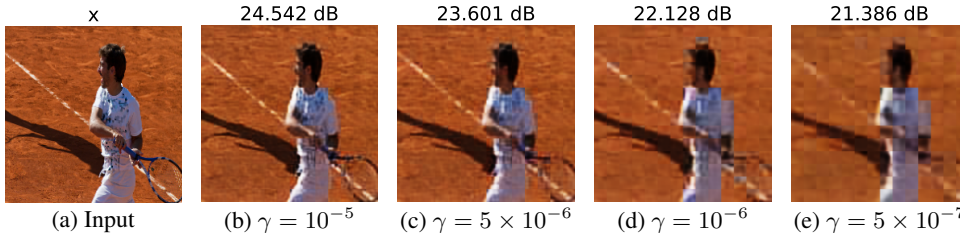


Figure 10: Input image and sample reconstructions with PSNR (in dB) values from the validation set for transformer-based decoder with different reconstruction loss weights γ .