Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels

Benoit Dufumier[1,2*] **Pietro Gori**[2] **Julie Victor**[1] **Antoine Grigis**[1] **Edouard Duchesnay**[1]
[1]NeuroSpin, CEA Saclay, Université Paris-Saclay, France
[2]LTCI, Télécom Paris, IPParis, France
benoit.dufumier@cea.fr

Abstract

Contrastive Learning has shown impressive results on natural and medical images, without requiring annotated data. However, a particularity of medical images is the availability of meta-data (such as age or sex) that can be exploited for learning representations. Here, we show that the recently proposed contrastive y-Aware InfoNCE loss, that integrates multi-dimensional meta-data, asymptotically optimizes two properties: *conditional alignment* and *global uniformity*. Similarly to [33], conditional alignment means that similar samples should have similar features, but conditionally on the meta-data. Instead, global uniformity means that the (normalized) features should be uniformly distributed on the unit hypersphere, independently of the meta-data. Here, we propose to define *conditional uniformity*, relying on the meta-data, that repel only samples with dissimilar meta-data. We show that direct optimization of both conditional alignment and uniformity improves the representations, in terms of linear evaluation, on both CIFAR-100 and a brain MRI dataset.

1 Introduction

Deep learning models have recently shown impressive results in medical imaging [2, 9, 10, 20], especially when large data-sets are available. However, in many medical applications (*e.g*, computer-aided diagnosis with MRI [34], chest X-ray on COVID-19 [30], etc.), data-sets are often small or heterogeneous and biased [32] (*e.g*, images coming from different hospitals) and annotations are costly.

Traditional transfer learning from natural images has been applied to several medical applications [1, 27], hypothesizing that features can be re-used or fine-tuned on the downstream tasks [35]. However, it does not always lead to better results [26], which might be due to the large domain gap between natural and medical images. On the contrary, self-supervised models do not rely on annotations and can be directly applied to big unlabeled medical data-sets, avoiding both the burden of expert annotations and the domain gap. These models are trained on a pretext task that incorporates prior information about the data and should estimate a relevant representation for the subsequent downstream tasks (*e.g*, the invariants [6], object shapes or colors [19, 21, 24], etc). Recent approaches are built upon Siamese networks [3], as in contrastive learning [6, 12], where an encoder is trained to map different views of the same image to the same point in the representation space while pushing away dissimilar images. In practice, the InfoNCE loss [6, 11, 22] is usually used and it has shown impressive results on both natural [6, 12] and medical [2, 4, 31] image classification and segmentation.

A particularity with medical images is the availability of meta-data (such as participant's age and sex) that often come freely with medical data-sets and that can be viewed as prior knowledge. As

*Corresponding author

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

observed in [4, 8], most of the methods in contrastive learning consider distinct images *equally* semantically different. However, two different images with close meta-data should probably share discriminative features, as opposed to images with very distinct meta-data. In [8], a new contrastive loss based on InfoNCE and integrating continuous meta-data has been proposed, namely y-Aware InfoNCE. It generalizes the SupCon loss [16] and has shown good results on brain MRI. We shall see its connection with the alignment and uniformity terms proposed in [33].

Contributions. First, we show that the y-Aware InfoNCE loss can be decomposed, asymptotically, into *conditional* alignment (pulling images with close meta-data y close together) and *global* uniformity (repelling all images independently of y, see Fig. 1). Second, remarking that images should not be uniformly repelled, and inspired by [7], we propose to re-define the uniformity term only between samples with dissimilar meta-data (ideally coming from different latent classes), calling it conditional uniformity. We empirically demonstrate that going from global to conditional uniformity allows the encoder to learn a better transferable representation on both CIFAR-100 and a brain MRI dataset.

2 Problem Formulation and y-Aware InfoNCE estimator

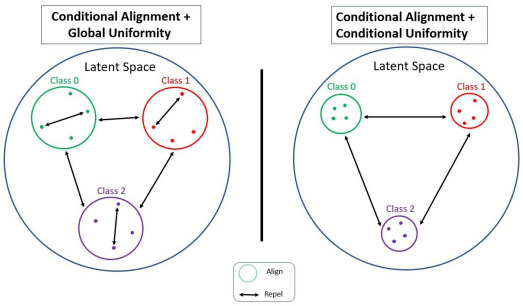

Figure 1: Proposed conditional alignment and uniformity terms with categorical meta-data. The conditional uniformity repel only points from different classes and not within the same class.

In contrastive learning, the goal is to learn an embedding $f_\theta : \mathcal{X} \mapsto \mathcal{S}^d$ that maps similar samples $(x, x^+) \in \mathcal{X} \times \mathcal{X}$ to close representations $f_\theta(x), f_\theta(x^+)$ but also different samples $(x, x^-) \in \mathcal{X} \times \mathcal{X}$ to distant representations $f_\theta(x), f_\theta(x^-)$, on a unit hyper-sphere $\mathcal{S}^d$. Similarly to [8], we assume that each input sample x has a proxy label $y \in \mathbb{R}^p$ giving prior knowledge about x (such as participant's age and/or sex) and coming from the joint distribution $p(x, y)$.

For an anchor $(x, y) \sim p(x, y)$ and N other samples $(x_k, y_k) \sim p(x, y)$, the y-Aware InfoNCE [8] is defined as:

$$\mathcal{L}^y_{NCE} = -\frac{1}{N}\sum_{k=1}^{N} \frac{w_\sigma(y, y_k)}{\hat{Z}_\sigma(y)} \log \frac{e^{f_\theta(x,x_k)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f_\theta(x,x_j)}} \tag{1}$$

where $\hat{Z}_\sigma(y) = \frac{1}{N}\sum_{j=1}^{N} w_\sigma(y, y_j)$, $f_\theta(x_1, x_2) = \frac{1}{\tau} f_\theta(x_1)^T f_\theta(x_2)$ and w_σ is a Gaussian kernel that measures similarity between the proxy labels y. Similarly to [33], this loss can be re-written as:

$$\mathcal{L}^y_{NCE} = \underbrace{-\frac{1}{N}\sum_{k=1}^{N} \frac{w_\sigma(y, y_k)}{\hat{Z}_\sigma(y)} f_\theta(x, x_k)}_{\text{Conditional Alignment}} + \underbrace{\log \frac{1}{N}\sum_{j=1}^{N} e^{f_\theta(x,x_j)}}_{\text{Global Uniformity}} \tag{2}$$

Lemma 1. *As the number of samples $N \to \infty$, the y-Aware InfoNCE loss $\mathcal{L}^y_{NCE}$ converges to:*

$$\lim_{N\to\infty} \mathcal{L}^y_{NCE} = -\underbrace{\mathbb{E}_{p_{pos}(x,x^+)}(f_\theta(x, x^+))}_{\mathcal{L}^y_{align}} + \underbrace{\mathbb{E}_{x\sim p(x)} \log \mathbb{E}_{x'\sim p(x')}(e^{f_\theta(x,x')})}_{\mathcal{L}_{unif}} \tag{3}$$

where $p_{pos}(x, x^+) = \int p(x|y)p(x^+|y^+)p^+_y(y^+)p(y)dydy^+$ is a positive distribution satisfying $p_{pos}(x) = p(x)$ and $p^+_y(y^+) = \frac{1}{Z_\sigma(y)} w_\sigma(y, y^+)p(y^+)$ quantifies the similarity between y and y^+. $Z_\sigma(y) = \mathbb{E}_{p(y^+)}(w_\sigma(y, y^+))$ is a normalizing constant. See the Appendix for a proof.

Similarly to [33], the alignment term brings all positive pairs (x, x^+) close together in $\mathcal{S}^d$ while the uniformity term is optimal when all points $f_\theta(x)$ for $x \sim p(x)$ are uniformly distributed on the hyper-sphere. In our formulation, we know this is hardly true since all points with close meta-data (measured by $p^+_y(y^+)$) should be close in $\mathcal{S}^d$. Instead of repelling uniformly all points from one another, we propose to *weight* the repulsion between points with the meta-data, as in the conditional alignment. To this end, we introduce a negative distribution $p_{neg}(x, x^-)$ (see below).

Conditional Uniformity. As proved in [33], the minimizers f_θ of $\mathcal{L}_{unif}$ minimize $\log \mathbb{E}_{x,x' \sim p(x)p(x')}(e^{f_\theta(x,x')})$. Instead of sampling independently x, x', we propose to draw them from $p_{neg}(x, x^-) := \int p(x|y)p(x^-|y^-)p_y^-(y^-)p(y)dydy^-$ where $p_y^-(y^-) = \frac{||w_\sigma||_\infty - w_\sigma(y,y^-)}{||w_\sigma||_\infty - Z_\sigma(y)}p(y^-)$. The infinite norm ensures the positiveness of our distribution and the kernel $w_\sigma(y, y^-)$ makes a direct connection with $p_y^+(y^+)$. We then define:

$$\mathcal{L}_{unif}^y = \log \mathbb{E}_{(x,x^-)\sim p_{neg}(x,x^-)}(e^{f_\theta(x,x^-)}) \tag{4}$$

that can be empirically estimated using: $\hat{\mathcal{L}}_{unif}^y = \log \frac{1}{N^2} \sum_{i,j} \frac{||w_\sigma||_\infty - w_\sigma(y_i,y_j)}{||w_\sigma||_\infty - \hat{Z}_\sigma(y_i)} e^{f_\theta(x_i,x_j)}$.

3 Experiments and Results

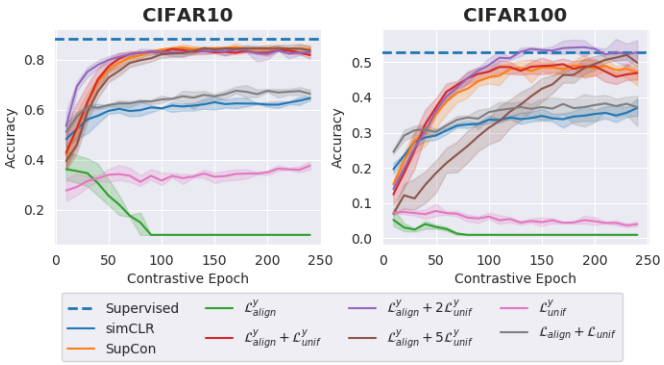

Figure 2: Linear Evaluation of ResNet18 pre-trained on CIFAR 10 and 100.

CIFAR. We pre-trained a ResNet18 [13] on CIFAR 10 and 100 [18] by considering the true labels as meta-data and we compared our approach with different baselines: SimCLR (unsupervised) [6], SupCon (supervised with labels) [16] and Alignment and Uniformity (unsupervised) [33] (see Appendix for experimental details). We reported in Fig. 2 the representation quality under the linear evaluation protocol. We observed a better convergence speed on CIFAR10 and a better representation on CIFAR100.

Brain MRI. In a real case scenario, following the experimental setup in [8], we pre-trained a DenseNet121 [15] on BHB-10K, a large multi-site brain MRI dataset comprising 10^4 3D scans of healthy controls. We used age and sex as meta-data with the associated kernel $\mathbf{w}_\sigma((age_1, sex_1), (age_2, sex_2)) = w_\sigma(age_1, age_2)\delta_{sex_1=sex_2}$. We then assessed the representation quality on BIOBD [14, 28], to discriminate between bipolar disorder patients and healthy controls.

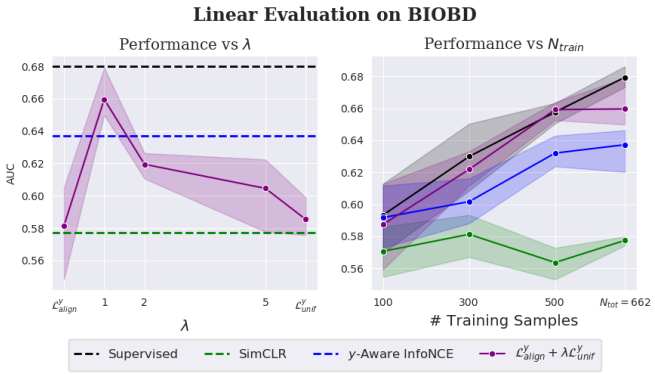

Figure 3: Linear Evaluation on BIOBD after pre-training on BHB-10K, varying λ (left image with $N_{train} = 662$) and the number of training samples (right image with $\lambda = 1$).

4 Conclusion

By decomposing the recently proposed y-Aware InfoNCE loss into conditional alignment and global uniformity, we established a connection with the framework recently proposed in [33]. We also demonstrated that a conditional uniformity term is preferable and lead to a better representation quality on both CIFAR 100 and a brain MRI dataset. Future work will consist in validating the proposed framework on various modalities (*e.g* CT and X-ray) and applying transfer learning from the learnt representations.

5 Broader Impact

Including continuous and categorical meta-data into a contrastive loss should enrich the learnt representation making it more discriminative for any downstream task. This should particularly benefit the applications with small data-sets such as rare diseases or private clinical data-sets. Furthermore,

the proposed loss could be also adapted for multi-modal data or for data-sets with missing imaging data, thus increasing and broadening its clinical impact and applicability.

References

[1] Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, J Santamaría, Ye Duan, and Sameer R Oleiwi. Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10(13):4523, 2020.

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

[3] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33, 2020.

[5] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, pages 416–422, 2001.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran Associates, Inc., 2020.

[8] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michel Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm McDonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International conference on medical image computing and computer-assisted intervention*. Springer, 2021.

[9] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[10] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[11] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Franz Hozer, Samuel Sarrazin, Charles Laidi, Pauline Favre, Melissa Pauling, Dara Cannon, Colm McDonald, Louise Emsell, Jean-François Mangin, Edouard Duchesnay, et al. Lithium prevents grey matter atrophy in patients with bipolar disorder: an international multicenter study. *Psychological medicine*, pages 1–10, 2020.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.

[20] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[21] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[26] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[27] Jaakko Sahlsten, Joel Jaskari, Jyri Kivinen, Lauri Turunen, Esa Jaanio, Kustaa Hietala, and Kimmo Kaski. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific reports*, 9(1):1–11, 2019.

[28] Samuel Sarrazin, Arnaud Cachia, Franz Hozer, Colm McDonald, Louise Emsell, Dara M Cannon, Michele Wessa, Julia Linke, Amelia Versace, Nora Hamdani, et al. Neurodevelopmental subtypes of bipolar disorder are related to cortical folding patterns: An international multicenter study. *Bipolar disorders*, 20(8):721–732, 2018.

[29] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019.

[30] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.

[31] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems*, volume 33, pages 18158–18172, 2020.

[32] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. *arXiv preprint arXiv:2103.02023*, 2021.

[33] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[34] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, page 101694, 2020.

[35] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

A Experimental Details

CIFAR-10 and 100. For all experiments, we used a batch size $b = 1024$ with a latent dimension $d = 128$. We optimized the ResNet18 with ADAM [17] and a learning rate $\alpha = 10^{-3}$. We used a small weight decay 0.00005. All images in a batch are transformed twice as in [6] through a set of transformations $\mathcal{T}$ implemented in PyTorch [23] and including: Random Crop, Horizontal Flip, Color Jittering and Random Grayscale, following [6]. We used the standard testing split as defined in [18] and we split the training set into training/validation with $n_{train} = 40000$ and $n_{val} = 10000$ randomly sub-sampled and stratified. The standard deviation in Fig. 2 is obtained by repeated all experiments 3 times with different random initialization.

Brain MRI. We followed the same experimental setting as [8] for pre-training with a batch size $b = 64$ and a learning rate $\alpha = 10^{-4}$ reduced by $\gamma = 0.9$ every 10 epochs. We trained the model for 50 epochs and evaluate the learned representation using a Logistic Regression implemented with scikit-learn [25]. For the transformations $\mathcal{T}$, we used random crop, random cutout, gaussian blur, gaussian noise and flip, as in [8]. To obtain the standard deviation Fig. 3, we performed 5-fold cross-validation on BIOBD.

BIOBD. This dataset contains $N = 662$ MRI scans acquired on 8 different sites with 356 HC and 306 patients with bipolar disorder, as described in [8].

B Proof of Lemma 1

In order to prove the lemma 1, we first give a natural way of introducing the distribution p_{pos}, quantifying the similarity between positive samples (x, x^+) and we then re-write the expectation under this distribution to derive the empirical estimator corresponding to the conditional alignment.

Discrete case. As in [29], we may define the notion of similarity through a distribution $p(c)$ defined over discrete latent classes $\mathcal{C}$ and a positive distribution $p^+(x, x^+)$: points (x, x^+) are similar if and only if they share the same latent class c, *i.e*, $x, x^+ \sim p^+(x, x^+) := \mathbb{E}_{p(c)}(p(x|c)p(x^+|c))$. In the fully supervised setting, the latent classes $\mathcal{C}$ are known and correspond to the available annotations $\mathcal{Y}$.

General case. In general, we do not know the latent classes $\mathcal{C}$ but we may replace them with the proxy labels $y \in \mathbb{R}^d$. Rewriting the positive distribution as:

$$p^+(x, x^+) = \int p(x|c)p(x^+|c^+)d\delta_c(c^+)p(c)dc$$

We may approximate this distribution by replacing the true latent class c with the *proxy* label y and the density $\delta_y(y^+)$ by an estimate $p_y^+(y^+)$ in the vicinity of y (following [5]):

$$p_{pos}(x, x^+) = \int p(x|y)p(x^+|y^+)p_y^+(y^+)p(y)dydy^+ \tag{5}$$

Equipped with this distribution, and by remarking that the marginal distribution $p_{pos}(x) = \int p_{pos}(x, x^+)dx^+ = p(x)$, we can derive a lower bound of the mutual information between x and x^+, $I(x, x^+) := KL(p_{pos}(x, x^+)||p(x)p_{pos}(x^+))$ with the InfoNCE estimator [11, 22]:

$$I_{NCE}^y = \mathbb{E}_{\substack{(x,x^+)\sim p_{pos}(x,x^+) \\ (x_i^-)_{i=1}^{N-1}\sim p(x^-)}} \log \frac{e^{f_\theta(x,x^+)}}{e^{f_\theta(x,x^+)} + \sum_{i=1}^{N-1} e^{f_\theta(x,x_i^-)}} + \log(N) \tag{6}$$

where $f_\theta(x_1, x_2) = \frac{1}{\tau} f_\theta(x_1)^T f_\theta(x_2)$ and $\tau > 0$ is the temperature.

y-Aware InfoNCE [8] We consider a Gaussian kernel function as a special case of density $p_y^+(y^+) = \frac{1}{Z_\sigma(y)} w_\sigma(y, y^+)p(y^+)$ (where $Z_\sigma(y)$ is a normalization constant and w_σ is the Radial Basis Function kernel) to draw a connection with the recently introduced y-Aware InfoNCE loss [8].

Let $g(x, x^+, (x_i^-)_{i=1}^{N-1}) = \log \frac{e^{f_\theta(x,x^+)}}{e^{f_\theta(x,x^+)} + \sum_{i=1}^{N-1} e^{f_\theta(x,x_i^-)}}$. We can re-write the previous I_{NCE}^y estimator as:

$$
\begin{aligned}
I_{NCE}^y &= \mathbb{E}_{\substack{(x,x^+)\sim p_{pos}(x,x^+) \\ (x_i^-)_{i=1}^{N-1}\sim p(x^-)}} g(x, x^+, (x_i^-)_{i=1}^{N}) + \log(N) \\
&= \mathbb{E}_{\substack{p(x,y)p(x^+|y^+)p_y^+(y^+) \\ (x_i^-)_{i=1}^{N-1}\sim p(x^-)}} g(x, x^+, (x_i^-)_{i=1}^{N}) + \log(N) \\
&= \mathbb{E}_{\substack{p(x,y)p(x^+,y^+) \\ (x_i^-)_{i=1}^{N-1}\sim p(x^-)}} \frac{1}{Z_\sigma(y)} w_\sigma(y, y^+) g(x, x^+, (x_i^-)_{i=1}^{N}) + \log(N) \\
&= \mathbb{E}_{\substack{p(x,y)p(x^+,y^+) \\ (x_i^-)_{i=1}^{N-1}\sim p(x^-)}} \frac{1}{Z_\sigma(y)} w_\sigma(y, y^+) \log \frac{e^{f_\theta(x,x^+)}}{\frac{1}{N}\left(e^{f_\theta(x,x^+)} + \sum_{i=1}^{N-1} e^{f_\theta(x,x_i^-)}\right)}
\end{aligned}
$$

This shows that y-aware InfoNCE loss $\mathcal{L}_{NCE}^y$ [8] is the empirical estimator of $-I_{NCE}^y$. Then, we can decompose I_{NCE}^y into 2 terms:

$$
I_{NCE}^y = \mathbb{E}_{p_{pos}(x,x^+)} f_\theta(x, x^+) - \mathbb{E}_{\substack{(x,x^+)\sim p_{pos} \\ (x_i^-)_{i=1}^{N}\sim p(x^-)}} \log \frac{1}{N}(e^{f_\theta(x,x^+)} + \sum_{i=1}^{N-1} e^{f_\theta(x,x_i^-)})
$$

Finally, by the same proof describe in [33] (Appendix A.2), using the Strong Law of Large Numbers, the Continuous Mapping Theorem and the marginal $p_{pos}(x) = p(x)$, we have:

$$
\lim_{N\to\infty} \mathbb{E}_{\substack{(x,x^+)\sim p_{pos} \\ (x_i^-)_{i=1}^{N}\sim p(x^-)}} \log \frac{1}{N}(e^{f_\theta(x,x^+)} + \sum_{i=1}^{N-1} e^{f_\theta(x,x_i^-)}) = \mathbb{E}_{x\sim p(x)} \log \mathbb{E}_{x'\sim p(x')}(e^{f_\theta(x,x')})
$$

This finally terminates the proof that:

$$
\lim_{N\to\infty} \mathcal{L}_{NCE}^y = -\lim_{N\to\infty} I_{NCE}^y = -\mathbb{E}_{p_{pos}(x,x^+)}(f_\theta(x, x^+)) + \mathbb{E}_{x\sim p(x)} \log \mathbb{E}_{x'\sim p(x')}(e^{f_\theta(x,x')})
$$