

MORE SYNERGY, LESS REDUNDANCY: EXPLOITING JOINT MUTUAL INFORMATION FOR SELF-SUPERVISED LEARNING

Salman Mohamadi, Gianfranco Doretto, Donald A. Adjeroh

Lane Department of Computer Science and Electrical Engineering, West Virginia University, USA

ABSTRACT

Self-supervised learning (SSL) is now a serious competitor for supervised learning, even though it does not require data annotation. Several baselines have attempted to make SSL models exploit information about data distribution, and less dependent on the augmentation effect. However, there is no clear consensus on whether maximizing or minimizing the mutual information between representations of augmentation views practically contribute to improvement or degradation in performance of SSL models. This paper is a fundamental work where, we investigate the role of mutual information in SSL, and reformulate the problem of SSL in the context of a new perspective on mutual information. To this end, we consider joint mutual information from the perspective of partial information decomposition (PID) as a key step in **reliable multivariate information measurement**. PID enables us to decompose joint mutual information into three important components, namely, unique information, redundant information and synergistic information. Our framework aims for minimizing the redundant information between views and the desired target representation while maximizing the synergistic information at the same time. Our experiments lead to a re-calibration of two redundancy reduction baselines, and a proposal for a new SSL training protocol. Experimental results on multiple datasets and two downstream tasks show the effectiveness of this framework.

1. INTRODUCTION

Self-supervised learning (SSL) is among very successful ML principles that are needless of huge labeled datasets [1]. While deep learning has shown tremendous success in many domains and applications including computer vision [2], biometrics [3], genomics [4], etc, data-efficiency has been the focus of few problem domains such as deep active learning [5, 6], and SSL [7]. Essentially, SSL frameworks consist of two key elements, namely, loss function, and pretext task [8]. Basically, the pretext task is a proxy task which is to be solved using a supervisory signal from the unlabeled data, guided by an objective (loss) function [8]. Loss functions on the other hand provide guidance in learning the representation of a given sample by comparing two or multiple augmented views of the same sample with each other or with views of other samples. In fact, early baselines known as contrastive baselines were developed around the idea of contrasting aug-

mented views of a sample with each other (positive pairs) and also with the views from other samples (negative pairs) [9, 10, 11, 7, 12]. This type of baselines, however, suffer from the problem of potential representation collapse, as well as the need for large negative batches for effective representation. Next generation of baselines emerged as non-contrastive or negative pair-free baselines [13, 14], essentially eliminating the need to contrast against negative views (negative pairs), and also almost with no risk of representation collapse. There is also a class of baselines known as clustering baselines (see [15]), primarily based on clustering views of samples in the latent space. Two most recent baselines are based on redundancy reduction in representation of augmented views of the samples [16, 17]. This class of approaches mainly suggests that whitening the latent/embedding space of the a pair of networks trained on augmented views of samples allows for reducing redundant information in representation of the sample [5]. Later theoretical work on whitening baselines showed that the prime reason for their success is eliminating another type of collapse, dimensional collapse [18, 19].

In this work, we assess how this whitening process unwittingly eliminates the synergistic information along with redundant information. This relates to a larger controversy on how mutual information relates to learning the target representation. Hence, in this paper, we start with investigating long-standing ambiguity about the role of mutual information in SSL. This eventually leads us to reconsider the problem of mutual information between two variables (two views of a sample) by reformulating it as joint mutual information between three variables (two views and the target representation). To elaborate on the controversy, the general idea is to maximize the mutual information between the encoder-representation of two augmented views for better representation; however some work [20, 21] suggested that more mutual information does not necessarily improve the representation. A recent work based on Info-Min principle suggests that, in fact, less mutual information between augmented views along with more task-associated information would improve the representation using a certain augmentation setting [22]. Another recent work [23] acknowledges the questionable role of mutual information, and suggests that decomposing the estimation of mutual information by adding an extra term representing the condition on the image with some blocked patches

would reinforce the role of mutual information. However, this work is different from our work as they decompose the estimation of two-variable mutual information, whereas we focus on three-variable joint mutual information decomposition [24]. In fact, we seek out the solution in the theory of partial information decomposition (PID). Eventually, this leads us to decompose the joint mutual information into its integral components, i.e., unique, redundant and synergistic components as was first introduced in [25]. In the following, we first state the problem and discuss the decomposition of joint mutual information, then re-define SSL in this new context. We elaborate on the SSL baselines that rely on redundancy reduction, and propose a new training protocol for such SSL models, then empirically evaluate the new protocol.

2. METHODS

2.1. Problem Statement

From an information theoretic perspective, the general, though controversial, idea is that SSL frameworks generally tend to maximize the mutual information between encoder representation $f(\cdot)$ of two augmented views x_1 and x_2 of sample data x upper bounded by $I(x_1; x_2)$, i.e., $I(f(x_1); f(x_2)) \leq I(x_1; x_2)$ [23, 26]. This objective comes with challenges including how to optimally generate x_1 and x_2 [22] for actionable mutual information, as well as how to reduce redundant information in the representation [17, 16]. To elaborate on the former challenge, Tian et al [22] suggested an heterodox idea, indicating that the augmentation process for generating views should be modified in a way that will enable reducing the mutual information between representation of positive views without affecting task-relevant information, i.e., mutual information is not necessarily task-relevant information. The later challenge, on the other hand, suggests that whitening the latent/embedding space would reduce redundant information. However, we argue that rather than focusing on mutual information between the representation of augmented views', the joint mutual information between views' representations and the target representation could provide a possible way to resolve this controversy. Hence, we take a different approach by formulating the core of SSL in terms of **joint mutual information between views and the target representation**. This leads us to the observation that, even though rigorous redundancy reduction through whitening such as in [16] drops redundant information, it also risks reduction of useful synergistic information. This motivates us to design experiments to assess this claim in Sec. 2.4, and then to offer a training protocol to alleviate this loss of the synergistic element in joint mutual information. Specifically, we find it necessary to revisit the SSL principle from the joint mutual information perspective. Therefore, we assess two most recent baselines, Barlow-Twins [16] and W-MSE [17] which aim for redundancy reduction. Below we elaborate on joint mutual information (in contrast with mutual information) and then we investigate two most recent baselines on whitening, which are also most

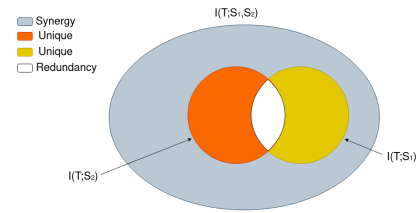


Fig. 1. Partial information decomposition in case of three variables.

relevant baseline to study redundancy and synergy.

2.2. Decomposing Joint Mutual Information

For the first time ever we consider the general SSL problem setting from the viewpoint of PID, which has diverse practical applications including in neuroscience, game theory and statistical learning. Hence, first we present the PID introduced in [25] and then reformulate the SSL accordingly. We note that PID is not the only approach to multivariate measurement of information. However, it has multiple advantages in our SSL context, including non-negative decomposition of information as well as separate and simultaneous measurement of redundancy and synergy as distinct quantities [27]. This new interpretation of SSL is primarily posed to address the ambiguity in the role of mutual information in SSL.

The PID is an approach to a non-overlapping decomposition of the joint mutual information between two sets of variables, a set of two or more source variables carrying information about a target, as well as the single target variable. This decomposition has been challenging as the proposed solutions mostly consisted of negative information terms, until a breakthrough work in [25] which introduced a non-negative decomposition in terms of quantifying three components, the unique, redundant, and synergistic information.

In its simplest form, suppose we have two source variables S_1 and S_2 carrying joint mutual information $I(T; S_1, S_2)$ about a target variable T . Hence each of the source variables has mutual information with the target variable. Decomposing the joint mutual information into some non-negative components, models information interaction to assess the contribution offered by each source variable and combination of sources. According to [25], as shown in Fig. 1 the joint mutual information between sources and target, could be decomposed as three elements, unique, redundant, and synergistic information. Unique information is the part provided by each source separately, redundant information is the minimum information provided by each source (aka common mutual information), and synergistic information is the information provided only by a combination of S_1 and S_2 about T , which neither alone can provide [24].

$$I(S_1, S_2 : T) = \text{Redundancy}(T; S_1, S_2) + \text{Synergy}(T; S_1, S_2) + \text{Unique}(T; S_1) + \text{Unique}(T; S_2) \quad (1)$$

Now consider the general setting of SSL, where at least two random augmented views of a sample are generated. The goal

is to contrast them in order to learn a representation that is maximally informative about the original sample distribution, while minimally informative about the augmentation. This contrast in essence creates an information interaction between the information of the variables which could be studied under the PID framework. Here, the two augmented views could be seen as source variables S_1 and S_2 , whereas the original sample distribution is the target variable T . In a more general sense, T could be considered the class distribution representing the invariant representation of the views of a given sample, i.e., the class the data sample belongs to. Here, as only redundant and synergistic information will be the results of interaction in contrasting views in SSL frameworks, unique information is not the subject matter of our study in this work. Unique information would be the subject of non-contrastive supervised learning on labeled data.

2.3. Redundancy Reduction Baselines

Interestingly, two most recent SSL baselines [17, 16] are redundancy reduction (aka hard/soft whitening) baselines. Both baselines take advantage of whitening (Cholskey whitening) of latent/embedding space of a cross-correlation matrix computed from augmented views of the same sample. Ermolov et al [17] proposed a hard whitening method based on a recent version of Cholesky decomposition [28, 29] for whitening the latent space vectors. At the same time, Zbontar et al [16] has gained more popularity by proposing a simpler process called soft whitening, which essentially forces the cross-correlation matrix of the embedding vectors of two networks to identity matrix. The later approach, known as Barlow-Twins, suggests that their whitening approach intuitively results in redundancy reduction embedded in off-diagonal elements of the cross-correlation matrix. We use both approaches for our investigation, and provide further insight on synergy versus redundancy. However due to the lack of space we only represent the theoretical reformulation of Barlow-Twins under our framework, as it is more popular. The following is the loss function of Barlow-Twins:

$$\mathcal{L}_{BT} \triangleq \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} (C_{ij})^2 \quad (2)$$

$$C_{ij} \triangleq \frac{\sum_m z_{m,i}^A z_{m,j}^B}{\sqrt{\sum_m (z_{m,i}^A)^2} \sqrt{\sum_m (z_{m,j}^B)^2}} \quad (3)$$

where C_{ij} are elements of the cross-correlation matrix C between the embedding vectors with element z of two networks (twins), as presented in Eq. 3. λ as a weighting factor, originally set to 5×10^{-3} .

2.4. Assessing synergy and redundancy

In order to provide a context for PID with respect to SSL, we find it necessary to design simple experiments around redundancy reduction and synergy in Barlow-Twins (BT). Note that as the augmented views for a sample generated under standard augmentation for SSL share lots of information in common (redundant or commonly known as mutual information),

BT attains desirable performance by implementing rigorous redundancy reduction. However we argue that if the redundant information was not as much, the performance would drop sharply. To assess this, we apply heavy augmentation on samples (such as [30]) to generate views with significantly less redundant information, and then test BT performance on these. The top-1 accuracy for CIFAR10 and CIFAR100 (under experimental settings in next section) drops by %5.69 and %5.13 respectively. Now under same heavy augmentation, we re-calibrate BT by setting $\lambda = 0.1$ and also forcing off-diagonal elements to a multivariate Gaussian $\mathcal{N}(0, 1)$ rather than zero to allow them to better affect the learned representation. We gain accuracy, +%0.91, and +%0.81 compared with the former case. This implies that the off-diagonal elements not only carry redundant information, but also some other type of information. Otherwise allowing more redundancy by using multivariate Gaussian off-diagonal elements would have degraded the performance. We argue that off-diagonal elements do not only represent redundant information, **but also synergistic information**. This is why when we reduce the redundant information by implementing heavy augmentation, BT's rigorous redundancy reduction constraint on off-diagonal elements of the cross-correlation matrix, degrades the performance by targeting synergistic information. Below, we propose a training protocol that works even better than forcing off-diagonal elements to multivariate Gaussian, and present our experimental results on two baselines BT and W-MSE in Sec. 4 to show the generality of our framework.

3. SYNERGY-BASED TRAINING PROTOCOL

We aim for re-calibrating the redundancy reduction in BT [16] and W-MSE [17] toward *protecting* the most synergistic information during the redundancy reduction process. In its current form, BT approach does not seem to optimally reduce redundancy, without significant loss in the synergistic component. Our approach consists of a serial pre-training with first phase of dropping redundancy and second phase of adding synergy. Hence, in this section, we define a new training protocol aiming for extracting more synergistic information during the process of redundancy reduction which will be implemented on both BT and W-MSE. We present this protocol aimed at more synergy and less redundancy via the use of engineered off-diagonal elements, to show the effectiveness of the joint mutual information decomposition in SSL. As the augmented views of a sample under standard augmentation share lots of mutual information, we find it practically more efficient to update/replace the loss function of BT and W-MSE after initial pre-training with the original loss function which solely aims at redundancy reduction. This is done under a new training protocol with two phases of pre-training in two different settings. First phase aims at reducing the redundancy, while the second phase aims at adding to synergy. Below we only present the new formulation for BT, however, we provide the experimental results for both BT and W-MSE.

A. Gaussian off-diagonal: After initial pre-training of original model, here BT, the network is fixed, to resume the training with an updated loss. For BT we set $\lambda = 0.1$ and replace the second term in Eq. 2 with $\lambda \sum_i \sum_{j \neq i} (C_{ij} - G_{ij})^2$ where G_{ij} are the multivariate Gaussian elements of a square matrix G of proper size. This allows the BT to better consider the off-diagonal elements of the cross-correlation matrix, which convey synergy and redundancy.

B. Reinforced off-diagonal: After initial pre-training of original model, here BT, the network is fixed and the average $C_{ij}^{Ave} = \frac{1}{n} \sum_n C_{ij}$ over all n samples will be computed. Then training resumes with new $\lambda = 0.1$ and the second term in Eq. 2 updated as $\lambda \sum_i \sum_{j \neq i} (C_{ij} - C_{ij}^{Ave})^2$ forcing each off-diagonal element to its corresponding average.

4. EXPERIMENTS AND RESULTS

4.1. Experiments

Baselines: Our modification on BT and W-MSE [16, 17] resulted in GSBT and RSBT, as well as GSW-MSE and RSW-MSE respectively. We perform experiments using our new training protocol under standard and heavy data augmentation. We contrast it with most recent baselines including Whitening-MSE ($d = 4$) [17], a non-contrastive baseline BYOL [13], and a clustering-based baseline SwAV [15]. Following [17], latent spaces of all methods are L_2 -normalized.

Dataset and augmentation: We use six datasets including ImageNet [31], CIFAR10, CIFAR100 [32], Tiny ImageNet [33], ImageNet-100, and VOC0712. We use two sets of augmentation protocols, standard and heavy. For standard augmentation including random grayscale, random crop, color jittering, aspect ratio adjustment, and horizontal mirroring, we follow the settings in [7], and for heavy augmentation we follow the settings in [30].

Network & implementation details: For CIFAR10/100, following the details of each baseline [7, 13, 14, 15, 17, 16], we use ResNet18, while for ImageNet, Tiny ImageNet, and VOC0712 we use ResNet50 [34], for the encoder and the same projector head as [16], with the same size of projector output in all baselines. For VOC0712 similar to [16], Faster R-CNN [35] is used. Optimization of all experiments were done using Adam optimizer [36]. Pre-training of RSBT, GSBT as well as RSW-MSE and GSW-MSE are performed in two phases, a phase one (redundancy reduction) consists of 500 epochs with batch size of 1024, which starts with a learning rate of 0.15 for some 20 epochs and drops to 0.001 for the remaining epochs. Phase two (synergy addition) also consists of another 500 epochs with the learning rate of 0.001, with their modified loss functions. The weight decay in both phases and all other experiments is 10^{-6} .

4.2. Evaluation and results

Similar to former methods, we perform the standard supervised linear evaluation for classification task as well as detection. Classification involves fixing the encoder weights after pre-training and replacing the projector with a linear

classifier (fully connected followed by softmax), and training the linear classifier for some 500 epochs on evaluation data, and then testing it. The classification results for ImageNet, CIFAR10/100, Tiny ImageNet, and ImageNet-100 with different settings of proposed training protocol are presented in the Tables 1, 2, and 3, whereas the detection results with VOC0712 is presented in Table 1. Results for modified BT using our protocol is presented in Table 1 and 2, whereas the results for modified W-MSE using our protocol are shown in Table 3. In both settings of data augmentation, our method outperforms prior approaches. While heavy augmentation degrade the performance of other approaches, it even improves the RSBT, GSBT, as well as RSW-MSE and GSW-MSE which shows robustness of our approach.

Table 1. Results of our methods on 2 downstream tasks – classification and object detection. Top-1 classification accuracy with supervised linear evaluation on ImageNet (Left), and Tiny ImageNet (Middle). Object detection results for VOC0712 (Right). Used both standard and heavy augmentations.

Framework	ImageNet (Class.)		Tiny ImageNet (Class.)		VOC0712 (Det.)	
	Standard	Heavy	Standard	Heavy	A _{All}	A ₅₀
BYOL	74.3	60.5 (↓)	51.43	47.16 (↓)	56.8	82.5
SwAV	71.8	58.9 (↓)	51.03	44.25 (↓)	56.1	82.6
W-MSE4	73.1	61.6 (↓)	50.59	48.11 (↓)	56.9	82.4
B-Twins	73.2	61.9 (↓)	50.63	47.49 (↓)	56.8	82.6
GSBT	75.4	75.5 (↑)	51.54	52.08 (↑)	57.3	82.6
RSBT	76.1	76.5 (↑)	51.94	52.46 (↑)	57.8	82.7

Table 2. Top-1 classification accuracy with supervised linear evaluation for CIFAR10/100 under both standard and heavy augmentations.

Framework	CIFAR10		CIFAR100	
	Standard	Heavy	Standard	Heavy
BYOL	91.81	84.11 (↓)	66.65	59.93 (↓)
SwAV	92.11	85.25 (↓)	67.77	60.89 (↓)
W-MSE4	91.89	87.74 (↓)	67.58	62.14 (↓)
B-Twins	92.33	86.64 (↓)	67.47	62.34 (↓)
GSBT	92.83	93.10 (↑)	67.81	68.23 (↑)
RSBT	93.03	93.47 (↑)	68.11	68.83 (↑)

Table 3. Experiments on another baseline, W-MSE, Top-1 classification accuracy with supervised linear evaluation for CIFAR100 and ImageNet-100.

Framework	CIFAR100		ImageNet100	
	Standard	Heavy	Standard	Heavy
W-MSE4	67.58	62.14 (↓)	79.02	71.14 (↓)
B-Twins	67.47	62.34 (↓)	77.93	72.57 (↓)
GSW-MSE	68.26	68.51 (↑)	79.58	79.91 (↑)
RSW-MSE	69.11	69.32 (↑)	79.93	80.66 (↑)

5. CONCLUSION

We address the ambiguity regarding how mutual information relates to better representation in SSL. To this end, we explore the use of PID in SSL and we re-define the formulation of SSL problem in terms of joint mutual information between three variables (two views of a sample and its original representation). This allows for recognition of synergistic information along with the redundant information and their role in boosting performance. We design and perform extensive experiments on the most recent redundancy reduction baselines, BT and W-MSE and instantiate the theoretical solution in practice under a new training protocol.

6. REFERENCES

- [1] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.
- [2] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al., "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.
- [3] Moktari Mostofa, Salman Mohamadi, Jeremy Dawson, and Nasser M Nasrabadi, "Deep gan-based cross-spectral cross-resolution iris recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 443–463, 2021.
- [4] Salman Mohamadi, Nasser M Nasrabadi, Gianfranco Doretto, and Donald A Adjeroh, "Human age estimation from gene expression data using artificial neural networks," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 3492–3497.
- [5] Salman Mohamadi, Gianfranco Doretto, and Don Adjeroh, "Deep active ensemble sampling for image classification," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4531–4547.
- [6] Salman Mohamadi and Hamidreza Amindavar, "Deep bayesian active learning, a brief survey on recent advances," *arXiv preprint arXiv:2012.08044*, 2020.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.
- [8] Salman Mohamadi, Gianfranco Doretto, and Donald A Adjeroh, "Fussl: Fuzzy uncertain self supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2799–2808.
- [9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [10] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive multiview coding," in *European Conference on Computer Vision*. Springer, 2020, pp. 776–794.
- [11] 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*, 2020, pp. 9729–9738.
- [12] Philip Bachman, R Devon Hjelm, and William Buchwalter, "Learning representations by maximizing mutual information across views," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [13] Jean-Bastien Grill, Florian Strub, Florent Althé, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent—a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.
- [14] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.
- [15] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.
- [16] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12310–12320.
- [17] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe, "Whitening for self-supervised representation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3015–3024.
- [18] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao, "On feature decorrelation in self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9598–9608.
- [19] Zixin Wen and Yuanzhi Li, "The mechanism of prediction head in non-contrastive self-supervised learning," *arXiv preprint arXiv:2205.06226*, 2022.
- [20] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh, "Tighter variational bounds are not necessarily better," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4277–4285.
- [21] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic, "On mutual information maximization for representation learning," *arXiv preprint arXiv:1907.13625*, 2019.
- [22] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, "What makes for good views for contrastive learning?," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6827–6839, 2020.
- [23] Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes, "Decomposed mutual information estimation for contrastive representation learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9859–9869.
- [24] Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh, "Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic," *Proceedings of the Royal Society A*, vol. 477, no. 2251, pp. 20210110, 2021.
- [25] Paul L Williams and Randall D Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.
- [26] Thomas M. Cover and Joy A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.
- [27] Nicholas Timme, Wesley Alford, Benjamin Flecker, and John M Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalist's perspective," *Journal of Computational Neuroscience*, vol. 36, no. 2, pp. 119–140, 2014.
- [28] Dariusz Dereniowski and Marek Kubale, "Cholesky factorization of matrices in parallel and ranking of graphs," in *International Conference on Parallel Processing and Applied Mathematics*. Springer, 2003, pp. 985–992.
- [29] Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe, "Whitening and coloring batch transform for gans," *arXiv preprint arXiv:1806.00420*, 2018.
- [30] Yalong Bai, Yifan Yang, Wei Zhang, and Tao Mei, "Directional self-supervised learning for heavy image augmentations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16692–16701.
- [31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.
- [32] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.
- [33] Ya Le and Xuan Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, pp. 3, 2015.
- [34] 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*, 2016, pp. 770–778.
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [36] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.