Self-supervised Learning from a Multi-view Perspective

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, Louis-Philippe Morency
Machine Learning Department, Carnegie Mellon University

Self-supervised learning (SSL) [12, 28, 35, 44] learns representations using a proxy objective (i.e., SSL objective) between inputs and self-defined signals. Despite great success in practice [8, 17], only a few work [3, 23, 38] provide theoretical insights into the learning efficacy of SSL. Our work formulates SSL from the perspectives of Information Theory [9] and multi-view representation learning. Specifically, we note that many previous approaches in self-supervised learning follow naturally from a multi-view perspective, where the input (e.g., original images) and the self-supervised signals (e.g., augmented images) can be seen as two redundant views of the data. Based on the multi-view perspective, our first contribution (Section 1) is to formally show that our learned representations can 1) extract all the task-relevant information (from the input) with a potential loss and 2) discard all the task-irrelevant information (from the input) with a fixed gap. As the second contribution (Section 2), our analysis 1) draws a connection between contrastive [4, 8, 28, 35] and predictive learning [12, 40, 42, 44] approaches for SSL; and 2) paves the way to broader interpretations on using SSL objectives to extract task-relevant and discard task-irrelevant information simultaneously.

1 A Multi-view Information-Theoretical Framework

Notations. For the input, we denote its random variable as X, sample space as $\mathcal{X}$, and outcome as x. We learn a representation (Z_X/ $\mathcal{Z}$/ z_x) from the input through a deterministic mapping F_X: $Z_X = F_X(X)$. For the self-supervised signal, we denote its random variable/ sample space/ outcome as S/ $\mathcal{S}$/ s. Two sample spaces can be different between the input and the self-supervised signal: $\mathcal{X} \neq \mathcal{S}$. The information required for downstream tasks is referred to as "task-relevant information": T/ $\mathcal{T}$/ t. Note that SSL has no access to the task-relevant information. Lastly, we use $I(A;B)$ to represent mutual information, $I(A;B|C)$ to represent conditional mutual information, $H(A)$ to represent the entropy, and $H(A|B)$ to represent conditional entropy for random variables A/B/C. We provide high-level takeaways for our main results in Figure 1. We defer all proofs to Appendix.

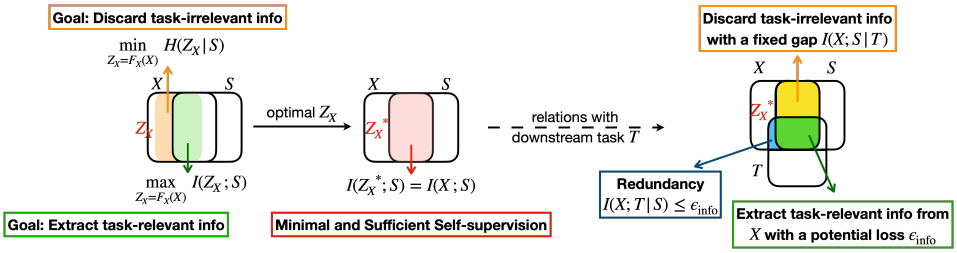

Figure 1: High-level takeaways for our main results using information diagrams. (a) We present to learn minimal and sufficient self-supervision: minimize $H(Z_X|S)$ for discarding task-irrelevant information and maximize $I(Z_X;S)$ for extracting task-relevant information. (b) The resulting learned representation $Z_X{}^*$ contains all task relevant information from the input with a potential loss ϵ_{info} and discards task-irrelevant information with a fixed gap $I(X;S|T)$. (c) Our core assumption: the self-supervised signal is approximately redundant to the input for the task-relevant information.

1.1 Multi-view Assumption

In our paper, we regard the input (X) and the self-supervised signals (S) as two views of the data. Here, we provide a table showing different X/S in various SSL frameworks:

Framework	BERT [12]	Look & Listen [2]	SimCLR [8]	Colorization [44]
Inputs (X)	Non-masked Words	Image	Image	Image Lightness
Self-supervised Signals (S)	Masked Words	Audio Stream	Same Image with Augmentation	Image Color

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

We note that not all SSL frameworks realize the inputs and the self-supervised signals as corresponding views, such as Jigsaw Puzzle [27] and Predicting Rotations [16]. However, SSL frameworks that regard X/S as two corresponding views [7, 8, 17] have a much better empirical downstream performance than the frameworks that do not [16, 27].

We then adopt the common *multi-view assumption* [34, 43] between input and self-supervised signal:

Assumption 1 (Multi-view, restating Assumption 1 in prior work [34])**.** *The self-supervised signal is approximately redundant to the input for the task-relevant information. In other words, there exist an* $\epsilon_{\text{info}} > 0$ *such that* $I(X;T|S) \leq \epsilon_{\text{info}}$.

Assumption 1 states that the task-relevant information lies mostly within the shared information between the input and the self-supervised signals, when ϵ_{info} is small. Such assumption is often satisfied in self-supervised visual contrastive learning [8, 20], where the input and the self-supervised signal are the same image with different augmentations. In addition, we point out the failure scenarios of Assumption 1 (i.e., when ϵ_{info} is large): the input and the self-supervised signal contain very different task-relevant information. For instance, a drastic image augmentation (e.g., adding large noise) may change the content of the image (e.g., the noise completely occludes the objects).

Similar to our approach, Tosh et al. [38] assumes strong independence between the downstream task and one view conditioning on the other view (i.e., $I(T;X|S) \approx 0$). On the contrary, Arora et al. [3] and Lee et al. [23] assume strong independence between the views conditioning on the downstream tasks (i.e. $I(X;S|T) \approx 0$). Previous studies [5, 13] have pointed out the latter assumption ($I(X;S|T) \approx 0$) is too strong and not likely to hold in practice.

1.2 Learning Minimal and Sufficient Representations for Self-supervision

We start by defining desirable representations for downstream supervision tasks. The Information Bottleneck (IB) method [1, 37] generalizes minimal sufficient statistics to the representations that are minimal (i.e., less complexity) and sufficient (i.e., better fidelity). To learn such representations for downstream supervision, we consider the following objectives:

Definition 1 (Minimal and Sufficient Representations for Downstream Supervision)**.** Let Z_X^{sup} be the sufficient supervised representation and $Z_X^{\text{sup}_{\text{min}}}$ be the minimal and sufficient representation:

$$Z_X^{\text{sup}} = \arg\max_{Z_X} I(Z_X;T) \text{ and } Z_X^{\text{sup}_{\text{min}}} = \arg\min_{Z_X} H(Z_X|T) \text{ s.t. } I(Z_X;T) \text{ is maximized.}$$

To reduce the complexity of the representation Z_X, our framework minimizes $H(Z_X|T)$ whereas prior methods [1, 37] minimize $I(Z_X;X)$. We note that minimizing $H(Z_X|T)$ leads to a more compressed representation (discarding redundant information)[1]. Specifically, minimizing $H(Z_X|T)$ reduces the randomness from T to Z_X which can be regarded as a form of incompressibility [6].

Next, we present SSL objectives to learn sufficient (and minimal) representations for self-supervision:

Definition 2 (Minimal and Sufficient Representations for Self-supervision)**.** Let Z_X^{ssl} be the sufficient self-supervised representation and $Z_X^{\text{ssl}_{\text{min}}}$ be the minimal and sufficient representation:

$$Z_X^{\text{ssl}} = \arg\max_{Z_X} I(Z_X;S) \text{ and } Z_X^{\text{ssl}_{\text{min}}} = \arg\min_{Z_X} H(Z_X|S) \text{ s.t. } I(Z_X;S) \text{ is maximized.}$$

Definition 2 defines our self-supervised representation learning strategy. Now, we are ready to associate the supervised and self-supervised learned representations:

Theorem 1 (Task-relevant information with a potential loss ϵ_{info})**.** *The supervised learned representations (i.e.,* Z_X^{sup} *and* $Z_X^{\text{sup}_{\text{min}}}$*) contain all the task-relevant information in the input (i.e.,* $I(X;T)$*). The self-supervised learned representations (i.e.,* Z_X^{ssl} *and* $Z_X^{\text{ssl}_{\text{min}}}$*) contain all the task-relevant information in the input with a potential loss* ϵ_{info}*. Formally,*

$$I(X;T) = I(Z_X^{\text{sup}};T) = I(Z_X^{\text{sup}_{\text{min}}};T) \geq I(Z_X^{\text{ssl}};T) \geq I(Z_X^{\text{ssl}_{\text{min}}};T) \geq I(X;T) - \epsilon_{\text{info}}.$$

[1] We do not claim $H(Z_X|T)$ minimization is better than $I(Z_X;X)$ minimization for reducing the complexity in the representations Z_X. In Appendix, we will show that $H(Z_X|T)$ minimization and $I(Z_X;X)$ minimization are interchangeable under our framework's setting.

Theorem 1 indicates that the learned representations can extract almost as much task-relevant information as the supervised one, when ϵ_{info} is small. On the other hand, when ϵ_{info} is non-trivial, the framework may not guarantee good downstream performance, as verified by [36, 39].

Theorem 2 (Task-irrelevant information with a fixed compression gap $I(X;S|T)$)**.** *The sufficient self-supervised representation (i.e., $I(Z_X^{\text{ssl}};T)$) contains more task-irrelevant information in the input than the sufficient and minimal self-supervised representation (i.e., $I(Z_X^{\text{ssl}_{\min}};T)$). The latter contains an amount of the information, $I(X;S|T)$, that cannot be discarded from the input. Formally,*

$$I(Z_X^{\text{ssl}};X|T) = I(X;S|T) + I(Z_X^{\text{ssl}};X|S,T) \geq I(Z_X^{\text{ssl}_{\min}};X|T) = I(X;S|T) \geq I(Z_X^{\text{sup}_{\min}};X|T) = 0.$$

Theorem 2 indicates that a compression gap (i.e., $I(X;S|T)$) exists when we discard the task-irrelevant information from the input, with $I(X;S|T)$ being the amount of the shared information between the input and the self-supervised signal excluding the task-relevant information. Hence, $I(X;S|T)$ would be large if the downstream tasks requires only a portion of the shared information.

2 Connections with Contrastive and Predictive Learning Objectives

We demonstrate that 1) popular SSL objectives, especially contrastive [4, 8, 17, 20, 28, 35] and predictive [12, 29, 30, 40, 42, 44] extract task-relevant information under our framework. We present 2) a new complementary *inverse predictive learning* objective to discard task-irrelevant information.

Contrastive Learning (is extracting task-relevant information). Contrastive learning objective [28] maximizes the dependency/contrastiveness between the learned representation Z_X and the self-supervised signal S, which suggests maximizing the the mutual information $I(Z_X;S)$. Theorem 1 suggests that maximizing $I(Z_X;S)$ results in Z_X containing (approximately) all the information required for the downstream tasks from the input X. The most common objective is the contrastive predictive coding (CPC) [28] , which is a mutual information lower bound with low variance [31, 32]:

$$\text{L}_{CL} := \max_{\substack{Z_S = F_S(S),\\ Z_X = F_X(X),\, G}} \mathbb{E}_{(z_{s1},z_{x1}),\cdots,(z_{sn},z_{xn})\sim P^n(Z_S,Z_X)} \left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{\langle G(z_{xi}),G(z_{si})\rangle}}{\frac{1}{n}\sum_{j=1}^{n} e^{\langle G(z_{xi}),G(z_{sj})\rangle}} \right], \quad (1)$$

where $F_S : \mathcal{S} \to \mathcal{Z}$ is a deterministic mapping and G is a project head that projects a representation in $\mathcal{Z}$ into a lower-dimensional vector. If the input and self-supervised signals share the same sample space, i.e., $\mathcal{X} = \mathcal{S}$, we can impose $F_X = F_S$ (e.g., self-supervised visual representation learning [8]). The projection head, G, can be an identity, a linear, or a non-linear mapping.

Forward Predictive Learning (is extracting task-relevant information). Forward predictive learning encourages the learned representation Z_X to reconstruct the self-supervised signal S, which suggests maximizing the log conditional likelihood $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$. By the chain rule, $I(Z_X;S) = H(S) - H(S|Z_X)$, where $H(S)$ is irrelevant to Z_X. Hence, maximizing $I(Z_X;S)$ is equivalent to maximizing $-H(S|Z_X) = \mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$, which is the predictive learning objective. Together with Theorem 1, if z_x can perfectly reconstruct s for any $(s,z_x) \sim P_{S,Z_X}$, then Z_X contains (approximately) all the information required for the downstream tasks from the input X. A common approach to avoid intractability in computing $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$ is assuming a variational distribution $Q_\phi(S|Z_X)$ with ϕ representing the parameters in $Q_\phi(\cdot|\cdot)$. Specifically, we present to maximize $\mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)]$, which is a lower bound of $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)]$[2]. For simplicity, let $Q_\phi(S|Z_X)$ be Gaussian $\mathcal{N}(S|R(Z_X),\sigma\mathbf{I})$ with $\sigma\mathbf{I}$ as a diagonal matrix[3], the objective (dropping the constants from the Gaussian distribution) becomes:

$$\text{L}_{FP} := \max_{Z_X = F_X(X),\, R} \mathbb{E}_{s,z_x\sim P_{S,Z_X}} \left[-\|s - R(z_x)\|_2^2 \right], \quad (2)$$

where $R : \mathcal{Z} \to \mathcal{S}$ is a deterministic mapping to reconstruct S from Z.

[2] $\mathbb{E}_{P_{S,Z_X}}[\log P(S|Z_X)] = \max_{Q_\phi} \mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)] + D_{\text{KL}}\big(P(S|Z_X) \,\|\, Q_\phi(S|Z_X)\big) \geq \max_{Q_\phi} \mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(S|Z_X)]$.

[3] The assumption of identity covariance in the Gaussian is only a particular parameterization of the distribution $Q(\cdot|\cdot)$. Other examples are MocoGAN [40], which assumes Q is Laplacian (i.e., ℓ_1 reconstruction loss) and ϕ is a deconvolutional network [25]. Transformer-XL [10] assumes Q is a categorical distribution (i.e., cross entropy loss) and ϕ is a Transformer network [41].

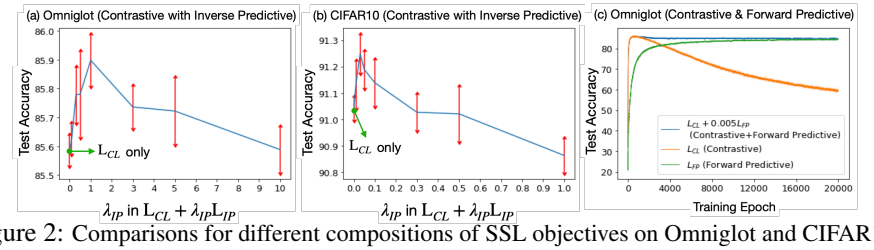

Figure 2: Comparisons for different compositions of SSL objectives on Omniglot and CIFAR10.

Inverse Predictive Learning (is discarding task-irrelevant information). Inverse predictive learning encourages the self-supervised signal S to reconstruct the learned representation Z_X, which suggests maximizing the log conditional likelihood $\mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$. Given Theorem 2 together with $-H(Z_X|S) = \mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$, we know if s can perfectly reconstruct z_x for any $(s, z_x) \sim P_{S,Z_X}$ under the constraint that $I(Z_X; S)$ is maximized, then Z_X discards the task-irrelevant information, excluding $I(X; S|T)$. Similar to the forward predictive learning, we use $\mathbb{E}_{P_{S,Z_X}}[\log Q_\phi(Z_X|S)]$ as a lower bound of $\mathbb{E}_{P_{S,Z_X}}[\log P(Z_X|S)]$. In our deployment, we take the advantage of the design in equation 1 and let $Q_\phi(Z_X|S)$ be Gaussian $\mathcal{N}\Big(Z_X|F_S(S), \sigma\mathbf{I}\Big)$:

$$\mathrm{L}_{IP} := \max_{Z_S = F_S(S), Z_X = F_X(X)} \mathbb{E}_{z_s, z_x \sim P_{Z_S, Z_X}} \Big[- \|z_x - z_s\|_2^2 \Big]. \tag{3}$$

Composed SSL Objective (extract task-relevant and discard task-irrelevant information). We consider a composite loss from all the objectives:

$$\mathrm{L}_{SSL} = \lambda_{CL} \mathrm{L}_{CL} + \lambda_{FP} \mathrm{L}_{FP} + \lambda_{IP} \mathrm{L}_{IP}, \tag{4}$$

where λ_{CL}, λ_{FP}, and λ_{IP} are hyper-parameters. We proceed to show experimentally that the composite loss learns better representation, by enabling us to extract task-relevant information and discard task-irrelevant information at the same time.

We experiment on Omniglot [22] and CIFAR10 [21]. For CIFAR10, we follow the same setting as SimCLR [8]. For Omgiglot, we regard image as input (X) and generate self-supervised signal (S) by first sampling an image from the same class and then applying translation/rotation to it. Furthermore, we represent task-relevant information (T) by the labels of the image. Under this self-supervised signal construction, the exclusive information in X or S are drawing styles (i.e., by different people) and image augmentations, and only their shared information contribute to T. To formally show the latter, if T representing the label for X/S, then $P(T|X)$ and $P(T|S)$ are Dirac. Hence, $T \perp\!\!\!\perp S|X$ and $T \perp\!\!\!\perp X|S$, suggesting Assumption 1 holds. See Appendix C for an experiment on the scenario in which Assumption 1 might fail.

We then train the feature mapping $F_X(\cdot)$ with SSL objectives (see equation 4), set $F_S(\cdot) = F_X(\cdot)$, let $R(\cdot)$ be symmetrical to $F_X(\cdot)$, and $G(\cdot)$ be an identity mapping. On the test set, we fix the mapping and randomly select 5 examples per character as the labeled examples. Then, we classify the rest of the examples using the 1-nearest neighbor classifier based on feature (i.e., $Z_X = F_X(X)$) cosine similarity. The random performance stands at $\frac{1}{659} \approx 0.15\%$. One may refer to Appendix for more details.

▷ *Results & Discussions.* In Figure 2, we evaluate the generalization ability on the test set for different SSL objectives. First, we examine how the introduced inverse predictive learning objective L_{IP} can help improve the performance along with the contrastive learning objective L_{CL}. We present the results for Omniglot in Figure 2 (a) and the results for CIFAR10 in Figure 2 (b), where $\lambda_{IP} = 0$ refers to the exact same setup as in SimCLR (which considers only L_{CL}). We find that adding L_{IP} in the objective can boost model performance, although being sensitive to the hyper-parameter λ_{IP}. According to Theorem 2, the improved performance suggests a more compressed representation results in better performance for the downstream tasks. Second, we add the discussions with the forward predictive learning objective L_{FP}. We present the results for Omniglot in Figure 2 (c). Comparing to L_{FP}, L_{CL} 1) reaches better test accuracy; 2) requires shorter training epochs to reach the best performance; and 3) suffers from overfitting with long-epoch training. Combining both of them ($\mathrm{L}_{CL} + 0.005\mathrm{L}_{FP}$) brings their advantages together.

3 Conclusion

This work studies self-supervised learning from both a theoretical and a empirical perspective. Under our framework, we show that the practical deployments of self-supervised leaning, such as contrastive

and predictive learning, can extract task-relevant information (with a potential loss) and discard task-irrelevant information (with a fixed gap). Our analysis also inspires a new loss term (inverse predictive learning) that discards irrelevant information for the downstream tasks. We believe this work sheds light on the advantages of self-supervised learning and can improve understanding on when and why self-supervised learning is likely to work.

Acknowledgments

This work was supported in part by the DARPA grants FA875018C0150 HR00111990016, NSF IIS1763562, NSF Awards #1750439 #1722822, National Institutes of Health, and Apple. We would also like to acknowledge NVIDIA's GPU support. We would like to thank Google Tensorflow Research Cloud program for their very generous TPU support.

References

[1] Achille, A. and S. Soatto (2018). Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research 19*(1), 1947–1980.

[2] Arandjelovic, R. and A. Zisserman (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617.

[3] Arora, S., H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi (2019). A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

[4] Bachman, P., R. D. Hjelm, and W. Buchwalter (2019). Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15509–15519.

[5] Balcan, M.-F., A. Blum, and K. Yang (2005). Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pp. 89–96.

[6] Calude, C. S. (2013). *Information and randomness: an algorithmic perspective*. Springer Science & Business Media.

[7] Chen, M., A. Radford, R. Child, J. Wu, and H. Jun. Generative pretraining from pixels.

[8] Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

[9] Cover, T. M. and J. A. Thomas (2012). *Elements of information theory*. John Wiley & Sons.

[10] Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

[11] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.

[12] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[13] Du, J., C. X. Ling, and Z.-H. Zhou (2010). When does cotraining work in real data? *IEEE Transactions on Knowledge and Data Engineering 23*(5), 788–799.

[14] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters 27*(8), 861–874.

[15] Federici, M., A. Dutta, P. Forré, N. Kushmann, and Z. Akata (2020). Learning robust representations via multi-view information bottleneck. International Conference on Learning Representation.

[16] Gidaris, S., P. Singh, and N. Komodakis (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

[17] He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick (2019). Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

[18] He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[19] Hénaff, O. J., A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord (2019). Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*.

[20] Hjelm, R. D., A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

[21] Krizhevsky, A. et al. (2009). Learning multiple layers of features from tiny images.

[22] Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum (2015). Human-level concept learning through probabilistic program induction. *Science 350*(6266), 1332–1338.

[23] Lee, J. D., Q. Lei, N. Saunshi, and J. Zhuo (2020). Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*.

[24] Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer.

[25] Long, J., E. Shelhamer, and T. Darrell (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

[26] Mukherjee, S., H. Asnani, and S. Kannan (2020). Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pp. 1083–1093. PMLR.

[27] Noroozi, M. and P. Favaro (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer.

[28] Oord, A. v. d., Y. Li, and O. Vinyals (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[29] Pathak, D., P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544.

[30] Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[31] Poole, B., S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker (2019). On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*.

[32] Song, J. and S. Ermon (2019). Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*.

[33] Sorower, M. S. A literature survey on algorithms for multi-label learning.

[34] Sridharan, K. and S. M. Kakade (2008). An information theoretic framework for multi-view learning.

[35] Tian, Y., D. Krishnan, and P. Isola (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

[36] Tian, Y., C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola (2020). What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*.

[37] Tishby, N., F. C. Pereira, and W. Bialek (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

[38] Tosh, C., A. Krishnamurthy, and D. Hsu (2020). Contrastive learning, multi-view redundancy, and linear models. *arXiv preprint arXiv:2008.10150*.

[39] Tschannen, M., J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic (2019). On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*.

[40] Tulyakov, S., M.-Y. Liu, X. Yang, and J. Kautz (2018). Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535.

[41] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

[42] Vondrick, C., H. Pirsiavash, and A. Torralba (2016). Generating videos with scene dynamics. In *Advances in neural information processing systems*, pp. 613–621.

[43] Xu, C., D. Tao, and C. Xu (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

[44] Zhang, R., P. Isola, and A. A. Efros (2016). Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer.

[45] Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

A Remarks on Learning Minimal and Sufficient Representations

In the main text, we discussed the objectives to learn minimal and sufficient representations (Definition 1). Here, we discuss the similarities and differences between the prior methods [1, 37] and ours. First, to obtain sufficient representations (for the downstream task T), all the methods presented to maximize $I(Z_X;T)$. Then, to maintain minimal amount of information in the representations, the prior methods [1, 37] presented to minimize $I(Z_X;X)$ and the ours presents to minimize $H(Z_X|T)$. Our goal is to relate $I(Z_X;X)$ minimization and $H(Z_X|T)$ minimization in our framework.

To begin with, under the constraint $I(Z_X;T)$ is maximized, we see that minimizing $I(Z_X;X)$ is equivalent to minimizing $I(Z_X;X|T)$. The reason is that $I(Z_X;X) = I(Z_X;X|T) + I(Z_X;X;T)$, where $I(Z_X;X;T) = I(Z_X;T)$ due to the determinism from X to Z_X (our framework learns a deterministic function from X to Z_X) and $I(Z_X;T)$ is maximized in our constraint. Then, $I(Z_X;X|T) = H(Z_X|T) - H(Z_X|X,T)$, where $H(Z_X|T)$ contains no randomness (no information) as Z_X being deterministic from X. Hence, $I(Z_X;X|T)$ minimization and $H(Z_X|T)$ minimization are interchangeable.

The same claim can be made from the downstream task T to the self-supervised signal S. In specific, when X to Z_X is deterministic, $I(Z_X;X|S)$ minimization and $H(Z_X|S)$ minimization are interchangeable. As discussed in the related work section, for reducing the amount of the redundant information, Federici et al. [15] presented to use $I(Z_X;X|S)$ minimization and ours presented to use $H(Z_X|T)$ minimization. We also note that directly minimizing the conditional mutual information (i.e., $I(Z_X;X|S)$) requires a min-max optimization [26], which may cause instability in practice. To overcome the issue, Federici et al. [15] assumes a Gaussian encoder for $X \to Z_X$ and presents an upper bound of the original objective.

B Proofs for Theorem 1 and 2

We start by presenting a useful lemma from the fact that $F_X(\cdot)$ is a deterministic function:

Lemma 1 (Determinism)**.** *If $P(Z_X|X)$ is Dirac, then the following conditional independence holds: $T \perp\!\!\!\perp Z_X|X$ and $S \perp\!\!\!\perp Z_X|X$, inducing a Markov chain $S \leftrightarrow T \leftrightarrow X \to Z_X$.*

Proof. When Z_X is a deterministic function of X, for any A in the sigma-algebra induced by Z_X we have $\mathbb{E}[\mathbf{1}_{[Z_X \in A]}|X,\{T,S\}] = \mathbb{E}[\mathbf{1}_{[Z_X \in A]}|X,S] = \mathbb{E}[\mathbf{1}_{[Z_X \in A]}|X]$, which implies $T \perp\!\!\!\perp Z_X|X$ and $S \perp\!\!\!\perp Z_X|X$. □

Theorem 1 and 2 in the main text restated:

Theorem 3 (Task-relevant information with a potential loss ϵ_{info}, restating Theorem 1 in the main text)**.** *The supervised learned representations (i.e., $I(Z_X^{\text{sup}};T)$ and $I(Z_X^{\text{sup}_{\min}};T)$) contain all the task-relevant information in the input (i.e., $I(X;T)$). The self-supervised learned representations (i.e., $I(Z_X^{\text{ssl}};T)$ and $I(Z_X^{\text{ssl}_{\min}};T)$) contain all the task-relevant information in the input with a potential loss ϵ_{info}. Formally,*

$$I(X;T) = I(Z_X^{\text{sup}};T) = I(Z_X^{\text{sup}_{\min}};T) \geq I(Z_X^{\text{ssl}};T) \geq I(Z_X^{\text{ssl}_{\min}};T) \geq I(X;T) - \epsilon_{\text{info}}.$$

Proof. The proofs contain two parts. The first one is showing the results for the supervised learned representations and the second one is for the self-supervised learned representations.

Supervised Learned Representations: Adopting Data Processing Inequality (DPI by [9]) in the Markov chain $S \leftrightarrow T \leftrightarrow X \to Z_X$ (Lemma 1), $I(Z_X;T)$ is maximized at $I(X;T)$. Since

both supervised learned representations (Z_X^{sup} and $Z_X^{\text{sup}_{\min}}$) maximize $I(Z_X;T)$, we conclude $I(Z_X^{\text{sup}};T) = I(Z_X^{\text{sup}_{\min}};T) = I(X;T)$.

Self-supervised Learned Representations: First, we have

$$I(Z_X;S) = I(Z_X;T) - I(Z_X;T|S) + I(Z_X;S|T) = I(Z_X;T;S) + I(Z_X;S|T)$$

and

$$I(X;S) = I(X;T) - I(X;T|S) + I(X;S|T) = I(X;T;S) + I(X;S|T).$$

By DPI in the Markov chain $S \leftrightarrow T \leftrightarrow X \rightarrow Z_X$ (Lemma 1), we know

- $I(Z_X;S)$ is maximized at $I(X;S)$
- $I(Z_X;S;T)$ is maximized at $I(X;S;T)$
- $I(Z_X;S|T)$ is maximized at $I(X;S|T)$

Since both self-supervised learned representations (Z_X^{ssl} and $Z_X^{\text{ssl}_{\min}}$) maximize $I(Z_X;S)$, we have $I(Z_X^{\text{ssl}};S) = I(Z_X^{\text{ssl}_{\min}};S) = I(X;S)$. Hence, $I(Z_X^{\text{ssl}};S;T) = I(Z_X^{\text{ssl}_{\min}};S;T) = I(X;S;T)$ and $I(Z_X^{\text{ssl}};S|T) = I(Z_X^{\text{ssl}_{\min}};S|T) = I(X;S|T)$. Using the result $I(Z_X^{\text{ssl}};S;T) = I(Z_X^{\text{ssl}_{\min}};S;T) = I(X;S;T)$, we get

$$I(Z_X^{\text{ssl}};T) = I(X;T) - I(X;T|S) + I(Z_X^{\text{ssl}};T|S)$$

and

$$I(Z_X^{\text{ssl}_{\min}};T) = I(X;T) - I(X;T|S) + I(Z_X^{\text{ssl}_{\min}};T|S).$$

Now, we are ready to present the inequalities:

1. $I(X;T) \geq I(Z_X^{\text{ssl}};T)$ due to $I(X;T|S) \geq I(Z_X^{\text{ssl}};T|S)$ by DPI.
2. $I(Z_X^{\text{ssl}};T) \geq I(Z_X^{\text{ssl}_{\min}};T)$ due to $I(Z_X^{\text{ssl}};T|S) \geq I(Z_X^{\text{ssl}_{\min}};T|S) = 0$. Since $H(Z_X|S)$ is minimized at $Z_X^{\text{ssl}_{\min}}$, $I(Z_X^{\text{ssl}_{\min}};T|S) = 0$.
3. $I(Z_X^{\text{ssl}_{\min}};T) \geq I(X;T) - \epsilon_{\text{info}}$ due to
$$I(X;T) - I(X;T|S) + I(Z_X^{\text{ssl}_{\min}};T|S) \geq I(X;T) - I(X;T|S) \geq I(X;T) - \epsilon_{\text{info}},$$
where $I(X;T|S) \leq \epsilon_{\text{info}}$ by the redundancy assumption.

□

Theorem 4 (Task-irrelevant information with a fixed compression gap $I(X;S|T)$, restating Theorem 2 in the main text)**.** *The sufficient self-supervised representation (i.e., $I(Z_X^{\text{ssl}};T)$) contains more task-irrelevant information in the input than then the sufficient and minimal self-supervised representation (i.e., $I(Z_X^{\text{ssl}_{\min}};T)$). The later contains an amount of the information, $I(X;S|T)$, that cannot be discarded from the input. Formally,*

$$I(Z_X^{\text{ssl}};X|T) = I(X;S|T) + I(Z_X^{\text{ssl}};X|S,T) \geq I(Z_X^{\text{ssl}_{\min}};X|T) = I(X;S|T) \geq I(Z_X^{\text{sup}_{\min}};X|T) = 0.$$

Proof. First, we see that

$$I(Z_X;X|T) = I(Z_X;X;S|T) + I(Z_X;X|S,T) = I(Z_X;S|T) + I(Z_X;X|S,T),$$

where $I(Z_X;X;S|T) = I(Z_X;S|T)$ by DPI in the Markov chain $S \leftrightarrow T \leftrightarrow X \rightarrow Z_X$.

We conclude the proof by combining the following:

- From the proof in Theorem 3, we showed $I(Z_X^{\text{ssl}};S|T) = I(Z_X^{\text{ssl}_{\min}};S|T) = I(X;S|T)$.
- Since $H(Z_X|S)$ is minimized at $Z_X^{\text{ssl}_{\min}}$, $I(Z_X^{\text{ssl}_{\min}};X|S,T) = 0$.
- Since $H(Z_X|T)$ is minimized at $Z_X^{\text{sup}_{\min}}$, $I(Z_X^{\text{sup}_{\min}};X|T) = 0$.

□

(a) MS COCO (Using L_{CL} as SSL objective)

Setting	Micro ROC-AUC	Subset Acc.
Cross-modality Self-supervised Learning		
Raw BERT + Raw ResNet	0.5963 ± 0.0034	0.0166 ± 0.0017
Pre-trained BERT + Raw ResNet	0.5915 ± 0.0035	0.0163 ± 0.0011
Raw BERT + Pre-trained ResNet	0.7049 ± 0.0040	0.2081 ± 0.0063
Pre-trained BERT + Pre-trained ResNet	0.7065 ± 0.0026	0.2123 ± 0.0040
Non Self-supervised Learning		
Only Pre-trained ResNet	0.6761 ± 0.0045	0.1719 ± 0.0015

(b) Raw BERT + Pre-trained ResNet (Contrastive with Inverse Predictive)

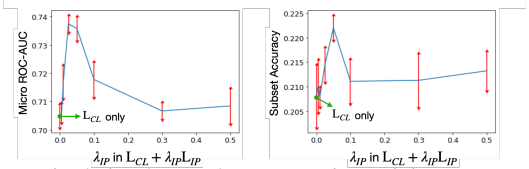

Figure 3: Comparisons for different settings on self-supervised visual-textual representation training. We report metrics on MS COCO validation set with mean and standard deviation from 5 random trials.

C Experiment II - Visual-Textual Representation Learning

We provide experiments using MS COCO dataset [24] that contains 328k multi-labeled images with 2.5 million labeled instances from 91 objects. Each image has 5 annotated captions describing the relationships between objects in the scenes. We regard image as input (X) and its textual descriptions as self-supervised signal (S). Since vision and text are two very different modalities, the multi-view redundancy may not be satisfied, which means ϵ_{info} may be large in Assumption 1.

We adopt L_{CL} ($+\lambda_{IP}\text{L}_{\text{IP}}$) as our SSL objective. We use ResNet18 [18] image encoder for $F_X(\cdot)$ (trained from scratch or fine-tuned on ImageNet [11] pre-trained weights), BERT-uncased [12] text encoder for $F_S(\cdot)$ (trained from scratch or BookCorpus [45]/Wikipedia pre-trained weights), and a linear layer for $G(\cdot)$. After performing self-supervised visual-textual representation learning, we consider the downstream multi-label classification over 91 categories. We evaluate learned visual representation (Z_X) using *downstream linear evaluation protocol* [4, 19, 20, 28, 35, 39]. Specifically, a linear classifier is trained from the self-supervised learned (fixed) representation to the labels on the training set. Commonly used metrics for multi-label classification are reported on MS COCO validation set: Micro ROC-AUC and Subset Accuracy.

- Subset Accuracy (A) [33], also know as the Exact Match Ratio (MR), ignores all partially correct (consider them incorrect) outputs and extend accuracy from the single label case to the multi-label setting.

$$MR = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i = H_i]}$$

- Micro AUC ROC score [14] computes the AUC (Area under the curve) of a receiver operating characteristic (ROC) curve.

▷ *Results & Discussions.* First, Figure 3 (a) suggests that the SSL strategy can still work when the input and self-supervised signals lie in different modalities. For example, pre-trained ResNet with BERT (either raw or the pre-trained one) outperforms pre-trained ResNet alone. We also see that the self-supervised learned representations benefit more if the ResNet is pre-trained but not the BERT. This result is in accord with the fact that object recognition requires more understanding in vision, and hence the pre-trained ResNet is preferrable than the pre-trained BERT. Next, Figure 3 (b) suggests that the self-supervised learned representations can be further improved by combining L_{CL} and L_{IP}, suggesting L_{IP} may be a useful objective to discard task-irrelevant information.

D More on Visual Representation Learning Experiments

In the main text, we design controlled experiments on self-supervised visual representation learning to empirically support our theorem and examine different compositions of SSL objectives. In this section, we will discuss 1) the architecture design; 2) different deployments of contrastive/ forward predictive learning; and 3) different self-supervised signal construction strategy. We argue that these three additional set of experiments may be interesting future work.

D.1 Architecture Design

The input image has size 105×105. For image augmentations, we adopt 1) rotation with degrees from $-10°$ to $+10°$; 2) translation from -15 pixels to $+15$ pixels; 3) scaling both width and height from 0.85 to 1.0; 4) scaling width from 0.85 to 1.25 while fixing the height; and 5) resizing the image to 28×28. Then, a deep network takes a 28×28 image and outputs a $1024-$dim. feature vector. The deep network has the structure: $\text{Conv} - \text{BN} - \text{ReLU} - \text{Conv} - \text{BN} - \text{ReLU} - \text{MaxPool} - \text{Conv} - \text{BN} - \text{ReLU} - \text{MaxPool} - \text{Conv}$

(a) Omniglot (Composing SSL Objectives with L_{FP} as MSE)

Objective	Trained for	Test Accuracy
L_{CL}	500 epochs	$85.59 \pm 0.05\%$
$\mathrm{L}_{CL} + \mathrm{L}_{IP}$	500 epochs	$85.90 \pm 0.09\%$
L_{FP}	20000 epochs	$84.83 \pm 0.07\%$
$\mathrm{L}_{FP} + 10\mathrm{L}_{IP}$	20000 epochs	$84.96 \pm 0.04\%$
$\mathrm{L}_{CL} + 10\mathrm{L}_{FP}$	9000 epochs	$86.13 \pm 0.21\%$
$\mathrm{L}_{CL} + 10\mathrm{L}_{FP} + \mathrm{L}_{IP}$	9000 epochs	$86.17 \pm 0.13\%$

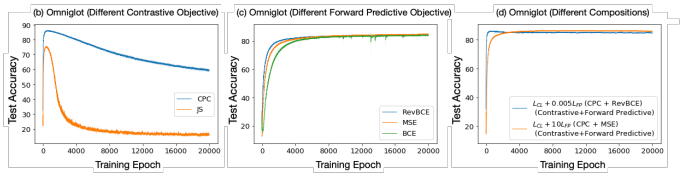

Figure 4: Comparisons for different objectives/compositions of SSL objectives on self-supervised visual representation training. We report mean and its standard error from 5 random trials.

$-\mathsf{BN} - \mathsf{ReLU} - \mathsf{MaxPool} - \mathsf{Flatten} - \mathsf{Linear} - \mathsf{L2Norm}$. Conv has 3x3 kernel size with 128 output channels, MaxPool has 2x2 kernel size, and Linear is a 1152 to 1024 weight matrix. $R(\cdot)$ is symmetric to $F_X(\cdot)$, which has $\mathsf{Linear} - \mathsf{BN} - \mathsf{ReLU} - \mathsf{UnFlatten} - \mathsf{DeConv} - \mathsf{BN} - \mathsf{ReLU} - \mathsf{DeConv} - \mathsf{BN} - \mathsf{ReLU} - \mathsf{DeConv} - \mathsf{BN} - \mathsf{ReLU} - \mathsf{DeConv}$. $R(\cdot)$ has the exact same number of parameters as $F_X(\cdot)$. Note that we use the same network designs in $I(\cdot,\cdot)$ and $H(\cdot|\cdot)$ estimations. To reproduce the results in our experimental section, please refer to our released code (Anonymous).

D.2 Different Deployments for Contrastive and Predictive Learning Objectives

In the main text, for practical deployments, we suggest Contrastive Predictive Coding (CPC) [28] for L_{CL} and assume Gaussian distribution for the variational distributions in L_{FP}/ L_{IP}. The practical deployments can be abundant by using different mutual information approximations for L_{CL} and having different distribution assumptions for L_{FP}/ L_{IP}. In the following, we discuss a few examples.

Contrastive Learning. Other than CPC [28], another popular contrastive learning objective is JS [4], which is the lower bound of Jensen-Shannon divergence between $P(Z_S, Z_X)$ and $P(Z_S)P(Z_X)$ (a variational bound of mutual information). Its objective can be written as

$$\max_{Z_S=F_S(S), Z_X=F_X(X), G} \mathbb{E}_{P(Z_S,Z_X)}[-\mathrm{softplus}(-\langle G(z_x), G(z_s)\rangle)] - \mathbb{E}_{P(Z_S)P(Z_X)}[\mathrm{softplus}(\langle G(z_x), G(z_s)\rangle)],$$

where we use softplus to denote $\mathrm{softplus}\,(x) = \log\,(1 + \exp\,(x))$.

Predictive Learning. Gaussian distribution may be the simplest distribution form that we can imagine, which leads to Mean Square Error (MSE) reconstruction loss. Here, we use forward predictive learning as an example, and we discuss the case when $\mathcal{S}$ lies in discrete $\{0, 1\}$ sample space. Specifically, we let $Q_\phi(S|Z_X)$ be factorized multivariate Bernoulli:

$$\max_{Z_X=F_X(X), R} \mathbb{E}_{P_{S,Z_X}} \left[\sum_{i=1}^{p} s_i \cdot \log\,[R(z_x)]_i + (1 - s_i) \cdot \log\,[1 - R(z_x)]_i \right]. \tag{5}$$

This objective leads to Binary Cross Entropy (BCE) reconstruction loss.

If we assume each reconstruction loss corresponds to a particular distribution form, then by ignoring which variatioinal distribution we choose, we are free to choose arbitrary reconstruction loss. For instance, by switching s and z in eq. equation 5, the objective can be regarded as Reverse Binary Cross Entropy Loss (RevBCE) reconstruction loss. In our experiments, we find RevBCE works the best among {MSE, BCE, and RevBCE}. Therefore, in the main text, we choose RevBCE as the example reconstruction loss as L_{FP}.

More Experiments. We provide an additional set of experiments by having {CPC, JS} for L_{CL} and {MSE, BCE, RevBCE} reconstruction loss for L_{FP} in Figure 4. From the results, we find different formulation of objectives bring very different test generalization performance. We argue that, given a particular task, it is challenging but important to find the best deployments for contrastive and predictive learning objectives.

D.3 Different Self-supervised Signal Construction Strategy

In the main text, we design a self-supervised signal construction strategy that the input (X) and the self-supervised signal (S) differ in {drawing styles, image augmentations}. This self-supervised signal construction strategy is different from the one that is commonly adopted in most self-supervised visual representation learning work [4, 8, 35]. Specifically, prior work consider the difference between input and the self-supervised signal only in image augmentations. We provide additional experiments in Fig. 5 to compare these two different self-supervised signal construction strategies.

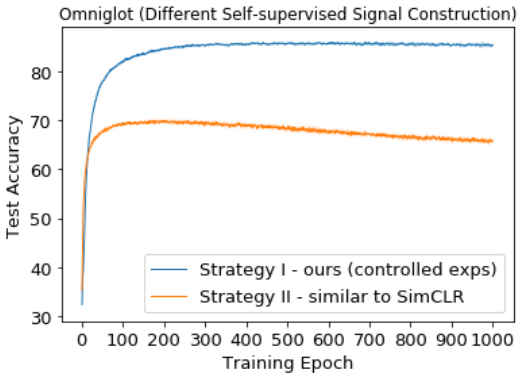

Figure 5: Comparisons for different self-supervised signal construction strategies. The differences between the input and the self-supervised signals are {drawing styles, image augmentations} for our construction strategy and only {image augmentations} for SimCLR [8]'s strategy. We choose L_{CL} as our objective, reporting mean and its standard error from 5 random trials.

We see that, comparing to the common self-supervised signal construction strategy [4, 8, 35], the strategy introduced in our controlled experiments has much better generalization ability to test set. It is worth noting that, although our construction strategy has access to the label information (i.e., we sample the self-supervised signal image from the same character with the input image), our SSL objectives do not train with the labels. Nonetheless, since we implicitly utilize the label information in our self-supervised construction strategy, it will be unfair to directly compare our strategy and prior one. An interesting future research direction is examining different self-supervised signal construction strategy and even combine full/part of label information into self-supervised learning.