Adversarial perturbation based latent reconstruction for domain-agnostic self-supervised learning

Kuilin Chen[*] **Sijie Tian**[*] **Chi-Guhn Lee**
University of Toronto
Toronto, Ontario, Canada
`{kuilin.chen, sophie.tian}@mail.utoronto.ca, cglee@mie.utoronto.ca`

Abstract

Most self-supervised learning (SSL) methods rely on domain-specific pretext tasks and data augmentations to learn high-quality representations from unlabeled data, which require expert domain knowledge to develop. Moreover, it is not clear why solving certain pretext tasks leads to useful representations. These two reasons hinder the wider application of SSL to different domains. To overcome such limitations, we propose adversarial perturbation based latent reconstruction (APLR) for domain-agnostic SSL. In APLR, a neural network is trained to generate adversarial noise to perturb the unlabeled training sample so that domain-specific augmentations are not required. The pretext task in APLR is to reconstruct the latent representation of a clean sample from a perturbed sample. We show that representation learning via latent reconstruction is closely related to multi-dimensional Hirschfeld-Gebelein-Rényi (HGR) maximal correlation and has theoretical guarantees on the linear probe error. We apply APLR to tabular data, images, and audios and the empirical results indicate that APLR not only outperforms existing domain-agnostic SSL methods but also closes the performance gap to domain-specific SSL methods. In many cases, APLR also outperforms training the full network in a supervised manner.

1 Introduction

Unsupervised deep learning has been highly successful in discovering useful representations in natural language processing (NLP) [1, 2] and computer vision (CV) [3, 4]. These methods define pretext tasks on unlabeled data so that unsupervised representation learning can be done in a self-supervised manner without explicit human annotations. The success of self-supervised learning (SSL) depends on domain-specific pretext tasks, as well as domain-specific data augmentations, which require extensive domain knowledge to be developed, and such knowledge may not be readily available for certain data types such as tabular data [5]. Furthermore, the theoretical understanding of why certain pretext tasks lead to useful representations remains fairly elusive [6]. Those two reasons hinder wider applications of SSL beyond the fields of NLP and CV.

Self-supervised algorithms benefit from inductive biases from domain-specific designs but they do not generalize across domains. For example, masked language models like BERT [1] are not directly applicable to untokenized data. Although contrastive learning does not require tokenized data, its success in CV cannot be easily transferable due to its sensitivity to image-specific data augmentations [3]. Furthermore, in contrastive learning, the quality of representations degrades significantly without those hand-crafted data augmentations [7]. Inspired by denoising auto-encoding [8, 9, 10], perturbation of natural samples with Gaussian, Bernoulli, and mixup noises [11, 12] has been utilized as domain-agnostic data augmentations applicable for self-supervised representation learning of

[*] Equal contributions

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

images, graphs, and tabular data. However, random noises may not be as effective since uniformly perturbing uninformative features may not lead to the intended augmentations. Specifically, convex combinations in mixup noises [13, 14] could generate out-of-distribution samples because there is no guarantee that the input data space is convex [5]. In this paper, we use generative adversarial perturbation as a semantic-preserving data augmentation method [15, 16, 17, 18] applicable to different domains of data. Adversarial perturbation is constrained by the ℓ_p norm distance to the natural sample so that it is semantic-preserving and does not change the label [19, 20].

With semantic-preserving perturbation, the pretext tasks in domain-agnostic SSL could be reconstruction of clean samples [12] or instance discrimination of perturbed samples [11]. Nevertheless, the reconstruction of clean samples in the input space is computationally expensive because of high input data dimensionality. Therefore, we present adversarial perturbation based latent reconstruction (APLR), a simple and intuitive domain-agnostic self-supervised pretext task derived from linear generative models, to learn representations from unlabeled data in a domain-agnostic manner. Contrary to the pretext task of instance discrimination, our method does not require comparison to a large number of negative samples to achieve good performance. The proposed APLR not only achieves strong empirical performance on SSL in various domains but also has theoretical guarantees on the linear probe error on downstream tasks.

2 Adversarial perturbation based latent reconstruction

Latent reconstruction Let $\mathbf{x}^1$ be a perturbed sample with some noise, which is adversarial noise in our case. Our pretext task in SSL is to reconstruct the latent representation of the clean sample $\mathbf{x}^2$ from the perturbed sample $\mathbf{x}^1$. We use deep neural networks $\psi(\cdot)$ and $\eta(\cdot)$ to project $\mathbf{x}^1$ and $\mathbf{x}^2$ into latent spaces, respectively. The reconstruction in the latent space can be achieved by maximizing the inner product between $\psi\left(\mathbf{x}^1\right)$ and $\eta\left(\mathbf{x}^2\right)$, when $\psi\left(\mathbf{x}^1\right)$ and $\eta\left(\mathbf{x}^2\right)$ have zero mean and unit variance. Based on discussions in Section A, latent reconstruction must be done with orthogonality constraints to avoid the trivial solution where $\psi(\cdot)$ and $\eta(\cdot)$ projects all input data into a constant vector. Latent reconstruction with orthogonality constraints is equivalent to finding the multi-dimensional Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [21, 22] between two random views. It is defined as follows

$$\rho(\mathbf{x}^1;\mathbf{x}^2) \quad \triangleq \sup_{\substack{\mathbb{E}\left[\psi(\mathbf{x}^1)\psi(\mathbf{x}^1)^\top\right]=\mathbb{E}\left[\eta(\mathbf{x}^2)\eta(\mathbf{x}^2)^\top\right]=\mathbf{I} \\ \mathbb{E}\left[\psi(\mathbf{x}^1)\right]=\mathbb{E}\left[\eta(\mathbf{x}^2)\right]=\mathbf{0}}} \mathbb{E}\left[\psi\left(\mathbf{x}^1\right)^\top \eta\left(\mathbf{x}^2\right)\right], \tag{1}$$

where zero mean constraints can be easily satisfied using a batch normalization layer [23] and the constraints on identity covariance matrices can be achieved by forcing the off-diagonal elements to be zero. In practice, the constrained optimization problem in Eq. (1) is solved by minimizing the following loss

$$\mathcal{L}_{\mathrm{LR}} = -\mathop{\mathbb{E}}_{\mathbf{x}^1,\mathbf{x}^2\in\mathcal{B}}\left[\psi\left(\mathbf{x}^1\right)^\top \eta\left(\mathbf{x}^2\right) + \frac{\gamma}{2}\left(\|\psi\left(\mathbf{x}^1\right)\psi\left(\mathbf{x}^1\right)^\top - \mathbf{I}\|_F^2 + \|\eta\left(\mathbf{x}^2\right)\eta\left(\mathbf{x}^2\right)^\top - \mathbf{I}\|_F^2\right)\right] \tag{2}$$

where γ is a Lagrange multiplier, $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathcal{B}$ is a mini-batch.

Adversarial perturbation Adversarial perturbation creates input samples that are almost indistinguishable from natural data but causes the deep learning models to make wrong predictions [24]. We use a generative model to generate adversarial perturbation because it is capable of creating diverse adversarial perturbations very quickly [15, 16, 17, 25].

A generator $\mathcal{G}$ is trained to produce an unbounded adversarial $\mathcal{G}(\mathbf{x}^2) = \delta$. The perturbation is then clipped to be within an ϵ bound of $\mathbf{x}^2$ under the ℓ_p norm. Let $\mathbf{x}^1 = \mathbf{x}^2 + \delta$ be the perturbed view of the clean sample $\mathbf{x}^2$. The vast majority of adversarial perturbation methods rely on the classification boundary of the attacked neural network ($\psi(\cdot)$ and $\eta(\cdot)$) to train the generator via maximizing a cross-entropy loss. However, it is not possible to obtain the generative adversarial perturbation via maximizing a cross-entropy loss in our case because no label is available. In addition, existing generative adversarial perturbation methods explicitly relying on the classification boundary of the attacked model tend to over-fit to the training data [18]. Instead of using a cross-entropy loss, we train $\mathcal{G}(\cdot)$ by maximizing the ℓ_2 distance between $\psi(\mathbf{x}^1)$ and $\eta(\mathbf{x}^2)$

$$\mathcal{L}_{\mathrm{adv}} = \mathop{\mathbb{E}}_{\mathbf{x}^1,\mathbf{x}^2\in\mathcal{B}}\|\eta(\mathbf{x}^2) - \psi(\mathbf{x}^1)\|^2, \tag{3}$$

where $\psi(\cdot)$ and $\eta(\cdot)$ are frozen.

Adversarial training As illustrated in Algorithm 1, our model is trained in an adversarial manner. Given a mini-batch, we train $\mathcal{G}(\cdot)$ by maximizing $\mathcal{L}_{\text{adv}}$ while freezing $\psi(\cdot)$ and $\eta(\cdot)$. Then the parameters in $\psi(\cdot)$ and $\eta(\cdot)$ are updated alternatively by minimizing $\mathcal{L}_{\text{LR}}$ while freezing $\mathcal{G}(\cdot)$.

Algorithm 1 Adversarial perturbation based latent reconstruction

```
Natural sample $\mathbf{x}$, encoders $\psi(\cdot)$ and $\eta(\cdot)$, noise generator $\mathcal{G}(\cdot)$, perturbation budget $\epsilon$, latent reconstruction loss $\mathcal{L}_{\text{LR}}$, adversarial loss $\mathcal{L}_{\text{adv}}$
for sampled minibatch do
    Train $\mathcal{G}(\cdot)$ (freeze $\psi(\cdot)$ and $\eta(\cdot)$)
        Generate an unbounded adversarial perturbation $\delta = \mathcal{G}(\mathbf{x})$    ▷ $\delta$ has the same shape as $\mathbf{x}$
        Clip adversarial perturbation $\delta = \epsilon\delta/|\delta|_p$
        Obtain the perturbed sample $\mathbf{x}^1 = \mathbf{x} + \delta$ and the clean sample $\mathbf{x}^2 = \mathbf{x}$
        Obtain latent representations $\psi(\mathbf{x}^1)$ and $\eta(\mathbf{x}^2)$
        Compute $\mathcal{L}_{\text{adv}}$ and update $\mathcal{G}(\cdot)$ using SGD
    Train $\psi(\cdot)$ and $\eta(\cdot)$ (freeze $\mathcal{G}(\cdot)$)
        Generate an unbounded adversarial perturbation $\delta = \mathcal{G}(\mathbf{x})$    ▷ $\delta$ has the same shape as $\mathbf{x}$
        Clip adversarial perturbation $\delta = \epsilon\delta/|\delta|_p$
        Obtain the perturbed sample $\mathbf{x}^1 = \mathbf{x} + \delta$ and the clean sample $\mathbf{x}^2 = \mathbf{x}$
        Obtain latent representations $\psi(\mathbf{x}^1)$ and $\eta(\mathbf{x}^2)$
        Compute $\mathcal{L}_{\text{LR}}$ and update $\psi(\cdot)$ using SGD
        Update $\eta(\cdot)$ using the exponentially moving average of parameters in $\psi(\cdot)$
end for
```

3 Theoretical Analysis

Let $\mathbf{x}$ be a data sample without perturbation and $y(\mathbf{x})$ be its downstream task label. The quality of the representation $\psi(\mathbf{x})$ is evaluated by the linear probe error, which is the linear classification error of predicting $y(\mathbf{x})$ from $\mathbf{z}$ using a linear model parameterized by $\mathbf{B} \in \mathbb{R}^{k \times r}$. Let $f_B(\mathbf{x}) = \arg\max_{i \in [r]} (\psi(\mathbf{x})^\top \mathbf{B})_i$ be the prediction of the linear model. The linear probe error on $\psi(\mathbf{x})$ is defined as

$$\text{Err}_\psi := \min_{\mathbf{B}} \Pr_{\mathbf{x} \sim P(\mathbf{x})} [y(\mathbf{x}) \neq f_B(\mathbf{x})], \quad (4)$$

where $P(\mathbf{x})$ is the data distribution.

We have to make two assumptions to bound the linear probe error on the learned representations. First, we assume that some universal minimizer of Eq. (1) can be realized by $\psi(\cdot)$ and $\eta(\cdot)$. When the nonlinear mapping to find multi-dimensional HGR maximal correlation is realizable by neural networks, we can analyze the quality of the learned representations using the properties in estimating the HGR maximal correlation.

Assumption 3.1 (Realizability)**.** *Let $\mathcal{H}$ be a hypothesis class containing functions $\psi : \mathcal{X}^1 \to \mathbb{R}^k$ and $\eta : \mathcal{X}^2 \to \mathbb{R}^k$. We assume that at least one of the global minima of $\mathcal{L}(\psi, \eta)$ belongs to $\mathcal{H}$.*

In addition, it is also reasonable to assume that an optimal classifier $f^*(\cdot)$ can predict the label of $\mathbf{x}$ almost deterministically under semantic-preserving perturbation. The assumption about the classification error of $f^*(\cdot)$ provides a baseline to quantify the linear probe error because part of the error is from approximating $f^*(\cdot)$ using a linear model.

Assumption 3.2 (α-bounded Error of the Optimal Classifier)**.** *Let $\mathbf{x}$ be an unperturbed data sample and $y(\mathbf{x})$ be its downstream task label. δ is semantic-preserving perturbation. Then, we assume that there is a classifier f^* such that $Pr(f^*(\mathbf{x}) \neq y(\mathbf{x})) \leq \alpha$ and $Pr(f^*(\mathbf{x} + \delta) \neq y(\mathbf{x})) \leq \alpha$.*

Given assumptions 3.1 and 3.2, we provide the following main theorem on the generalization bound when learning a linear classifier with finite labeled samples on the representations learned by maximizing the HGR maximal correlation.

Theorem 3.3. *Let $\psi^*, \eta^* \in \mathcal{H}$ be a minimizer of $\mathcal{L}(\psi, \eta)$. The linear classification parameter $\hat{\mathbf{B}}$ is estimated with n_2 i.i.d. random samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_2}$. With probability at least $1 - \zeta$ over the randomness of data, we have*

$$\Pr_{\mathbf{x}\sim P(\mathbf{x})}[y(\mathbf{x}) \neq f_{\hat{B}}(\mathbf{x})] \leq \widetilde{O}\left(\frac{\alpha}{1-\lambda_{k+1}} + \sqrt{\frac{k}{n_2}}\right), \tag{5}$$

where λ_{k+1} is the $k+1$-th Hirschfeld-Gebelein-Rényi maximal correlation between $\mathcal{X}^1$ and $\mathcal{X}^2$.

We hide universal constant factors and logarithmic factors of k in $\widetilde{O}(\cdot)$. The first term on the right-hand side of Theorem 3.3 guarantees the existence of a linear classifier that achieves a small downstream classification error. It indicates whether the downstream label is linearly separable by the learned representation, thus measuring the expressivity of the learned representation. The second term on the right-hand side reveals the sample complexity of learning $\mathbf{B}$ from finite labeled samples in the downstream task. It measures the data efficiency of learning the downstream task using the learned representation. The proof is presented in Appendix C.

4 Experiments

We apply APLR on three different data domains: tabular data, images, and audios. We describe the datasets used in Appendix D, and include additional details in Appendix E. For all datasets, we follow the widely used linear evaluation protocol in SSL as a proxy to examine the quality of the learned representations [4, 26].

The results in Tables 1, 2, and 3 indicate that the proposed APLR not only outperforms existing domain-agnostic SSL methods, but also achieves comparable performance with SOTA domain-specific SSL methods. Details analyses and ablation studies can be found in Appendix E.

Table 1: Linear evaluation accuracy on tabular data

	MNIST	Fashion	Gas	Gesture
Tabular-specific				
VIME-Self [12]	96.62	**87.26**	93.17	38.99
Domain-agnostic				
DACL [11]	94.70	79.78	95.39	38.56
Ours	**97.11**	87.12	**97.98**	**42.97**
Supervised	98.67	90.00	94.91	42.32

Table 2: Linear evaluation accuracy on audio data

	ESC-10	ESC-50	LibriSpeech-100
Audio-specific			
CLAR [27]	68.70	40.40	62.14
Domain-agnostic			
DACL [11]	77.75	48.50	37.30
Viewmaker [28]	70.00	35.75	88.30
Ours	**81.25**	**57.75**	**96.29**
Supervised	76.25	59.14	N/A

Table 3: Linear evaluation accuracy on image data

	CIFAR-10	CIFAR-100	STL-10	Tiny-ImageNet
Image-specific				
SimCLR [3]	**86.47**	54.86	85.49	**43.27**
Domain-agnostic				
DACL [11]	60.49	35.28	57.34	22.69
Viewmaker [28]	84.51	52.28	82.73	40.51
Ours	85.92	**55.83**	**86.21**	42.93

5 Conclusions

In this paper, we introduce APLR, a domain-agnostic SSL method by reconstruction of adversarial perturbed samples in the latent space. The adversarial perturbation is created by a generative network, which is trained concurrently with the feature encoder in an adversarial manner. Our empirical results show that the proposed method is better than the existing domain-agnostic SSL methods and achieves comparable performance with SOTA domain-specific SSL methods. In many cases, APLR also outperforms training the same architecture in a fully supervised manner, demonstrating its strong ability to learn useful latent representations. In addition, the proposed latent reconstruction is linked to Hirschfeld-Gebelein-Rényi maximal correlation and thus has theoretical guarantees of downstream classification tasks. We believe that the proposed method can be applied to applications beyond classification, such as reinforcement learning.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] 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.

[4] 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.

[5] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10268–10278. PMLR, 2021.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the 33nd International Conference on Neural Information Processing Systems*, 2020.

[8] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[9] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[10] 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.

[11] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541. PMLR, 2021.

[12] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[14] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[15] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Thirty-second aaai conference on artificial intelligence*, 2018.

[16] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[17] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Krishna Kanth Nakkal and Mathieu Salzmann. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[21] A. Renyi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

[22] Anuran Makur, Fabián Kozynski, Shao-Lun Huang, and Lizhong Zheng. An efficient algorithm for information decomposition and extraction. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 972–979. IEEE, 2015.

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[25] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan L. Yuille. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356, pages 795–813. Springer, 2020.

[26] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[27] Haider Al-Tahan and Yalda Mohsenzadeh. Clar: contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR, 2021.

[28] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning. In *International Conference on Learning Representations*, 2021.

[29] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. *Technical Report, Department of Statistics, University of California*, 2005.

[30] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

[31] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[32] H Hotelling. Relations between two sets of variates. *Biometrika*, 1936.

[33] Herman Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142, 1975.

[34] S. Wold, A. Ruhe, H. Wold, and W.J. Dunn, III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Comput.*, 5:735–743, 1984.

[35] Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

[36] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[37] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[38] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013*, 2013.

[40] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[41] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[43] 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.

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

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[46] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[47] 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.

[48] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021.

[49] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021.

[50] 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*, pages 9598–9608, 2021.

[51] Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.

[52] Kaiwen Yang, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Identity-disentangled adversarial augmentation for self-supervised learning. In *International Conference on Machine Learning*, pages 25364–25381. PMLR, 2022.

[53] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[54] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020.

[55] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

[56] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 2019.

[57] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[58] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021.

[59] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[60] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

[61] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022.

[62] Hans S Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

[63] Shao-Lun Huang and Xiangxiang Xu. On the sample complexity of hgr maximal correlation functions for large datasets. *IEEE Transactions on Information Theory*, 67(3):1951–1980, 2020.

[64] William H Greene. *Econometric analysis*. Prentice Hall, 2003.

[65] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[66] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[67] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[68] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A. Ryan, Margie L. Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166-167:320–329, 2012.

[69] Irene Rodriguez-Lujan, Jordi Fonollosa, Alexander Vergara, Margie Homer, and Ramon Huerta. On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. *Chemometrics and Intelligent Laboratory Systems*, 130:123–134, 2014.

[70] Priscilla Wagner, Sarajane Peres, Renata Madeo, Clodoaldo Lima, and Fernando Freitas. Gesture unit segmentation using spatial-temporal information and machine learning. 01 2014.

[71] A Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, University of Toronto, 2009.

[72] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *AISTATS*, 2011.

[73] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015.

[74] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.

[75] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[76] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943. Curran Associates, Inc., 2021.

[77] 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.

[78] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[79] Anurag Kumar and Vamsi Krishna Ithapu. A sequential self teaching approach for improving generalization in sound event recognition. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

A Background

Learning representations from two views of an input, $\mathbf{x}^1$ and $\mathbf{x}^2$, is appealing if the learned representations do not contain the noises in different views. This assumption can be explicitly encoded into the following **generative model** [29] with one shared latent variable $\mathbf{z}$:

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ p(\mathbf{x}^1 \mid \mathbf{z}) &= \mathcal{N}\left(\boldsymbol{\psi}^\top \mathbf{z}, \Sigma^1\right) \\ p(\mathbf{x}^2 \mid \mathbf{z}) &= \mathcal{N}\left(\boldsymbol{\eta}^\top \mathbf{z}, \Sigma^2\right), \end{aligned} \tag{6}$$

where the model parameters $\boldsymbol{\psi}$, $\boldsymbol{\eta}$, Σ^1 and Σ^2 can be learned by maximum likelihood estimation. Reconstruction of input data via maximum likelihood estimation is computationally expensive when the dimension of the input data is high. Instead, the probabilistic generative model can be reinterpreted as **latent reconstruction**, which has the benefit of direct representation learning while retaining the properties of generative modeling.

To convert generative modelling into latent reconstruction, two assumptions need to be met. First, the assumption in generative modeling is that both datasets have similar low-rank approximations. In latent reconstruction, this can be achieved by correlating the pair of latent embeddings $\boldsymbol{\psi}\mathbf{x}^1$ and $\boldsymbol{\eta}\mathbf{x}^2$. Secondly, it is assumed in generative modelling that the latent variables follow an isotropic Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. In latent reconstruction, the covariance matrix of the latent variables is diagonal, meaning that there is no covariance between different dimensions of the latent variable. This orthogonality constraint is equivalent to the assumption of isotropic Gaussian prior and avoids the trivial solution. As a result, by correlating the latent embeddings $\boldsymbol{\psi}\mathbf{x}^1$ and $\boldsymbol{\eta}\mathbf{x}^2$ and enforcing a diagonal covariance matrix, the properties of generative modelling can be retained in latent reconstruction.

The key principle behind latent reconstruction is that the latent representation of $\mathbf{x}^1$ is a good predictor for that of $\mathbf{x}^2$. Given two datasets $\mathbf{X}^1$ and $\mathbf{X}^2$ of N observations, the projection directions are found by maximizing the regularized correlation between the latent scores of $\mathbf{x}^1$ and $\mathbf{x}^2$

$$\max_{\boldsymbol{\psi}_i, \boldsymbol{\eta}_i} \frac{\mathrm{Cov}(\mathbf{X}^1\boldsymbol{\psi}_i, \mathbf{X}^2\boldsymbol{\eta}_i)^2}{\sqrt{\gamma + (1-\gamma)\mathrm{Var}(\mathbf{X}^1\boldsymbol{\psi}_i)}\sqrt{\gamma + (1-\gamma)\mathrm{Var}(\mathbf{X}^2\boldsymbol{\eta}_i)}}, \tag{7}$$

where $\boldsymbol{\psi}_i$ and $\boldsymbol{\eta}_i$ are the i-th directions of the projection matrices and $0 \leq \gamma \leq 1$ is a regularization coefficient [30, 31]. When $\gamma = 0$, it is unregularized canonical correlation analysis (CCA) [32]. When $\gamma = 1$, it corresponds to partial least squares (PLS) [33, 34]. Solving the optimization problem in Eq. (7) requires singular value decomposition or non-linear iterative methods (NIPLAS algorithm) to make projection directions orthogonal to each other. The computation costs of both methods are prohibitively expensive when the data dimension is high or the number of samples is large. Therefore, it is more desirable to alternatively update $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ in an iterative manner [35].

B Related methods

Unsupervised representation learning can be categorized into two classes based on the type of the pretext task: generative or discriminative. Generative approaches learn to generate or reconstruct unlabeled data in the input space [36, 37, 38]. Reconstructing the masked portion of data is highly successful in discovering useful representations in natural language processing [39, 2, 1]. Before the success of masked language models, variants of denoising or masked autoencoders are developed for CV tasks [8, 9, 10] but the performance is worse than discriminative SSL methods. It was not until recently that masked image models are revived in unsupervised visual representation learning by discretizing image patches via tokenizers [40, 41]. MAE [42] further simplifies masked image models by directly inpainting masked images without tokenizers or image-specific augmentations. Although masking is a simple data augmentation that can be flexibly applied to different domains of data, computationally expensive generation or reconstruction in the input space may not be necessary for representation learning. Our method is derived from a generative model and shares the idea of reconstructing the corrupted samples in masked autoencoding. Instead of reconstructing discrete tokens or raw inputs, our method reconstructs the continuous latent representation, which is related to discriminative SSL methods using data augmentations.

Augmentation-based discriminative SSL methods learn representation by *comparing* (including but not limited to contrastive learning) augmented views of unlabeled data in the latent space. This line of work involves a contrastive framework with variants of InfoNCE loss [43, 44] to pull together representations of augmented views of the same training sample (positive sample) and disperse representations of augmented views from different training samples (negative samples) [45, 46, 47]. Typically, contrastive learning methods require a large size of negative samples to learn high-quality representations from unlabeled data [3, 4]. Meanwhile, non-contrastive methods train neural networks to match the representations of augmented positive pairs without comparison to negative samples or cluster centers. However, non-contrastive methods suffer from trivial solutions where the model maps all inputs to the same constant vector, known as a collapsed representation. Various methods have been proposed to avoid a collapsed representation on an ad hoc basis, such as asymmetric network architecture [7], stop gradient [26], and feature decorrelation [48, 49, 50]. Interestingly, our method also includes a feature decorrelation constraint, which is adapted from a generative model. Recently, adversarial perturbations are combined with image augmentations to create more challenging positive and negative samples in SSL [51, 52]. APLR does not require domain-specific augmentations and can be applied to different domains of data.

Learning augmentations has been investigated in supervised learning to obtain data-dependent augmentation policies for better generalization [53, 54]. In parallel, adversarial perturbation can be treated as a special form of learnable augmentations to enhance the robustness of models with adversarial training [19, 20]. The domain-agnostic augmentations in our method are closely related to generative adversarial perturbation, where data augmentations are obtained through a forward pass of learnable generative models [25, 15, 16, 17]. The vast majority of adversarial perturbation methods rely on the classification boundary of the attacked neural network to train the generator via maximizing a cross-entropy loss. Those ideas have been extended to SSL to get adversarial perturbation by maximizing the InfoNCE loss in SimCLR [55, 28]. However, existing generative adversarial perturbation methods rely explicitly on the classification boundary or the instance discrimination boundary of the attacked model and tend to make them over-fit to the source data [18]. Instead of maximizing a cross-entropy loss, we maximize the ℓ_2 distance between mid-level feature maps to obtain generative adversarial perturbations.

Theoretical understanding of SSL has been studied under the assumption that augmented views of the same raw sample are somewhat conditionally independent given the label or a hidden variable [56, 57, 58, 59]. However, those assumptions do not hold in practice because augmented views of a natural sample are usually highly correlated. Augmented views are unlikely to be independent given the hidden label. Recent studies in contrastive learning provide theoretical guarantees of the learned representation without the assumption of conditional independence [60, 61]. In parallel, [6] investigates the training dynamics of non-contrastive SSL methods to show how feature collapse is avoided but lacks guarantees for solving downstream tasks. Note that our proposed method does not involve an explicit comparison between positive and negative samples. Our theoretical analysis relies on the divergence transition matrix without the assumption of conditional independence.

C Proof of the main theorem

The HGR maximal correlation can be estimated from divergence transition matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{X}^1| \times |\mathcal{X}^2|}$, whose entries are defined by the joint and marginal distributions of $\mathbf{x}^1$ and $\mathbf{x}^2$ [62]. Let P_{x^1} and P_{x^2} be the marginal distribution and $P_{x^1 x^2}$ be the joint distribution. $P_{x^1}(\mathbf{x}_i^1)$ can be viewed as the probability mass of $\mathbf{x}_i^1$ being randomly sampled from $\mathcal{X}^1$. Then, each entry of $\mathbf{A}$ is given by

$$\mathbf{A}_{ij} = \frac{P_{x^1 x^2}(\mathbf{x}_i^1, \mathbf{x}_j^2)}{\sqrt{P_{x^1}(\mathbf{x}_i^1) P_{x^2}(\mathbf{x}_j^2)}} \tag{8}$$

The solution to Eq. (1) is the sum of the top k singular values of $\mathbf{A}$ with left singular vectors $\mathbf{Z}^1 \in \mathbb{R}^{N \times k}$ and right singular vectors $\mathbf{Z}^2 \in \mathbb{R}^{N \times k}$ defined as

$$\begin{aligned} \mathbf{z}_i^1 &= \sqrt{P_{x^1}(\mathbf{x}_i^1)}\psi(\mathbf{x}_i^1), \quad i = 1, ..., N \\ \mathbf{z}_i^2 &= \sqrt{P_{x^2}(\mathbf{x}_i^2)}\eta(\mathbf{x}_i^2), \quad i = 1, ..., N \end{aligned} \tag{9}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ are the i-th row of embedding matrices $\mathbf{Z}^1$ and $\mathbf{Z}^2$, respectively [63]. This is essentially a rank-k approximation of $\mathbf{A}$ via minimizing $\|\mathbf{A} - \mathbf{Z}^1 {\mathbf{Z}^2}^\top\|_F^2$. Note that $\mathbf{x}^2$ is a clean sample and $\mathbf{z}^2$ is the representation of a clean sample. We use clean samples in downstream tasks. We drop the superscription to avoid cluttered notation.

The first term on the right hand side of the main theorem (theorem 3.3) measures the approximation error of the optimal classifier f^* by a linear classifier parameterized by $\mathbf{B}$. It amounts to the residual of the least squares problem $\|\mathbf{f}^* - \mathbf{Z}\mathbf{B}\|^2$ in Fig. 1, where the representation matrix $\mathbf{Z} \in \mathbb{R}^{N \times K}$ contains the top-k left singular vectors of $\mathbf{A}$ and $\mathbf{f}^* \in \{0, 1\}^N$ is the vector that contains the predicted labels of all the data by the optimal classifier f^*. The approximation error is bounded if $\mathbf{f}^*$ has limited projection into the residual subspace that is perpendicular to the column space of the representation.

In the first step, we construct a quadratic form of $\mathbf{f}^*$ to quantify its projection into the residual space based on singular value decomposition of $\mathbf{A}$. The largest singular value of $\mathbf{A}$ is 1, with constant left and right singular vectors being $\mathbf{1}$ and $\mathbf{1}$ [63]. Therefore, it is more convenient to subtract the top singular mode and introduce $\widetilde{\mathbf{A}} = \mathbf{I} - \mathbf{A}$. $\widetilde{\mathbf{A}}$ can be factorized as $\widetilde{\mathbf{A}} = \sum_{i=1}^N \gamma_i \mathbf{u}_i \mathbf{v}_i^\top$ via singular value decomposition, where γ_i is the i-th singular value of $\widetilde{\mathbf{A}}$ with the left singular vector $\mathbf{u}_i$ and the

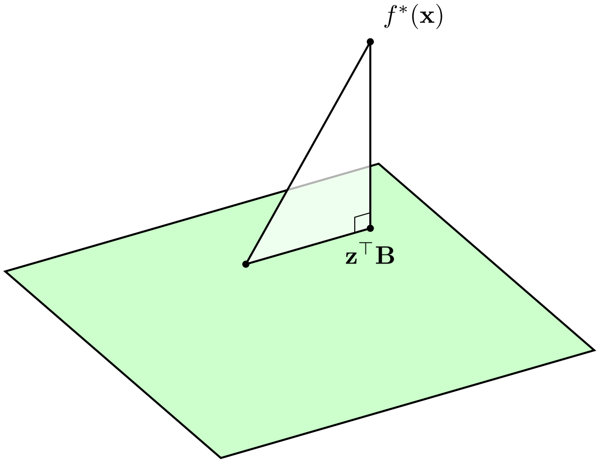

Figure 1: Geometric interpretation of least squares. $f^*(\mathbf{x}) : \mathcal{X} \to \{0, 1\}$ is the Bayes optimal classifier for predicting the label given $\mathbf{x}$ with error at most α according to Assumption 3.2. The green panel is the subspace spanned by the columns of the representation $\mathbf{z}$. $\mathbf{B}$ is the parameters of the linear classification model.

right singular vector $\mathbf{v}_i$. The quadratic form is given as follows

$$
\begin{aligned}
\mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^* &= \mathbf{f}^{*\top}\left(\sum_{i=1}^{k}\gamma_i\mathbf{u}_i\mathbf{v}_i^\top + \sum_{i=k+1}^{N}\gamma_i\mathbf{u}_i\mathbf{v}_i^\top\right)\mathbf{f}^* && (10)\\
&= \mathbf{f}^{*\top}\left(\sum_{i=1}^{k}\gamma_i\mathbf{u}_i\mathbf{u}_i^\top + \sum_{i=k+1}^{N}\gamma_i\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{f}^* && (11)\\
&\geq \mathbf{f}^{*\top}\left(\sum_{i=k+1}^{N}\gamma_i\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{f}^* && (12)\\
&\geq \mathbf{f}^{*\top}\left(\gamma_{k+1}\sum_{i=k+1}^{N}\mathbf{u}_i\mathbf{u}_i^\top\right)\mathbf{f}^* && (13)\\
&= \gamma_{k+1}\mathbf{f}^{*\top}\mathbf{P}\mathbf{f}^* = \gamma_{k+1}\mathbf{f}^{*\top}\mathbf{P}^\top\mathbf{P}\mathbf{f}^* = \gamma_{k+1}\|\mathbf{P}\mathbf{f}^*\|^2 && (14)
\end{aligned}
$$

where Eq. (11) is due to the fact that the left and right singular vectors are the same in the symmetric matrix $\widetilde{\mathbf{A}}$, the inequality in Eq. (12) is because of dropping a quadratic term, and Eq. (13) is due to $\gamma_{k+1} \leq \gamma_{k+2} \leq \ldots \gamma_N$. $\mathbf{P} \triangleq \sum_{i=k+1}^{N}\mathbf{u}_i\mathbf{u}_i^\top$ defines a projection matrix that projects $\mathbf{f}^*$ into a residual subspace spanned by singular vectors $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_N$. Eq. (14) is obtained because $\mathbf{P}$ is an idempotent matrix ($\mathbf{P}^2 = \mathbf{P}$) [64]. In addition, $(\mathbf{I} - \mathbf{P})\mathbf{f}^*$ is in the subspace spanned by singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$, which is the column space of $\mathbf{Z}$. Based on the geometric interpretation of the least squares problem $\|\mathbf{f}^* - \mathbf{Z}\mathbf{B}\|^2$, there exists $\mathbf{B}$ that such that $(\mathbf{I} - \mathbf{P})\mathbf{f}^* = \mathbf{Z}\mathbf{B}$ is the projection of $\mathbf{f}^*$ onto the column space of $\mathbf{Z}$.

In the second step, we upper bound $\gamma_{k+1}\|\mathbf{P}\mathbf{f}^*\|^2$. Based on Eq. (14), we have $\gamma_{k+1}\|\mathbf{P}\mathbf{f}^*\|^2 \leq \mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^*$. It is more convenient to upper bound $\mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^*$

$$\begin{aligned}
\mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^* &= \mathbf{f}^{*\top}\mathbf{I}\mathbf{f}^* - \mathbf{f}^{*\top}\mathbf{A}\mathbf{f}^* \\
&= \sum_i^N \mathbf{f}^*_{x_i}\mathbf{f}^*_{x_i} - \sum_{i,j=1}^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j)\Big(\frac{\mathbf{f}^*_{x_i}}{\sqrt{P_{x^1}(\mathbf{x}_i)}}\frac{\mathbf{f}^*_{x_j}}{\sqrt{P_{x^2}(\mathbf{x}_j)}}\Big) \\
&= \frac{1}{2}\Big(\sum_i^N \mathbf{f}^*_{x_i}\mathbf{f}^*_{x_i} - 2\sum_{i,j=1}^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j)\Big(\frac{\mathbf{f}^*_{x_i}}{\sqrt{P_{x^1}(\mathbf{x}_i)}}\frac{\mathbf{f}^*_{x_j}}{\sqrt{P_{x^2}(\mathbf{x}_j)}}\Big) + \sum_i^N \mathbf{f}^*_{x_j}\mathbf{f}^*_{x_j}\Big) \\
&= \frac{1}{2}\Big(\sum_i^N P_{x^1}(\mathbf{x}_i)\left(\frac{\mathbf{f}^*_{x_i}}{\sqrt{P_{x^1}(\mathbf{x}_i)}}\right)^2 - 2\sum_{i,j=1}^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j)\Big(\frac{\mathbf{f}^*_{x_i}}{\sqrt{P_{x^1}(\mathbf{x}_i)}}\frac{\mathbf{f}^*_{x_j}}{\sqrt{P_{x^2}(\mathbf{x}_j)}}\Big) \qquad (15) \\
&+ \sum_j^N P_{x^2}(\mathbf{x}_i)\left(\frac{\mathbf{f}^*_{x_j}}{\sqrt{P_{x^2}(\mathbf{x}_j)}}\right)^2\Big) \\
&= \frac{1}{2}\sum_i^N\sum_j^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j)\Big(\frac{\mathbf{f}^*_{x_i}}{\sqrt{P_{x^1}(\mathbf{x}_i)}} - \frac{\mathbf{f}^*_{x_j}}{\sqrt{P_{x^2}(\mathbf{x}_j)}}\Big)^2,
\end{aligned}$$

where $\mathbf{f}^*_x = f^*(\mathbf{x})$, $P_{x^1}(\mathbf{x}_i) = \sum_j^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j) = 1/N$ and $P_{x^2}(\mathbf{x}_j) = \sum_i^N P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j) = 1/N$. Note that we only sample a pair of samples $(\mathbf{x}_i,\mathbf{x}_j)$ where $\mathbf{x}_i$ is created from semantic-preserving perturbation of $\mathbf{x}_j$ to train the model. The probability mass $P_{x^1x^2}(\mathbf{x}_i,\mathbf{x}_j) > 0$ only if $(\mathbf{x}_i,\mathbf{x}_j)$ are generated from a shared latent variable. Let $(\mathbf{x},\mathbf{x}^+)$ be a positive pair to denote a pair of samples created from semantic-preserving perturbation. We can rewrite equation (15) as

$$\mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^* = \frac{N}{2}\mathbb{E}_{x,x^+}[(\mathbf{f}^*_x - \mathbf{f}^*_{x^+})^2], \tag{16}$$

where $\mathbb{E}_{x,x^+}[\left(\mathbf{f}^*_x - \mathbf{f}^*_{x^+}\right)^2]$ quantifies the probability that the optimal classifier $f^*(\cdot)$ predicts different labels for $(\mathbf{x},\mathbf{x}^+)$. When $\mathbf{f}^*_x \neq \mathbf{f}^*_{x^+}$, there must be $f^*(\mathbf{x}) \neq y(\mathbf{x})$ or $f^*(\mathbf{x}^+) \neq y(\mathbf{x})$. With Assumption 3.2, we have $\Pr(f^*(\mathbf{x}) \neq f^*(\mathbf{x}^+)) = 2\alpha$. As such, the quadratic form in Eq. (16) can be upper bounded:

$$\mathbf{f}^{*\top}\widetilde{\mathbf{A}}\mathbf{f}^* \leq N\alpha. \tag{17}$$

With Eq. (14) and (17), we have

$$\|\mathbf{f}^* - \mathbf{Z}\mathbf{B}\|^2 = \|\mathbf{P}\mathbf{f}^*\|^2 \leq \frac{N\alpha}{\gamma_{k+1}} \tag{18}$$

$$\frac{1}{N}\|\mathbf{f}^* - \mathbf{Z}\mathbf{B}\|^2 \leq \alpha/\gamma_{k+1} \tag{19}$$

The k-th singular value of $\mathbf{A}$ is $\lambda_k = 1 - \gamma_k$, which is also the k-th Hirschfeld-Gebelein-Rényi maximal correlation by definition. Therefore, we have $\Pr_{\mathbf{x}\sim P(\mathbf{x})}[y(\mathbf{x}) \neq f_B(\mathbf{x})] \leq \widetilde{O}\left(\frac{\alpha}{1-\lambda_{k+1}}\right)$.

The second term on the right hand side of the main theorem is the estimation error that measures sample complexity of learning $\mathbf{B}$ with access to n_2 i.i.d. training samples in the downstream task. It can be upper bounded using the Rademacher complexity of linear models. Let $\mathcal{H}_1 = \{\mathbf{z} \rightarrow \mathbf{z}^\top\mathbf{B} : \|\mathbf{B}\|_F \leq C_b\}$. We have the Rademacher complexity of the linear model

$$R_{n_2}(\mathcal{H}_1) = \frac{C_b\sqrt{C_z}}{\sqrt{n_2}} \tag{20}$$

where $\mathbb{E}[\|\mathbf{z}\|^2] \leq C_z$. By definition of Eq. (1), $\mathbb{E}[\|\mathbf{z}\|^2]$ captures the summation of first k HGR maximal correlation. $\mathbb{E}[\|\mathbf{z}\|^2] \leq k$ because the HGR maximal correlation less equal than 1. Therefore, we have

$$\Pr_{\mathbf{x}\sim P(\mathbf{x})}[y(\mathbf{x}) \neq f_{\hat{B}}(\mathbf{x})] \leq \widetilde{O}\left(\frac{\alpha}{1-\lambda_{k+1}} + \sqrt{\frac{k}{n_2}}\right).$$

D Datasets

D.1 Tabular Data

For tabular data, we follow existing works [11, 12] and use MNIST and Fashion-MNIST as proxy datasets by flattening the images into 1-dimensional vectors. In addition, we use two real tabular datasets from the UCI repository to evaluate the proposed method [65].

MNIST/Fashion-MNIST are two image datasets of handwritten digits and Zalando's article images, respectively [66, 67]. The images of size 28×28 are flattened into vectors with 784 features. Both datasets have 10 classes, and contain 60,000 training examples and 10,000 test examples.

Gas Concentrations is a dataset containing chemical sensor measurements of 128 features when exposed to 6 different gases [68, 69]. The classification task is to identify the target gas. We perform a 80/20 train/test split to obtain 11,128 training examples and 2,782 test examples.

Gesture Phase is a dataset containing 32 features extracted from videos of people in 5 different gestures [70, 65]. We perform a 80/20 train/test split to obtain 7,898 training examples and 1,975 test examples.

D.2 Iamge Datasets

CIFAR-10/100 are two datasets of tiny natural images with a size of 32×32 [71]. CIFAR-10 and CIFAR-100 have 10 and 100 classes, respectively. Both datasets contain 50,000 training images and 10,000 test images.

STL-10 is a 10-class image recognition dataset for unsupervised learning [72]. Each class contains 500 labeled training images and 800 test images. In addition, it also contains 100,000 unlabeled training images. Both labeled and unlabeled training images are used for feature extractor pretraining without using labels. The linear classifier is learned using the labeled training images.

Tiny-ImageNet is a subset of the ILSVRC-2012 classification dataset [73]. It consists of 200 classes, with 500 training images, 50 validation images, and 50 test images in each class. The size of each image is 64×64.

D.3 Audio Data

ESC-10/50 are two environmental sound classification datasets containing 5-seconds of environmental recordings [74]. ESC-10 and ESC-50 have 10 and 50 classes, and contain 400 and 2000 examples, respectively. We use the original fold settings from the authors [74], and follow the experimental setup in [27] to use the first fold for testing and the rest for training.

LibriSpeech-100 is a corpus of read English speech [75]. We use speaker identification as the downstream classification task. We follow the experimental setup from [28] to pretrain with the LibriSpeech-100 hour corpus which contains 28,539 examples, and perform linear evaluation on the LibriSpeech development set which contains 2,703 examples.

E Additional Details in Experiments

For tabular and audio experiments, we search the perturbation budget hyperparameter ϵ from the set $\{0.05, 0.1, 0.15\}$. For image experiments, we fix ϵ to 0.05 for a direct comparison with Viewmaker networks [28]. We find that constraining the perturbations to an ℓ_1 norm distance achieves the best results. For all experiments, we train the feature extractor and the adversarial generator in an alternating fashion. The feature extractor $\psi(\cdot)$ is trained with the SGD optimizer with momentum of 0.9 and weight decay of 5e-4. The learning rate is 0.03 without decay. The momentum coefficient in exponential moving average is set to 0.99 when updating $\eta(\cdot)$. The generator is trained with the Adam optimizer with an initial learning rate of 1e-3 and its architecture is described in the Appendix. Both the feature extractor and the generator are trained for 200 epochs with a batch size of 256. After self-supervised training on unlabeled data, a linear classifier is trained using SGD with a batch size of 256 and no weight decay for 200 epochs. The learning rate starts at 30.0 and is decayed to 0 after 200 epochs with a cosine schedule.

Tabular Data For tabular data, we follow existing works [11, 12] to use flattened MNIST [66] and Fashion-MNIST [67] as proxy datasets. In addition, we use two real tabular datasets from the UCI repository to evaluate the proposed method [65] on Gas [68, 69] and Gesture [65, 70] classification (Appendix D). We adopt a 10-layer MLP with residual connections [76] as the feature extractor, and adapt the generator from [18] by replacing convolutional layers with linear layers. The linear evaluation results on test datasets are reported in Table 1. APLR outperforms VIME-Self [12], which corrupts tabular data and uses mask vector estimation and feature vector estimation as pretext tasks, on three out of four datasets. It aligns with the empirical observations that reconstructing high-dimensional data in the input space is not necessary for learning high-quality representations. APLR outperforms the domain-agnostic benchmark DACL [11], which uses mixup noise, on all datasets. Mixup noise is less effective than adversarial noise because it perturbs informative and uninformative dimensions in the input space uniformly. Furthermore, convex combinations in the input space via mixup may result in augmented views off the data manifold. Interestingly, our proposed APLR also outperforms training the full architecture in a supervised manner on the two real tabular datasets, Gas and Gesture.

Image Data We use four benchmark image datasets to evaluate the effectiveness of the proposed method, including CIFAR-10/100 [71], STL-10 [72] and Tiny-ImageNet [73] (Appendix D). ResNet18 [77] is adopted as the backbone network in the feature extractor. We adopt the generator in [18] for image data. We present the results for self-supervised representation learning on image data in Table 3. It is observed that APLR outperforms DACL [11] by a large margin, indicating that adversarial noise is a more effective semantic-preserving perturbation than mix-up noise in DACL. Interpolation of input samples via mix-up could lead to out-of-distribution training samples because the input data space may not be convex. Our method also achieves better performance than Viewmaker [28], which is a domain-agnostic SSL method by discriminating adversarially perturbed data. The adversarial noise in APLR is more robust because the training process of adversarial noise in APLR does not rely on the classification boundary between augmented samples [18]. Furthermore, we also compare APLR against methods that use image augmentations (e.g. cropping, rotation, horizontal flipping), such as SimCLR [3]. It is found in previous studies [7, 3] that random crop is a crucial data augmentation towards learning high-quality representations for image data. However, it is impossible to create cropped views of images using adversarial perturbation because the adversarial noise is additive to the natural sample. Given the importance of random crop and the inability to create cropped views with adversarial perturbations, achieving comparable accuracies between APLR and SimCLR indicates that adversarial noise is a highly effective data augmentation method.

Audio Data We use three audio datasets to evaluate APLR: ESC-10, ESC-50 [74] and LibriSpeech-100 [75] (Appendix D). For audio experiments, we use 1-D ResNet18 [77] as the feature extractor and adopt the generator in [18] with one input channel. The time-frequency representation is a 2D log mel spectrogram, normalized to zero mean and variance. We report the results on audio data benchmarks in Table 2, and visualize examples from LibriSpeech-100 in Appendix E.4. APLR performs significantly better than CLAR [27], which experimented extensively with combinations of audio-specific augmentations and uses fade in/out and time masking as their best-performing augmentations. Compared to image augmentations, data augmentations for audio data are relatively underexplored. Our results demonstrate the advantage of learning audio augmentations over manually designed augmentations. Our proposed method also outperforms both domain-agnostic methods, DACL and Viewmaker. DACL performs close to APLR on the simple yet small ESC-10 and ESC-50 datasets. However, it is unable to learn effective representations on LibriSpeech-100 which is larger and significantly more complex. Even though both APLR and Viewmaker use adversarial noise, APLR outperforms Viewmaker by a large margin across the benchmarks, indicating the effectiveness of learning augmentations by maximizing the discrepancy between latent representations. In Table 2, we also report results on training the full architecture in a supervised manner. We find that linear classifiers trained on top of the representations learned by APLR outperforms the supervised model on ESC-10, and closes the gap to ESC-50 compared to other benchmarks, demonstrating the ability for APLR to learn useful latent representations. Current state-of-the-art supervised approaches report high accuracies (over 94%) on the ESC-50 dataset [78, 79]. However, they perform pretraining using large datasets such as AudioSet and ImageNet, and use multiple audio-specific data augmentations. With the supervised training experiments, we do not perform pretraining with large datasets, and we use time masking and frequency masking as augmentations. Our goal is to simply compare APLR against training the same architecture in a supervised manner.

To understand the effectiveness of adversarial perturbations within APLR, we perform several additional experiments. First, we compare perturbation by adversarial noise against perturbation by Gaussian noise and random masking. For image datasets, we additionally compare the proposed adversarial perturbations against common image augmentations used in supervised learning, including CutMix [14], RandAugment [53], and Random Erasing. Next, we explore the sensitivity of APLR to different perturbation strengths and Lagrange multipliers. Lastly, we compare our framework against SOTA SSL methods on image data.

E.1 Ablation Study

First, for all datasets, we perform ablations to compare perturbations with adversarial noise against Gaussian noise and masking. To obtain a sample augmented with Gaussian noise, we use $\mathbf{x}^1 = \mathbf{x}^2 + \delta$, where $\delta \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. For each dataset, we search the standard deviation σ from the set $\{1, 3, 5, 10\}$ and report the best linear evaluation accuracy. For experiments with masking, we randomly mask a proportion of the clean sample $\mathbf{x}^2$. We search the proportion of masking from the set $\{20\%, 40\%, 50\%, 60\%, 70\%\}$ and report the best linear evaluation accuracy.

Tables 4 - 6 summarize the results. The adversarial noise outperforms the Gaussian noise and random masking on all datasets, except MNIST. Random noises may not be as effective since uniformly perturbing uninformative features may not lead to the intended goal of augmentations. That is why APLR leads to significant improvement over random perturbations on complex data, such as images and audios.

Table 4: Ablation study on tabular data.

	MNIST	Fashion-MNIST	Gas	Gesture
Gaussian noise	97.43	85.77	96.25	41.46
Masking	**97.58**	86.95	95.70	42.30
APLR	97.11	**87.12**	**97.98**	**42.97**

Table 5: Ablation study on image data.

	CIFAR-10	CIFAR-100	STL-10	Tiny-ImageNet
Gaussian Noise	53.58	28.43	52.76	12.07
Masking	48.79	27.43	50.39	11.29
APLR	**85.92**	**55.83**	**86.21**	**42.93**

Table 6: Ablation study on audio data.

	ESC-10	ESC-50	LibriSpeech-100
Gaussian noise	75.00	41.75	78.52
Masking	77.50	45.00	76.76
APLR	**81.25**	**57.75**	**96.29**

Additionally, we compare adversarial noise against image augmentations in supervised learning, namely CutMix [14], RandAugment [53], and Random Erasing . The results are summarized in Table 7. Random Erasing results in the worst performance among all methods, while CutMix is on par with Mixup in SSL. This is expected because CutMix performs slightly better or similar to MixUp in supervised learning. RandAugment leads to better performance than CutMix and MixUp because RandAugment contains a wide range of image augmentations. However, RandAugment does not outperform SimCLR. The studies in SimCLR show that careful selection of image augmentations is necessary for good performance in SSL. Some effective image augmentations in supervised learning do not lead to good performance in SSL.

E.2 Sensitivity Analysis

We perform experiments to understand how sensitive APLR is to different strengths of the adversarial perturbation and Lagrange multiplier.

Table 7: Additional ablation study on image data.

	CIFAR-10	CIFAR-100	STL-10	Tiny-ImageNet
SimCLR	**86.47**	54.86	85.49	**43.27**
CutMix	61.27	35.14	58.33	22.83
RandAugment	84.34	51.92	84.37	40.68
Random Erasing	56.71	28.89	55.91	18.45
Ours	85.92	**55.83**	**86.21**	42.93

We experiment with perturbation strengths of $\epsilon \in \{0.05, 0.1, 0.15\}$, and report the results in Table 8. The sensitivity analysis indicates that our method is robust to the adversarial perturbation strengths.

Table 8: Sensitivity to adversarial perturbation strengths.

	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.15$
Tabular Data			
MNIST	97.11	96.15	93.73
Fashion-MNIST	87.04	86.40	84.14
Gas	97.50	97.19	97.98
Gesture	42.97	40.90	41.46
Image Data			
CIFAR-10	85.92	84.66	85.26
CIFAR-100	55.83	54.37	54.77
STL-10	86.21	85.04	85.64
Tiny-ImageNet	42.93	42.42	41.47
Audio Data			
ESC-10	78.75	81.25	77.50
ESC-50	54.25	54.50	57.75
LibriSpeech-100	93.55	96.29	96.29

For the Lagrange multiplier, we experiment with $\gamma \in \{0.1, 0.5, 0.1\}$ and report the results in Table 9. We find that our method is robust to γ and achieves strong performance. For APLR, we selected γ = 1 as the default value since it performed well consistently.

Table 9: Sensitivity to Lagrange multiplier.

	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1$
Tabular Data			
MNIST	96.54	96.93	97.11
Fashion-MNIST	86.83	86.69	87.12
Gas	84.61	97.97	97.98
Gesture	40.35	40.35	42.97
Audio Data			
ESC-10	80.00	75.00	81.25
ESC-50	47.75	44.50	57.75
LibriSpeech-100	89.45	87.30	96.29

Our sensitivity analyses indicate that our method is robust to hyperparameters such as ϵ and γ. The proposed APLR achieves strong performance as long as the hyperparameter values are within reasonable ranges.

E.3 APLR Against SOTA Image-Specific SSL Methods

We perform an analysis to compare the proposed framework against SOTA SSL methods on images, namely SimCLR [3], Barlow Twins [49], and BYOL [7]. For this experiment, we use the image augmentations described in SimCLR [3] for a fair comparison against image-specific SSL methods.

We train each model for 200 epochs and summarize the results in Table 10. Our method achieves comparable performance to BYOL and Barlow Twins.

Table 10: APLR vs. SOTA SSL methods on image data

	CIFAR-10	CIFAR-100	STL-10	Tiny-ImageNet
SimCLR	86.47	54.86	85.49	43.27
Barlow Twins	89.02	**62.84**	85.43	**45.33**
BYOL	88.54	61.76	85.59	42.75
Ours	**89.63**	62.55	**86.41**	44.76

E.4 Visualizations of Original and Perturbed Spectrograms

In Figure 2, we visualize random spectrograms from LibriSpech-100 and the deltas between the original and perturbed spectrograms. The perturbations are indistinguishable and thus semantic-preserving.

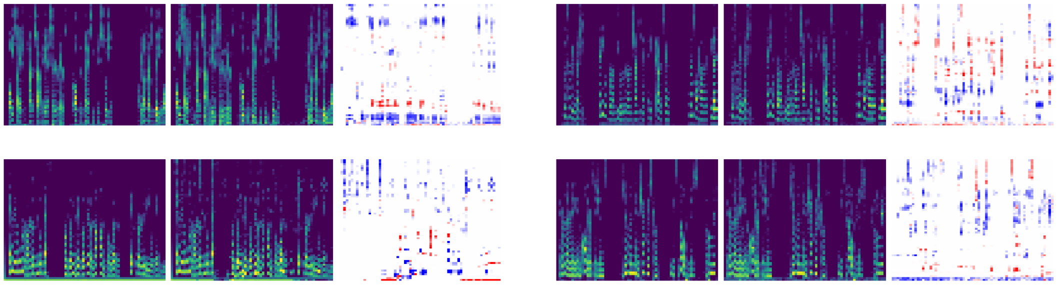

Figure 2: Examples triplets of original spectrograms (left), perturbed spectrograms (middle) and their differences (right) from LibriSpeech-100. The color scales for original and perturbed spectrograms are set to the scale of the original spectrogram. The color scale for the differences is set to -2.5 (red) to + 2.5 (blue), though some values exceed this range. Best viewed when zoomed.

E.5 Adversarial Generator Architecture

The architecture of the generator is described in Table 11. For experiments on tabular data, we replace the convolution layers with fully connected layers.

Table 11: Architecture of the adversarial generator.

Layer	**Number of Filters**	**Kernel Size**
Convolution Layer	32	9
Convolution Layer	64	3
Convolution Layer	128	3
Residual Block	128	3
Residual Block	128	3
Residual Block	128	3
Upsampling Convolution Layer (Upsample = 2)	64	3
Upsampling Convolution Layer (Upsample = 2)	32	3
Convolution Layer		