Supplementary Material:
Concurrent Discrimination and Alignment for Self-Supervised Feature Learning

Anjan Dutta[1], Massimiliano Mancini[2], Zeynep Akata[2]
[1]University of Exeter, [2]University of Tübingen

Experiments with β on STL-10: Since our mutual information estimator is trained together with other parametric models, the hyperparameter β is slowly increased during training, starting from a very small value 10^{-6} to its final value with an exponential scheduling. We set the number of epochs by which β reaches to $\beta_{\text{end\_value}}$ from $\beta_{\text{start\_value}}$ equal to 100. We experiment with two parameters $\beta_{\text{start\_epoch}}$ and $\beta_{\text{end\_value}}$ which are related to the β hyperparameter. First, we experiment with the start epoch ($\beta_{\text{start\_epoch}}$) which indicates the epoch at which β starts increasing to the final value $\beta_{\text{end\_value}} = 1.0$ during 100 epochs. We consider $\beta_{\text{start\_epoch}} = 10, 30, 50, 70, 90$ and the obtained results are shown in Table 1 (top), where we observe that for $\beta_{\text{start\_epoch}} = 10$, the obtained results across all the convolutional layers of AlexNet are consistently better that the other $\beta_{\text{start_epoch}}$s. Next, we experiment with $\beta_{\text{end\_value}} = 0.00001, 0.0001, 0.01, 1.0$ and set $\beta_{\text{start\_epoch}} = 10$ (as revealed the best), and the obtained results are presented in Table 1 (bottom), where we can see that with $\beta_{\text{end\_value}} = 1.0$, we steadily achieve superior results across different convolutional layers.

	STL-10				
$\beta_{\text{start\_epoch}}$	`c1`	`c2`	`c3`	`c4`	`c5`
10	**60.5**	71.5	**74.3**	**75.3**	75.4
30	59.7	70.7	73.6	74.7	75.2
50	59.6	**71.9**	74.0	74.5	74.4
70	59.8	70.4	74.0	74.6	75.3
90	60.3	71.3	73.4	74.6	**75.5**
$\beta_{\text{end\_value}}$	`c1`	`c2`	`c3`	`c4`	`c5`
0.00001	60.1	71.3	73.2	74.4	74.7
0.0001	60.3	71.3	73.6	74.6	74.9
0.01	**60.5**	**71.6**	73.7	75.1	75.1
1.0	**60.5**	71.5	**74.3**	**75.3**	**75.4**

Table 1. Ablating different $\beta_{\text{start_epoch}}$s (top): We report test set performance of our pre-text model trained with different $\beta_{\text{start\_epoch}}$. Ablating different $\beta_{\text{end_value}}$s (bottom): We report the test set performance of our final model by varying the $\beta_{\text{end_value}}$s. (STL-10 with CNN backbone AlexNet with conv layers (`c1-5`).

ResNet Experiments on STL-10: We also perform additional experiments with a more modern network architecture on STL-10. For doing so, we follow [7, 6] and consider the ResNet-34 [2] framework instead of the AlexNet [8]. We train our model to solve our hybrid discriminating and aligning pretext task for 200 epochs on 100K unlabeled training samples of STL-10. Once pretrained, we use those weights to initialize the network for downstream classification task on STL-10, and fine tune the model for 300 epochs on the 5K labeled training images and evaluate on the 8K test images.

Method	Accuracy
MultTaskBayes [10]	70.1%
DiscUFL [1]	74.2%
StackedAE [11]	74.3%
DiscAttr [4]	76.8%
ScaleScatter [9]	87.6%
SpotArtifacts [5]	80.1%
InfoMax [3]	77.0%
IIC [7]	88.8%
GlobStat [6]	91.8%
Ours	**92.1%**

Table 2. Comparison of test set accuracy on STL-10 with other published results.

We compare our CODIAL model with nine existing works. Among them, MultTaskBayes [10] proposes multi-task Gaussian processes to the Bayesian optimization framework; DiscUFL [1] trains the network to discriminate between a set of surrogate classes; StackedAE [11] presents an architecture, called *stacked what-where auto-encoders*, which integrates discriminative and generative pathways and uses a convolutional net to encode the input, and employs a deconvolutional net to produce the reconstruction; DiscAttr [4] trains a CNN coupled with unsupervised discriminative clustering, which uses the cluster membership as a soft supervision to discover shared attributes from the clusters while maximizing their separability; ScaleScatter [9] uses the scattering transform in combination with convolutional architectures; SpotArtifacts [5] learns self-supervised knowledge by spotting synthetic artifacts in images; InfoMax [3] learns representations by maximizing mutual information between two transformed views of the same image which are formed by applying a variety of transformations to a randomly sampled 'seed' image patch; and GlobStat [6] distinguishes diverse image transformations, such as rotation angles, warping and limited context inpainting. Table 2 presents the results obtained by our CODIAL model and compares it with other methods mentioned above, which shows that our CODIAL achieves highest results with ResNet-34 backbone as well and surpasses the closest model GlobStat by a margin of 0.3%. This proves that our model is effective with other backbone network as well and can be benefited from our joint discriminating and aligning pretext task.

References

[1] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *IEEE TPAMI*, 2015. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 1

[3] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 1, 2

[4] Chen Huang, Chen Change Loy, and Xiaoou Tang. Unsupervised learning of discriminative attributes and visual representations. In *CVPR*, 2016. 1

[5] Simon Jenni and Paolo Favaro. Self-Supervised Feature Learning by Learning to Spot Artifacts. In *CVPR*, 2018. 1, 2

[6] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics. In *CVPR*, 2020. 1, 2

[7] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *ICCV*, 2019. 1

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 1

[9] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. In *ICCV*, 2017. 1

[10] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-Task Bayesian Optimization. In *NIPS*, 2013. 1

[11] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann LeCun. Stacked What-Where Auto-encoders. In *ICLRW*, 2016. 1