Supplementary Material
Contrastive Learning Relies More on Spatial Inductive Bias Than Supervised Learning: An Empirical Study

Yuanyi Zhong[1†] Haoran Tang[2†] Jun-Kun Chen[1†] Yu-Xiong Wang[1]
[1]University of Illinois at Urbana-Champaign [2]University of Pennsylvania [†]Equal Contribution
{yuanyiz2, junkun3, yxw}@illinois.edu thr99@seas.upenn.edu

This document contains additional descriptions and extra experiments. The content of this document is summarized as below:

A. Additional Results

A.1. Variance of pre-training results

We repeat MoCo-v2 on the original CIFAR-10 200ep three times: The KNN evaluation mean and std is 82.44 ± 0.18. Repeating MoCo-v2 on the global 8x8 shuffling corrupted CIFAR-10 gives KNN evaluation mean and std 59.24 ± 0.40. The linear evaluation variance is similar. The randomness has a smaller order than the gap between MoCo and Sup results.

A.2. Analysis of the feature space

In addition to feature uniformity, we also utilize feature distance to evaluate the learning dynamics of CL and SL. Denoting $\mathcal{D}_i$ and $\mathcal{D}_j$ as feature matrices of two classes, the feature distance is calculated as: $d(f_t, \mathcal{D}_i, \mathcal{D}_j) = \mathbb{E}_{x_0 \sim \mathcal{D}_i, x_1 \sim \mathcal{D}_j} \left[\|f_t(x_0) - f_t(x_1)\|_2^2 \right]$. Note that if $\mathcal{D}_i = \mathcal{D}_j$, it actually measures the intra-class variance of class i. We plot curves of feature distance in Figure A.1. We are also interested in SupCon, because it bridges CL and SL by leveraging a similar contrastive loss. As illustrated, the overall feature uniformity of MoCo-v2 [1] is greater than

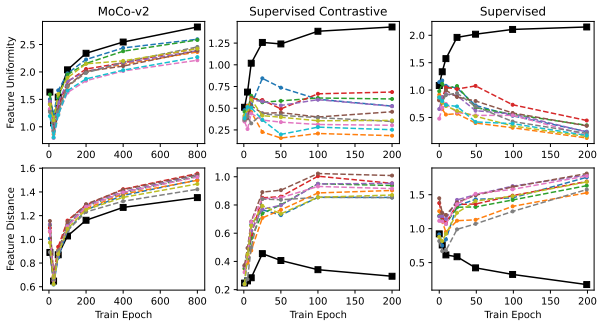

Figure A.1: **Above:** Solid black line – uniformity of the overall feature space. Dashed lines – class-wise feature uniformities of the 10 classes. While the overall uniformity of all methods grows, the uniformity of each class of Sup or SupCon is shrinking as training progresses. In the end, the overall uniformity of MoCo is the largest. **Below:** Solid black line – $d(f_t, \mathcal{D}_0, \mathcal{D}_0)$, *i.e.*, the intra-class variance of class 0. Dashed lines – feature distances between $\mathcal{D}_i (i \neq 0)$ and $\mathcal{D}_0$. The intra-class variance behavior of MoCo (increasing) is the opposite to that of Sup or SupCon (decreasing).

2.5 and approaching 3, while the overall uniformity of supervised contrastive learning (SupCon) [2] and supervised methods range from 1.25 to 2.2. This means that features from CL methods are more uniformly distributed on the unit sphere, or captures more information.

By looking at the class-wise feature uniformity, we notice that SL tends to compress (and maybe over-compress) the features of each class, while CL steadily increases the overall and class-wise uniformity. SupCon has the smallest uniformity values and similar patterns with SL due to the same way of supervision, and the class-wsie uniformity stabilizes with longer training. We hypothesize that such dynamics is related to its contrastive objectives. By looking at the feature distance, we can observe the similar trends for overall and class-wise measurements. While CL has a steadily increasing overall intra-class variance and class-wise feature distance w.r.t class 0, SupCon and SL have a shrinking overall intra-class variance and increasing class-wise feature distance. On the other hand, SupCon has a similar class-wise dynamics as that of CL in the sense of

stable class-wise relations: the differences between classes stabilize with time, while those of SL approach to a closer range.

B. Additional Visualization

B.1. Visualizing Grad-CAM attention maps

Figure 2 in our main paper visualizes the Grad-CAM [3] attention maps of ResNet-18 models pre-trained and linearly fine-tuned on either uncorrupted or global patch-shuffled images.

To further understand qualitatively how different corruption strategies impact the model's ability to learn semantic concepts, we draw the CAMs of models trained under different corruption settings on the corrupted versions of two ImageNet validation images in Figures B.1 and B.2. Global shuffling and defocus blur especially hinder the ability of MoCo to learn meaningful semantics.

B.2. Visualizing corrupted images

Please check Figure B.3 for more visual examples of the pixel-level gamma distortion and patch-level shuffling corruptions we used.

C. Using Uniformity As Performance Estimation Metrics

[5] derives uniformity from contrastive loss and shows that improving uniformity helps minimize contrastive loss (thus higher accuracy). As their proof does not rely on specific assumptions about data cleanliness, we hypothesize that such relationship extends to scenarios where images have spatial corruptions. Note, as uniformity is a metric derived specifically from the CL objective, there may not be an explicit relationship with SL.

We present uniformity scores of pre-trained models with data from another domain in the **table below**. We observe within the same type of corruption (global or local shuffling), higher uniformity implies higher accuracy. A comprehensive investigation on utilizing uniformity as metric in various scenarios is beyond the scope of this paper, and we leave it as interesting future work.

Regarding other calibration metrics, we pick E.C.E [4] and it does not show a clear relationship with model performance (Table C.1), indicating identifying a good metric is non-trivial.

D. Limitations

In this paper, we aim to observe, define, and analyze the dependency of CL on spatial inductive bias. We regard such dependency an intrinsic property of CL, rather than a weakness, constraint, or issue that needs to be solved or mitigated. Therefore, we did not propose a workaround or a new method to decrease such dependency.

This paper mainly focuses on CL methods among all SSLs. It remains an interesting future work to investigate the dependency of such inductive biases in non-CL SSLs.

References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[2] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 1

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[4] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization, 2022. 2

[5] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *CoRR*, abs/2005.10242, 2020. 2

Table C.1: Experiments shows that E.C.E. does not work as a performance estimation indicator like uniformity.

Item	Ori / G.1	Glo.4	Glo.8	G.32 / L.1	Loc.4	Loc.8	L.32 / Ori.
Acc.	96.80	89.60 (7.4%↓)	81.50 (15.8%↓)	29.44 (69.6%↓)	77.40 (20.0%↓)	89.40 (7.6%↓)	96.80
Uni.	2.59	2.16 (16.6%↓)	2.11 (18.3%↓)	2.10 (19.0%↓)	2.13 (17.8%↓)	2.18 (15.8%↓)	2.59
E.C.E.	1.97	2.80 (42.1%↑)	0.06 (97.0%↓)	1.18 (40.1%↓)	0.69 (65.0%↓)	1.64 (16.8%↓)	1.97

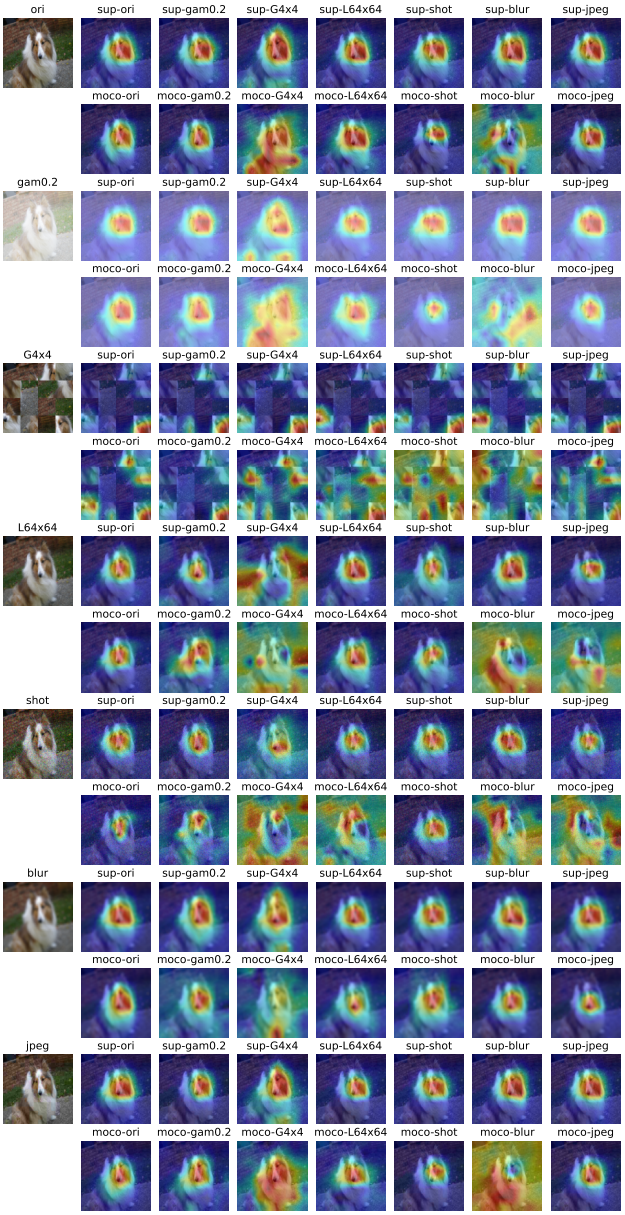

Figure B.1: GradCAM on corrupted versions of a dog image of sup/MoCo models trained under 7 corruptions.

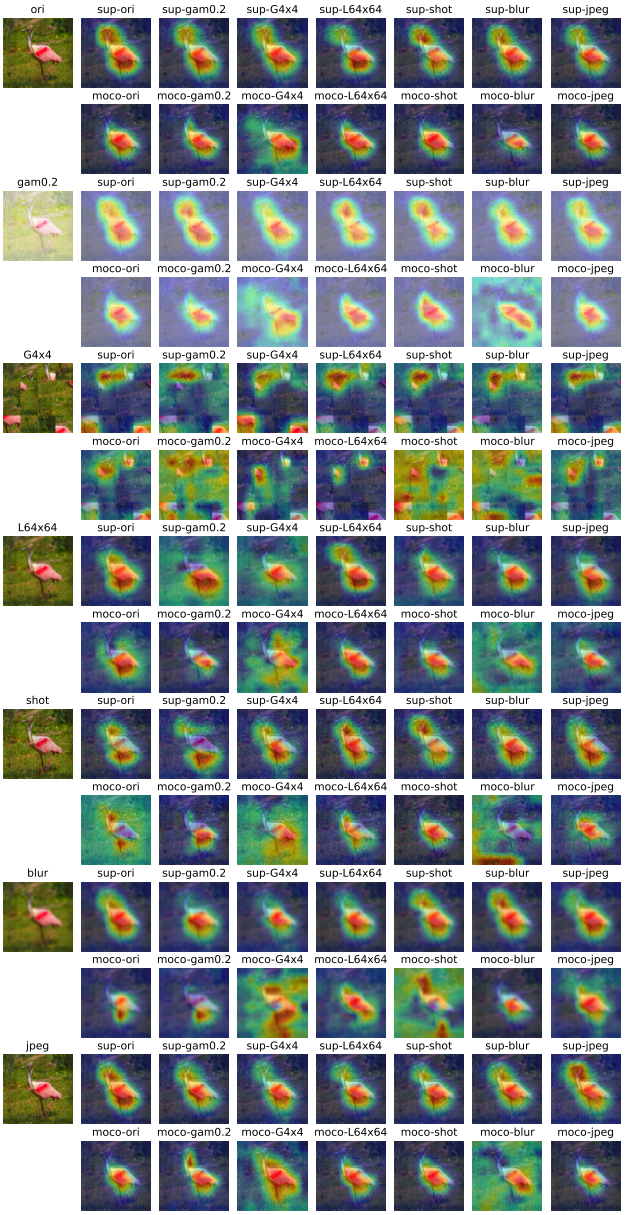

Figure B.2: GradCAM on corrupted versions of a bird image of sup/MoCo models trained under 7 corruptions.

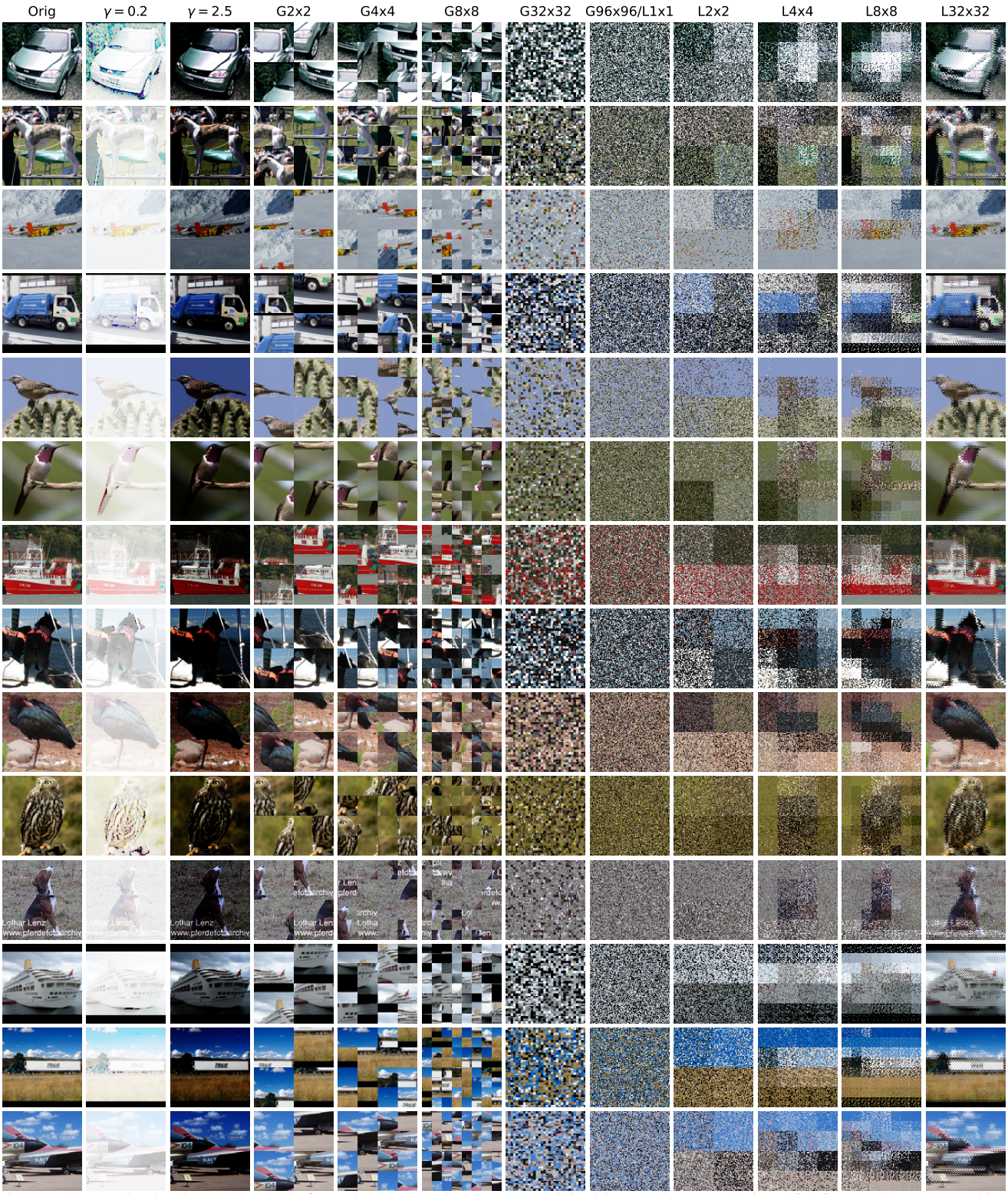

Figure B.3: Randomly chosen examples from the STL-10 dataset. The original images have resolution 96x96. We show the resulting images of gamma distortion ($\gamma = 0.2, 2.5$), global patch shuffling, and local patch shuffling. G1x1 and L96x96 revert to the original, while G96x96 and L1x1 are the most random ones (and have similar effect). Gamma distortion reduces information in pixel intensity. Global shuffling destroys global but preserves local structure, while local shuffling is the opposite.