A DERIVATION FROM EQ. (8) TO (9)

$$\arg\min_f \sum_{i=1}^{n} -\log\left(\sum_{j=1}^{M} \frac{e^{-\frac{1}{2}\left\|f(x_i)-f(x_j^+)\right\|^2}}{\sum_{k\neq i} e^{-\frac{1}{2}\|f(x_i)-f(x_k)\|^2}}\right)$$

$$=\arg\min_f \sum_{i=1}^{n} -\log\left(\sum_{j=1}^{M} \frac{e^{f(x_i)^T f(x_j^+)-\frac{1}{2}\|f(x_i)\|^2-\frac{1}{2}\left\|f(x_j^+)\right\|^2}}{\sum_{k\neq i} e^{f(x_i)^T f(x_k)-\frac{1}{2}\|f(x_i)\|^2-\frac{1}{2}\|f(x_k)\|^2}}\right) \tag{S1}$$

$$=\arg\min_f \sum_{i=1}^{n} -\log\left(\sum_{j=1}^{M} \frac{e^{f(x_i)^T f(x_j^+)-1}}{\sum_{k\neq i} e^{f(x_i)^T f(x_k)-1}}\right) \tag{S2}$$

$$=\arg\min_f \sum_{i=1}^{n} -\log\left(\frac{\sum\limits_{j=1}^{M} e^{f(x_i)^T f(x_j^+)}}{\sum_{k\neq i} e^{f(x_i)^T f(x_k)}}\right)$$

$$=\arg\min_f \sum_{i=1}^{n} -\log\left(\frac{\sum\limits_{j=1}^{M} e^{f(x_i)^T f(x_j^+)}}{\sum_{k\neq i, x_k\in\{x_j^+\}} e^{f(x_i)^T f(x_k)}+\sum_{k\neq i, x_k\notin\{x_j^+\}} e^{f(x_i)^T f(x_k)}}\right) \tag{S3}$$

$$=\arg\min_f \mathbb{E}_{x\sim\mathcal{D}}\left[-\log\left(\frac{\sum\limits_{j=1}^{M} e^{f(x)^T f(x_j^+)}}{\sum\limits_{j=1}^{M} e^{f(x)^T f(x_j^+)}+\sum\limits_{i=1}^{N} e^{f(x)^T f(x_i^-)}}\right)\right] \tag{S4}$$

$$=\arg\min_f \mathbb{E}_{x\sim\mathcal{D}}\left[-\log\left(\frac{\sum\limits_{j=1}^{M} e^{f(x)^T f(x_j^+)}}{\sum\limits_{j=1}^{M} e^{f(x)^T f(x_j^+)}+Ng_0(x,\{x_i^-\}^N)}\right)\right],$$

where we go from Eq. equation S1 to Eq. equation S2 based on the fact that $\|f(x)\|=1$, and from Eq. equation S3 to Eq. equation S4 assuming that set $\{x_k : k\neq i\}=\{x_j^+ : 1\leq j\leq M\}\cup\{x_i^- : 1\leq i\leq N\}$.

B COMPLETE TABLES OF RESULTS

We give the full table of results in Section 5 in the following. Notably, we gather the standard accuracy, adversarial accuracy, transfer accuracy, and transfer adversarial accuracy for each specification.

Table S1: The effectiveness evaluation of NaCl on SimCLR (i.e. $\alpha = 0, g^1 = g_0$). The best performance within each loss type is in boldface. We color the overall best performance in blue.

M	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{VAR}}(g_0, M)$			
	CIFAR100 Acc.	FGSM Acc.	CIFAR10 Acc.	FGSM Acc.
1	53.69±0.25	25.17±0.55	76.34±0.28	43.50±0.41
2	56.04±0.17	27.19±0.79	77.32±0.14	**44.61±0.33**
3	57.11±0.21	27.39±0.36	78.02±0.27	44.23±0.39
4	57.27±0.14	27.63±0.78	77.91±0.29	42.97±0.61
5	**57.91±0.12**	**28.37±0.56**	**78.09±0.29**	44.51±0.44
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{BIAS}}(g_0, M)$			
1	53.69±0.25	25.17±0.55	76.34±0.28	43.50±0.41
2	55.72±0.15	27.04±0.45	77.40±0.14	44.58±0.41
3	56.67±0.12	**28.41±0.24**	77.53±0.24	**45.21±0.89**
4	57.09±0.26	28.20±0.81	77.75±0.22	45.13±0.44
5	**57.32±0.17**	28.33±0.59	**77.93±0.40**	44.46±0.53
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_0, M, 0.5)$			
1	53.69±0.25	**25.17±0.55**	76.34±0.28	**43.50±0.41**
2	54.76±0.29	23.66±0.27	76.78±0.26	40.76±0.66
3	55.21±0.17	24.46±0.44	77.45±0.18	41.78±0.80
4	55.68±0.27	24.19±0.46	77.40±0.24	41.33±0.34
5	**55.85±0.16**	24.01±0.91	**77.50±0.16**	40.77±0.66
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_0, M, 0.6)$			
1	53.69±0.25	25.17±0.55	76.34±0.28	**43.50±0.41**
2	54.84±0.35	25.94±0.81	77.11±0.15	42.81±0.83
3	55.49±0.13	**26.25±0.89**	76.95±0.32	42.99±0.96
4	55.65±0.24	25.41±0.53	**77.39±0.37**	42.69±1.20
5	**55.66±0.22**	26.01±0.60	77.26±0.48	43.06±0.79
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_0, M, 0.7)$			
1	53.69±0.25	25.17±0.55	76.34±0.28	43.50±0.41
2	55.57±0.32	27.67±0.60	77.09±0.27	44.68±0.71
3	55.83±0.25	27.72±0.59	77.23±0.28	43.68±0.72
4	56.29±0.25	**27.92±0.60**	77.33±0.29	44.69±0.82
5	**56.37±0.32**	27.78±0.54	**77.40±0.20**	**45.07±0.98**
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_0, M, 0.8)$			
1	53.69±0.25	25.17±0.55	76.34±0.28	43.50±0.41
2	55.75±0.21	29.30±0.86	76.80±0.20	46.56±1.02
3	56.27±0.26	**29.96±0.29**	77.11±0.37	46.52±0.50
4	**56.39±0.26**	29.49±0.65	77.34±0.31	46.79±0.93
5	56.23±0.13	29.47±0.95	**77.40±0.14**	**47.36±0.69**
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_0, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_0, M, 0.9)$			
1	53.69±0.25	25.17±0.55	76.34±0.28	43.50±0.41
2	56.20±0.33	30.95±0.36	76.96±0.15	**48.85±0.75**
3	56.41±0.13	**30.98±0.90**	77.10±0.21	48.76±0.63
4	56.00±0.42	29.90±0.63	**77.11±0.40**	48.16±0.40
5	**56.63±0.31**	30.58±0.52	77.04±0.19	47.96±0.46

Table S2: The effectiveness evaluation of NaCl on Debised+HardNeg (i.e. $\alpha = 0, g^1 = g_2$). The best performance within each loss type is in boldface. We color the overall best performance in blue.

M	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{VAR}}(g_2, M)$			
	CIFAR100 Acc.	FGSM Acc.	CIFAR10 Acc.	FGSM Acc.
1	56.83±0.20	31.03±0.41	77.24±0.29	**48.38±0.70**
2	58.17±0.39	31.92±0.45	77.43±0.18	48.05±0.38
3	59.08±0.29	32.63±0.74	77.87±0.29	47.58±0.57
4	59.29±0.16	32.48±0.62	77.92±0.17	47.08±0.53
5	**59.67±0.38**	**33.10±0.71**	**78.04±0.09**	46.90±0.91
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{BIAS}}(g_2, M)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	57.87±0.15	32.50±0.48	77.43±0.11	48.14±0.31
3	58.42±0.23	**33.19±0.60**	77.41±0.17	48.09±0.93
4	**58.86±0.18**	32.65±1.07	77.46±0.29	**48.43±0.94**
5	58.81±0.21	32.86±0.47	**77.58±0.23**	48.30±0.39
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.5)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	**60.69±2.43**	**32.22±0.35**	79.36±0.65	48.86±0.34
3	59.81±0.25	32.04±0.67	79.41±0.17	48.91±0.81
4	59.75±0.33	32.03±0.34	79.42±0.18	**49.05±0.71**
5	59.85±0.30	32.06±0.72	**79.45±0.20**	48.32±0.70
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.6)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	58.94±0.29	32.65±0.36	78.67±0.15	**49.86±0.59**
3	59.43±0.35	32.91±0.40	78.94±0.19	48.84±1.09
4	**59.54±0.28**	33.02±0.62	78.92±0.29	49.64±0.74
5	59.52±0.28	**33.10±0.50**	**79.29±0.21**	49.39±1.02
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.7)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	58.24±0.19	33.24±0.90	78.30±0.31	**50.40±0.83**
3	58.74±0.26	33.12±0.59	78.49±0.30	49.85±0.38
4	58.79±0.38	**33.63±0.53**	78.51±0.29	49.88±0.75
5	**58.99±0.18**	32.93±0.81	**78.57±0.12**	49.53±1.55
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.8)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	57.60±0.15	34.14±0.22	**77.96±0.07**	**51.82±0.68**
3	58.04±0.28	33.93±0.45	77.55±0.18	50.30±0.81
4	58.05±0.16	**34.16±0.54**	77.90±0.21	50.40±0.43
5	**58.43±0.27**	33.87±0.62	77.90±0.17	50.78±0.95
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.9)$			
1	56.83±0.20	31.03±0.41	77.24±0.29	48.38±0.70
2	57.16±0.15	34.25±0.55	77.19±0.09	**51.42±0.45**
3	57.08±0.10	33.96±0.19	77.21±0.26	51.30±1.05
4	57.36±0.19	**34.29±0.15**	**77.34±0.34**	51.16±0.55
5	**57.38±0.16**	34.25±0.30	77.13±0.16	50.68±0.74

Table S3: The effectiveness evaluation of NaCl ($M \neq 1$) on IntCl ($M = 1$) when $\alpha = 1, g^1 = g^2 = g_2$. The best performance within each loss type is in boldface. We color the overall best performance in blue.

M	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{VAR}}(g_2, M)$			
	CIFAR100 Acc.	FGSM Acc.	CIFAR10 Acc.	FGSM Acc.
1	56.22±0.15	40.05±0.67	76.39±0.10	**59.33±0.94**
2	57.51±0.12	41.01±0.36	76.88±0.49	58.77±0.67
3	58.08±0.18	41.02±0.83	76.95±0.19	58.28±0.50
4	58.31±0.23	**41.49±0.51**	77.30±0.30	58.61±0.80
5	**58.64±0.24**	40.50±0.23	**77.42±0.17**	58.11±0.72
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{BIAS}}(g_2, M)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	**59.33±0.94**
2	56.71±0.11	39.80±0.57	76.55±0.27	58.44±0.31
3	57.13±0.26	40.53±0.29	**76.67±0.22**	58.47±0.31
4	57.06±0.19	40.85±0.31	76.34±0.22	58.91±0.62
5	**57.46±0.04**	**41.00±0.86**	76.60±0.37	57.98±0.47
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.5)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	59.33±0.94
2	58.97±0.19	40.25±0.52	78.61±0.20	58.41±0.59
3	59.26±0.18	40.96±0.58	**78.83±0.22**	59.20±1.25
4	59.32±0.21	40.82±0.54	78.83±0.27	59.03±0.52
5	**59.43±0.23**	**41.01±0.34**	78.80±0.21	**59.51±0.93**
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.6)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	59.33±0.94
2	58.55±0.34	**40.85±0.62**	78.34±0.22	**59.56±0.88**
3	59.05±0.21	40.83±0.44	78.41±0.12	59.14±0.78
4	59.06±0.25	40.80±0.89	78.61±0.22	58.41±1.00
5	**59.10±0.23**	40.68±0.50	**78.63±0.21**	58.92±0.76
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.7)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	59.33±0.94
2	58.00±0.18	40.35±0.34	77.73±0.24	59.40±1.27
3	58.23±0.18	40.94±0.75	77.91±0.25	**59.57±0.81**
4	58.20±0.25	40.95±0.45	77.89±0.20	59.49±0.49
5	**58.37±0.14**	**41.15±0.48**	**78.27±0.26**	59.17±0.94
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.8)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	59.33±0.94
2	57.07±0.24	41.29±0.57	77.27±0.28	60.16±0.51
3	**57.62±0.22**	40.93±0.49	77.54±0.27	59.47±0.52
4	57.61±0.25	**41.36±0.41**	77.50±0.34	**60.28±0.68**
5	57.56±0.18	40.71±0.34	**77.58±0.42**	59.99±0.30
	$\alpha \neq 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.9)$			
1	56.22±0.15	40.05±0.67	76.39±0.10	59.33±0.94
2	56.54±0.33	40.85±0.13	76.81±0.22	60.40±0.46
3	56.69±0.11	41.23±0.66	**76.98±0.22**	60.13±0.56
4	56.43±0.26	**41.56±0.56**	76.97±0.20	**61.21±0.49**
5	**56.86±0.11**	41.09±0.31	76.91±0.21	60.09±0.39

C ADVERSARIAL ACCURACY

For a more comprehensive study of adversarial robustness, we extend Table S2 to include PGD attack results with the same strength as FGSM attacks ($\epsilon = 0.002$). One can readily see from Table S4 that the adversarial accuracy under PGD attacks of the same magnitude is slightly lower (roughly 2-3% lower) as PGD is a stronger attack. Nevertheless, the trend is consistent – the models that exhibit better adversarial robustness w.r.t. FGSM attacks also demonstrate superior adversarial robustness w.r.t. PGD attacks.

Table S4: The complete Table S2 (Table 1 right column) with additional PGD accuracy.

M	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{VAR}}(g_2, M)$					
	CIFAR100 Acc.	FGSM Acc.	PGD Acc.	CIFAR10 Acc.	FGSM Acc.	PGD Acc.
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	**48.38±0.70**	**46.24±0.77**
2	58.17±0.39	31.92±0.45	29.77±0.43	77.43±0.18	48.05±0.38	45.63±0.50
3	59.08±0.29	32.63±0.74	30.33±0.84	77.87±0.29	47.58±0.57	45.02±0.62
4	59.29±0.16	32.48±0.62	30.12±0.73	77.92±0.17	47.08±0.53	44.52±0.54
5	**59.67±0.38**	**33.10±0.71**	**30.87±0.88**	**78.04±0.09**	46.90±0.91	44.20±1.08
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{BIAS}}(g_2, M)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	**46.24±0.77**
2	57.87±0.15	32.50±0.48	30.25±0.60	77.43±0.11	48.14±0.31	45.81±0.43
3	58.42±0.23	**33.19±0.60**	**30.93±0.59**	77.41±0.17	48.09±0.93	45.67±0.93
4	**58.86±0.18**	32.65±1.07	30.22±1.09	77.46±0.29	**48.43±0.94**	45.99±1.15
5	58.81±0.21	32.86±0.47	30.57±0.55	**77.58±0.23**	48.30±0.39	45.80±0.48
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.5)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	46.24±0.77
2	**60.69±2.43**	**32.22±0.35**	**30.11±0.43**	79.36±0.65	48.86±0.34	46.67±0.40
3	59.81±0.25	32.04±0.67	29.87±0.65	79.41±0.17	48.91±0.81	46.61±0.86
4	59.75±0.33	32.03±0.34	29.85±0.36	79.42±0.18	**49.05±0.71**	**46.70±0.80**
5	59.85±0.30	32.06±0.72	29.99±0.76	**79.45±0.20**	48.32±0.70	45.89±0.82
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.6)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	46.24±0.77
2	58.94±0.29	32.65±0.36	30.16±0.27	78.67±0.15	**49.86±0.59**	**47.38±0.70**
3	59.43±0.35	32.91±0.40	30.36±0.52	78.94±0.19	48.84±1.09	46.24±1.32
4	**59.54±0.28**	33.02±0.62	**30.68±0.72**	78.92±0.29	49.64±0.74	47.15±0.88
5	59.52±0.28	**33.10±0.50**	30.63±0.48	**79.29±0.21**	49.39±1.02	46.89±1.12
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.7)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	46.24±0.77
2	58.24±0.19	33.24±0.90	30.40±1.06	78.30±0.31	**50.40±0.83**	**47.50±0.89**
3	58.74±0.26	33.12±0.59	29.94±0.62	78.49±0.30	49.85±0.38	46.69±0.32
4	58.79±0.38	**33.63±0.53**	**30.70±0.60**	78.51±0.29	49.88±0.75	47.01±0.96
5	**58.99±0.18**	32.93±0.81	29.89±0.99	**78.57±0.12**	49.53±1.55	46.41±1.91
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.8)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	46.24±0.77
2	57.60±0.15	34.14±0.22	31.35±0.25	**77.96±0.07**	**51.82±0.68**	**48.81±0.85**
3	58.04±0.28	33.93±0.45	31.31±0.62	77.55±0.18	50.30±0.81	47.41±0.76
4	58.05±0.16	**34.16±0.54**	**31.41±0.61**	77.90±0.21	50.40±0.43	47.58±0.47
5	**58.43±0.27**	33.87±0.62	31.23±0.76	77.90±0.17	50.78±0.95	47.96±1.12
	$\alpha = 0,\ \mathcal{L}_{\text{Na}}(g_2, M, \lambda) = \mathcal{L}_{\text{MIXUP}}(g_2, M, 0.9)$					
1	56.83±0.20	31.03±0.41	28.80±0.48	77.24±0.29	48.38±0.70	46.24±0.77
2	57.16±0.15	34.25±0.55	31.83±0.57	77.19±0.09	**51.42±0.45**	**49.09±0.53**
3	57.08±0.10	33.96±0.19	31.56±0.34	77.21±0.26	51.30±1.05	48.60±1.28
4	57.36±0.19	**34.29±0.15**	**31.93±0.32**	**77.34±0.34**	51.16±0.55	48.64±0.61
5	**57.38±0.16**	34.25±0.30	31.89±0.26	77.13±0.16	50.68±0.74	48.14±0.83

In Figure S1, we show the adversarial accuracy as a function of the FGSM attack strength ϵ. Specifically, we range the attack strength from 0.002 to 0.032 and give the adversarial accuracy of our proposals (IntCl & IntNaCl) together with baselines under all attacks. From Figure S1, one can see that among all baselines, AdvW demonstrates the best adversarial robustness, whereas our proposals still consistently win over it by a noticeable margin.

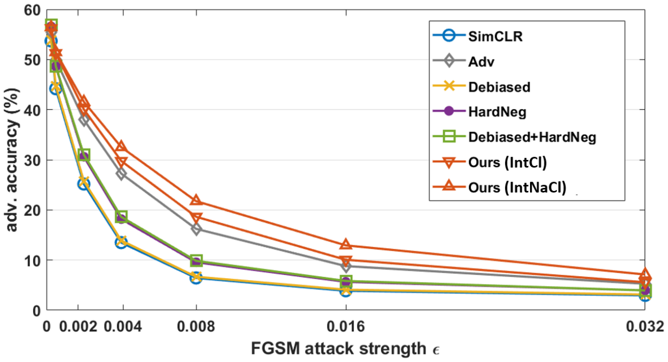
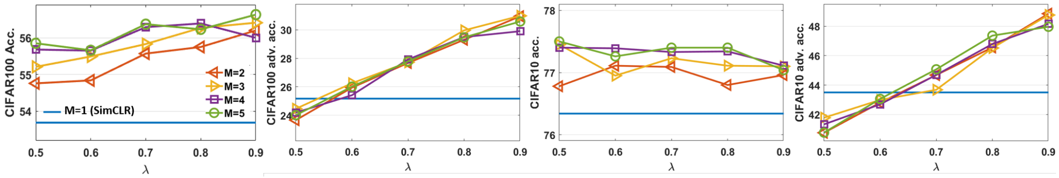


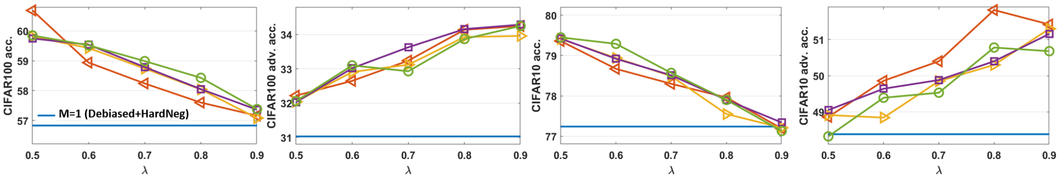
Figure S1: The adversarial accuracy under FGSM attacks of different strength on CIFAR100.

D THE EFFECT OF λ

To investigate the effect of λ on different metrics, we include in Figure S2 the standard and adversarial accuracy on CIFAR100 and CIFAR10 as functions of λ. Intriguingly, we see that all the accuracy curves in Figure S2(a) have tended to increase over λ. Comparatively, two of the accuracy curves in Figure S2(b), the standard accuracy on CIFAR100 and CIFAR10, show downward trends. One plausible reason is related to the room for improvements of individual baselines. Since Debiased+HardNeg is a much stronger baseline than SimCLR, it is closer to the robustness-accuracy trade-off.



(a) NaCl on SimCLR, i.e. $\alpha = 0, \mathcal{L}_{\text{Na}} = \mathcal{L}_{\text{MIXUP}}, g^1 = g_0$ in Equation 10



(b) NaCl on Debiased+HardNeg, i.e. $\alpha = 0, \mathcal{L}_{\text{Na}} = \mathcal{L}_{\text{MIXUP}}, g^1 = g_2$ in Equation 10

Figure S2: The standard and adversarial accuracy (%) on CIFAR100 and CIFAR10 as functions of λ.

E EXTENDED RUNTIME

#epoch	100	200	400	600	800	1000
$\mathcal{L}_{\text{SimCLR}}$	53.69	57.45	60.06	60.96	61.27	**61.90**
$\mathcal{L}_{\text{VAR}}(g_0, 2)$	56.04	59.44	61.42	**62.37**	62.06	-
$\mathcal{L}_{\text{BIAS}}(g_0, 2)$	55.72	59.31	61.19	61.66	**62.49**	-
$\mathcal{L}_{\text{MIXUP}}(g_0, 2, 0.9)$	56.20	58.98	61.81	62.43	**62.46**	-
$\mathcal{L}_{\text{Debiased+HardNeg}}$	56.83	59.35	61.77	**62.74**	62.68	-
$\mathcal{L}_{\text{VAR}}(g_2, 2)$	58.17	60.66	62.38	63.43	**63.51**	-
$\mathcal{L}_{\text{BIAS}}(g_2, 2)$	57.87	60.06	62.36	62.58	**62.86**	-
$\mathcal{L}_{\text{MIXUP}}(g_2, 2, 0.5)$	60.69	62.14	64.06	**65.59**	65.53	-

Table S5: The CIFAR100 linear evaluation results (%) after different numbers of training epochs.

F EXPERIMENTAL DETAILS

Architecture. We follow [4,16] to incorporate an MLP projection head during the contrastive learning on resnet18.

Optimizer. Adam optimizer with a learning rate of $3e-4$.

Batching. A batch size of 256 is used across all the experiments.

Methodological hyperparameters. Throughout out experiments, we use $\tau^{+}=0.01$ and $\beta=1.0$ for $\mathcal{L}_{\text{Debiased}}$ [5] and $\mathcal{L}_{\text{Debiased+HardNeg}}$ [16], $\alpha=1$ for $\mathcal{L}_{\text{Adv}}$ [12]. The same set of hyperparameters are used in our IntCl and IntNaCl.

Data augmentation. Our data augmentation includes random resized crop, random horizontal flip, random grayscale, and color jitter. Specifically, we implement the color jitter by calling $torchvision.transforms.ColorJitter(0.8*s, 0.8*s, 0.8*s, 0.2*s)$ and execute with probability 0.8. Random grayscale is performed with probability 0.2.

Adversarial hyperparameters. When evaluating the adversarial robustness using the codebase provided in [32], we use a PGD step size of $1e-2$, 10 iterations, and 2 random restarts.

Error bar. We run five independent trials for each of the experiments and report the mean and standard deviation for all tables and figures. The error bars in Figure S1 is omitted for better visual clarity.

G SUPERVISED LEARNING BASELINE

We give in the following the standard and adversarial accuracy of a supervised learning baseline with the same network architecture, optimizer, and batch size. In our self-supervised representation learning experiments, we train the representation network for 100 epochs and train the downstream fully-connected classifying layer for 1000 epochs. Therefore, to obtain a fair supervised learning baseline, we train the complete network end-to-end for 1000 epochs. We follow the same procedures in evaluating the transfer standard accuracy and adversarial accuracy as described in Section 5.

CIFAR100 (std. acc., FGSM acc., PGD acc.): 65.16±0.32, 35.89±0.23, 32.62±0.23.

Transfer CIFAR10 (std. acc., FGSM acc., PGD acc.): 77.45±0.21, 44.39±0.47, 40.35±0.52.