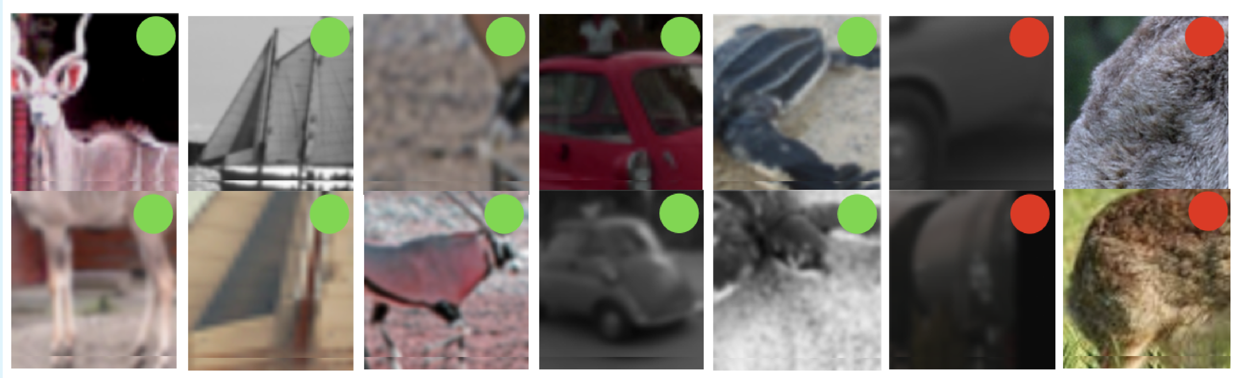
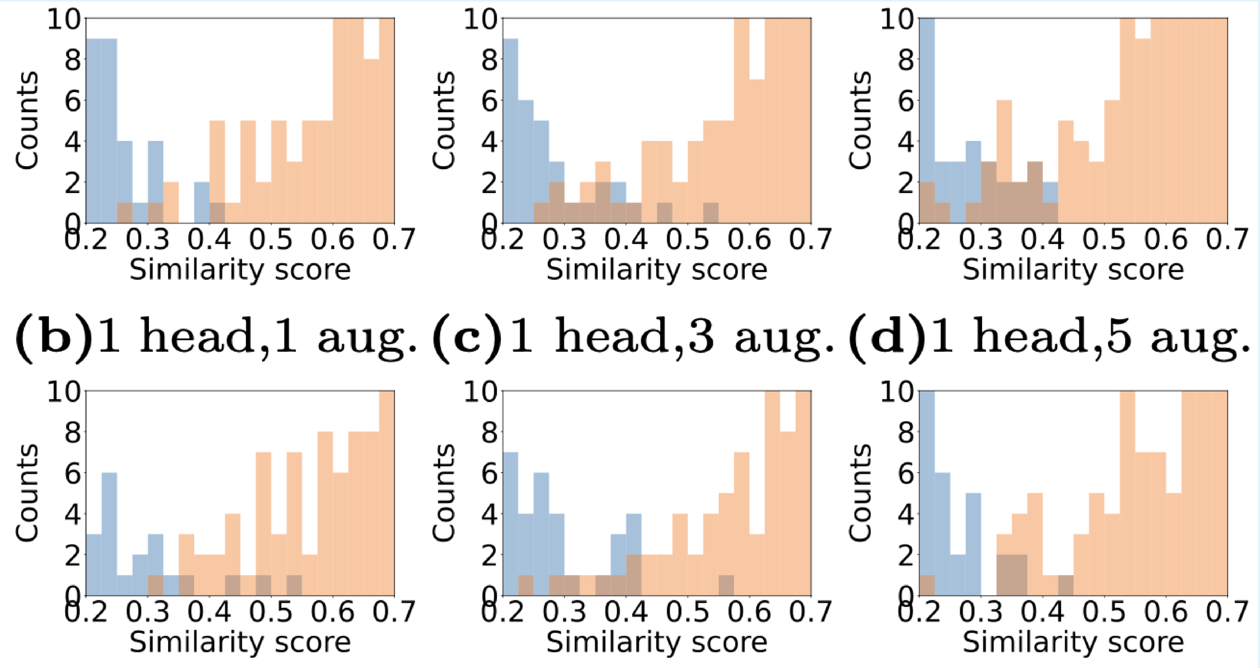
Adaptive Multi-head Contrastive Learning

Lei Wang[1,2] Piotr Koniusz[2,1] Tom Gedeon[3] Liang Zheng[1]
[1]Australian National University [2] Data61/CSIRO [3]Curtin University

Motivation and key ideas



0.48	0.60	0.35	0.42	0.58	0.75	0.82
0.80	**0.95**	**0.76**	**0.83**	**0.81**	**0.42**	**0.53**



(b) 1 head, 1 aug. **(c)** 1 head, 3 aug. **(d)** 1 head, 5 aug.

(a) Examples of augmented image pairs (STL-10). **(e)** Ours, 1 aug. **(f)** Ours, 3 aug. **(g)** Ours, 5 aug.

- Diverse augmentation strategies in contrastive learning and varying intra-sample similarity cause views from the same image may not always be similar.
- Owing to inter-sample similarity, views from different images may be more akin than those from the same image.
- The table in (a) shows the original (gray) and our method's similarity scores (black).
- (b)-(d): for traditional contrastive learning methods, when increasing the number of augmentations from 1 to 5, similarities of more positive pairs drop below 0.5, causing more significant overlapping regions between histograms of positive (orange) and negative (blue) pairs.
- In comparison, our multi-head approach (e)-(g) yields better separation of positive and negative sample pairs as more augmentation types are used, e.g., (g) vs (d).

Standard contrastive learning methods and their loss functions:

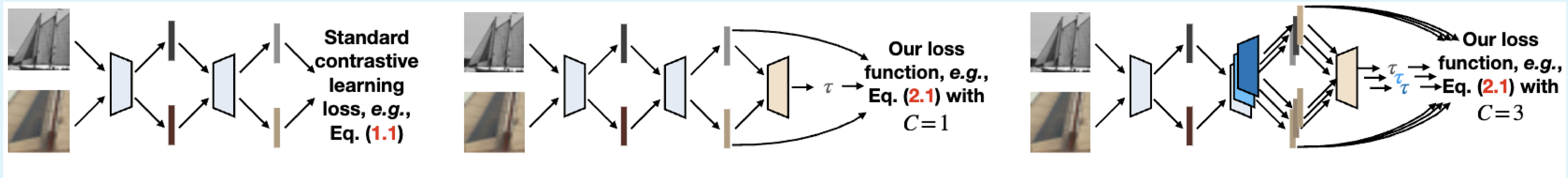
Method	Loss name	Loss function
SimCLR, MoCo	NT-Xent	$\ell_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_i^+)/\tau)}{\sum_{n=1}^N \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{in}^-)/\tau)}$
SimSiam	Negative cos.	$\ell_{\text{SymNegCos}} = -\frac{1}{2}\text{sim}\big(\boldsymbol{z}_i, [\boldsymbol{h}_i^+]_{\text{sg}}\big) - \frac{1}{2}\text{sim}\big(\boldsymbol{z}_i^+, [\boldsymbol{h}_i]_{\text{sg}}\big)$
Barlow Twins	Cross-corr.	$\ell_{\text{Cross-Corr}} = \sum_{l=1}^{d'}(1-\mathcal{C}_{ll})^2 + \lambda \sum_{l=1}^{d'} \sum_{m\neq l}^{d'} \mathcal{C}_{lm}^2$
LGP, CAN	InfoNCE	$\ell_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_i^+)/\tau)}{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_i^+)/\tau) + \sum_{n=1}^N \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{in}^-)/\tau)}$

- We propose adaptive multi-head contrastive learning (AMCL): it better captures the diverse image content and gives similarity scores that better separate positive and negative pairs.
- Within AMCL, we design an adaptive temperature which depends on both the projection head and the similarity of the current pair.

Loss functions for applying AMCL to widely used contrastive learning frameworks involve introducing C heads and a regularization term (highlighted in green):

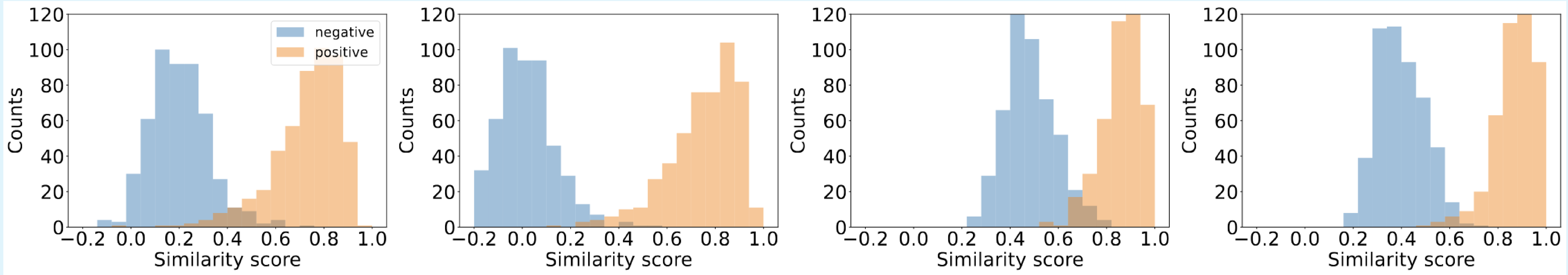
Method	Loss function	Regularization
SimCLR, MoCo	$\ell^{\ddagger}_{\text{NT-Xent}} = \sum_{c=1}^{C}\Big(-\frac{1}{\tau_i^{c+}}\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_i^{c+}) + \frac{1}{\tau_{in^*}^{c-}} \max_{n=1,\cdots,N} \text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_{in}^{c-})$	$+\beta\Omega(\tau_i^{c+}) - \beta\Omega(\tau_{in^*}^{c-})\Big)$
SimSiam	$\ell^{\ddagger}_{\text{SymNegCos}} = \sum_{c=1}^{C}\Big(-\frac{1}{2\tau_i^{c+}}\text{sim}\big(\boldsymbol{z}_i^c, [\boldsymbol{h}_i^+]_{\text{sg}}\big) - \frac{1}{2\tau_i^{c\tilde{+}}}\text{sim}\big(\boldsymbol{z}_i^{c+}, [\boldsymbol{h}_i]_{\text{sg}}\big)$	$+\beta\Omega(\tau_i^{c+}) + \beta\Omega(\tau_i^{c\tilde{+}})\Big)$
Barlow Twins	$\ell^{\ddagger}_{\text{Cross-Corr}} = \sum_{c=1}^{C}\Big(\sum_{l=1}^{d'}(1-\frac{1}{\tau_l^{c+}}\mathcal{C}_{ll})^2 + \lambda \sum_{l=1}^{d'} \sum_{m\neq l}^{d'} \frac{1}{\tau_{lm}^{c-}}\mathcal{C}_{lm}^2$	$+\beta\sum_{l=1}^{d'}\Omega(\tau_l^{c+}) - \beta\sum_{l=1}^{d'}\sum_{m\neq l}^{d'}\Omega(\tau_{lm}^{c-})\Big)$
LGP, CAN	$\ell^{\ddagger}_{\text{InfoNCE}} = \sum_{c=1}^{C}\Big(-\frac{1}{\tau_i^{c+}}\text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_i^{c+}) + \frac{1}{\tau_{in^*}^{c-}} \max_{n=1,\cdots,N+1} \text{sim}(\boldsymbol{z}_i^c, \boldsymbol{z}_{in}^{c\pm})$	$+\beta\Omega(\tau_i^{c+}) - \beta\Omega(\tau_{in^*}^{c\pm})\Big)$

The pipeline: further details



(a) Constant temperature, single-head. (b) Adaptive temperature, single-head. (c) Adaptive temperature, multi-head.

- The baseline uses one projection head and constant temperature, while our method has multiple projection heads and adaptive temperature.
- We use SimCLR for pre-training with ResNet-18 on STL-10. After pre-training, we choose 500 positive pairs and 500 negative pairs from the validation to compute the cosine similarity.
- In (a) and (b), similarity score (temperature scaled) is computed between the 128-dim features extracted from the projection head(s).
- In (c) and (d), cosine similarity score is computed between the 512-dim features extracted from the backbone after removing the project heads.



(a) head sim. (baseline) **(b)** head sim. (ours) **(c)** feat. sim. (baseline) **(d)** feat. sim. (ours)

Results