FALCON : Fairness Learning via Contrastive Attention Approach to Continual Semantic Scene Understanding

Thanh-Dat Truong[1], Utsav Prabhu[2], Bhiksha Raj[3,4], Jackson Cothren[5], Khoa Luu[1]
[1]CVIU Lab, University of Arkansas, USA [2]Google DeepMind, USA
[3]Carnegie Mellon University, USA [4]Mohammed bin Zayed University of AI, UAE
[5]Dep. of Geosciences, University of Arkansas, USA
{tt032, jcothre, khoaluu}@uark.edu, bhiksha@cs.cmu.edu, utsavprabhu@google.com
http://uark-cviu.github.io/projects/FALCON

1. Proof of Propositions 1 and 2

1.1. Proof of Proposition 1

Proposition 1: *If the contrastive clustering loss $\mathcal{L}_{Cont}(;, \mathbf{c})$ achieve the optimal value, the enforcement ℓ_i between the feature and the cluster will converges to $\ell_i = L^{-1}$.*

Proof: Let us consider the optimization of the Eqn. (4) in the paper as follows:

$$\begin{aligned} &\min -\sum_{i=1}^{L} \log \frac{\exp(\mathbf{f}_i^t \times \mathbf{c})}{\sum_{\mathbf{f}'} \exp(\mathbf{f}' \times \mathbf{c})} = -\sum_{i=1}^{L} \log \ell_i \\ &\text{subject to} \quad \sum_{i=1}^{L} \ell_i = \ell \end{aligned} \tag{1}$$

where ℓ is the total enforcement between features $\mathbf{f}_i^t$ and cluster $\mathbf{c}$. Then, the optimization of Eqn. (4) in the paper can be rewritten by using Lagrange multiplier as follows:

$$\mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right) = -\sum_{i=1}^{L} \log \ell_i + \lambda(\sum_{i=1}^{L} \ell_i - \ell) \tag{2}$$

where λ is the Lagrange multiplier. Then, the contrastive clustering loss in Eqn. (4) in the paper achieves minimum if and only if:

$$\begin{aligned} \frac{\partial \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)}{\partial \ell_i} &= -\ell_i^{-1} + \lambda = 0 \\ \frac{\partial \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)}{\partial \lambda} &= \sum_{i=1}^{L} \ell_i - \ell = 0 \\ \Rightarrow \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right) &= -L \log \frac{\ell}{L} \end{aligned} \tag{3}$$

As the total enforcement between features and the cluster is normalized, i.e., $\ell \in [0..1]$, the contrastive clustering loss $\mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)$ achieves minimum when $\log \ell = 0 \Rightarrow \ell = 1$. Then, the enforcement between a single feature and the cluster will be equal to $\ell_i = \frac{\ell}{L} = L^{-1}$.

1.2. Proof of Proposition 2

Proposition 2: *If the fairness contrastive clustering loss $\mathcal{L}^{\alpha}_{Cont}(;, \mathbf{c})$ achieve the optimal value, the enforcement ℓ_i between the feature and the cluster will converges to $\ell_i = (\alpha^{-1} + L)^{-1}$.*

Proof: We first define the the enforcement between transitive vector $\mathbf{v}$ and the cluster $\mathbf{c}$ as $\ell_{\mathbf{v}} = \frac{\exp(\mathbf{v} \times \mathbf{c})}{\sum_{\mathbf{f}'} \exp(\mathbf{f}' \times \mathbf{c})}$. Then, let us consider the optimization of Eqn. (5) in the paper as follows:

$$\begin{aligned} &\min -\sum_{i=1}^{L} \alpha \log \ell_i - \log \ell_{\mathbf{v}} \\ &\text{subject to} \quad \sum_{i=1}^{L} \ell_i + \ell_{\mathbf{v}} = \ell \end{aligned} \tag{4}$$

Similar to Eqn. (1), Eqn. (4) can be reformulated via Lagrange multiplier as follows:

$$\mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right) = -\sum_{i=1}^{L} \alpha \log \ell_i - \log \ell_{\mathbf{v}} + \lambda(\sum_{i=1}^{L} \ell_i + \ell_{\mathbf{v}} - \ell) \tag{5}$$

Then, the fairness contrastive loss $\mathcal{L}^{\alpha}_{Cont}$ achieves minimum if and only if:

$$\begin{aligned} \frac{\partial \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)}{\partial \ell_i} &= -\alpha \ell_i^{-1} + \lambda = 0 \\ \frac{\partial \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)}{\partial \ell_{\mathbf{v}}} &= -\ell_{\mathbf{v}}^{-1} + \lambda = 0 \\ \frac{\partial \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)}{\partial \lambda} &= \sum_{i=1}^{L} \ell_i + \ell_{\mathbf{v}} - \ell = 0 \\ \Rightarrow \mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right) &= -\alpha L \log \frac{\alpha \ell}{1 + \alpha L} - \log \frac{\ell}{1 + \alpha L} \end{aligned} \tag{6}$$

As in Eqn. (6), the fairness contrastive learning loss $\mathcal{L}\left(\{\ell_i\}_{i=1}^{L}, \lambda\right)$ archives minimum when $\log \ell = 0 \rightarrow \ell = 1$.

Thus, the enforcement between the single feature the cluster will be re-balanced as $\ell_i = \frac{\alpha}{1+\alpha L} = (\alpha^{-1} + L)^{-1}$.

2. Implementation

Implementation Our framework is implemented in PyTorch and trained on four 40GB-VRAM NVIDIA A100 GPUs. The contrastive loss in our implementation is normalized with respect to the number of samples. These models are optimized by the SGD optimizer [1] with momentum 0.9, weight decay 10^{-4}, and a batch size of 16. The learning rate of the first learning step and the continual steps is set to 10^{-4} and 5×10^{-5} respectively. To update the cluster vectors $\mathbf{c}$, following prior work [9–11], we maintain a set of 500 features for each cluster and update the clusters after $K = 100$ steps with a momentum $\eta = 0.99$. In our domain incremental experiments, all clusters are updated at each learning step by momentum update. The number of features selected for each cluster in the visual grammar model is set to $M = 128$. The balanced weight of CSS objective λ_{CL} and the cluster regularizer λ_C is set to 1. Following the common practices [10, 11], the margin between clusters ∇ is set to 10.

Unknown Cluster Initialization As mentioned in the main paper, we adopt the DB-SCAN algorithm to initialize the clusters for unknown samples. In addition, to reduce the noise clusters and isolated clusters, we also merge several close clusters, i.e., if the distance between two clusters is less than the margin 2∇, these will be merged into a single cluster where the new cluster center will be the means of these two merging cluster centers. By empirical observation, we have noticed that the number of unknown clusters initialized at each learning step, i.e., N_U at the current learning step t, is not greater than $1.5\times$ times of the remaining classes (i.e., $|\mathcal{C}^{t+1..T}|$) in the dataset, e.g., in our ADE20K 100-50 experiments, at the first learning step of 100 classes, there are 68 unknown clusters that have been initialized while there are 50 remaining unknown classes in the dataset.

Cluster Assignment In our approach, we use our visual grammar model to assign the cluster for each feature representation. Theoretically, although there is a possibility that a feature could not be assigned to a cluster via the visual grammar model, we have empirically observed that this issue rarely happens in our approach. Indeed, since we initialize the known clusters via the DB-SCAN, it guarantees that for each feature, there is at least one cluster nearby that the feature representation should belong to. However, to preserve the integrity of our approach, for the outlier features in cases that cannot be assigned clusters via the visual grammar model, these outliers will be heuristically assigned to their closest clusters as similar to [10, 11].

Continual Learning Procedure Algorithm 1 illustrates the training procedure of our CSS approach.

Algorithm 1 CSS Procedure At Learning Step t

```
Require: Learning Step $t$, Dataset $\mathcal{D}^t$, Visual Grammar
    $\phi(;\Theta_{t-1})$, and Segmentation Model $F(;\Theta_{t-1})$
Step 0: Extract features on $\mathcal{D}^t$ by $F(;,\theta_{t-1})$
Step 1: Initialize new known clusters for $\mathcal{C}^t$ of features extracted in Step 0
Step 2: Initialize potential unknown clusters of features extracted in Step 0
Step 3: Train CSS Model $F(;\theta_t)$ on $\mathcal{D}^t$
Step 4: Extract features on $\mathcal{D}^t$ by $F(;,\theta_t)$
Step 5: Train Visual Grammar Model $\phi(;,\Theta_t)$ on current
    known clusters $\mathbf{c}$ and features extracted in Step 4
return $F(;\theta_t)$ and $\phi(;\Theta_t)$
```

3. Additional Experimental Results

3.1. Experiment Results of ADE20K 50-50 Benchmark

Table 1 presents the results of our method on the ADE20K 50-50 benchmark compared to prior methods. For fair comparisons, we use the DeepLab-V3 and Transformer in this experiment. As shown in the results, our proposed FALCON approach significantly outperforms prior methods. The results of our approach have reduced the gap with the upper bound result.

Table 1. Experimental results on ADE20K 50-50 Benchmark

	ADE20K 50-50 (3 steps)			
Network	Method	0-50	50-150	all
DeepLab-V3	MiB [2]	45.6	21.0	29.3
	PLOP [8]	48.8	21.0	30.4
	LGKD+PLOP [13]	49.4	29.4	36.0
	RCIL [14]	47.8	23.0	31.2
	RCIL+LGKD [13]	49.1	27.2	34.4
	FairCL [11]	49.7	26.8	34.6
	FALCON	**50.6**	**31.2**	**37.6**
	Upper Bound	51.1	33.25	38.9
Transformer	FairCL [11]	49.6	27.8	35.6
	FALCON	**53.0**	**36.8**	**42.2**
	Upper Bound	54.9	40.8	45.5

3.2. Ablation Study

Effectiveness of Choosing Margin ∇ Table 2 studies the effectiveness of the value of margin ∇ to the performance of our approach on ADE20K 100-50 and ADE20K 100-10 benchmarks. As shown in the results, the change of ∇ also slightly influences the performance of the model. Since the margin defines the distance between two clusters, while the smaller value of the margin ∇ could cause the incorrect cluster assignment of the features, the larger value of the margin ∇ could produce the less compact clusters.

Table 2. Effectiveness of Choosing Margin ∇

(a) ADE20K 100-50					
	0-100	101-150	all	Major	Minor
$\nabla = 5$	44.4	21.8	36.9	51.9	29.4
$\nabla = 10$	**44.6**	**24.5**	**37.9**	**52.1**	**30.8**
$\nabla = 20$	44.7	22.2	37.2	51.7	29.9
(b) ADE20K 100-10					
	0-100	101-150	all	Major	Minor
$\nabla = 5$	43.2	18.7	35.0	50.5	27.3
$\nabla = 10$	**44.4**	**20.4**	**36.4**	**51.8**	**28.7**
$\nabla = 20$	43.5	19.9	35.7	51.2	27.9

Effectiveness of Choosing Number of Features M We study the impact of choosing the number of features M in the visual grammar model. As in shown Table 3, the optimal performance of our approach is $M = 128$. When the number of features selected is small ($M = 96$), it does not have enough number of features to form the visual grammar so the model is hard to exploit the correlation among features and the cluster. Meanwhile, when we increase the number of selected features ($M = 256$), the clusters will consist of many outlier features (the ones that do not belong to the cluster), thus being challenging for the visual grammar model to exploit the topological structures of the feature distribution.

Table 3. Effectiveness of Number of Features M in a Cluster of Visual Grammar Model.

(a) ADE20K 100-50					
	0-100	101-150	all	Major	Minor
$M = 96$	43.0	19.6	35.2	50.5	27.5
$M = 128$	**44.6**	**24.5**	**37.9**	**52.1**	**30.8**
$M = 256$	43.6	21.6	36.3	51.0	28.9
(b) ADE20K 100-10					
	0-100	101-150	all	Major	Minor
$M = 96$	42.2	16.4	33.6	50.2	25.3
$M = 128$	**44.4**	**20.4**	**36.4**	**51.8**	**28.7**
$M = 256$	42.7	17.1	34.2	50.6	26.0

Effectiveness of Different Segmentation Networks To illustrate the flexibility of our proposed approach, we evaluate our proposed approach with different network backbones. Table 4 illustrates the results of our approach using DeepLab-V3 [5], SegFormer [12] with different backbones, i.e., ResNet-50, ResNet-101, MiT-B2, and MiT-B3. As shown in the performance, the more powerful the segmentation model is, the better performance of the model is. In particular, our approach has shown its flexibility since it consistently improves the performance of the segmentation model and achieves the SOTA performance on two different benchmarks, i.e., the performance of Transformer models achieves 41.9%, and 40.3% on ADE20K 100-50, ADE20K 100-10, respectively.

Table 4. Effectiveness of Different Backbones on ADE20K.

(a) ADE20K 100-50						
	Backbone	0-100	101-150	all	Major	Minor
DeepLab-V3	R-50	44.3	15.2	34.7	51.5	26.4
	R-101	**44.6**	**24.5**	**37.9**	**52.1**	**30.8**
Transformer	MiT-B2	44.5	27.4	38.8	52.4	32.2
	MiT-B3	**47.5**	**30.6**	**41.9**	**53.8**	**35.8**
(b) ADE20K 100-10						
	Backbone	0-100	101-150	all	Major	Minor
DeepLab-V3	R-50	43.5	16.5	34.5	51.1	26.2
	R-101	**44.4**	**20.4**	**36.4**	**51.8**	**28.7**
Transformer	MiT-B2	45.4	22.7	37.8	52.6	30.4
	MiT-B3	**47.3**	**26.2**	**40.3**	**54.0**	**33.4**

4. Relation to Knowledge Distillation

Knowledge Distillation is a common approach to continual semantic segmentation [3, 4, 8, 14]. Prior work in clustering [11] has shown that the clustering loss is an upper bound of the knowledge distillation loss. Formally, the knowledge distillation loss can be formed as follows:

$$\mathcal{L}_{distill}(\mathbf{x}^t, F, \theta_t, \theta_{t-1}) = \mathcal{L}(\mathbf{F}^{t-1}, \mathbf{F}^t) \tag{7}$$

where $\mathbf{F}^t$ and $\mathbf{F}^{t-1}$ are the feature representations extracted from the model at learning step t and step $t-1$, respectively, and the metric $\mathcal{L}$ measure the knowledge gap between $\mathbf{F}^t$ and $\mathbf{F}^{t-1}$. Then, given a set of cluster $\mathbf{c}$, we consider the following triangle inequality of the metric $\mathcal{L}$ as follows:

$$\begin{aligned}
\forall \mathbf{c}: \quad & \mathcal{L}(\mathbf{F}^t, \mathbf{F}^{t-1}) \leq \mathcal{L}(\mathbf{F}^t, \mathbf{c}) + \mathcal{L}(\mathbf{c}, \mathbf{F}^{t-1}) \\
\Leftrightarrow & \underbrace{\mathcal{L}(\mathbf{F}^t, \mathbf{F}^{t-1})}_{\mathcal{L}_{distill}} \leq \frac{1}{|\mathcal{C}^{1..T}|} \sum_{\mathbf{c}} \left[\underbrace{\mathcal{L}(\mathbf{F}^t, \mathbf{c})}_{\mathcal{L}_{Cont}} + \mathcal{L}(\mathbf{c}, \mathbf{F}^{t-1}) \right]
\end{aligned} \tag{8}$$

At the computational time of Contrastive Clustering loss, the set of cluster vectors $\mathbf{c}$ is fixed (could be considered as constants). In addition, the features extracted at learning step $t-1$, i.e., $\mathbf{F}^{t-1}$, are constant due to the fix pre-trained model θ_{t-1}. Therefore, without a strict argument, the distance $\mathcal{L}(\mathbf{c}, \mathbf{F}^{t-1})$ could be considered as constant. Therefore, Eqn. (8) can be further derived as follows:

$$\begin{aligned}
\underbrace{\mathcal{L}(\mathbf{F}^t, \mathbf{F}^{t-1})}_{\mathcal{L}_{distill}} &= \mathcal{O}\left(\underbrace{\mathcal{L}\frac{1}{|\mathcal{C}^{1..T}|}}_{Constant} \sum_{\mathbf{c}} \left[\underbrace{\mathcal{L}(\mathbf{F}^t, \mathbf{c})}_{\mathcal{L}_{Cont}} + \underbrace{\mathcal{L}(\mathbf{c}, \mathbf{F}^{t-1})}_{\text{Constant}} \right] \right) \\
&= \mathcal{O}\left(\underbrace{\sum_{\mathbf{c}} \mathcal{L}(\mathbf{F}^t, \mathbf{c})}_{\mathcal{L}_{Cont}} \right) \\
\Rightarrow \mathcal{L}_{distill}(\mathbf{F}^{t-1}, \mathbf{F}^t) &= \mathcal{O}\left(\mathcal{L}_{Cont}(\mathbf{F}^t, \mathbf{c})\right)
\end{aligned} \tag{9}$$

where $\mathcal{O}$ is the Big-O notation. Hence, from Eqn. (9), without lack of generality, we can observe that the Contrastive Clustering Loss is the upper bound of the Knowledge Distillation loss. Therefore, by minimizing the Contrastive Clustering Loss, the constraint of Knowledge Distillation is also maintained due to the property of the upper bound.

5. Discussion of Limitations and Broader Impact

Limitations. In our paper, we choose a specific set of hyperparameters and learning approaches to support our hypothesis. However, our work could contain several limitations. First, choosing the scaling factor α could be considered as one of the potential limitations of our approach. In practice, when data keeps continuously growing, the pre-defined scaling factor α could not be good enough to control the fairness among classes. Our work focuses on investigating the effectiveness of our proposed losses to fairness, catastrophic forgetting, and background shift problems. Thus, the investigation of balance weights among losses has not been fully exploited, and we leave this experiment as our future work. Third, initializing the unknown clusters at each training step could potentially be room for improvement since the bad initial clusters could result in difficulty during training and updating these clusters and linking the unknown clusters learned in previous steps and new initial unknown clusters at the current learning steps have been yet fully exploited in our method. In addition, while our approach is designed for the DeepLab-V3 and Transformer segmentation networks [5, 12], the extensions of FALCON to mask-based segmentation networks [3, 6, 7] could be a potential next research for further performance improvement. These limitations could motivate new studies to further improve Fairness Learning via the Contrastive Attention Approach to continual learning in the future.

Broader Impact. Our paper investigates and addresses the fairness problem in continual learning. Our contribution is a step toward the fairness and transparency of continual semantic segmentation. Our study highlights the significance of fairness in continual semantic segmentation learning and presents a novel approach to address fairness issues, enhancing the robustness and credibility of the segmentation model.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.

[2] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2020.

[3] Fabio Cermelli, Matthieu Cord, and Arthur Douillard. Comformer: Continual learning in semantic and panoptic segmentation. *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2023.

[4] Sungmin Cha, beomyoung kim, YoungJoon Yoo, and Taesup Moon. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. In *Advances in Neural Information Processing Systems*, pages 10919–10930. Curran Associates, Inc., 2021.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 2018.

[6] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[8] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021.

[9] 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.

[10] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5840, 2021.

[11] Thanh-Dat Truong, Hoang-Quan Nguyen, Bhiksha Raj, and Khoa Luu. Fairness continual learning approach to semantic scene understanding in open-world environments. In *NeurIPS*, 2023.

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[13] Ze Yang, Ruibo Li, Evan Ling, Chi Zhang, Yiming Wang, Dezhao Huang, Keng Teck Ma, Minhoe Hur, and Guosheng Lin. Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18601–18612, 2023.

[14] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022.