CL-MAE: Curriculum-Learned Masked Autoencoders – Supplementary

Neelu Madan[1,◇], Nicolae-Cătălin Ristea[2,3,◇], Kamal Nasrollahi[1,4],
Thomas B. Moeslund[1], Radu Tudor Ionescu[3,5,*]
[1]Aalborg University, Denmark, [2]University Politehnica of Bucharest, Romania,
[3]University of Bucharest, Romania, [4]Milestone Systems, Denmark, [5]SecurifAI, Romania

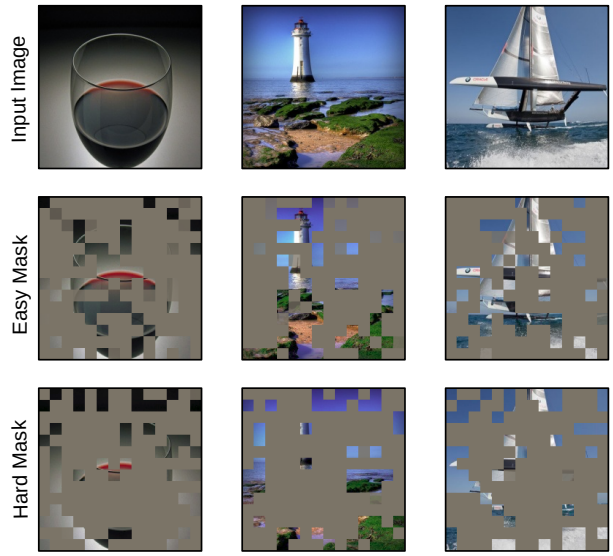


Figure 1. Masks generated by our masking module at two different moments during training, when all losses are in place, for the images on the top row. The masks on the second row are generated halfway during training, when the masking module is still acting as a partner to the MAE. In contrast, the masks on the bottom row are generated in the last epoch, when the masking module is behaving as an adversary to the MAE. Our module shifts its preference from masking non-salient tokens to masking tokens situated on edges and object contours, generating an easy-to-hard curriculum for the MAE.

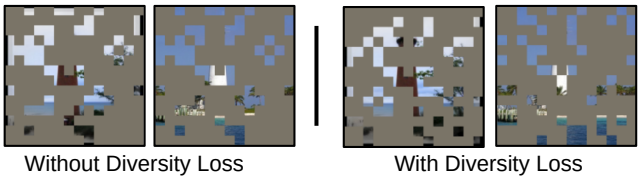


Figure 2. Masks generated by the proposed masking module without (left) and with (right) adding the diversity loss ($\mathcal{L}_{div}$). If the diversity loss ($\mathcal{L}_{div}$) is not included, the masking module can enter mode collapse and produce nearly identical masks. This can lead to overfitting CL-MAE on reconstructing certain patch configurations. The effect is no longer observed when the proposed diversity loss is employed.

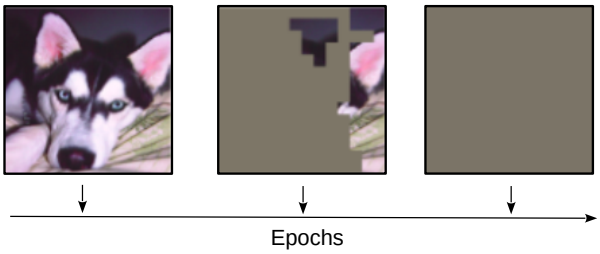


Figure 3. Masks generated by the proposed masking module during training, without the Kullback-Leibler loss. The masks evolve from leaving all patches visible (to reduce the reconstruction error for the MAE) to hiding all patches (to increase the reconstruction error for the MAE). The Kullback-Leibler loss is required to make sure the model always masks the desired number of patches.

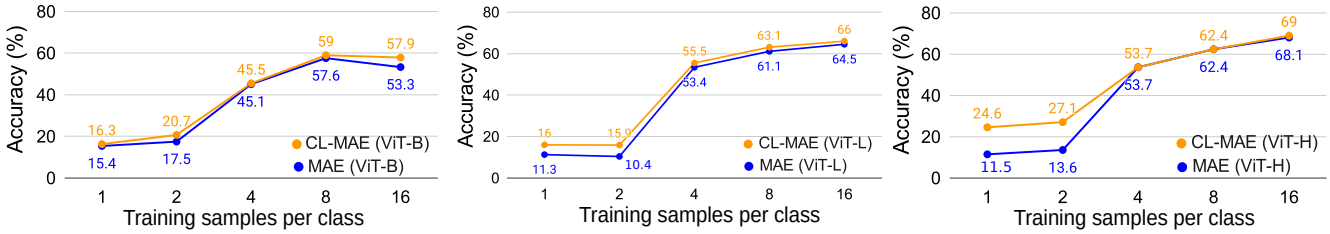
1. Additional Experiments and Discussion

In the supplementary, we analyze the behavior of our loss functions from a qualitative perspective, and present few-shot linear probing results on five downstream tasks.

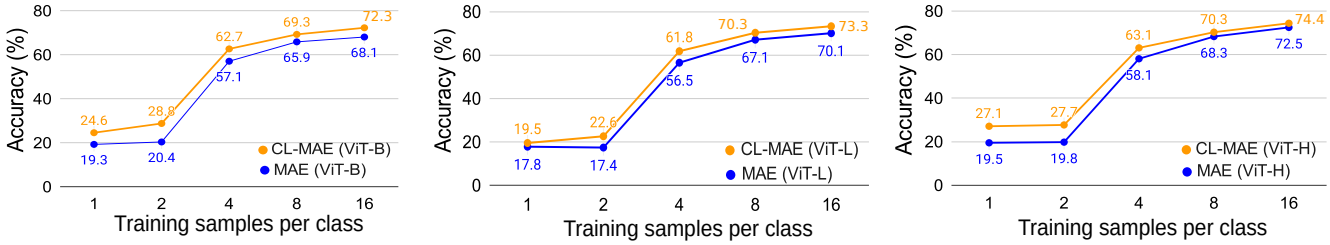
Effect of curriculum learning on masking. In Figure 1, we provide some masks generated by our learnable masking module at two different states during the training process on ImageNet [5]. In the first state (second row), the masking module acts as a partner to the MAE, helping to ease the reconstruction task. In this state, the module seems to mask tokens from non-salient or plain texture regions, which can be easily inferred from surrounding patches. We observe that the object contours are mostly visible in the easy masks. In the second state (third row), our masking module acts as an adversary to the MAE, aiming to make the reconstruction task harder for the MAE. In this state, we observe that the module prefers to mask patches located on edges, object contours and salient regions. These patches are much harder to reconstruct based on the visible information. In summary, the examples presented in Figure 1 confirm that our learnable masking module behaves as expected, creating an easy-to-hard curriculum for the MAE.

Effect of diversity loss on masking. Figure 2 shows the

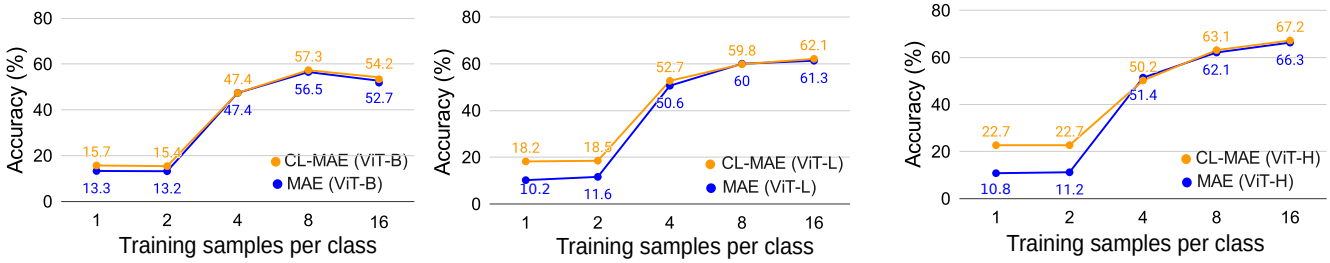
*corresp. author: raducu.ionescu@gmail.com; ◇equal contribution.



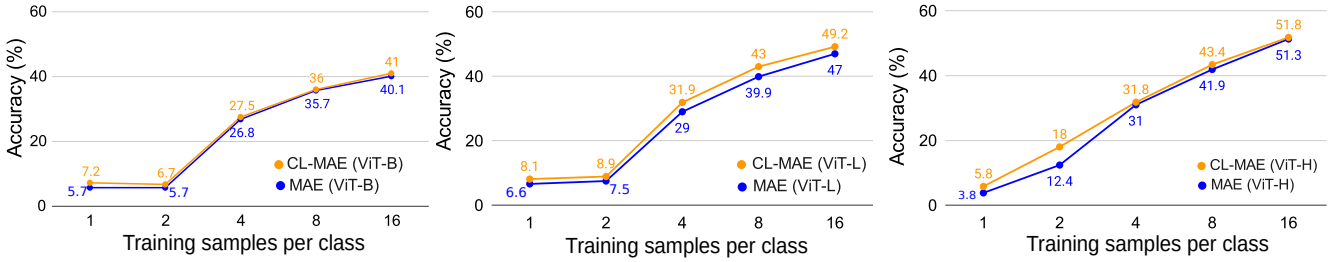
(a) Few-shot linear probing results on Aerial Images [7] with ViT-B (left), ViT-L (center) and ViT-H (right).



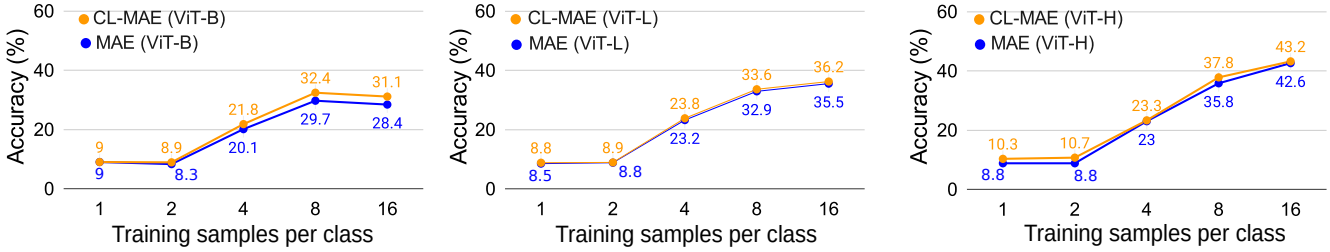
(b) Few-shot linear probing results on Airbus Wind Turbines [2] with ViT-B (left), ViT-L (center) and ViT-H (right).



(c) Few-shot linear probing results on Architectural Heritage Elements [4] with ViT-B (left), ViT-L (center) and ViT-H (right).



(d) Few-shot linear probing results on Sea Animals [6] with ViT-B (left), ViT-L (center) and ViT-H (right).



(e) Few-shot linear probing results on Sport Balls [1] with ViT-B (left), ViT-L (center) and ViT-H (right).

Figure 4. Few-shot linear probing results for MAE [3] and CL-MAE (ours) based on various backbones (ViT-B, ViT-L, ViT-H). The number of training samples per class varies between 1 and 16. The reported accuracy rates are averaged over three runs. Best viewed in color.

results before and after introducing the diversity loss ($\mathcal{L}_{div}$) as an objective of our learnable masking module. When the masking module is trained without adding the diversity loss, the module enters mode collapse, producing nearly identical masks, regardless of the input sample. This can be problematic, as CL-MAE might overfit to the task and learn to reconstruct only a certain configuration of patches. In contrast, integrating the diversity loss helps our module

Few-shot scenario	Method	Aerial Images		Airbus Wind Turbines		Architectural Heritage Elements		Sea Animals		Sport Balls	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
1-shot	MAE (Vit-B) [3]	15.4	48.6	19.3	54.3	13.3	47.2	5.7	20.6	**9.0**	32.6
	CL-MAE (Vit-B)	**16.3**	**58.9**	**24.6**	**63.9**	**15.7**	**61.7**	**7.2**	**26.7**	**9.0**	**34.5**
	MAE (Vit-L) [3]	11.3	39.8	17.8	56.2	10.2	40.5	6.6	27.7	8.5	34.9
	CL-MAE (Vit-L)	**16.0**	**50.7**	**19.5**	**63.8**	**18.2**	**55.8**	**8.1**	**31.6**	**8.8**	**36.9**
	MAE (Vit-H) [3]	11.5	54.1	19.5	57.2	10.8	51.6	3.8	**22.8**	8.8	34.8
	CL-MAE (Vit-H)	**24.6**	**58.2**	**27.1**	**59.0**	**22.7**	**56.0**	**5.8**	22.4	**10.3**	**35.7**
2-shot	MAE (Vit-B) [3]	17.5	53.1	20.4	65.3	13.2	51.4	5.7	24.6	8.3	34.9
	CL-MAE (Vit-B)	**20.7**	**60.3**	**28.8**	**69.9**	**15.4**	**60.2**	**6.7**	**25.8**	**8.9**	**37.2**
	MAE (Vit-L) [3]	10.4	39.6	17.4	60.1	11.6	42.9	7.5	**33.6**	8.8	32.9
	CL-MAE (Vit-L)	**15.9**	**51.3**	**22.6**	**66.2**	**18.5**	**53.5**	**8.9**	31.9	**8.9**	**34.7**
	MAE (Vit-H) [3]	13.6	50.7	19.8	61.7	11.2	51.5	12.4	43.0	8.8	34.9
	CL-MAE (Vit-H)	**27.1**	**58.8**	**27.7**	**69.2**	**22.7**	**55.8**	**18.0**	**52.2**	**10.7**	**35.0**
4-shot	MAE (Vit-B) [3]	45.1	83.3	57.1	89.8	**47.4**	87.3	26.8	62.5	20.1	58.0
	CL-MAE (Vit-B)	**45.5**	**89.9**	**62.7**	**91.5**	**47.4**	**91.6**	**27.5**	**63.2**	**21.8**	**58.8**
	MAE (Vit-L) [3]	53.4	87.2	56.5	88.1	50.6	88.7	29.0	65.3	23.2	58.4
	CL-MAE (Vit-L)	**55.5**	**91.8**	**61.8**	**90.4**	**52.7**	**89.9**	**31.9**	**66.4**	**23.8**	**59.4**
	MAE (Vit-H) [3]	**53.7**	90.8	58.1	89.9	**51.4**	89.3	31.0	65.7	23.0	**58.5**
	CL-MAE (Vit-H)	**53.7**	**90.9**	**63.1**	**92.0**	50.2	**91.4**	**31.8**	**67.0**	**23.3**	**58.5**
8-shot	MAE (Vit-B) [3]	57.6	95.9	65.9	97.8	56.5	**96.6**	35.7	70.6	29.7	65.5
	CL-MAE (Vit-B)	**59.0**	**96.5**	**69.3**	**98.2**	**57.3**	96.3	**36.0**	**71.5**	**32.4**	**69.4**
	MAE (Vit-L) [3]	61.1	97.2	67.1	97.9	**60.0**	97.4	39.9	74.7	32.9	68.1
	CL-MAE (Vit-L)	**63.1**	**97.4**	**70.3**	**98.7**	59.8	**97.7**	**43.0**	**76.3**	**33.6**	**69.4**
	MAE (Vit-H) [3]	**62.4**	**96.3**	68.3	97.9	62.1	95.7	41.9	76.2	35.8	70.1
	CL-MAE (Vit-H)	**62.4**	96.2	**70.3**	**98.6**	**63.1**	**96.6**	**43.3**	**76.5**	**37.8**	**72.1**
16-shot	MAE (Vit-B) [3]	53.3	93.7	68.1	97.4	52.7	**93.1**	40.1	73.9	28.4	67.6
	CL-MAE (Vit-B)	**57.9**	**96.8**	**72.3**	**99.0**	**54.2**	92.7	**41.0**	**76.4**	**31.1**	**69.0**
	MAE (Vit-L) [3]	64.5	95.1	70.1	97.8	61.3	95.4	47.0	81.1	35.5	73.7
	CL-MAE (Vit-L)	**66.0**	**97.3**	**73.3**	**99.1**	**62.1**	**95.5**	**49.2**	**81.9**	**36.2**	**75.3**
	MAE (Vit-H) [3]	68.1	96.7	72.5	98.2	66.3	**96.7**	51.3	82.9	42.6	76.9
	CL-MAE (Vit-H)	**69.0**	**98.1**	**74.4**	**99.1**	**67.2**	96.6	**51.8**	**82.7**	**43.2**	**78.1**

Table 1. Few-shot linear probing results on five benchmarks: Aerial Images, Airbus Wind Turbines, Architectural Heritage Elements, Sea Animals, and Sport Balls. The results are reported for MAE [3] and CL-MAE (ours) based on various backbones (ViT-B, ViT-L, ViT-H). The reported accuracy rates are averaged over three runs. The top scores for each backbone on each data set are in bold.

to escape mode collapse and generate diverse masks, which solves the problem of overfitting to certain mask configurations.

Effect of Kullback-Leibler loss on masking. Figure 3 illustrates three masks generated by our masking module during training, illustrating the effect of removing the Kullback-Leibler loss. At the beginning of the training process, when the module aims to make the reconstruction task easy for the MAE, it learns to leave all patches visible. As the training progresses, the module starts behaving like an adversary, aiming to make the reconstruction task hard for the MAE. From this point on, our module starts masking a number of tokens until it converges to masking all tokens. At this point, the MAE has literally no chance at reconstructing the input, being unable to further learn any useful information. To mitigate this issue, we add the Kullback-Leibler loss, which ensures the number of masked tokens complies with the desired ratio given as hyperparameter, irrespective of the complexity of the task.

Few-shot linear probing results. In Figure 4, we present few-shot linear probing results on Aerial Images [7], Airbus Wind Turbines [2], Architectural Heritage Elements [4], Sea Animals [6], and Sport Balls [1] data sets. We complement the graphs illustrated in Figure 4 with the results reported in Table 1. These experiments are meant to quantify the transferability of the self-supervised representations learned by MAE [3] and CL-MAE in the few-shot learning scenario. In 138 out of 150 cases, CL-MAE outperforms MAE, with absolute gains varying between $+0.1\%$ and $+13.5\%$. We generally observe that CL-MAE tends to bring higher gains for the 1-shot and 2-shot scenarios, especially when the ViT-L and ViT-H backbones are applied on the Aerial Images [7], Airbus Wind Turbines [2], Architectural Heritage Elements [4], and Sea Animals [6] data sets. Overall, the few-shot experiments confirm the observations on the nearest neighbor and linear probing experiments pre-

sented in the main article. We thus conclude that our curriculum learning approach represents a useful addition to the MAE framework.

References

[1] Samuel Cortinhas. Sport Balls - Multiclass Image Classification. https://www.kaggle.com/datasets/samuelcortinhas/sports-balls-multiclass-image-classification, 2022. Accessed: 2023-08-28. 2, 3

[2] Airbus DS GEO. Airbus Wind Turbines Patches. https://www.kaggle.com/datasets/airbusgeo/airbus-wind-turbines-patches, 2022. Accessed: 2023-08-28. 2, 3

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009, 2022. 2, 3

[4] Ivan Kobzev and Vasiliev Roman. Architectural Heritage Elements Image Dataset. https://www.kaggle.com/datasets/ikobzev/architectural-heritage-elements-image64-dataset, 2021. Accessed: 2023-08-28. 2, 3

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[6] Lanz Vencer. Sea animals image dataset. https://www.kaggle.com/datasets/vencerlanz09/sea-animals-image-dataste, 2023. Accessed: 2023-08-28. 2, 3

[7] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 2, 3