

# Minimizing Embedding Distortion for Robust Out-of-Distribution Performance

Tom Shaked\*    Yuval Goldman\*    Oran Shayer  
AppsFlyer

{tom.shaked,yuval.goldman,oran.shayer}@appsflyer.com

## Abstract

Foundational models, trained on vast and diverse datasets, have demonstrated remarkable capabilities in generalizing across different domains and distributions for various zero-shot tasks. Our work addresses the challenge of retaining these powerful generalization capabilities when adapting foundational models to specific downstream tasks through fine-tuning. To this end, we introduce a novel approach we call "similarity loss", which can be incorporated into the fine-tuning process of any task. By minimizing the distortion of fine-tuned embeddings from the pre-trained embeddings, our method strikes a balance between task-specific adaptation and preserving broad generalization abilities. We evaluate our approach on two diverse tasks: image classification on satellite imagery and face recognition, focusing on open-class and domain shift scenarios to assess out-of-distribution (OOD) performance. We demonstrate that this approach significantly improves OOD performance while maintaining strong in-distribution (ID) performance.

## 1. Introduction

One of the primary challenge in machine learning is achieving strong generalization capabilities [10, 30], particularly in real-world applications where models must handle unpredictable scenarios. This is crucial in diverse fields such as autonomous driving [1, 29], robotics [13], and aerial imagery analysis [17, 22], where systems must adapt to varied conditions. The need to handle this long-tail of possible scenarios [25] demands vast, comprehensive datasets. However, creating such datasets is often infeasible, as many scenarios are too rare or unpredictable to be adequately represented in training data, posing a significant challenge for model development and deployment.

Foundational models [12, 23], trained on massive and diverse datasets [3], have shown promise in addressing these generalization challenges with remarkable zero-shot and



Figure 1. Images sampled from the ID and OOD datasets for face recognition (top - domain shift) and image classification (bottom - unseen classes, for example the rightmost image is labeled "solar panel", not present in train set).

few-shot learning capabilities. However, adapting them to specific downstream tasks through fine-tuning [14, 21] often leads to a trade-off between task-specific performance and preservation of broad generalization capabilities, potentially compromising the model's ability to handle OOD scenarios.

In this paper, we take a step towards "enjoying both worlds" - tuning the embedding space for our specific task while leveraging the powerful generalized embeddings from pretrained models. We propose a novel "similarity loss" approach that can be incorporated into the fine-tuning process of any task. Our method aims to minimize the distortion of fine-tuned embeddings from their pre-trained counterparts, striking a balance between task-specific adaptation and the preservation of broad generalization abilities. This approach allows us to train on downstream tasks without discarding the strong generalizability capabilities of foundation models.

We evaluate our approach on two challenging tasks: image classification on satellite imagery and face recognition. These domains are well suited for assessing OOD performance due to their inherent variability and the potential for significant distribution shifts (fig. 1). Our experiments focus on open-class and domain shift scenarios, providing a comprehensive assessment of our method's effectiveness in improving OOD performance. We demonstrate that our method significantly improves OOD performance while incurring only a small reduction in ID performance across different tasks. We provide extensive experiments and analy-

\*Equal contribution.

<sup>1</sup>Accepted to ECCV 2024 Workshop.

sis, showcasing the effectiveness of our approach in real-world scenarios with potential distribution shifts.

## 2. Related Work

Adapting pre-trained foundation models to downstream tasks while maintaining robustness is an ongoing challenge. [14] showed that fine-tuning can distort pretrained features and underperform linear probing on OOD data, proposing LP-FT as a solution. Several approaches followed: WiSE-FT [27] used weight-space ensembling, CLIPood [24] employed margin metric softmax and Beta moving average, and FLYP [4] continued using the contrastive loss from pretraining. These methods have shown varying success in improving OOD performance while maintaining in-distribution accuracy.

Other works have explored lightweight fine-tuning approaches to preserve pretrained features, such as prompt-based methods [26]. Concurrent work by [28] showed that ensembling the weights of zero-shot and fine-tuned models can help balance ID and OOD performance. These studies highlight the importance of carefully considering the fine-tuning process to maintain the generalization capabilities of foundation models like CLIP [23].

## 3. Approach

### 3.1. Similarity Loss

One of the key advantages of leveraging pre-trained foundation models for downstream tasks is the fact that these models are usually trained on massive amounts of data, covering a diverse set of domains and achieving impressive generalization on them. This broad generalization is achieved in terms of both the coverage of open-world vocabulary, and a wide range of data domains. Our goal is to perform fine-tuning of such foundation models on a downstream task while preserving these generalization capabilities.

Our key finding is that while these models achieve impressive zero-shot generalization across many domains and tasks, fine-tuning on a specific downstream task distorts the embedding space, fitting it to the dataset’s domain and hurting generalization to other domains. We therefore argue that additional constraints during fine-tuning are necessary to prevent this distortion and preserve the pre-trained model’s generalization capabilities. At the heart of our approach lies the similarity loss. It is a simple constraint which can also be viewed as a form of regularization, that can be clipped onto any loss function and it is task independent. It is formalized as follows:

$$L_{sim}(x) = \|f_{\theta}(x) - f_{\theta_0}(x)\|_2^2 \quad (1)$$

where  $x$  is the input,  $f_{\theta_0}$  is the original pre-trained model, kept frozen through the entire fine-tuning phase, and  $f_{\theta}$  is

the model we train, initialized with the pre-trained weights from  $f_{\theta_0}(x)$ .

The similarity constraint can be weighted in the total loss by a coefficient  $\alpha$ , which balances task-specific specialization with the preservation of the pre-trained model’s semantic properties. Notice that when  $\alpha = 0$ , standard unconstrained fine-tuning occurs, while  $\alpha \rightarrow \infty$  essentially maintains the original pre-trained model. We can then control the value of  $\alpha$  in training according to how we want to balance between ID and OOD performance.

In this paper, we focus on the CLIP model [9, 23] to demonstrate our approach, but this can be applied to different foundation models, depending on the task at hand.

### 3.2. Similarity Loss for Image Classification

**Framework.** We utilize CLIP in a manner similar to that described in [4] and [24]. Specifically, we fine-tune the model on the downstream task in a contrastive manner, without introducing an additional linear classification layer. During training and inference, we convert class labels into a caption format, following the template "a photo of a {class}".

**Loss Function.** We imply the following loss:

$$\mathcal{L} = L_{clip}(I_{1:B}, T_{1:B}) + \alpha \cdot \frac{1}{B} \sum_{i=1}^B L_{sim}(I_i), \quad (2)$$

where  $B$  is the batch size, and  $I, T$  are the image-text pairs.

This loss is composed of the standard CLIP loss [23] referred as  $L_{clip}$  and our similarity loss from 1.

For this task, we imply the similarity loss over the vision encoder embeddings. In our experiments, we found that extending this similarity constraint to the text encoder as well did not yield additional improvements. We hypothesize that this is because CLIP loss aligns the vision and text embeddings, hence the similarity loss over the vision encoder indirectly constrains the text encoder as well.

### 3.3. Similarity Loss for Face Recognition

**Framework.** For the face recognition task, we leverage CLIP as the pretrained model to fine-tune. While CLIP’s architecture and pre-training data are not specifically tailored for face recognition, as opposed to common approaches in this field [2, 8, 11, 16] that tailor the architecture for this task, we see value in examining our method in this context. Although there is currently no foundation model specifically designed for facial data, and we acknowledge that our in-distribution performance will not match recent state-of-the-art models, we believe there is room to show the benefits of this approach for this task, and that the results we will show in sec. 4 on extreme domain shifts will serve as an additional validation source of our claim.

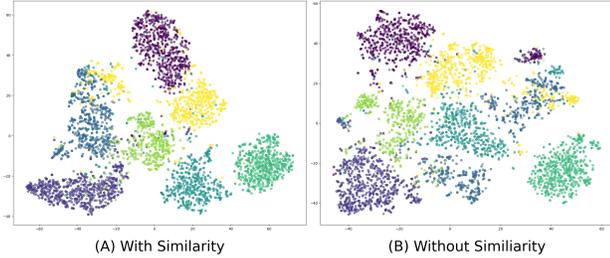


Figure 2. Image embeddings of the OOD EuroSAT dataset, color-coded by class. (A) Model trained with similarity loss (Avg. cluster variance: 1.87e-04) (B) Model trained without similarity loss (Avg. cluster variance: 4.00e-04)

Lacking paired text or semantic class labels, we use only the vision encoder, tuning it to capture discriminative facial features. Unlike standard CLIP training, which uses image-text pairs for contrastive loss, we train on image-image pairs. For each example in the batch we sample a pair of images  $U_i, V_i$  that match the same identity, and will serve as the positive pair.

**Loss function.** Following the suggested framework, our loss function takes the following form:

$$\mathcal{L} = L_{\text{clip}}(U_{1:B}, V_{1:B}) + \alpha \cdot \frac{1}{B} \sum_{i=1}^B (L_{\text{sim}}(U_i) + L_{\text{sim}}(V_i)) \quad (3)$$

We imply the similarity loss over both facial images in each pair,  $U_i$  and  $V_i$ .

**Training procedure** is outlined in algorithm 1. The contrastive nature of the training process necessitates careful batch construction. We build each batch by sampling face images such that all identities within the batch are distinct. The overall algorithm for face recognition training.

---

**Algorithm 1** Face Recognition Training Procedure

---

**Require:** Pre-trained CLIP model  $f_{\theta_0}$

**Require:** Facial images dataset  $\mathcal{D}$  with  $N$  unique identities

- 1: Initialize the fine-tuned model  $\theta \leftarrow \theta_0$
  - 2: **for**  $k = 1$  **to**  $K$  **do**
  - 3:   Create batch of size  $B$  by sampling  $(U, V)$  image pairs from  $B$  unique identities
  - 4:   Calculate  $\mathcal{L}$  as in equation 3
  - 5:   Update model parameters  $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$
  - 6: **end for**
- 

## 4. Experiments

To evaluate the effectiveness of our similarity loss approach, we conduct experiments across two distinct OOD scenarios, as mentioned in sec. 1 and detailed in [24]: unseen-classes in the test set and extreme domain shift.

All of our models were initialized with weights of ViT-B/32-laion2b pre-trained model from OpenCLIP [9]. Further implementation details are provided in the supplement.

### 4.1. Image Classification

For the task of image classification we focus on the satellite imagery domain. We examine OOD performance by testing on a distribution shift with a possible open-class scenario, where the test set might contain classes not seen in the fine-tuning train set.

#### 4.1.1 Datasets

**RSICD** [19] is comprised of 10,921 remote sensing images collected from Google Earth, Baidu Map, MapABC, and Tianditu. The dataset is split into 8,734 training images and 1,093 test images. Each image is paired with 5 text captions, together they comprise 54,605 image-text pairs. This dataset will serve as our ID dataset.

**EuroSAT** [7] contains 27,000 geo-referenced samples of Sentinel-2 satellite imagery, labeled into 10 land use and land cover classes. **RS-ICB128** [15] consists of 36,707 128x128 pixel images of 45 scene categories, sourced globally to represent China’s land use classification standards. It’s valuable for land cover and land use analysis. **Pattern-Net** [32] includes 30,400 high-resolution (256x256 pixel) images across 38 classes, with 800 images per class. These images are sourced from Google Earth and Map API. These datasets will serve as our OOD datasets, showcasing a distribution shift as they are gathered from different data sources than RSICD, and present an open-class case study as well, where some classes and content differ significantly from our ID dataset (fig. 1).

#### 4.1.2 Evaluation protocol

For the ID test set (RSICD), we utilize the Retrieval@K (RET@K) metric, where  $K = 1, 5, \text{ and } 10$ . This metric is the common metric for image-caption paired data [17]. RET@1 effectively measures the model’s accuracy in exact caption retrieval, while RET@5 and RET@10 provide insights into broader retrieval performance. For OOD datasets, we employ the standard classification accuracy metric, which is more suitable for the multi-class classification nature of these datasets.

#### 4.1.3 Results

We used RemoteCLIP [17] as our primary comparison baseline, as it has shown strong performance on this task and trained in a similar manner. We began by evaluating the pretrained CLIP model without fine-tuning, followed by our fine-tuned model with and without the custom similarity component.

Method	RSICD (ID)				EuroSAT (OOD)	RS-ICB128 (OOD)	PatternNet (OOD)
	RET@1	RET@5	RET@10	Mean RET	Accuracy	Accuracy	Accuracy
RemoteCLIP [17]	<b>17.02</b>	<b>37.97</b>	<b>51.51</b>	<b>35.26</b>	35.9	24.18	57.81
CLIP pre-trained [9]	4.75	18.3	28.6	17.2	43.26	26.62	<b>64.3</b>
Ours baseline (no similarity loss)	12.5	31.9	51.0	32.3	26.81	19.3	38.49
Ours + similarity loss	11.4	35.1	51.3	32.6	<b>51.2</b>	<b>34.83</b>	63.41

Table 1. Evaluation results on ID and OOD datasets for the satellite image task.

Method	IJB-C (ID)			iCartoonFace (OOD)	
	TAR @ 1e-6	TAR @ 1e-5	TAR @ 1e-4	TAR @ 0.01	TAR @ 0.05
ArcFace [2]	<b>89.97</b>	<b>94.34</b>	<b>96.18</b>	4.85	13.04
CLIP pre-trained [9]	5.11	13.58	27.44	32.97	56.55
Ours - CLIP, contrastive, no similarity loss	18.49	34.55	55.21	20.07	37.85
Ours - CLIP, contrastive, with similarity loss	13.12	23.64	40.96	<b>39.67</b>	<b>62.29</b>
Ours - CLIP, ArcLoss, no similarity loss	39.01	47.14	60.14	10.09	25.13
Ours - CLIP, ArcLoss, with similarity loss	7.35	18.94	34.7	36.14	58.69

Table 2. Evaluation results on ID and OOD datasets for the face recognition task.

As shown in Table 1, our custom similarity component significantly improved performance on the OOD datasets, as reflected in accuracy metrics, while introducing only a minor decline in ID performance. Figure 4 illustrates the embeddings from the OOD dataset EuroSAT, where the model that trained with our similarity loss exhibits substantially lower average cluster variance compared to the model trained without it.

While we observed a slight reduction in ID performance compared to the results reported by RemoteCLIP, our model consistently outperformed on all OOD datasets. Notably, for EuroSAT and RS-ICB128, our model not only exceeded RemoteCLIP’s OOD performance but also outperformed the pretrained CLIP baseline. This indicates that our fine-tuned model achieved superior generalizability, which we attribute to the integration of our custom similarity component. In contrast, RemoteCLIP showed a performance drop below the pretrained CLIP baseline on these datasets, further highlighting the advantages of our proposed approach.

## 4.2. Face Recognition

In our face recognition experiments, we evaluate our model’s OOD performance by testing its ability to generalize across different domains. We train on natural face images and test on cartoon and animated faces, presenting an extreme domain shift.

### 4.2.1 Datasets

Following recent work in face recognition [6, 8, 16], we fine-tune our model on the MS1MV2 dataset [5], comprising approximately 5.8 million natural facial images of 85,000 real identities. For ID evaluation, we test our tuned model on the common and standard evaluation dataset IJB-C [20], composed of 3,531 real subjects with 31,334 natural

images and 117,542 video frames, resembling the domain of the training data.

To assess OOD performance on extreme domain shifts, we perform our OOD evaluation on the iCartoonFace dataset [31]. This dataset, the largest for animated and cartoon face recognition, consists of 389,678 images of 5,013 identities. The test split is composed of 20,000 images of 2,000 identities. iCartoonFace provides a challenging testbed for evaluating the ability to recognize face identities in a domain vastly different from the training data, as can be seen in fig. 1.

### 4.2.2 Evaluation protocol

For IJB-C and iCartoonFace benchmarks, we follow standard face recognition evaluation procedures [2, 8, 11, 16]. We use the 1:1 Verification protocol, measuring the True Accept Rate (TAR) at different False Accept Rates (FAR). For IJB-C, we report TAR@FAR=[1e-6, 1e-5, 1e-4]. For iCartoonFace, due to the extreme domain shift, we evaluate at higher FAR values of 1e-2 and 5e-2 to better assess the model’s generalization capabilities on cartoon face identities.

### 4.2.3 Results

We use ArcFace [2] as our primary baseline, as it holds as one of the strongest baselines to date, and has provided the training framework widely used in face recognition models to date. For our CLIP evaluations, we start with evaluating the pretrained CLIP without fine-tuning. Then we apply our CLIP-based implementation with contrastive learning (alg. 1) with and without similarity loss. For additional comparison with ArcFace we adapt our method to use the ArcLoss (with and without similarity loss).

Table 2 presents our experimental results, which demonstrate several key findings. For ID, as expected and acknowledged in sec. 3.3, ArcFace remains superior, given its specialization for natural face recognition. Our main observation is the results and trend for the OOD dataset. While ArcFace struggles with this dataset, our CLIP-based method with contrastive learning and similarity loss significantly outperforms all other approaches.

Notably, the key finding is that the addition of similarity loss consistently improves OOD performance across all variants. This observation demonstrates the effectiveness of our proposed loss not only on preserving generalization capabilities that are hurt in standard fine-tuning, but also improves upon that. ArcLoss variants show a similar trend, but with a significant drop in ID performance, suggesting that CLIP is better suited for contrastive fine-tuning.

## 5. Conclusion

In this paper, we introduced a simple but novel "similarity loss" approach to preserve the generalization capabilities of foundational models during fine-tuning for downstream tasks. Our method demonstrated significant improvements OOD performance while maintaining strong ID results across two diverse tasks: satellite imagery classification and face recognition. Notably, the trade-off between ID and OOD performance can be controlled during training through the weighting of the similarity loss, allowing for flexible adaptation to different requirements.

## References

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1
- [2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 4
- [3] Bleys Goodson. Fine flan: Seqio to parquet so you don't have to, 2023. 1
- [4] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023. 2
- [5] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 4
- [6] Mingjie He, Jie Zhang, Shiguang Shan, and Xilin Chen. Enhancing face recognition with self-supervised 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4062–4071, 2022. 4
- [7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 3
- [8] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 2, 4
- [9] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 2, 3, 4, 7
- [10] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017. 1
- [11] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020. 2, 4
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1
- [13] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013. 1
- [14] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *ArXiv*, abs/2202.10054, 2022. 1, 2
- [15] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowd-sourced data. *Sensors*, 20(6):1594, 2020. 3
- [16] Pengyu Li. Bionet: A biologically-inspired network for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10354, 2023. 2, 4
- [17] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiacong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1, 3, 4
- [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

- [19] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 3
- [20] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018. 4
- [21] Jishnu Mukhoti, Yarin Gal, Philip HS Torr, and Puneet K Dokania. Fine-tuning can cripple your foundation model; preserving features may be the solution. *arXiv preprint arXiv:2308.13320*, 2023. 1
- [22] Mark Pritt and Gary Chern. Satellite image classification with deep learning. In *2017 IEEE applied imagery pattern recognition workshop (AIPR)*, pages 1–7. IEEE, 2017. 1
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2
- [24] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023. 2, 3
- [25] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017. 1
- [26] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. *CVPR*, pages 139–149, 2021. 2
- [27] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *CVPR*, pages 7949–7961, 2021. 2
- [28] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, pages 23965–23998. PMLR, 2022. 2
- [29] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020. 1
- [30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1
- [31] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2264–2272, 2020. 4
- [32] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 3

# Supplementary Material

## 6. Implementation Details

### 6.1. Image Classification

We initialize both our fine-tuned model  $f_\theta$  and the frozen baseline model  $f_{\theta_0}$  with the weights of the ViT-B/32 pre-trained model from OpenCLIP [9]. We allow the entire model layers to be trained in order to adapt to the downstream task. We used the loss function outlined in eq. 2.

For our optimization process, we employed the AdamW optimizer [18] with an initial learning rate of 5e-5 and implemented a linear decay learning rate schedule. We maintained a consistent batch size of 128 across all experiments. To find the right balance between our custom similarity loss and the CLIP loss, we conducted experiments with various values of the hyper-parameter  $\alpha=[0.1, 1, 100, 1000]$ . After careful evaluation, we determined that  $\alpha=100$  yielded the most favorable results when looking at our training progress and our ID metrics, striking an optimal balance between the magnitude of the similarity component and the CLIP loss. We recommend conducting similar exploration when utilizing our solution as each data and loss can behave differently.

### 6.2. Face Recognition

We initialize both our fine-tuned model  $f_\theta$  and the frozen baseline model  $f_{\theta_0}$  with the weights of the ViT-B/32-laion2b pretrained model from OpenCLIP [9]. We follow algorithm 1 as our training procedure.

For training, we use the AdamW optimizer [18] with a learning rate of 1e-5 and a linear decay learning rate schedule. The batch size is set to 256 across all experiments. We set the weight  $\alpha$  for the similarity loss to 1. In addition, we found that increasing the softmax temperature to  $\tau = 0.1$  (in the original implementation  $\tau = 0.01$ ) improved the results.