UCoL: Unsupervised Learning of Discriminative Facial Representations via Uncertainty-Aware Contrast

Hao Wang[*1], **Min Li**[*1], **Yangyang Song**[1], **Youjian Zhang**[2], **Liying Chi**[1]
[1]ByteDance Inc.
[2]The University of Sydney
haoww.wh@gmail.com

Abstract

This paper presents Uncertainty-aware Contrastive Learning (UCoL): a fully unsupervised framework for discriminative facial representation learning. Our UCoL is built upon a momentum contrastive network, referred to as Dual-path Momentum Network. Specifically, two flows of pairwise contrastive training are conducted simultaneously: one is formed with intra-instance self augmentation, and the other is to identify positive pairs collected by online pairwise prediction. We introduce a novel uncertainty-aware consistency K-nearest neighbors algorithm to generate predicted positive pairs, which enables efficient discriminative learning from large-scale open-world unlabeled data. Experiments show that UCoL significantly improves the baselines of unsupervised models and performs on par with the semi-supervised and supervised face representation learning methods.

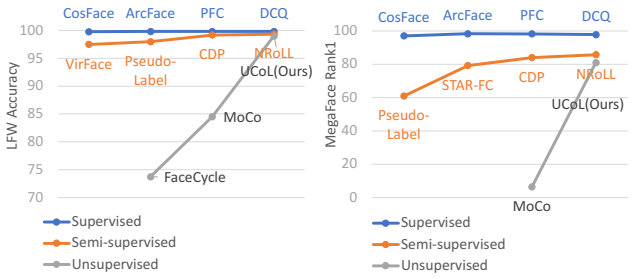

Figure 1: The face recognition performance of supervised/semi-supervised/unsupervised learning on benchmarks (LFW and MegeFace). UCoL improves the baseline of MoCo by a large margin and closes the gap between unsupervised and supervised methods.

1 Introduction

Unsupervised visual representation learning has led to efficient training on many downstream vision tasks, driven by recent advances such as the Contrastive Learning based (Wu et al. 2018; Oord, Li, and Vinyals 2018; Ye et al. 2019; Tian, Krishnan, and Isola 2020; Chen et al. 2020a,b; He et al. 2020) and the Masked Image Modeling (MIM) based (He et al. 2022; Xie et al. 2022; Liu et al. 2022; Li et al. 2022) methods. However, most techniques cannot generalize well to face-related areas without further supervised fine-tuning (Bulat et al. 2022). On the other hand, the performance gains of face recognition are usually achieved at prohibitive annotation costs (Guo et al. 2016; An et al. 2021). Scaling up the current annotation size to more identities tends to suffer from annotation noises (Zhan et al. 2018; Wang et al. 2018a) and may raise privacy concerns (Erkin et al. 2009). Thus, the fundamental question on *"How to learn effective and discriminative facial representation without any supervision?"* remains unsolved.

Several prior studies on supervised face clustering (Yang et al. 2019; Wang et al. 2019; Yang et al. 2020; Shen et al. 2021) stand the dominant approach for solving this problem, where the goal is to apply a graph convolutional network (GCN) to learn the cluster patterns based on identity-prior. However, they still depend on face identity labels to capture biases in facial affinity graphs (Yang et al. 2020). Clustering requires processing the entire dataset for pseudo label generation, making online end-to-end learning challenging. Additionally, assuming no overlap between pseudo labels and original identities is not always valid in supervised face clustering methods (RoyChowdhury et al. 2020).

In this work, we present a simple framework to approach unsupervised discriminative facial representation learning, namely *Uncertainty-aware Contrastive Learning* (UCoL). The objective of contrastive loss is incorporated to separate positive and negative pairs to improve the discriminative power of face representations. We show that the contrastive-based model enables online self-training and efficiently exploits massive unlabeled face images for open-world large-scale unsupervised representation learning.

Following the central idea of vallina MoCo (He et al. 2020), we first build a similar momentum contrast framework to train a self-supervised model on face datasets without identity labels. However, it fails to generate promising results on face verification tasks. The reasons may lie in the preprocessing of faces by strict alignment. Thus, we remove the alignment and compose stronger data augmentations. Accordingly, the performance can be boosted, yet still lags behind the supervised models (See the performance of MoCo in Figure 1). From this perspective, we hypothesize

[*]These authors contributed equally.

that the discriminative clue beyond self-augmented views is desirable for improvement. To this end, we attempt to incorporate the information of pseudo-labels, i.e., the joint training of the self-supervised representation with pairwise self-labeling (Asano, Rupprecht, and Vedaldi 2019; Rebuffi et al. 2021). While some contrastive learning methods (Dwibedi et al. 2021; Van Gansbeke et al. 2021) also utilize the nearest-neighbors strategy, the appropriate selection of pairwise instances has not been extensively explored. Our approach aims to enhance fully unsupervised contrastive learning by introducing a novel, uncertainty-aware, and consistent pairwise labeling technique.

Our UCoL is a unified contrastive learning framework with a simple positive pair selection mechanism. We first construct the contrastive learning network upon a MoCo model, wherein the contrastive learning is formulated as dictionary look-up with *dual paths*. Two types of paired samples are constructed as positive samples for discriminative training, i.e., one is from intra-instance with self augmentation, and the other is from inter-instance with pairwise prediction. Similar to the dictionary keys, we maintain an additional dynamic queue to store the predicted positive pairs. Furthermore, we introduce a novel *uncertainty-aware consistency K-nearest neighbors* algorithm to select positive pairs from the query and dictionary keys. Specifically, by looking up the dictionary keys with queries, the nearest neighbors are predicted as positive pairs and placed in the positive queue for inter-instance contrastive learning. We further exploit the uncertainty-aware strategy of multi-view and MC-Dropout (Gal and Ghahramani 2016) to enhance the consistency of labeling. In contrast to the GCN-based offline clustering, our kNN-based pairwise self-labeling is simple and effective for online contrastive learning. We show that UCoL achieves superior results that are close to the state-of-the-art semi-supervised/supervised methods in various benchmarks (Figure 1).

The contributions of this paper are summarized as follows:

- UCoL is a simple framework for fully unsupervised facial representation learning. To the best of our knowledge, the proposed model is the first end-to-end unsupervised method towards large-scale open-world face verification learning.
- We introduce an uncertainty-aware consistency k-nearest neighbors method, which enables fast and accurate online pairwise self-labeling. The joint task of contrastive learning and self-labeling allows learning from predicted pairs and effectively improves the discriminative information of feature embeddings.
- UCoL significantly improves the baseline of self-supervised learning methods without any human annotations for face verification. Notably, our unsupervised model achieves competitive evaluation results on several benchmarks compared with state-of-the-art supervised methods, closing the gap between supervised and unsupervised training.

2 Related Works

Facial representation learning. Learning a compact and generic facial representation benefits numerous facial analysis tasks like face detection (Yang et al. 2016), alignment (Kowalski, Naruniec, and Trzcinski 2017), and recognition (Cao, Li, and Zhang 2018; Wang et al. 2018b), as evidenced by research on supervised training methods (Li et al. 2021; Wang et al. 2018b; Liu et al. 2017; Deng et al. 2019a; Schroff, Kalenichenko, and Philbin 2015; Zhang et al. 2018; An et al. 2021) using large-scale datasets (Zhu et al. 2021; Guo et al. 2016). Some of these methods enforce better clustering on feature embedding from different identities by introducing margin penalties in the loss function. Specifically, Sphereface (Liu et al. 2017) uses a multiplicative angular margin penalty, CosFace (Wang et al. 2018b) adds a cosine margin penalty to the target logit, and ArcFace (Deng et al. 2019a) calculates the angle between the current feature and the target weight using an arc-cosine function. While these methods improve feature embedding compactness and class separation, they require annotated pairs of training data, limiting their use with massive unlabeled face images.

To exploit unlabeled data, face clustering methods (Wang et al. 2019; Yang et al. 2019, 2020; Shen et al. 2021) cluster unlabeled face images into "pseudo pairs" using Graph Convolutional Networks (GCN). However, these methods still require partially annotated data and cannot perform fully unsupervised clustering. Despite these supervised clustering methods, there are also works exploring unsupervised face clustering (Otto, Wang, and Jain 2017; Zhan et al. 2018) with deep features, yet they mainly focus on designing new similarity metrics.

Self-supervised representation learning. Recently, contrastive learning (Hadsell, Chopra, and LeCun 2006; Wu et al. 2018; Oord, Li, and Vinyals 2018; Ye et al. 2019; Tian, Krishnan, and Isola 2020; Chen et al. 2020a,b; He et al. 2020; Zhang et al. 2021) provides a feasible solution to obtain a generic and practical feature representation from large-scale unlabelled data. This technique employs contrastive loss, such as InfoNCE, to maximize mutual information, bringing together associated samples and separating those in different classes. To improve contrastive learning, MoCo (He et al. 2020) proposes the use of a memory bank to store previous representations. Our work follows the MoCo framework, as it allows for easy reformulation of the momentum encoder dictionary in our pipeline. Bulat et al. (2022) conducts a comprehensive evaluation on facial dataset to compare supervised pretraining with unsupervised pretraining based on SwAV (Caron et al. 2020). Zheng et al. (2022) combine image-text contrastive learning with masked image modeling (He et al. 2022; Xie et al. 2022) to explore low-level and high-level information simultaneously. Differ from these works which need to be finetuned before being applied on downstream tasks, our proposed method manages to conduct face recognition directly with the learned facial representation.

Pseudo-labeling. The goal of Pseudo-labeling (Asano, Rupprecht, and Vedaldi 2019) is to generate pseudo-labels for unlabeled samples with a model trained on labeled data.

Face clustering (Wang et al. 2019; Yang et al. 2019, 2020) also uses this approach by using pre-trained networks to generate pseudo-labels. Noroozi et al. (2018) extends this idea by training a large network with a pretext task in an unsupervised setting. Meanwhile, a unique form of Pseudo-labeling, i.e., pairwise pseudo-labeling, has been widely used in the literature for dimension reduction or clustering (Van der Maaten and Hinton 2008). A few methods provide pseudo labels to train deep convolutional neural networks (Rebuffi et al. 2021; Shaham et al. 2018; Hsu et al. 2019). In our method, we utilize the representation learned from unsupervised learning (i.e., MoCo), and select positive pairs on-the-fly for pairwise contrastive learning.

Calibration and uncertainty. Much evidence shows that the deep learning models are not well-calibrated and often fail by providing over-confident incorrect predictions (Nguyen, Yosinski, and Clune 2015; Hendrycks and Gimpel 2016). To mitigate poor calibration, various works (Guo et al. 2017; Xing et al. 2019) propose post-processing techniques of model prediction to guarantee that probability associated with the predicted class label reflect its ground truth correctness likelihood. Such calibration can be interpreted as the overall prediction uncertainty of network from a frequentist view (Lakshminarayanan, Pritzel, and Blundell 2017). On the other hand, estimating the individual prediction uncertainty is also of importance in many domains (Henne et al. 2020; Rizve et al. 2020; Debnath et al. 2021). Monte Carlo Dropout (MC-Dropout) (Gal and Ghahramani 2016) estimates uncertainty by applying Dropout at the test phase for multiple times as approximate Bayesian inference. In this work, we leverage MC-Dropout to guide the *uncertainty-aware* pairwise labeling.

3 UCoL

Inspired by contrastive training methods (Hadsell, Chopra, and LeCun 2006; Wu et al. 2018; Oord, Li, and Vinyals 2018; Ye et al. 2019; Tian, Krishnan, and Isola 2020; Chen et al. 2020a,b; He et al. 2020), we construct a pipeline with a momentum framework (similar to MoCo (He et al. 2020)) and reformulate the contrastive training as a *dictionary look-up* task. That is, given a query sample q, a positive key sample k^+ and a set of negative key samples $\{k^-\} = \{k_0, k_1, k_2, \cdots\}$, the contrastive loss, namely InfoNCE (Oord, Li, and Vinyals 2018), is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)} \quad (1)$$

where τ denotes a temperature hyper-parameter. Here, the InfoNCE can be thought of as a softmax formed measure between the similarity of positive and its corresponding negatives. It can be further replaced with margin-based losses like CosFace (Wang et al. 2018b) or ArcFace (Deng et al. 2019a), as described in the following subsection.

We introduce the two major designs of our method in the following sub-sections: the dual-path momentum contrastive network (Section 3.1) and the uncertainty-aware pairwise labeling mechanism (Section 3.2).

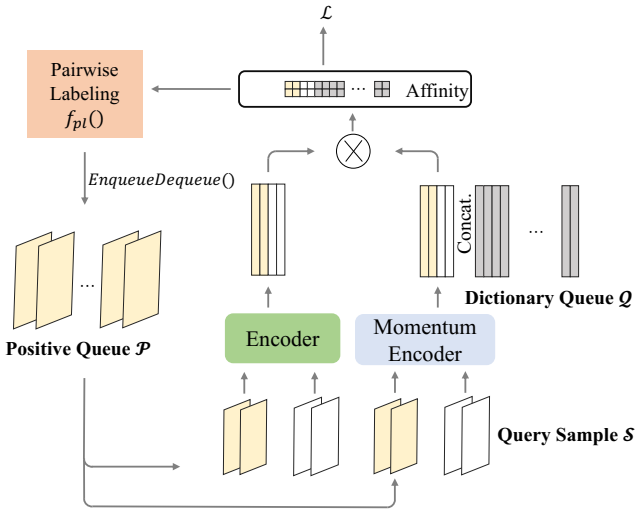

Figure 2: The dual-path contrastive network of UCoL. The blocks filled with yellow refers to the images/features selected by pairwise labeling. The blocks filled with white refers to the images/features of query samples.

3.1 Dual-path Momentum Contrastive Network

For conventional contrastive learning, the *instance discrimination* pretext task is leveraged, where the augmented views of identical instances comprise the positive pairs while the cross-instance pairs are negative. Nevertheless, the discriminating information is insufficient to identify actual pairs with ground truth labels. Motivated by this, we consider two sources of positive pairs in our contrastive training, one is from intra-instance with self augmentation (same way as MoCo (He et al. 2020) and SimCLR (Chen et al. 2020a)), and the other is inter-instance pairs predicted by online *pairwise labeling*. Thus, the positive pairs of cross-instance collected by predictions could provide additional clues to enhance the discrimination. Here, we implement this dual-path contrastive learning with a Dictionary Queue and a Positive Queue. The overview of the proposed dual-path momentum framework of UCoL is visualized in Figure 2.

Dictionary Queue. We start with a brief review of MoCo's dictionary queue: given the query samples, MoCo relies on a dynamic queue of dictionary key samples to perform the dictionary look-up. In the dictionary, the augmented counterpart samples of queries are positive, and the rest are negative. In our framework, we maintain an identical *Dictionary Queue* (Figure 2 right) as same to MoCo for the training part of intra-instance self-augmentation. The size of the dictionary is set to be much larger than the training mini-batch size, thus scaling up the size of negative pairs to improve the results.

Positive Queue. In parallel with the Dictionary Queue that is for intra-instance self-contrast, we keep another dynamic queue to store the collected inter-instance contrastive pairs, which are actually positive pairs in the dataset, referred to as *Positive Queue*. That is, with an efficient pairwise labeling

strategy, it continually retrieves neighboring pairs as positive and caches the results in the Positive Queue alongside the online training. Formally, given a set of query samples $\mathcal{S}$ and a Dictionary Queue $Q = \{k^-\}$, the pairwise labeling function $f_{pl}()$ outputs the predicted Positive Pair samples, which are then placed in the Positive Queue P in a $DequeueEnqueue$ process:

$$\mathcal{P} = DequeueEnqueue(\mathcal{P}, f_{pl}(\mathcal{S}, \mathcal{Q})) \tag{2}$$

As the positive pairs with high confidence collected by f_{pl} could be small-sized and inaccurate in the initial training state, it is difficult to bootstrap learning from unstable and insufficient data discrimination. Therefore, the proposed Positive Queue is beneficial to facilitate the training and improve the discrimination power. Here, the positive queue size can equal the mini-batch size for balance, or be set as a hyper-parameter.

Dual-path Momentum Network. We follow MoCo and exploit the momentum updating mechanism to train an encoder network with the contrastive loss. Specifically, the key encoder is trained with standard forward-backward propagations to update the network parameters, while the query encoder is a counterpart of the key encoder and is updated with momentum. With the proposed Dictionary Queue and Positive Queue described above, we have two contrastive flows from both intra-instances and inter-instances. The intra-instance flow is formed under the same settings as MoCo, wherein various augmentations are adopted. For the inter-instance flow, the contrastive scheme further learns to encourage discrimination from actual paired samples, i.e., the positive query-key pairs are from Positive Queue while Dictionary Queue provides negatives.

Margin-based InfoNCE. Conventionally, the InfoNCE loss is implemented as the log loss of $(N+1)$-way softmax-based classification loss that tries to classify q as k^+. For face recognition, the loss function plays an important role. In this work, we adopt the Large Margin Cosine (lmc) loss function used in CosFace (Wang et al. 2018b), and reformulate the InfoNCE as:

$$\mathcal{L}_{\text{lmc}} = -\log \frac{\exp(s(q \cdot k^+ - m))}{\exp(s(q \cdot k^+ - m)) + \sum_{k^-} \exp(s(q \cdot k^-))} \tag{3}$$

where m and s denote the cosine margin and scale, respectively. Note that the loss takes the L_2 normalized encoded query and keys.

Formally, given a batch $\mathcal{S}$ of unlabeled face instances and two queues described above $(\mathcal{Q}, \mathcal{P})$, the combined loss $\mathcal{L}$ for unsupervised face representation learning in UCoL is defined as:

$$\mathcal{L}_{\mathcal{Q}} = -\log \frac{\exp((q_a \cdot k_a^+ - m)/\tau)}{\exp((q_a \cdot k_a^+ - m)/\tau) + \sum_{k^-} \exp(q_a \cdot k^-/\tau)}, \quad \text{where} \quad (q_a, k_a^+) = Aug(x), x \in \mathcal{S}, k^- \in \mathcal{Q}. \tag{4}$$

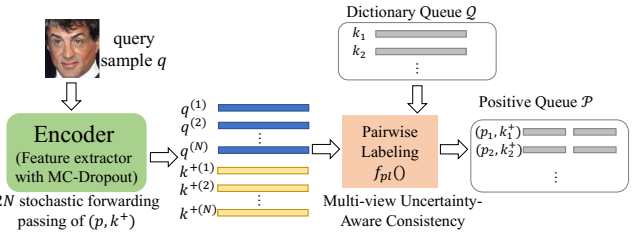

Figure 3: Demonstration of uncertainty-aware pairwise self-labeling. The predictions of positive pairs are the intersection of multiple stochastic forward passes of the augmented views. Then the generated positive pairs are placed into the positive queue.

$$\mathcal{L}_{\mathcal{P}} = -\log \frac{\exp((q_p \cdot k_{p+}) - m)/\tau)}{\exp((q_p \cdot k_p^+ - m)/\tau) + \sum_{k^-} \exp(q_p \cdot k^-/\tau)}, \quad \text{where} \quad (q_p, k_p^+) \in \mathcal{P}. \tag{5}$$

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathcal{Q}} + \lambda \mathcal{L}_{\mathcal{P}} \tag{6}$$

where $Aug()$ is a stochastic image transformation and outputs two versions of augmented x, and λ is the hyperparameter weighting the contribution of the two contrastive losses. Note that two losses share the same dictionary keys $\{k^-\}$.

3.2 Uncertainty-aware Pairwise Self-labeling

The pairwise self-labeling is performed onto the latent embeddings of paired samples, which are sampled from the input query set $\mathcal{S}$ and the key dictionary $\mathcal{Q}$ respectively. The general purpose of self-labeling is to predict and assign pseudo labels. In UCoL, we search the paired neighboring samples as the actual positive pairs, and the negative pairs are from inter-instance, processed by thresholding and subset sampling. Figure 3 demonstrates our uncertainty-aware pairwise self-labeling, which is elaborated below.

K-NN Search for Pairwise Retrieval and Labeling. To generate accurate positive and negative labels of paired samples, the mainstream techniques (Wang et al. 2019; Yang et al. 2019, 2020; Shen et al. 2021) incorporate supervised clustering like GCNs to construct the facial affinity graphs. However, it is time-consuming, and the training of GCN also needs hand-craft supervision. Instead, to keep it simple, we leverage a K-nearest neighbors strategy for fast and accurate pairwise predictions. It is assumed query-key pairs with higher similarity are more likely to be positive, while lower similarities are considered negative. Therefore, we always rank the keys by similarity scores to find out the head ones as positive, and the tail ones as negative.

Thresholding and Negative Sampling. Furthermore, to reduce labeling noise, we set thresholds of confidence onto the ranked pairs to output label assignments. Specifically, for the positive label assignments, a initially high threshold is set and then decreased progressively alongside the training. For the negative part, an adaptive threshold is empirically set to $T_n = \mu - 2\sigma$, where $\mu = \sum_i p_i y_i$ and $\sigma = \sum_i p_i (y_i - \mu)^2$

are calculated over the pairwise per-sample similarity distribution with the posterior probabilities obtained by softmax of Eq. (5).

Positive Selection with Multi-view Uncertainty-aware Consistency. Although the aforementioned thresholding and sampling strategy reduce the pairwise prediction error, the network may still output incorrect pseudo-labels with high confidence, due to the poor calibrations and high uncertainties. The relationship between network calibration and prediction uncertainties has been investigated in (Rizve et al. 2020). It is verified that pseudo-labels with more certain predictions would lead to fewer calibration errors. Following this core analysis, we propose to incorporate an uncertainty-aware consistency criterion into the positive selection process: we apply MC-Dropout (Gal and Ghahramani 2016) to obtain the predictions of positive pair labels by multiple stochastic forward passes, and then calculate the intersection set of all the predictions. In addition, we also use the augmented views to conduct the multiple forwards, referred to as Multi-view Uncertainty-aware Consistency. Thus, if N stochastic forwards are conducted, we obtain $2N$ predictions in total for inter-section merging. With the criterion of Multi-view Uncertainty-aware Consistency, the accuracy of obtained pairwise pseudo-labels can be largely improved, providing adequate correct pairs for the unsupervised training.

Formally, according to the similarity measured by $\text{sim}(q,k) = q \cdot k, k \in \mathcal{Q}$, we denote $\mathcal{N}_K(q,Q)$ as K nearest neighbors of a probe v in set $\mathcal{Q}$.Given the above, we denote K nearest neighbors of i-th stochastic forward pass and its multi-view variant as $\mathcal{N}_K^{(i)}(q^{(i)},Q)$ and $\mathcal{N}_K^{+(i)}(k^{+(i)},Q)$, respectively. Then $\mathcal{R}(q,\mathcal{Q})$ is defined as inter-section of $2N$ K-nearest neighbors of q and k^+ in the set $\mathcal{S}$:

$$\begin{aligned}&\mathcal{R}(q,\mathcal{Q})\\&=\{k_i|k_i\in\mathcal{N}_K^{(0)}(q^{(0)},Q)\wedge\mathcal{N}_K^{+(0)}(k^{+(0)},Q)\wedge\cdots\}\end{aligned}\quad(7)$$

3.3 Learning from Open-world Unlabeled Data

The proposed UCoL learns from unlabeled data by constructing contrastive pairs, which is intuitive and natural for open-set face evaluations in practice. In general, traditional supervised training on face verification usually relies on complete annotations of identities onto the face dataset for supervision. As the training is limited in the labeled dataset, it is unable to utilize data from the out-of-set. For example, even if the out-of-set data samples are well-annotated, one should still merge the data into the existing training dataset to add additional annotations and remove conflicts, and rebuild a larger closed dataset. Our UCoL provides an effective solution to address this and allows training under the open-set settings. By conducting online pairwise self-labeling and contrastive learning, UCoL is capable of learning from massive data in the open world, and the discriminative ability could be enhanced consistently driven by the increasing large-scale of data.

4 Experiments

4.1 Setup

Datasets. MS-Celeb-1M (Guo et al. 2016) dataset is adopted as the training dataset, and the ground-truth labels are eliminated under the unsupervised setting. MS-Celeb-1M is a large-scale face recognition dataset. We adopt the version of MS1M-RetinaFace (Deng et al. 2019b), which consists of 5.1M images from 93K classes. All the face images are preprocessed with alignment and cropping in the same way with arcface (Deng et al. 2019a).

To demonstrate the effectiveness of the learned facial representation, we conduct experiments on several standard face recognition benchmarks, including LFW (Huang and Learned-Miller 2014), MegaFace (Kemelmacher-Shlizerman et al. 2016) and LJB-C (Nech and Kemelmacher-Shlizerman 2017), to test the face verification accuracy.

Baseline Model. We adopt the standard vision transformer (ViT) architecture as the backbone network, since ViT has been demonstrated to have a strong representation ability in MoCo-v3 (Chen, Xie, and He 2021) and MAE (He et al. 2022). To make the architecture more suitable for face recognition tasks, we make some minor tweaks to the original ViT-B/16 architecture proposed in (Dosovitskiy et al. 2020). First, the patch size in the patch embedding layer is changed from 16 to 8, such that the number of patches for a 112×112 image is still 14×14. We observed that the original patch-size 16 performed slightly worse, as the inadequate number of image patches tends to limit the fitting power of self-attention in vision transformers. Second, the classification head in ViT is replaced with a linear layer, which projects the "hidden" embedding to the face embedding of dimension 512. We denote the modified version as ViT-B/8.

There are also some changes we made in the MoCo framework: 1) The projection head introduced in MoCo-v2 is removed, and the loss function is directly computed over the output embedding of the encoder. This will lead to faster convergence. 2) We incorporate CosFace (Wang et al. 2018b) margin into the InfoNCE loss, which improves the discriminative ability of the learned feature representation.

Implementation Details. We follow a common recipe and employ the AdamW optimizer to train vision transformers steadily. Our training approach starts with 5 epochs of linear warmup and employs a cosine learning rate decay. For improved stability, we utilize the same setting of layerwise learning rate decay as MoCo-v3. Our models are trained for 20 epochs on 8 Tesla V100s, with a batch size of 512, learning rate of 0.004, and weight decay of 5×10^{-4}. During the first 4 epochs, we exclusively use intra-instance contrastive learning ($\lambda = 0$), then introduce inter-instance self-labeled pairs by increasing the coefficient λ to 0.5. To perform margin-based InfoNCE loss, we set positive and dictionary queues to sizes 512 and $204,800$, respectively, with hyper-parameters $\tau = 1/80$ and $m = 0.3$.

Method	LFW	MegaFace Rank 1	MegaFace Veri	IJB-C TAR@1e-04	Model	Training-set
Supervised						
CosFace* (Wang et al. 2018b)	99.77	97.02	97.51	95.57	ViT-B/8	MS1M-RetinaFace
ArcFace (Deng et al. 2019a)	99.82	98.35	98.48	95.6	ResNet-100	MS1MV2
PFC (An et al. 2021)	99.83	98.25	98.03	95.8	ResNet-100	MS1MV2
DCQ* (Li et al. 2021)	99.83	97.81	97.77	95.76	ViT-B/8	MS1M-RetinaFace
Semi-supervised						
STAR-FC (Shen et al. 2021)	-	79.26	-	-	-	MS1M
CDP (Zhan et al. 2018)	-	81.88	-	-	Inception-ResNet V2	MS1M
CDP(NRoLL) (Liu et al. 2021)	99.20	84.02	87.19	-	MobileFaceNet	MS1M
NRoLL (Liu et al. 2021)	99.35	85.77	88.14	-	MobileFaceNet	MS1M
Unsupervised						
MoCo* (He et al. 2020)	84.53	6.53	5.66	32.90	ViT-B/8	MS1M-RetinaFace
MAE* (He et al. 2022)	84.60	-	-	-	ViT-B/8	Glint360k
FaceCycle (Chang, Chen, and Chiu 2021)	73.72	-	-	-	16-layer CNN	VoxCeleb1 + VoxCeleb2
UCoL* (ours)	**99.00**	**81.02**	**82.92**	**84.82**	ViT-B/8	MS1M-RetinaFace

Table 1: Comparison of evaluation results on several face recognition benchmarks. Note the * denotes those results are obtained by our implementations under the same settings with UCoL, and the rest are from the authors.

4.2 Evaluation on Face Recognition Benchmarks

We evaluate our UCoL model on several face recognition benchmark datasets (LFW, MegaFace and LJB-C) to compare against state-of-the-art methods in different training scheme, covering supervised (CosFace (Wang et al. 2018b), Arcface (Deng et al. 2019a), PFC (An et al. 2021), DCQ (Li et al. 2021)), semi-supervised (STAR-FC (Shen et al. 2021)) and unsupervised settings (MoCo (He et al. 2020), MAE (He et al. 2022), FaceCycle (Chang, Chen, and Chiu 2021)). Here, the MoCo model without further pseudo-labeled pairs is the baseline of our UCoL, which sets up a lower-bound. Meanwhile, as the DCQ builds a similar MoCo-based framework with contrastive pairs generated with ground-truth labels, it can be regarded as the supervised upper-bound of UCoL. The state-of-the-art of semi-supervised models mostly relies on the GCN to generate pseudo labels and train a face recognition model with both labeled and pseudo-labeled data.

The evaluation results and benchmark comparisons are reported in table 1. The proposed UCoL outperforms the prior unsupervised methods, especially improving the baseline of MoCo by a large margin. The MoCo leads to poor performance on the challenging large-scale evaluations of MegaFace and IJB-C, mainly due to the insufficient variations in instance self-augmentation. In contrast, benefiting from the extra discriminative information enhanced by online pairwise labeling, our UCoL shows effectiveness in improving unsupervised representation learning. In the meanwhile, UCoL achieves competitive results compared with the state-of-the-art semi-supervised and supervised models in the community, and the gap between unsupervised and supervised training has been significantly reduced.

4.3 Ablation Study

We conducted experiments to verify the effect of the ablative variants of our method. The results are presented in Table 2. First, we incorporate the kNN search with multi-view consistency without thresholdinghe model fails to converge and leads to degraded training. Then, we adopt the proposed positive and adaptive negative thresholding; the model improves the baseline by a remarkable margin. The negative sampling (with a sample rate of 0.3) provides more than 24% improvement. The uncertainty-aware consistency (with 4 stochastic MC-Dropout forward) further improves the result to 81.02%. A more comprehensive analysis of each component is presented below.

	kNN	Thres-holding	Negative sampling	Uncertainty	Mega-Face
A					6.53
B	✓				-
C	✓	✓			53.68
D	✓	✓	✓		78.19
E	✓	✓	✓	✓	**81.02**

Table 2: Effects of components of UCoL. For negative sampling, the sample rate is 0.3. For uncertainty, the number of MC-Dropout forward is 4. A refers to MoCo; E refers to UCoL.

Positive Thresholding. For the positive thresholding, an initial high threshold of confidence is used to guarantee the accuracy of positive selections at the beginning of the training procedure. The threshold is progressively decreased to a fixed value within 2 training epochs. We vary this threshold hyper-parameter, and the experimental results are presented in Figure 4. With a strictly high threshold 0.6, the result is inferior, since the generated positives with high confidence are easy samples for training. By relaxing the restriction with a lower threshold, the evaluation performance can be boosted consistently. The threshold around $0.45 - 0.5$ yields the op-

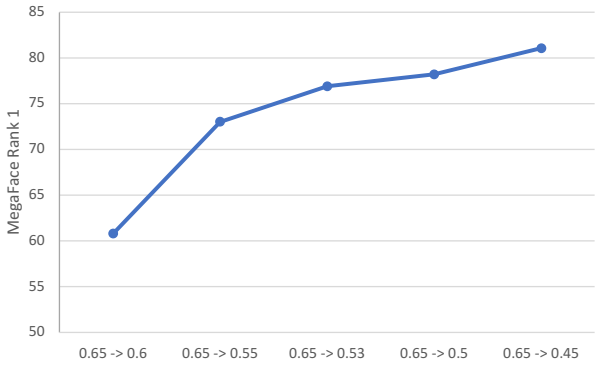

Figure 4: Ablation on positive thresholding. The x-axis denotes the range of decreasing of the positive threshold. The y-axis represents MegaFace rank1 accuracy.

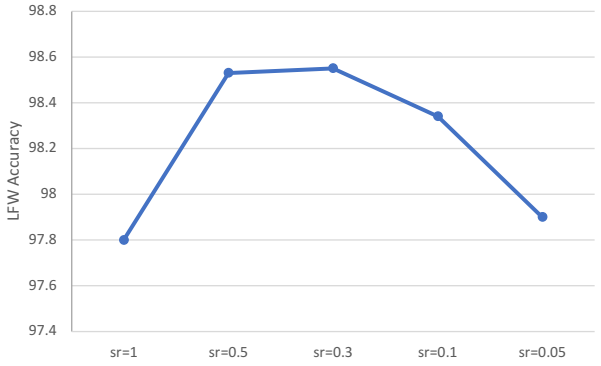

Figure 5: Ablation on negative sampling. The x-axis is the negative sampling rate and the y-axis is the LFW accuracy performance.

timal performance in our experiments.

Negative Sampling. We further analyze the effect of negative sampling. The purpose of negative sampling is to prevent the underlying negative noise in the pairwise pseudo-labels. Figure 5 demonstrate the results with varying negative sample rates on LFW. As the size of negative pairs provided by the dictionary queue is much larger than the training mini-batch size, the experiments show that a relatively lower negative sample rate would lead to superior performance. A lower sample rate would lead to less noisy labels, but the effective correctly labeled negative pairs are also reduced. Hence, if we further reduce the sample rate(lower than 0.1 in this ablative experiment), it would degrade the performance.

Multi-view Uncertainty-aware Consistency. For the Multi-view Uncertainty-aware Consistency, we verify the effect of MC-Dropout with multiple stochastic forward. In detail, as the two augmented views are involved to improve the consistency as well, it would output $2N$ positive pseudo-label predictions if we conduct $2N$ forward passes, and the intersection of all the predictions is used as the eventual prediction. The number of MC-Dropout forward N is set to different values from $N = 4$ to $N = 64$ under the setting of positive threshold 0.5 and negative sampling rate 0.3, and the results are presented in Table 3. We can see that all the multi-view and MC-Dropout contribute to the improvement, as these two strategies can both improve prediction consistency. Increasing forward numbers to 4 led to consistent improvement on MegaFace and LFW, while larger numbers led to comparable and slightly lower performance, probably due to the exclusion of hard samples with too strict consistency conditions.

	Multi-view	Dropout Forward	MegaFace rank1	LFW
Exp1			75.32	97.45
Exp2	✓		78.19	98.55
Exp3	✓	$T = 4$	**81.02**	**99.00**
Exp4	✓	$T = 16$	81.11	98.85
Exp5	✓	$T = 64$	79.54	98.87

Table 3: Ablation on multi-view uncertainty-aware consistency. The positive threshold is set to 0.5, and the negative sampling rate is 0.3.

Pretraining Dataset	LFW	MegaFace Rank1	MegaFace Veri	IJB-C
MS1M	99.00	81.02	82.92	84.82
WebFace260M	**99.40**	**89.71**	**91.22**	**89.45**

Table 4: Comparison of evaluation results obtained by pretraining on different datasets.

4.4 Discussion about Large-scale Pretraining

Apart from the standard end-to-end training, a two-stage training is also an option for our proposed method, where the encoder is initialized by MoCo pretraining on a larger dataset, and then "finetuned" by UCoL on a relatively small dataset. In this way, before the self-labeling even begins, the discriminative power of the encoder is initially improved. As a consequence, Ucol can recognize positive samples more precisely. In table 4, we show the additional result of model pre-trained on a larger-scale dataset, *i.e.* WebFace260M (Zhu et al. 2021). It can be observed that pretraining on WebFace260M has a significant performance improvement over the end-to-end training on MS1M.

5 Conclusion

We propose the novel uncertainty-aware contrast for unsupervised facial representation learning. A dual-path momentum framework is constructed with two types of contrastive learning: one is identical to pre-text training with self-augmentation, and the other is to identify the intra-instance positive pairs collected by online predictions. Also, we incorporate a multi-view uncertainty-aware mechanism in pair-labeling to guarantee accurate selection. We demonstrate that UCoL improves the performance of unsupervised models and closes the gap between supervised and unsupervised training.

References

An, X.; Zhu, X.; Gao, Y.; Xiao, Y.; Zhao, Y.; Feng, Z.; Wu, L.; Qin, B.; Zhang, M.; Zhang, D.; et al. 2021. Partial fc: Training 10 million identities on a single machine. In *IEEE Int. Conf. Comput. Vis.*, 1445–1449.

Asano, Y.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. In *Int. Conf. Learn. Represent.*

Bulat, A.; Cheng, S.; Yang, J.; Garbett, A.; Sanchez, E.; and Tzimiropoulos, G. 2022. Pre-training strategies and datasets for facial representation learning. In *Eur. Conf. Comput. Vis.*, 107–125.

Cao, J.; Li, Y.; and Zhang, Z. 2018. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 4873–4882.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.*

Chang, J.-R.; Chen, Y.-S.; and Chiu, W.-C. 2021. Learning Facial Representations from the Cycle-consistency of Face. In *IEEE Int. Conf. Comput. Vis.*, 9680–9689.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *IEEE Int. Conf. Comput. Vis.*, 9640–9649.

Debnath, B.; Coviello, G.; Yang, Y.; and Chakradhar, S. 2021. UAC: An uncertainty-aware face clustering algorithm. In *IEEE Int. Conf. Comput. Vis.*, 3487–3495.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 4690–4699.

Deng, J.; Guo, J.; Zhang, D.; Deng, Y.; Lu, X.; and Shi, S. 2019b. Lightweight face recognition challenge. In *IEEE Int. Conf. Comput. Vis. Workshops*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Represent.*

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *IEEE Int. Conf. Comput. Vis.*, 9588–9597.

Erkin, Z.; Franz, M.; Guajardo, J.; Katzenbeisser, S.; Lagendijk, I.; and Toft, T. 2009. Privacy-preserving face recognition. In *Int. Sym. on Priv. Enhanc. Tech.*

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Int. Conf. Mach. Learn.*, 1050–1059.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. 2017. On calibration of modern neural networks. In *Int. Conf. Mach. Learn.*, 1321–1330.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Eur. Conf. Comput. Vis.*, 87–102.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 1735–1742.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 9729–9738.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Henne, M.; Schwaiger, A.; Roscher, K.; and Weiss, G. 2020. Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics. In *SafeAI@ AAAI*, 83–90.

Hsu, Y.-C.; Lv, Z.; Schlosser, J.; Odom, P.; and Kira, Z. 2019. Multi-class classification without multi-class labels. *arXiv preprint arXiv:1901.00544*.

Huang, G. B.; and Learned-Miller, E. 2014. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, 14(003).

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4873–4882.

Kowalski, M.; Naruniec, J.; and Trzcinski, T. 2017. Deep alignment network: A convolutional neural network for robust face alignment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops*, 88–97.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.*, 30.

Li, B.; Xi, T.; Zhang, G.; Feng, H.; Han, J.; Liu, J.; Ding, E.; and Liu, W. 2021. Dynamic Class Queue for Large Scale Face Recognition In the Wild. In *IEEE Conf. Comput. Vis. Pattern Recognit.Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3763–3772.

Li, X.; Wang, W.; Yang, L.; and Yang, J. 2022. Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. *arXiv preprint arXiv:2205.10063*.

Liu, J.; Huang, X.; Liu, Y.; and Li, H. 2022. MixMIM: Mixed and Masked Image Modeling for Efficient Visual Representation Learning. *arXiv preprint arXiv:2205.13137*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 212–220.

Liu, Y.; Shi, H.; Du, H.; Zhu, R.; Wang, J.; Zheng, L.; and Mei, T. 2021. Boosting semi-supervised face recognition with noise robustness. *IEEE Transactions on Circuits and Systems for Video Technology*, 778–787.

Nech, A.; and Kemelmacher-Shlizerman, I. 2017. Level playing field for million scale face recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 7044–7053.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 427–436.

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting self-supervised learning via knowledge transfer. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 9359–9367.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

Otto, C.; Wang, D.; and Jain, A. K. 2017. Clustering millions of faces by identity. *IEEE transactions on pattern analysis and machine intelligence*, 40(2): 289–303.

Rebuffi, S.-A.; Ehrhardt, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2021. Lsd-c: Linearly separable deep clusters. In *IEEE Int. Conf. Comput. Vis.*, 1038–1046.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *Int. Conf. Learn. Represent.*

RoyChowdhury, A.; Yu, X.; Sohn, K.; Learned-Miller, E.; and Chandraker, M. 2020. Improving face recognition by clustering unlabeled faces in the wild. In *Eur. Conf. Comput. Vis.*, 119–136. Springer.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 815–823.

Shaham, U.; Stanton, K.; Li, H.; Nadler, B.; Basri, R.; and Kluger, Y. 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*.

Shen, S.; Li, W.; Zhu, Z.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021. Structure-Aware Face Clustering on a Large-Scale Graph with 107 Nodes. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 9085–9094.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Eur. Conf. Comput. Vis.*, 776–794.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; and Gool, L. V. 2021. Revisiting contrastive methods for unsupervised learning of visual representations. *Adv. Neural Inf. Process. Syst.*, 34: 16238–16250.

Wang, F.; Chen, L.; Li, C.; Huang, S.; Chen, Y.; Qian, C.; and Loy, C. C. 2018a. The devil of face recognition is in the noise. In *Eur. Conf. Comput. Vis.*, 765–780.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 5265–5274.

Wang, Z.; Zheng, L.; Li, Y.; and Wang, S. 2019. Linkage based face clustering via graph convolution network. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 1117–1125.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 3733–3742.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 9653–9663.

Xing, C.; Arik, S.; Zhang, Z.; and Pfister, T. 2019. Distance-Based Learning from Errors for Confidence Calibration. In *International Conference on Learning Representations*.

Yang, L.; Chen, D.; Zhan, X.; Zhao, R.; Loy, C. C.; and Lin, D. 2020. Learning to cluster faces via confidence and connectivity estimation. In *IEEE Conf. Comput. Vis. Pattern Recognit.*

Yang, L.; Zhan, X.; Chen, D.; Yan, J.; Loy, C. C.; and Lin, D. 2019. Learning to cluster faces on an affinity graph. In *IEEE Conf. Comput. Vis. Pattern Recognit.*

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2016. Wider face: A face detection benchmark. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 5525–5533.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 6210–6219.

Zhan, X.; Liu, Z.; Yan, J.; Lin, D.; and Loy, C. C. 2018. Consensus-driven propagation in massive unlabeled data for face recognition. In *Eur. Conf. Comput. Vis.*, 568–583.

Zhang, X.; Yang, L.; Yan, J.; and Lin, D. 2018. Accelerated training for massive classification via dynamic class selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, Y.; Wang, C.; Maybank, S. J.; and Tao, D. 2021. Exposure trajectory recovery from motion blur. *IEEE Trans. Pattern Anal. Mach. Intell*, 44(11): 7490–7504.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *IEEE Conf. Comput. Vis. Pattern Recognit.Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 18697–18709.

Zhu, Z.; Huang, G.; Deng, J.; Ye, Y.; Huang, J.; Chen, X.; Zhu, J.; Yang, T.; Lu, J.; Du, D.; et al. 2021. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 10492–10502.