VCL-PL:Semi-Supervised Learning from Noisy Web Data with Variational Contrastive Learning

Mehmet Can Yavuz and Berrin Yanikoglu
Faculty of Engineering and Natural Sciences
Center of Excellence in Data Analytics (VERİM)
Sabancı University, Istanbul, Türkiye 34956
{mehmetyavuz,berrin}@sabanciuniv.edu

***Abstract*—We address the problem of web supervised learning, in particular for face attribute classification. Web data suffers from image set noise, due to unrelated images that may be retrieved in response to the query. We propose a semi-supervised pseudo-labeling approach where the embedding space distribution is learnt via variational contrastive learning. We use 40 Gaussian sampling heads for the 40 attributes in the CelebA dataset and apply supervised contrastive learning over a limited amount of labelled data, to address the multi-label face attribute classification problem. Soft pseudo-labeling is then used to label the unlabelled data at attribute level, followed by two-stage domain adaptation. We show that the proposed method using noisy web data brings improvements in accuracy over supervised multi-label face attribute classification in all experimental settings (over 2% points for very low-data setting). We suggest that learning the embedding distribution and the subsequent soft pseudo-labeling according to the nearest neighbors help in overcoming the noise in the unlabeled data.**

I. Introduction

Unsupervised and semi-supervised learning paradigms are expected to have a great potential for progress in machine learning, as it is possible to collect images, audio or video from nearly limitless data sources on Internet. For example, a web search can be used to collect images to be used in training a visual concept. Unfortunately, the weakly labelled data found on the web in response to the query, often contains large amounts of irrelevant or noisy images. In the domain of face images, a particular Internet search may return images that are unrelated or that only loosely correspond to the query (e.g. images of makeup for "rosy cheek"). In this paper, we propose a semi-supervised learning approach and evaluate its performance on classifying the 40 face attributes depicted in the CelebA dataset, using the internet as the source of the unlabelled data.

Several different approaches are suggested in the literature to leverage unlabelled data. Among these, we can distinguish two broad categories. In the first category, we see unsupervised or self-supervised methods that are used to learn good feature representations. Among these approaches, one group of algorithms including including SimCLR [8], Context Encoders[47], Selfaugment [50], Deeppermnet [54], Clusterfit [71], use a *pretext* task to learn features using self-supervision [76, 4, 6, 7, 9, 10, 15, 17, 22, 24, 26, 30, 29, 32, 31, 36, 37, 40, 45, 67, 72, 78]. Another group of algorithms, including such as Deep Cluster [5], Clustergan[39], SCAN [61], aim to learn good feature representations that lead to *good clusters* [3, 13, 18, 20, 57, 61, 72, 73].

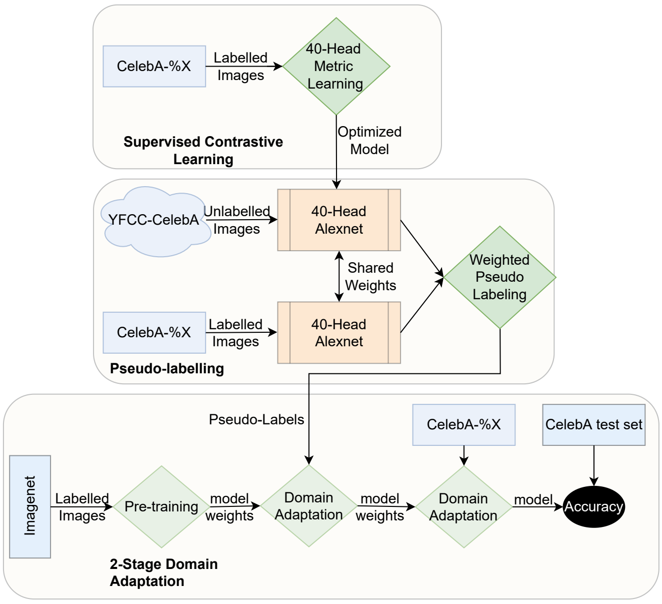

Fig. 1. Proposed pipeline.

In the second category, there are semi-supervised approaches, such as MixMatch [1], FixMatch[56] and FlexMatch[77], and others [1, 41, 48, 52, 69] that use pseudo-labeling or self-labeling, where unlabelled data is assigned *pseudo-labels*. Generative or teacher-student or ensemble models can also be listed among the semi-supervised approaches [12, 16, 23, 33, 38, 49, 53, 55, 59, 64, 65].

Among the first category, SimCLR [8] and SimCLRv2 [9] are the current state-of-the-art self-supervised methods, based on the contrastive learning approach where learning aims to reduce the distance between the embedded representations of the two augmentations by the same image. The effectiveness of these algorithms have been demonstrated on non-noisy benchmark datasets such as Imagenet [14], CIFAR10 and CIFAR100 [28]; however their applicability to multi-label data and noisy web images are yet unaddressed issues. Another successful algorithm is SCAN [61], which aims to form semantic clusters, by using a multi-step learning scheme that starts with the pretext task of SimCLR and continues with novel clustering loss functions.

Our aim in this paper is to increase the accuracy of the existing multi-label face recognition systems by using the visual data collected from the Internet. To this end, we propose a semi-supervised algorithm called *VCL-PL*, consisting of (i) a representation learning step using supervised variational contrastive learning, inspired by variational auto-encoders [25]; (ii) a pseudo-labeling step based on the nearest neighbor mining used in [61]; and a domain adaptation step where the general deep features learned using ImageNet is adapted to the target domain in two-steps. The algorithm is illustrated in Figure 1.

The feature learning component of the proposed method resembles SimCLR [8], but differs from it by the variational approach that aims to learn the underlying distribution of the latent space. Furthermore, unlike SimCLR, we apply contrastive learning to a fraction of the labelled data and construct a separate embedding space for each attribute in order to address the multi-attribute classification, which would not have been possible with unlabelled data. The pseudo-labeling component is inspired by the SCAN [61] and SPICE [42] algorithms that use neighborhood mining in the embedding space, but we use a distance weighting and obtain soft pseudo-labels.

For repeatable experiments, we use the YFCC100M dataset as the data collected from the Internet and the YFCC-CelebA subset obtained by filtering YFCC100M with keywords related to the 40 facial attributes present in CelebA [74]. Note that YFCC100M is an uncurated dataset with only weak labels and is used without labels in this work.

We demonstrate the effectiveness of the proposed algorithm by using varying amounts of labelled data from the CelebA dataset (%1,%10, or %100) and the YFCC-CelebA dataset as the unlabeled dataset. Our main contributions are learning the embedding space distribution using a variational approach and extending the contrastive learning framework to multi-label problems by using 40 Gaussian heads and a limited amount of labelled data. Our system also benefits from a weighted nearest neighbor pseudo-labelling, as well as a two-step domain adaptation.

The paper is structured as follows. In Subsection II-A and II-B, we discuss the backbone network and the Gaussian sampling heads and the supervised metric learning with the variational approach. In Section II-C and II-D, the pseudo-labeling algorithm and the two-stage domain adaptation are presented, respectively. Last two sections are the Experimental Evaluation and the Conclusion sections.

II. Methodology

The proposed algorithm has three consecutive stages and is illustrated in Figure 1.

i Supervised Contrastive Metric learning (Section II-A and II-B). We use the available labelled data (%1 or %10 or %100 of CelebA) and apply contrastive metric learning in a supervised fashion, to learn each of the 40 embedding spaces.

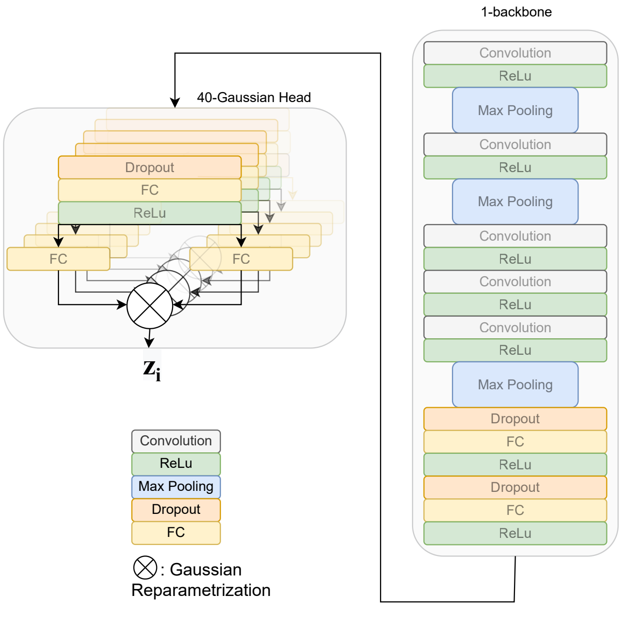

Fig. 2. Learning embeddings: AlexNet is used as the backbone with 40 Gaussian heads for sampling.

ii Nearest neighborhood based weighted pseudo-labeling of the noisy web data (Section II-C).

iii Domain adaptation of the backbone network in two stages. We fine-tune the Imagenet pretrained Alexnet network using the pseudo-labeled YFCC-CelebA and then apply a second domain adaptation with the avaliable labelled CelebA subset (Section II-D).

The supervised metric learning and weighted pseudo-labeling is accomplished in the multi-label domain of face attributes (each image has 40 face attribute labels) with Gaussian embeddings.

A. Feature Extraction and Gaussian Sampling

In this step of the proposed method, we use the labelled set to learn useful embedding distributions, separately for each binary attribute label. The backbone feature extractor is a standard convolutional network, which is followed by sampling heads, as shown in Figure 2.

An input $\boldsymbol{x}$ is applied a stochastic transformation (t) and is fed to the feature extractor network that extracts the embedding representation f_θ. The feature extractor is followed by Gaussian sample heads (g_W) that outputs the parameters of the distribution of the learned embedding space. The process is explained in Eq. 1:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) = g_W(f_\theta(t(\boldsymbol{x}))) \tag{1}$$

We then sample from this distribution using the parametrization trick as used in variational autoencoders [25]:

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^2 \odot \boldsymbol{\xi} \tag{2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\odot$ is the element-wise multiplication. The parametrization trick makes it possible to use backpropagation despite the sampling process.

The network that learns the embeddings consists of two blocks, shown in Figure 2. The backbone is the feature extractor network and its the output vector is shared between 40 Gaussian heads. In our implementation the backbone network is Alexnet architecture [27] with the dropped FC and softmax layers. The activation vectors obtained from the last layer of Alexnet are 4096 dimensional (for 224x224 pixels input) and shared by 40 Gaussian sampling heads which corresponds to 40 attributes of CelebA.

Each embedding space is modelled by a 128-dimensional multi-variate Gaussian distribution with diagonal covariance matrix and sampled with a Gaussian sampling head that has a non-linear layer followed by a linear layer to get the deterministic values of mean and variance. The output of mean and variance embeddings are the inputs for Gaussian reparametrization. This view corresponds to one branch of contrastive network shown in Figure 3.

B. The Supervised Variational Contrastive Learning

A simple contrastive learning algorithm, based on reducing the distance between augmentations of the same image, is run in each 40 embedding spaces independently, with an objective function consisting of three terms, explained below. The algorithm for the variational contrastive learning is given in Alg. 1 and illustrated in Fig. 3.

Large Margin Cosine Loss. The first loss terms is the Large Margin Cosine Loss (LMCL) [62] whose effectiveness has been demonstrated in comparison to softmax [11], center loss [66], large margin softmax loss [35], and angular loss [63] in face recognition domain. Given an input x_i with binary label y_i, LMCL is derived starting from the cross-entropy loss, requiring the weights and input to have unit norm and using the large margin formulation. Specifically, given input x_i and the corresponding ground-truth, y_i:

$$L_{xent} = \frac{1}{N}\sum_{i=1}^{N} - \log p_{y_i} = \sum_{i=1}^{N} - \log \frac{e^{f_{i,y_i}}}{\sum_{j=1}^{K} e^{f_{i,j}}} \tag{3}$$

where N is the number of training samples, p_{y_i} is the posterior probability of the correct class and $f_{i,j}$ is the output of the jth class for the i-th input sample. Denoting the weight vector of the j-th output node as W_j, we have:

$$f_{i,j} = W_j^T x_i = ||W_j|| ||x_i|| \cos\theta_{ij} \tag{4}$$

Then, using normalized weight and input vectors, the cosine loss is derived first. Finally, LMCL is obtained with the large margin formulation:

$$\mathcal{L}_{LMC} = \frac{1}{N}\sum_{i=1}^{N} - \log \frac{e^{\{s(\cos\theta_{ij} - m)\}}}{e^{\{s(\cos\theta_{iy_i} - m)\}} + \sum_{j \neq y_i} e^{\{s(\cos\theta_{ij})\}}} \tag{5}$$

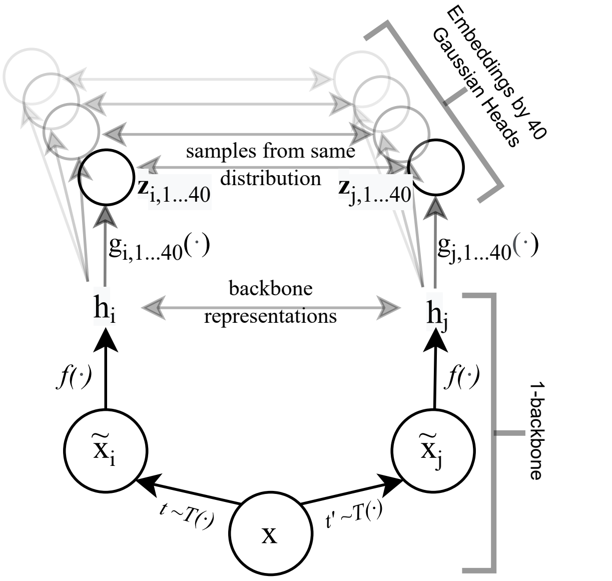

Fig. 3. Variational Contrastive Learning. Two random augmentations of an image x are input to the network to obtain the backbone representations, followed by the 40-Gaussian sampling heads. The two samples are then compared to reduce the total loss (Eq. 8).Figure inspired by SimCLR [8]

where θ_{ij} is the angle between W_j and x_i; s is a constant and m is the margin parameter.

Distribution Similarity Loss. The second loss term encourages the augmentations of the same image (x_i, x_j) being drawn from similar distributions q and p respectively, by penalizing the divergence between the two using the Kullback-Leibler divergence [46].

$$\begin{aligned}\mathcal{L}_S &= -\frac{1}{N}\sum_{i=1}^{N} D_{KL}(q_1(z_i|x_i)||q_2(\tilde{z}_i|\tilde{x_i})) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\sigma_{q_1,i}}{\sigma_{q_2,i}}\right) - \frac{\sigma_{q_1,i}^2 + (\mu_{q_1,i} - \mu_{q_2,i})^2}{2\sigma_{q_2,i}^2} + \frac{1}{2}\end{aligned} \tag{6}$$

Distribution Normalizing Loss. This loss encourages the learned distributions to have zero mean and unit variance, as per [46].

$$\begin{aligned}\mathcal{L}_D &= -\frac{1}{N}\sum_{i=1}^{N} D_{KL}(q_\theta(z_i|x_i)||\mathcal{N}(0,1)) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left[1 + \log(\sigma_{q}^{2}{}_{i}) - \sigma_{q}^{2}{}_{i} - \mu_{q}^{2}{}_{i}\right]\end{aligned} \tag{7}$$

Total Loss. The embedding representations are learned for each binary face attribute, using a portion of the labelled dataset. The optimization is done based on the *total* loss:

$$\mathcal{L}_{total} = \frac{1}{40}\sum_{att=1}^{40} \{\mathcal{L}_{LMC} + \mathcal{L}_S + \mathcal{L}_D\}_{att} \tag{8}$$

Algorithm 1: Contrastive learning design for supervised metric learning, as in [8].

```
input: batch size N, networks $f, g$ and augmentation function distribution $T$
for each sampled minibatch $\{x_k\}_{k=1}^{N}$ do
    for each image $k \in 1, ..., N$ do
        Draw two augmentation functions $t \sim T$, $t' \sim T$
        $\tilde{\mathbf{x}}_{2k-1}$ = t($\mathbf{x}_k$) # first augmentation
        $\mathbf{h}_{2k-1}$ = f($\tilde{\mathbf{x}}_{2k-1}$) # its representation
        for $c \in 1, ..., 40$ do
            $\mathbf{z}_{2k-1,c}$ = g($\mathbf{h}_{2k-1,c}$) # first sample
        end for
        $\tilde{\mathbf{x}}_{2k}$ = $t'$($\mathbf{x}_k$) # second augmentation
        $\mathbf{h}_{2k}$ = f($\tilde{\mathbf{x}}_{2k}$) # its representation
        for $c \in 1, ..., 40$ do
            $\mathbf{z}_{2k,c}$ = g($\mathbf{h}_{2k,c}$) # second sample
        end for
    end for
    Compute loss for each different head using Eq. 8
    Update the network $f, g$ to minimize $\mathcal{L}_{total}$
end for
return: base and sampler networks $f(\cdot)$ and $g(\cdot)$
```

Algorithm 2: Weighted Pseudo-labeling

```
Data: $D_{Labeled}$, $D_{Unlabeled}$
Init: $\{W, \theta\} \leftarrow$ pretrained $\{W^*, \theta^*\}$
Obtain the representation foreach $x_l$ in $D_{Labeled}$
foreach sample $x_u \in D_{Unlabelled}$ do
    Sample augmentation function $t \sim T$
    $\tilde{\mathbf{x}}$ = t($x_u$) # an augmentation
    $\mathbf{h}$ = f($\tilde{\mathbf{x}}$) # representation
    # Gaussian projections in 40 embedding space
    for $c \in 1, ..., 40$ do
        $\mathbf{z}_c$ = g$_c$(h)
        ($distance$, $label$) = mine 1-NN($\mathbf{z}_c$) in $D_{Labeled}$
        $pseudoLabel = label * e^{-distance}$
    end for
    Normalize the labels into [-1,1] range.
end foreach
return: YFCC-CelebA with soft pseudo-labels.
```

C. Weighted Pseudo-labeling

We use the k-nearest neighbor (k-NN) algorithm to pseudo-label the elements of the unlabeled YFCC-CelebA dataset, by the labels of their closest neighbor(s) in the CelebA subset.

For an unlabelled image u, we find the k nearest neighbors in the labelled dataset and obtain the confidence-weighted pseudo-label, according to the labels and distance of each neighbor:

$$pseudoLabel(u) = \sum_{i=1}^{k} \frac{\{label_i * e^{-d_i}\}}{k} \tag{9}$$

where d_i is the distance from u to nearest neighbor i with label $label_i \in \{-1, +1\}$.

The pseudo-labels are normalized into the $[-1, 1]$ range after the pseudo-labelling process. Note that an image can be confident in some labels and less confident in some others. We give the algorithm for $k = 1$ in Alg. 2, as it gave the best results.

D. Two-Step Domain Adaptation

The domain adaptation is done in two steps. The Imagenet pretrained network is fine-tuned with the pseudo-labelled YFCC-CelebA set first; and then to the labelled CelebA set.

We have found doing the adaptation in two steps brings roughly 1% point in accuracy, compared to a single step adaptation (either directly to CelebA or with both data sets combined.

III. Experimental Evaluation

We evaluate the proposed approach on the problem of classifying the 40 facial attributes in the CelebA dataset and compare its performance to : i) standart supervised learning where we fine-tune the ImageNet pretrainet network with the available labelled dataset; ii) DeepCluster [5]; iii) SimCLR [8]; iv) SCAN [61] v) CL-PL (which is the proposed system, only lacking the variational component) vi) VCL-PL (proposed system).

The experiments are run with a portion (%100, %10 or %1) of the CelebA dataset being used as the labelled dataset and YFCC-CelebA dataset as the unlabelled dataset. We used AlexNet [27] in all of the experiments, for simplicity.

Datasets As labelled data, we use the CelebA dataset which is resized and cropped into 128 by 128 pixels, along with its ground truth labels. As unlabelled data, we use a subset of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M), which is the largest public multimedia collection that have approximately 99.2 million are photos and 0.8 million are videos. The subset YFCC-CelebA [74] consists of approximately 1 million photos that are found when searching in English for the 40 face features that exist in the CelebA set (“attractive“, “eyeglasses“ etc). In addition to the face attribute words, the word "face" was added in these searches (e.g. "chubby face").

CelebA has 40 face attributes, but it is also possible to express the opposite concept when the attribute is an adjective, which was determined using ConceptNet [58]). The opposite concept was then obtained from Wiktionary [75] and used in enriching the query (e.g. "wide eyes" along with "narrow eyes"). The search and downloads during this process were done automatically. When multiple queries returned the same photo, repetitions were eliminated.

A total of 392K *non-repetitive* images were obtained using 58 query words obtained by the above process and after eliminating low resolution images. As a last operation, the images were aligned and scaled similar to CelebA. For this, the photos are padded from the edges so as to center the faces and then scaled to obtain 128 × 128 images.

Image Transformation For the augmentations needed in the contrastive metric learning task, we use Resize, Crop, Horizontal Flip, Grayscale, Color Jitter augmentations and sample a random transformation as a combination of these augmentations within the allowed parameter range. The stochastic data augmentation consists of resizing (scale between [0.2, 1.0]), cropping (128 random crops), grayscale transformation (with probability 0.2), and color jitter (with probability 0.8, brightness in [0.6,1.4], contrast in [0.6,1.4], saturation in [0.6,1.4] and hue in [0.9,1.1]).

Training Details. We use AlexNet for the feature extractor (backbone) part of the network, where the embedded representation is 2304 dimensional (for 128x128 pixel input) and the output z of a sample head is 128-dimensional. For training, the network weights are updated using SGD, with a learning rate of 1e-3, momentum coefficient of 0.9 and weight decay of 1e-5. We run pre-training experiments with 400 epochs with Cosine Annihilation Scheduler and fine-tuning experiments with 100 epochs with early validation stopping. The batch size is 128 for pre-training and 64 for fine-tuning. The code is available at https://github.com/verimsu/VCL-PL.

IV. Results

Table I gives the comparison between the proposed VCL-PL algorithm to other well-known approaches, as well as the standart supervised training and the CL-PL approach (proposed system, only without the variational component). Here, supervised learning refers to fine-tuning the ImageNet pretrained AlexNet with the available labelled dataset. DeepCluster [5], SimCLR [8], and SCAN [61] are well-known, state-of-art unsupervised algorithms that are implemented with the code provided in their official repositories.

We see that VCL-PL outperforms state-of-the-art self-supervised learning schemes and standard supervised learning, for all settings (100% , 10% , 1% of CelebA), showing the effectiveness of the proposed method.

The improvements over the best approach from the literature (SimCLR) are 0.49, 0.93, 0.61% points, respectively for 100%, 10% and 1% settings. Note that the scale of these improvements is on par with those observed between other state-of-art methods.

It is also worth noting that VCL-PL and CL-PL are the only two systems that can outperform supervised training in 100% CelebA settings. Furthermore, VCL-PL consistently outperforms CL-PL, showing the benefit of using the variational approach. The accuracies of four of the methods are plotted in Figure 4 for clarity.

In Table II, we observe the k-NN from the pseudo-labeling stage is best for $k = 1$. In fact, the SCAN algorithm [61] also uses 1-NN approach in its nearest neighbor pseudo-labelling. We further observe that that the algorithm benefits from soft labelling (Eq. 9) as compared to using hard labels.

We also evaluated other well-known algorithms, namely Semi-supervised Label Propagation [21] and MixMatch [1]. These methods were observed to underperform compared to supervised learning with the available labelled data. We presume that the main reason for the degradation is that our problem deals with the data noise in the unlabelled set.

TABLE I
Average Accuracy for different algorithms and settings. Bold results indicates the best results while underlined results show the best results from the literature

Method	Imagenet	YFCC	CelebA	**Accuracy**
Supervised	yes	no	100%	90.24%
DeepCluster [5]	no	yes	100%	89.40%
SimCLR [8]	yes	yes	100%	89.98%
SCAN [61]	yes	yes	100%	89.11%
CL-PL	yes	yes	100%	90.32%
VCL-PL	yes	yes	100%	**90.47%**
Supervised	yes	no	10%	88.65%
DeepCluster [5]	no	yes	10%	87.53%
SimCLR [8]	yes	yes	10%	88.75%
SCAN [61]	yes	yes	10%	88.35%
CL-PL	yes	yes	10%	89.43%
VCL-PL	yes	yes	10%	**89.68%**
Supervised	yes	no	1%	85.90%
DeepCluster [5]	no	yes	1%	84.12%
SimCLR [8]	yes	yes	1%	87.51%
SCAN [61]	yes	yes	1%	85.85%
CL-PL	yes	yes	1%	87.69%
VCL-PL	yes	yes	1%	**88.12%**

TABLE II
Average Accuracy for Different k values and using hard labels, at low-data regime

CelebA 1%	1-NN	3-NN	5-NN	Hard labels
Accuracy	88.12%	88.07%	88.03%	87.43%

V. Conclusion

We study the problem of improving the performance of existing supervised systems by the use of weakly labelled data collected from the internet. The specific problem addressed in this work is face attribute classification, where we obtained performance improvements over the supervised learning framework (over 2% points for very low-data setting), and two existing baselines (DeepCluster and SimCLR), with the proposed method.

The main contributions are to use a variational approach to learn the underlying distribution of the embedding space and extending the contrastive learning framework to multi-label problems.

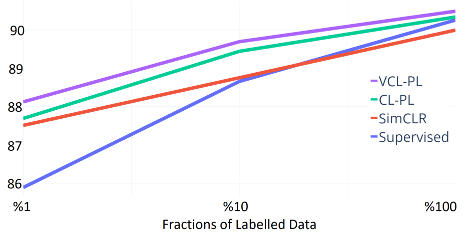

Fig. 4. Accuracy values for varying fractions of available labelled data.

TABLE III
Detailed comparison of supervised training and our proposed systems, VCL-PL and CL-PL.

Attributes	Supervised	CL-PL	VCL-PL	Attributes	Supervised	CL-PL	VCL-PL
5 o'Clock Shadow	89.60%	90.56%	91.06%	Male	90.40%	93.54%	93.96%
Arched Eyebrows	76.06%	77.87%	78.58%	Mouth Slightly Open	69.12%	81.15%	82.62%
Attractive	75.59%	77.49%	77.87%	Mustache	96.04%	96.06%	96.07%
Bags Under Eyes	79.13%	81.01%	81.20%	Narrow Eyes	84.92%	84.93%	85.16%
Bald	97.91%	97.83%	97.92%	No Beard	87.36%	90.09%	91.26%
Bangs	91.69%	94.04%	94.37%	Oval Face	70.09%	71.38%	72.12%
Big Lips	67.38%	68.46%	68.93%	Pale Skin	95.67%	95.78%	95.79%
Big Nose	79.52%	80.43%	80.83%	Pointy Nose	70.07%	71.53%	72.33%
Black Hair	81.90%	84.89%	85.84%	Receding Hairline	91.35%	91.56%	91.65%
Blond Hair	93.70%	94.33%	94.53%	Rosy Cheeks	92.74%	92.89%	93.19%
Blurry	95.11%	94.99%	94.94%	Sideburns	95.13%	95.93%	96.30%
Brown Hair	83.69%	84.95%	85.12%	Smiling	76.64%	85.95%	87.46%
Bushy Eyebrows	86.32%	87.27%	87.74%	Straight Hair	77.39%	79.56%	79.98%
Chubby	94.13%	94.09%	94.47%	Wavy Hair	77.49%	79.69%	80.27%
Double Chin	95.38%	95.28%	95.38%	Wearing Earrings	81.26%	83.82%	84.58%
Eyeglasses	95.95%	97.18%	97.31%	Wearing Hat	97.43%	97.93%	97.79%
Goatee	94.88%	95.26%	95.64%	Wearing Lipstick	87.30%	90.45%	90.64%
Gray Hair	96.80%	97.19%	96.99%	Wearing Necklace	85.42%	85.95%	86.30%
Heavy Makeup	84.08%	87.04%	87.66%	Wearing Necktie	95.05%	94.78%	95.20%
High Cheekbones	75.69%	81.74%	82.63%	Young	80.63%	82.58%	83.30%
				Average	85.90%	87.69%	88.12%

VI. Acknowledgements

This study was supported by TUBITAK (Project number: 119E429).

References

[1] David Berthelot et al. "MixMatch: A Holistic Approach to Semi-Supervised Learning". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 5049–5059.

[2] David Berthelot et al. "ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring". In: *International Conference on Learning Representations*. 2019.

[3] Deyu Bo et al. "Structural Deep Clustering Network". In: *Proc. of The Web Conf. 2020*. 2020, pp. 1400–1410.

[4] Piotr Bojanowski and Armand Joulin. "Unsupervised Learning by Predicting Noise". In: *Int. Conf. on Machine Learning*. PMLR. 2017, pp. 517–526.

[5] Mathilde Caron et al. "Deep Clustering for Unsupervised Learning of Visual Features". In: *Proc. of the European Conf. on Computer Vision (ECCV)*. 2018, pp. 132–149.

[6] Mathilde Caron et al. "Unsupervised Pre-training of Image Features on Non-curated Data". In: *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*. 2019, pp. 2959–2968.

[7] Pengguang Chen, Shu Liu, and Jiaya Jia. "Jigsaw Clustering for Unsupervised Visual Representation Learning". In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2021, pp. 11526–11535.

[8] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Int. Conf. on Machine Learning*. PMLR. 2020, pp. 1597–1607.

[9] Ting Chen et al. "Big Self-Supervised Models are Strong Semi-Supervised Learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22243–22255.

[10] Ting Chen et al. "Self-supervised GANs via Auxiliary Rotation Loss". In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2019, pp. 12154–12163.

[11] Edwin KP Chong and Stanislaw H Zak. *An introduction to optimization*. John Wiley & Sons, 2004.

[12] Zihang Dai et al. "Good Semi-supervised Learning That Requires a Bad GAN". In: *Advances in Neural Information Processing Systems* 30 (2017).

[13] Zhiyuan Dang et al. "Nearest Neighbor Matching for Deep Clustering". In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2021, pp. 13693–13702.

[14] Jia Deng et al. "Imagenet: A Large-scale Hierarchical Image Database". In: *2009 IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised Visual Representation Learning by Context Prediction". In: *Proc. of the IEEE Int. Conf. on Computer Vision*. 2015, pp. 1422–1430.

[16] Hao-Zhe Feng et al. "SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 7413–7421.

[17] Zeyu Feng, Chang Xu, and Dacheng Tao. "Self-supervised Representation Learning by Rotation Feature Decoupling". In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2019, pp. 10364–10374.

[18] Jhosimar Arias Figueroa and Adın Ramırez Rivera. “Is Simple Better?: Revisiting Simple Generative Models for Unsupervised Clustering”. In: *NIPS Workshop on Bayesian Deep Learning*. 2017.

[19] Eren Golge and Pinar Duygulu. “ConceptMap: Mining Noisy Web Data for Concept Learning”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 439–455.

[20] Xifeng Guo et al. “Deep Embedded Clustering with Data Augmentation”. In: *Asian Conf. on Machine Learning*. PMLR. 2018, pp. 550–565.

[21] Ahmet Iscen et al. “Label Propagation for Deep Semi-supervised Learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5070–5079.

[22] Simon Jenni and Paolo Favaro. “Self-supervised Feature Learning by Learning to Spot Artifacts”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2018, pp. 2733–2742.

[23] Zhanghan Ke et al. “Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 6727–6735.

[24] Dahun Kim et al. “Learning Image Representations by Completing Damaged Jigsaw Puzzles”. In: *2018 IEEE Winter Conf. on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 793–802.

[25] Diederik P Kingma and Max Welling. “Auto-encoding Variational Bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[26] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. “Revisiting Self-supervised Visual Representation Learning”. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2019, pp. 1920–1929.

[27] Alex Krizhevsky. *One Weird Trick for Parallelizing Convolutional Neural Networks*. 2014. arXiv: 1404.5997.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning Multiple Layers of Features from Tiny Images”. In: (2009).

[29] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. “Colorization as a Proxy Task for Visual Understanding”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2017, pp. 6874–6883.

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. “Learning Representations for Automatic Colorization”. In: *European Conf. on Computer Vision*. Springer. 2016, pp. 577–593.

[31] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. “Self-supervised Label Augmentation via Input Transformations”. In: *Int. Conf. on Machine Learning*. PMLR. 2020, pp. 5714–5724.

[32] Hsin-Ying Lee et al. “Unsupervised Representation Learning by Sorting Sequences”. In: *Proc. of the IEEE Int. Conf. on Computer Vision*. 2017, pp. 667–676.

[33] Chongxuan Li et al. “Triple Generative Adversarial Networks”. In: *IEEE transactions on pattern analysis and machine intelligence* PP (2021).

[34] Jiaxin Liu et al. “A Survey of Image Clustering: Taxonomy and Recent Methods”. In: *2021 IEEE Int. Conf. on Real-time Computing and Robotics (RCAR)*. IEEE. 2021, pp. 375–380.

[35] Weiyang Liu et al. “Large-margin softmax loss for convolutional neural networks.” In: *ICML*. Vol. 2. 3. 2016, p. 7.

[36] Matthias Minderer et al. “Automatic Shortcut Removal for Self-supervised Representation Learning”. In: *Int. Conf. on Machine Learning*. PMLR. 2020, pp. 6927–6937.

[37] Ishan Misra and Laurens van der Maaten. “Self-supervised Learning of Pretext-invariant Representations”. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[38] Takeru Miyato et al. “Virtual Adversarial Training: A Regularization Method for Supervised and Semi-supervised Learning”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 1979–1993.

[39] Sudipto Mukherjee et al. “Clustergan: Latent Space Clustering in Generative Adversarial Networks”. In: *Proc. of the AAAI Conf. on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 4610–4617.

[40] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. “Improvements to Context Based Self-supervised Learning”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2018, pp. 9339–9348.

[41] Islam Nassar et al. “All Labels Are Not Created Equal: Enhancing Semi-supervision via Label Grouping and Co-training”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 7237–7246.

[42] Chuang Niu, Hongming Shan, and Ge Wang. “Spice: Semantic Pseudo-labeling for Image Clustering”. In: *arXiv preprint arXiv:2103.09382* (2021).

[43] David A Nix and Andreas S Weigend. “Estimating the Mean and Variance of the Target Probability Distribution”. In: *Proc. of 1994 IEEE Int. Conf. on Neural Networks (ICNN’94)*. Vol. 1. IEEE. 1994, pp. 55–60.

[44] Mehdi Noroozi and Paolo Favaro. “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. In: *European Conf. on Computer Vision*. Springer. 2016, pp. 69–84.

[45] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. “Representation Learning by Learning to Count”. In: *Proc. of the IEEE Int. Conf. on Computer Vision*. 2017, pp. 5898–5906.

[46] Stephen G. Odaibo. “Tutorial: Deriving the Standard Variational Autoencoder (VAE) Loss Function”. In: *ArXiv* abs/1907.08956 (2019).

[47] Deepak Pathak et al. “Context Encoders: Feature Learning by Inpainting”. In: *Proc. of the IEEE Conf.*

on Computer Vision and Pattern Recognition. 2016, pp. 2536–2544.

[48] Hieu Pham et al. “Meta Pseudo Labels”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 11552–11563.

[49] Antti Rasmus et al. “Semi-supervised Learning with Ladder Networks”. In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 3546–3554.

[50] Colorado J Reed et al. “Selfaugment: Automatic Augmentation Policies for Self-supervised Learning”. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2021, pp. 2674–2683.

[51] Douglas A Reynolds. “Gaussian Mixture Models”. In: *Encyclopedia of biometrics* 741 (2009), pp. 659–663.

[52] Mamshad Nayeem Rizve et al. “In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning”. In: *International Conference on Learning Representations*. 2020.

[53] Tim Salimans et al. “Improved Techniques for Training GANs”. In: *Advances in neural information processing systems* 29 (2016), pp. 2234–2242.

[54] Rodrigo Santa Cruz et al. “Deeppermnet: Visual Permutation Learning”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2017, pp. 3949–3957.

[55] Philip Sellars, Angelica I. Avilés-Rivero, and Carola-Bibiane Schönlieb. “LaplaceNet: A Hybrid Energy-Neural Model for Deep Semi-Supervised Classification”. In: *ArXiv* abs/2106.04527 (2021).

[56] Kihyuk Sohn et al. “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[57] Chunfeng Song et al. “Auto-encoder Based Data Clustering”. In: *Iberoamerican Congress on Pattern Recognition*. Springer. 2013, pp. 117–124.

[58] Robyn Speer, Joshua Chin, and Catherine Havasi. “Conceptnet 5.5: An Open Multilingual Graph of General Knowledge”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[59] Antti Tarvainen and Harri Valpola. “Mean Teachers Are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results”. In: *NIPS*. 2017.

[60] Bart Thomee et al. “YFCC100M: The New Data in Multimedia Research”. In: *Commun. ACM* 59.2 (Jan. 2016), pp. 64–73. ISSN: 0001-0782.

[61] Wouter Van Gansbeke et al. “Scan: Learning to Classify Images without Labels”. In: *European Conf. on Computer Vision*. Springer. 2020, pp. 268–285.

[62] Hao Wang et al. “Cosface: Large Margin Cosine Loss for Deep Face Recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5265–5274.

[63] Jian Wang et al. “Deep metric learning with angular loss”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2593–2601.

[64] Qin Wang, Wen Li, and Luc Van Gool. “Semi-Supervised Learning by Augmented Distribution Alignment”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 1466–1475.

[65] Xiao Wang et al. “EnAET: A Self-Trained Framework for Semi-Supervised and Supervised Learning With Ensemble Transformations”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 1639–1647.

[66] Yandong Wen et al. “A discriminative feature learning approach for deep face recognition”. In: *European conference on computer vision*. Springer. 2016, pp. 499–515.

[67] Zhirong Wu et al. “Unsupervised Feature Learning via Non-parametric Instance Discrimination”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2018, pp. 3733–3742.

[68] Zhenda Xie et al. “Propagate Yourself: Exploring Pixel-level Consistency for Unsupervised Visual Representation Learning”. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2021, pp. 16684–16693.

[69] Yi Xu et al. “Dash: Semi-Supervised Learning with Dynamic Thresholding”. In: *ICML*. 2021.

[70] Ozge Yalcinkaya, Eren Golge, and Pinar Duygulu. “I-ME: Iterative Model Evolution for Learning from Weakly Labeled Images and Videos”. In: *Machine Vision and Applications* 31.5 (2020), pp. 1–20.

[71] Xueting Yan et al. “Clusterfit: Improving Generalization of Visual Representations”. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 2020, pp. 6509–6518.

[72] Jianwei Yang, Devi Parikh, and Dhruv Batra. “Joint Unsupervised Learning of Deep Representations and Image Clusters”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2016, pp. 5147–5156.

[73] Linxiao Yang et al. “Deep Clustering by Gaussian Mixture Variational Autoencoders with Graph Embedding”. In: *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*. 2019, pp. 6440–6449.

[74] Mehmet Can Yavuz et al. “YFCC-CelebA Face Attributes Datasets”. In: *2021 29th Signal Processing and Communications Applications Conf. (SIU)*. 2021, pp. 1–4.

[75] Torsten Zesch, Christof Müller, and Iryna Gurevych. “Using Wiktionary for Computing Semantic Relatedness.” In: *AAAI*. Vol. 8. 2008, pp. 861–866.

[76] Xiaohua Zhai et al. “S4l: Self-supervised Semi-supervised Learning”. In: *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*. 2019, pp. 1476–1485.

[77] Bowen Zhang et al. “FlexMatch: Boosting Semi-supervised Learning with Curriculum Pseudo Labeling”. In: *Neural Information Processing Systems*.

[78] Richard Zhang, Phillip Isola, and Alexei A Efros. “Colorful image colorization”. In: *European Conf. on Computer Vision*. Springer. 2016, pp. 649–666.