Caption-Supervised Face Recognition: Training a State-of-the-Art Face Model without Manual Annotation

Qingqiu Huang[1][0000−0002−6467−1634], Lei Yang[1][0000−0002−0571−5924], Huaiyi Huang[1][0000−0003−1548−2498], Tong Wu[2][0000−0001−5557−0623], and Dahua Lin[1][0000−0002−8865−7896]

[1] The Chinese University of Hong Kong
[2] Tsinghua Univerisity
{hq016, yl016, hh016, dhlin}@ie.cuhk.edu.hk
wutong16.thu@gmail.com

Abstract. The advances over the past several years have pushed the performance of face recognition to an amazing level. This great success, to a large extent, is built on top of millions of annotated samples. However, as we endeavor to take the performance to the next level, the reliance on annotated data becomes a major obstacle. We desire to explore an alternative approach, namely using captioned images for training, as an attempt to mitigate this difficulty. Captioned images are widely available on the web, while the captions often contain the names of the subjects in the images. Hence, an effective method to leverage such data would significantly reduce the need of human annotations. However, an important challenge along this way needs to be tackled: the names in the captions are often noisy and ambiguous, especially when there are multiple names in the captions or multiple people in the photos. In this work, we propose a simple yet effective method, which trains a face recognition model by progressively expanding the labeled set via both selective propagation and caption-driven expansion. We build a large-scale dataset of captioned images, which contain $6.3M$ faces from $305K$ subjects. Our experiments show that using the proposed method, we can **train a state-of-the-art face recognition model without manual annotation** (99.65% in LFW). This shows the great potential of caption-supervised face recognition.

1 Introduction

Recent years have seen remarkable advances in face recognition [44,41,7,53,51,49]. However, state-of-the-art face recognition models are primarily trained on large-scale annotated datasets [13,5,24], which is becoming a major problem as we pursue further improvement. Obtaining massive amount of accurately annotated data has never been a trivial task. As the scale increases, the cost of annotation, the difficulty in quality control, and the ambiguities faced by the annotators gradually approaches a prohibitive level.

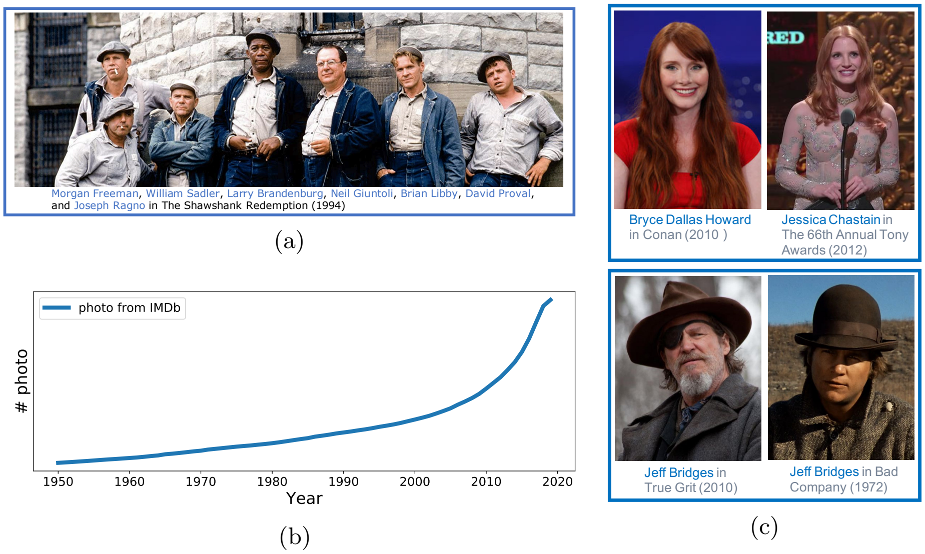

Morgan Freeman, William Sadler, Larry Brandenburg, Neil Giuntoli, Brian Libby, David Proval, and Joseph Ragno in The Shawshank Redemption (1994)
(a)
photo from IMDb
# photo
1950
1960
1970
1980
1990
2000
2010
2020
Year
(b)
Bryce Dallas Howard in Conan (2010)
Jessica Chastain in The 66th Annual Tony Awards (2012)
Jeff Bridges in True Grit (2010)
Jeff Bridges in Bad Company (1972)
(c)

Fig. 1: Captions are often provided by those people who are familiar with the subjects in the photos. The first row shows that captions are often accurate even for difficult cases, *e.g.* different subjects that look similar or an subject that looks differently. The second row shows a key challenge of caption-supervised learning, namely multi-to-multi associations.

An effective way to mitigate this heavy reliance on manual annotations has therefore become a common quest of the community. Semi-automatic schemes have been explored in the development of some large-scale datasets, *e.g.* using search engines [13,5] and clustering with a trained model [24]. However, it has been observed that the noises and bias introdued by these schemes significantly hurt the performance [40].

In this paper, we explore an alternative approach to addressing this problem, namely, to exploit the tremendous amount of captioned images available on the web. This is motivated by the observation that the captions of the photos with people often contain the names of the subjects. These names can provide valuable supervisory signals for training face recognition models. It is also worth noting that in addition to the large quantity, captioned images have another important advantage – the names in the captions are often very accurate even for images that are very difficult to be distinguished visually, as illustrated in Figure 1. This is partly ascribed to the fact that the captions are usually provided by "experts", *i.e.* those people who are familiar with the subjects in the photos or the underlying stories.

While it sounds appealing, training a face recognition model based on caption images is indeed a very challenging task. The key challenge lies in *inexact labels*, *i.e.* a label may be corresponding to one of the several instances in a photo or none of them. Inexact labels would arise when a photo contains more than one faces or a caption contains more than one names. As we are exploring the setting without manual annotation, the associations between faces and names need to be resolved in a certain way, explicitly or implicitly. On the other hand, it is also noteworthy that this is not the same as a multi-instance learning (MIL) problem [45,32,42], as for a considerable portion of the cases, we have exactly one face in the photo and one name in the caption. Figure 2 how caption-supervised face recognition differs from other widely studied learning paradigms.

To tackle the challenges caused by inexact labels while fully exploiting the portion of samples with one-to-one correspondence, we propose a simple method that combines selective propagation with caption-driven expansion. Specifically, our method begins with those samples with one-to-one correspondence as initial labeled seeds, and iteratively expand the labeled set by propagating the labels to neighbors with selective criteria and reasoning about co-existing associations based on captions. We found that by leveraging both the learned feature space and the caption-based supervision, the labeled set can significantly grow while maintaining high accuracy in the inferred labels.

To facilitate this study, we construct a large-scale dataset named *MovieFace* by collecting movie photos and their captions. This dataset contains $6.3M$ faces with $305K$ identities, and the faces exhibit large variations in scale, pose, lighting, and are often subject to partial occlusion. Our model trained on this dataset **without any manual annotation** achieves competitive performance on MS1M[13], a widely used testbed for face recognition techniques. For example, a network with the ResNet-50 backbone [14] trained thereon achieves the accuracy of 99.65% in LFW [15].

Our contributions consist in three aspects: (1) We explore a new paradigm to train face recognition model without manual annotation, namely, caption-supervied training. (2) We develop a simple yet effective method for this, which exploits both the learned feature space and the caption-based supervision in an iterative label expansion process. (3) We construct a large dataset *MovieFaces* without manual annotation to support this study, and manage to train a state-of-the-art model thereon. Overall, this work demonstrates the great potential of caption-supervised face recognition and provides a promising way towards it.

2 Related Work

Semi-Supervised Face Recognition Some of the researchers are also concerned about the unaffordable annotation cost in face recognition and try to alleviate the challenges with the help of semi-supervised learning [35,53,50]. Roli *et al* [35] employed a self-training strategy with multiple PCA-based classifiers, where the labels of unlabeled samples were inferred with an initial classifier and then added to augment the labeled set. Zhao *et al* [55] took LDA [2] as the classifier under a similar self-training scheme. Gao *et al* [11] developed a semi-supervised sparse representation-based approach by modeling both linear and non-linear variation between the labeled and unlabeled samples. Zhan *et al* [53] proposed a consensus-driven propagation algorithm to assign pseudo labels with the help of

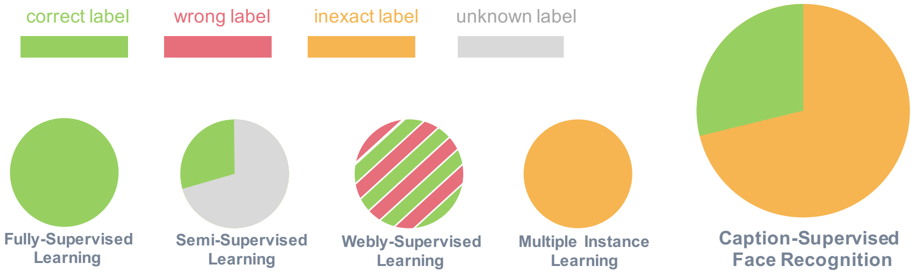

correct label
wrong label
inexact label
unknown label
Fully-Supervised Learning
Semi-Supervised Learning
Webly-Supervised Learning
Multiple Instance Learning
Caption-Supervised Face Recognition

Fig. 2: Comparison of different learning problems and our Caption-Supervised Face Recognition (CSFR). In CSFR, photos contain only one face and one mentioned name can be taken as samples with correct labels and others are with inexact labels.

a constructed relational graph. Although some of these methods are claimed to achieve great performance with only a few labeled samples [53], they are usually tested on some artificial benchmarks modified from fully-labeled datasets, the distribution of which is not natural. The caption-supervised face recognition proposed in this work is much more practical and effective, which would be shown in Sec. 5.

Webly-Supervised Face Recognition Webly-Supervised Learning(WSL) leverages raw data from the Internet and needs no human annotation [12,22]. While the scale of training sources can be easily expanded in this case, massive data noise has become the bottleneck to the classification performance [47]. Efforts have been devoted to addressing the problem from different angles. Some proposed robust algorithms to learn directly against noisy data, as Patrini *et al* [31] proposed a robust loss correction procedure and Rolnick *et al* [36] explored the robustness of the DNN itself when enough examples are available. Others aimed to remove or correct mislabeled data as [38,3,22], while they suffer from distinguishing mislabelled examples from hard training examples. In the specific scenario of face recognition, where noise exists in nearly all the existing large-scale databases [40], a widely accepted solution is to adopt a cleaning procedure to improve the quality of large-scale face datasets [30] Gallo *et al* [10]proposed a pipeline to improve face recognition systems based on Center loss. Jin *et al* [23]proposed a graph-based cleaning method that employed the community detection algorithm to delete mislabeled images. Similar to us, Chen *et al* [6] also made use of web sources and avoided human annotation, but they focused more on dealing with data noise by distinguishing the misclassifications with modification signal. In comparison to most WSL methods, our caption-supervised setting takes full advantage of the web data by transferring the issue of data noise to a multiple instance problem, leading to a breathtaking performance even with a simple approach.

Multiple Instance Learning Multiple Instance Learning (MIL) has an especial yet practical setting that the instances and labels are provided in groups, respectively. It provides more information than semi-supervised manner yet lacks a accurate one-to-one mapping compared with fully-supervised manner. It was originally proposed for drug activity prediction [9] and are now widely applied to many domains [1]. Since a complete survey of MIL is out of the scope of this paper, here we only introduced some recent works based on deep networks. Most of the MIL works focus on how to aggregate the scores or the features of multiple instances [45,32,42,21]. Wu *et al* [45] proposed to use max pooling for score aggregation, which aimed to find the positive instances or patches for image classification. Pinheiro *et al* [32] used log-sum-exp pooling in a CNN for weakly supervised semantic segmentation. Wang *et al* [42] summarized the aggregation module of previous works as MIL Pooling on instance scores. It then proposed MI-Net, which applied MIL Pooling to instance features with a deeply-supervised fashion. Instead of pooling, Ilse *et al* [21] proposed an attention-based MIL Network, which used learnable weights for feature aggregation. Suffering from the same drawback that the information of the web data is provided by users, directly adopting MIL methods to data with an unstable quality would achieve much worse results compared to our approach, which would be shown in Sec. 5.

3 Methodology

To take full advantage of the caption supervision, we propose a framework named caption-supervised face recognition by progressively expanding the labeled samples. Specifically, we maintain a labeled set containing samples with correct labels and an unlabeled set with inexact samples during training. The labeled set would be iteratively enlarged by selective propagation and caption-driven expansion. The former aims to enlarge the number of instances with the help of a trained model. The latter would increase both identities and instances with by means of the caption supervision. Specifically, our framework consists of three stages, namely labeled set initialization, selective propagation and caption-driven expansion, as shown in Fig. 3. The last two stages would be run iteratively until converge, *i.e.* no extra samples can be added to the labeled set.

I. Labeled Set Initialization Suppose that there are n faces $\{f_1, \cdots, f_n\}$ and m identites mentioned in the caption $\{y_1, \cdots, y_m\}$ in the a photo. Here $y_i \in \{1, \cdots, N\}, i \in \{1, \cdots, m\}$ and N is the total number of identities in the dataset. As we mentioned before, some photos contain just one face and one mentioned identity in its caption, which we name as "one2one" samples. Considering the high quality of the captions, we take the faces in "one2one" photos as labeled samples, namely

$$\mathcal{I}(f_n) = y_m, \quad \text{if} \quad n = 1, m = 1$$

These samples would be used to initialize the labeled set.

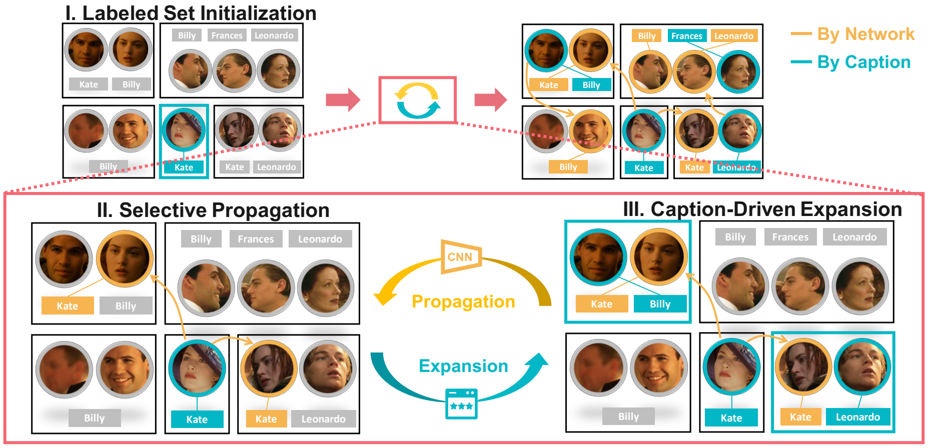

I. Labeled Set Initialization
Kate
Billy
Frances
Leonardo
By Network
By Caption
II. Selective Propagation
III. Caption-Driven Expansion
CNN
Propagation
Expansion

Fig. 3: Our framework for caption-supervised face recognition. It consists of three stages: (I). Initialize a labeled set with those samples containing just one face detected and one identity mentioned in the caption. (II). Then a network is trained on the labeled set and applied to the unlabeled samples. Samples (in orange) would be selected and added to the labeled set following such criterions: (1) with a high prediction score, and (2) the predicted identity is mentioned in the caption (III). We get more labels with the help of caption (in blue), *i.e.* we assign the identity to the face if only one face left in the photo and one identity left in the caption. By running stage II and III iteratively, we would finally propagate the labels to almost all the samples.

II. Selective Propagation With the labeled set, we train a neural network in a fully-supervised manner, which would then be applied to the unlabeled faces. Here we denote the predition score of an unlabeled face f_i as $\mathbf{p}_i \in \mathcal{R}^{\bar{N}}$, where $\bar{N}$ is the number of identities in the labeled set and $\bar{N} \leq N$. At the propagation stage, a face would be labeled under the following criterions: (1) the predicted identity is mentioned in the caption, and (2) the prediction score is higher than a threshold τ, as shown in Eq. 1

$$\mathcal{I}(f_i) = k, \quad \text{if} \begin{cases} \text{argmax}(\mathbf{p}_i) = k \\ k \in \{y_1, \cdots, y_m\} \\ p_{ik} > \tau \end{cases} \tag{1}$$

Since the trained model is incapable of predicting unseen persons, only the number of samples in the labeled set would be increased in a selective manner while the number of identities would remain constant at this stage.

III. Caption-driven Expansion Here we make a reasonable assumption that if there are only one unlabeled face and one unassigned identity in a photo, then the label of the face should be the left identity. After some of the faces are labeled at stage II, there would be some photos with only one unlabeled face and one mentioned identity left. Base on the assumption, the face would be labeled, as shown in Eq. 2, where $\mathcal{U}$ denotes the filter to get the unlabeled ones from a face or identity set.

$$\mathcal{I}(f_i) = y_j, \quad \text{if} \begin{cases} \mathcal{U}(\{f_1, \cdots, f_n\}) = f_i \\ \mathcal{U}(\{y_1, \cdots, y_n\}) = y_j \end{cases} \tag{2}$$

At this stage, the number of the identities as well as the number of the samples in the labeled set would increase, and the driving force comes from the information extracted from caption. After new identities are added, i.e. the labeled set is enlarged, we would finetune the model with the whole labeled set, which contains both old samples and newly added identities.

The proposed framework is so simple that can be reimplemented easily. More importantly, it works well surprisingly on the caption-supervised datasets, even outperforming a model trained on a fully-supervised dataset like MS1M [13], the results of which would be demonstrated in Sec. 5. However, there are also some imperfections with the proposed framework, which would also be discussed in Sec. 5 to benefit the further explorations.

4 Dataset

Dataset	# ID	# face	# annotation
LFW[15]	5K	13K	automatic
CelebFaces[39]	10K	202K	manually
IMDb-Face[40]	59K	1.7M	manually
CASIA[52]	10K	500K	semi-automatic
VGGFace2[5]	9K	3.3M	semi-automatic
MS1Mv2[13,8]	85K	5.8M	semi-automatic
MegaFace[24]	670K	4.7M	automatic
MovieFace	305K	6.3M	caption

Table 1: Comparison between datasets for face recognition.

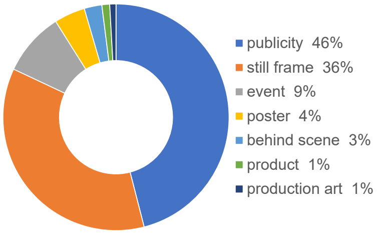
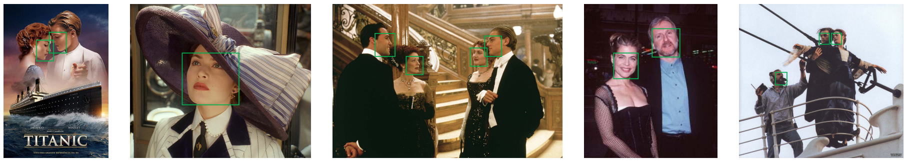


Fig. 4: Different Types of Photo in MovieFace.

Datasets play an important role in most of the researches in computer vision [26,4,16]. Since there is no large-scale dataset to support caption-supervised face recognition, we build a dataset, namely MovieFace, in this paper. A comparison of some popular datasets for face recognition is shown in Tab. 1, from which we can see that our proposed dataset is competitive to the existing largest datasets, for both identities and faces. But our datasets would continuously grow without any manual efforts as shown in Fig. 1. More details of MovieFace would be introduced below. And note that MovieFace is a part of MovieNet [19], which is a holistic dataset that support various of research topics in person recogntion [46,28,18,17,46], video analysis [20,34,33] and story understanding [37,48].



Leonardo DiCaprio and Kate Winslet in Titanic (1997)

Kate Winslet in Titanic (1997)

Leonardo DiCaprio, Kate Winslet, Billy Zane, and Frances Fisher in Titanic (1997)

James Cameron and Linda Hamilton at an event for Titanic (1997)

James Cameron, Leonardo DiCaprio, and Kate Winslet in Titanic (1997)

Fig. 5: Here we show some samples from MovieFace

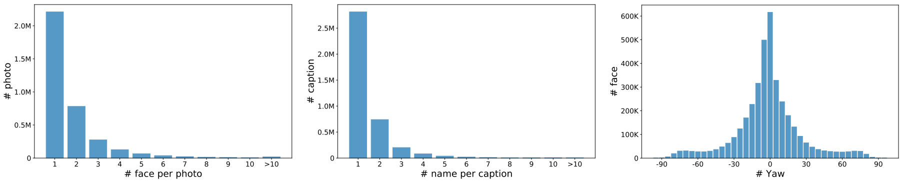


Fig. 6: Here we show some statistics of MovieFace including (a) number of faces per photo, (b) number of name entities per caption, and (c) the yaw distribution of faces.

Face Processing We get different types of photos from IMDb [3] including "still frame", "poster", "publicity", "event", "behind the scenes", "product" and "production art", the definition of which can be seen in the description page [4]. Totally 3.9M photos with name entities in caption are downloaded. Then we detect all the faces in the photos with MTCNN [54], resulting in 6.3M faces detected.

Identity Processing We download the caption of each photo. For each name mentioned in the caption, there would be a hyperlink to the person's homepage, which is created by the users. So it is easy for us to get the identities of the mentioned persons by the hyperlinks. There are 5.8M name entities appeared in the captions, belonging to 305K unique identities.

Dataset Statistic We show the percentage of each type of photos in Fig. 4. Different types of photos would capture faces of a person under different situations, which would highly raise the diversity of the dataset. Some photos and captions from *Titanic* are shown in Fig. 5, from which we can also see the high quality and diversity of MovieFace. We further calculate the yaw of each face, the distributions of which are shown in Fig. 6. And one of the most critical factors for the caption-supervised setting is probably the number of faces per photo and the number of names per caption. If all the photos contain just one face and one mentioned name, then the caption-supervised problem would degenerate into a simple fully-supervised one. The less the faces per photo, the easier for us to

[3] https://www.imdb.com/

[4] https://help.imdb.com/article/contribution/images-videos/imdb-image-faqs/G64MGN2G43F42PES#

train a powerful model. The distributions of the number of faces and the number of names are shown in Fig. 6. We can see that more than 50% of the photos contain just one face and more than 60% of the captions contain just one name, which would highly benefit the training process.

5 Experiments

5.1 Experiment Setting

We test our method on three benchmarks on face recognition/verification, which is the application that motivates this work. We not only compare it with various methods, but also investigate important design choices via a series of ablation studies.

Training set. Following the convention in face recognition, we train networks on large training sets that are *completely disjoint* from the testing sets, namely the identities (*i.e.* classes) used for testing are excluded from the training set. Specifically, six large datasets below are used for training: **(1) MS-Celeb-1M** [13]. This dataset consists of $100K$ identities, each with about 100 facial images on average. In total, the dataset contains $10M$ images. As the original identity labels were extracted *automatically* from webpages and thus are very noisy. We clean up the annotations according to [7], resulting in a subset that contains $5.8M$ images from $86K$ classes. **(2) Megaface2** [24]. It contains $4.7M$ images from $672K$ identities. This dataset is automatically collected from the Internet and the distribution is very long-tail. **(3) IMDb-Face** [40], collects large-scale images for the IMDb website. It develops an effective way to clean the dataset and produces a noise-controlled dataset with $1.7M$ images from $59K$ identities. **(4) CASIA** [52]. This dataset uses the same source as IMDb-Face for data collection. In addition to images, it also collects tags for semi-automatic clean. Applying tag-constrained similarity clustering, it cleans the collected image prudently and result in a dataset contains $494,414$ images of $10,575$ subjects. **(5) MovieFace**. To facilitate the study in caption supervised face recognition, we also collect a large-scale face dataset from IMDb website. This dataset comprises $3.9M$ photos with corresponding captions. We detect $6.3M$ face images from photos and extract $305K$ identities from captions. Note that our collected dataset does not involve any manual annotations.

Testing set. The trained networks are then evaluated on three testing sets: **(1) LFW** [15], the *de facto* standard testing set for face verification under unconstrained conditions, which contains $13,233$ face images from $5,749$ identities. **(2) IJB-A** [25], which contains $5,712$ face images from 500 identities. It partitions all pairs of face images into 10 disjoint sets, and the final result is the average of those obtained from individual partitions. **(3) Megaface & Facescrub**, the largest and most challenging public benchmark for face recognition, which combines the gallery sets from both Megaface [24] (with $1M$ images from $690K$ identities), and Facescrub [29] (with $100K$ images from 530 identities). Specifically, the evaluation is done as follows. In each testing, one image from each celebrity in Facescrub will be mixed into the Megaface gallery to form an augmented gallery set, while the remaining images will be used as queries. The task is to identify the ones from the corresponding classes that were mixed into the gallery, among a large number of distractors from Megaface.

Metrics. We assess the performance on two tasks, namely *face identification* and *face verification*. Face identification is to select top k images from the gallery, where the performance is measured by the top-k hit rate, *i.e.*, the fraction of predictions where the true identity occurs in the top-k list. Face verification is to determine whether two given face images are from the same identity. We use a widely adopted metric [25,24] namely the true positive rate under the condition that the false positive rate is fixed to be 0.001.

Networks. We conducted two series of experiments, with different network architectures. First, to experiment over different training sets and loss functions within a reasonable budget, we use a modified ResNet-50 [14] with input size reduced to 112x112. To further study how different methods work with very deep networks, we conducted another series of experiments for selected methods using R-100 and ArcLoss [8], which achieves the state-of-the-art in face recognition benchmarks. For all settings, the networks are trained using SGD with momentum. The mini-batch sizes are set to $2,048$ and $1,024$ respectively for ResNet-50 and R-100.

5.2 Comparison to Fully Supervised Training

	Softmax Loss			Cosine Loss			ArcFace		
Dataset	LFW	IJBA	MegaFace	LFW	IJBA	MegaFace	LFW	IJBA	MegaFace
MS1M	**99.52**	**88.24**	**84.44**	99.63	**91.93**	94.33	99.85	96.82	97.92
MegaFace2	98.35	55.48	53.47	98.75	79.94	66.81	99.28	86.60	84.75
IMDb-Face	98.70	73.21	73.02	99.37	84.17	79.99	99.65	94.25	94.81
CASIA	98.08	55.05	58.63	98.28	60.79	71.33	99.00	72.05	78.90
MovieFace	99.10	77.75	83.34	**99.65**	88.95	**95.44**	99.83	**96.96**	96.96

Table 2: Comparion of the Performance between Webly-Supervised and Fully-Supervised Face Recognition

The results are shown in Tab. 2. MovieFace is trained by supervision of captions, which can be automatically collected from the web; While other datasets are trained under the supervision of labels, which are usually obtained by massive human annotations. Comparing the performance of models under different settings, we observe that: (1) The model trained on MovieFace yields comparable identification/verification accuracies with trained MS1M, for different loss functions and the ArcFace method; (2) Under all different settings, it consistently outperforms models trained on other three datasets, namely MegaFace2, IMDb-Face, and CASIA, by a large margin; (3) By applying the state-of-the-art method ArcFace on IJBA, it can further produce performance gain over MS1M by 0.14 percent, despite the fact that no explicit annotation is offered when learned caption-supervised.

5.3 Comparison to SSL and MIL Methods

SSL Methods We collect images that only have one face and one name item in their captions to form a training set $\mathcal{S}_1$/One2One, where the face is labeled with the name item. To employ semi-supervised methods in our setting, we first regard the $\mathcal{S}_1$ data as labeled data and use it to train a feature extractor. With the trained feature extractor, we extract features for all unlabeled images except $\mathcal{S}_1$ and apply it in our scenario. To avoid the overlap between $\mathcal{S}_1$ labels and pseudo labels, we adopt a multitask scheme for training, *i.e.*, there are two classifiers on top of the network for $\mathcal{S}_1$ labels and pseudo labels respectively.

As unlabeled face images are likely to belong to an unseen identity, clustering are widely adopted to exploit unlabeled face data [53,43,51]. We study two clustering methods in our settings, namely K-means [27] and LTC [51]. K-means is the most widely used unsupervised clustering methods, while the recent proposed LTC introduces supervised clustering and shows its effectiveness in exploiting unlabeled face images. For K-means, we set the number of clusters to the total number of identities extracted from captions. For LTC [51], we use $\mathcal{S}_1$ as the labeled set to train the clustering models.

The results in Table. 3a shows that: (1) Compared with the model trained on $\mathcal{S}_1$, K-means achieves comparable performance over three benchmarks. Relying on simple assumptions that all samples are distributed around a center, K-means may fail to handle the complex distribution in large-scale dataset in real-world setting, especially when the number of clusters is inaccurate. (2) LTC outperforms the model trained on $\mathcal{S}_1$ consistently. Although it is more effective than K-means in exploiting large-scale unlabeled data, the improvement is limited. As a supervised method, LTC assumes the distribution between training set and testing set is similar. In our scenario, we take $\mathcal{S}_1$ as the labeled data for training, but there is no guarantee that the remained unlabeled data has a similar distribution to $\mathcal{S}_1$.

MIL Methods Multiple Instance Learning (MIL) aims to train a model with samples annotated by a bag-level label. Comparing to fully-supervised learning where every instance is labeled with its category, a bag of instances is annotated with just one category in MIL, which means that at least one of the instance in the bag belongs the labeled category. A bag of faces are fed to a network and their features are aggregated in the last but one layer with MIL pooling. Existing method for MIL using neural network can be formulated as different kinds of MIL pooling [42]. We try 3 kinds of MIL pooling in this paper, namely average pooling, max pooling, and log-exp-sum pooling.

In addition to methods designed specifically for multiple instance learning, we also compare with an intuitive baseline. For an image with K labels, each instance on the image is assigned a soft label over the K classes, with the groundtruth probability on each instance setting to $\frac{1}{K}$. It is similar to mean-pooling but do not require instances on an image appear in a batch during training.

As shown in Tab. 3a, max-pooling achieves the best results. As the training proceeds, the features are more discriminative and thus the max-pooling may select the most prominent feature for supervision. As for mean-pooling, it eliminates the variance between different instances on an image, weakening the discriminative powers of features. Compared with SSL methods, the inferior performance of MIL-based approach indicates the importance of correctly predicting the unknown labels, especially in a fine-grained feature learning scenario like face recognition.

	method	LFW	IJBA	MegaFace
SSL	$\mathcal{S}_1$	99.05	68.67	79.81
	K-means	99.01	67.57	79.9
	LTC	99.07	71.34	80.55
MIL	$\mathcal{S}_1 + \mathcal{S}_4$	97.83	45.31	70.2
	mean-pooling	97.85	45.22	69.62
	max-pooling	98.32	49.06	73.11
	LES-pooling	98.17	48.25	72.06
CSFR	ours	99.10	77.50	83.34

(a)

Data/Method	LFW	IJBA	MegaFace
$\mathcal{S}_1$	99.05	68.67	79.81
$\mathcal{S}_1 + \mathcal{S}_2$	99.07	75.06	82.41
$\mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$	99.10	77.50	83.34
$\mathcal{S}_2 + \mathcal{S}_3$	98.57	69.99	69.52

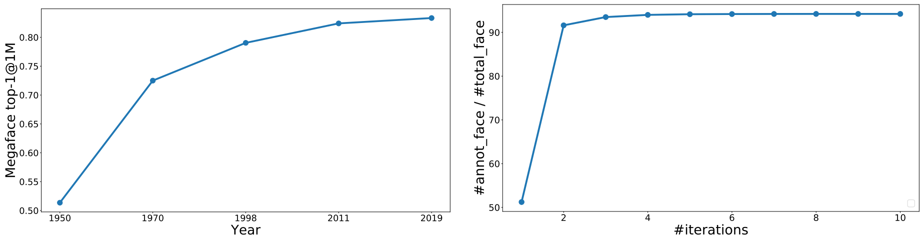
(b)

Table 3: (a). Comparison on the performance between our framework and some poplar methods in SSL and MIL in MovieFace. $\mathcal{S}_1$ represents data of One2One. $\mathcal{S}_4$ represents data with inexact labels. (b). Ablation of Different Stages in our Framework. $\mathcal{S}_1/\mathcal{S}_2/\mathcal{S}_3$ repersents training data of One2One/Propogation/Expansion, respectively.

5.4 Ablation Study and Discussion

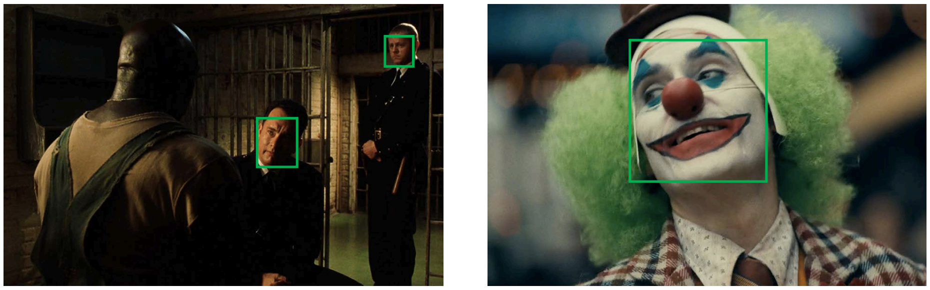
The quality of data propogation and expansion As illustrated in Tab. 3b, by adding propagation data where more faces for existing identities are labeled (*i.e.* $\mathcal{S}_1 + \mathcal{S}_2$), the model brings a performance gain from 79.81 to 82.41 in MegaFace; by further considering expansion data where faces of new identities are tagged (*i.e.* $\mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$), the model receives further performance gain. As shown in Fig. 7b, the annotated data increases from 51% to 90% with one round of label propagation and label expansion. To evaluate the performance of annotated data, we only use the annotated data to train a face recognition model. As Tab. 3b illustrates, the annotated data itself (*i.e.* $\mathcal{S}_2 + \mathcal{S}_3$) achieves comparable result as $\mathcal{S}_1$.

Relation between face recognition model and the year. We investigate the relation between the performance of face recognition model and the year. The



(a) (b)

Fig. 7: (a). MegaFace top-1 Identification@1M. vs. year. As the increase of photos with captions, the performance face recognition has been remarkably boosted. (b). Ratio of annotated face images vs. iterations. After the second round of iteration, around 90% of face has been assigned a label, indicating the effectiveness of our label expansion algorithm.



Tom Hanks and Michael Clarke Duncan in The Green Mile (1999)

Joaquin Phoenix in Joker (2019)

Fig. 8: Some noisy cases in caption-supervised face recognition.

key variant is the number of collected images. As shown in Figure. 1, this data source continuously growing every year, the performance of face recognition benefits from the increase of the images. With the proposed method, we effectively leverage the photos with captions and greatly boost the performance of the face recognition model. Figure. 1 illustrates the data source showing an exponential growth in recent years, indicating the potential improvement space of the proposed method.

Noisy cases Since the caption supervision is not specially designed for face recognition, it may sometimes introduce noise. Some noisy cases of MovieFace are shown in Fig. 8. 1) Usually, a website user would only mention the persons that he pays attention to. For example, the policeman in the background is ignored in this photo. What's worse, the prisoner is annotated even though his face is invisible. Therefore, it is easy for our model to wrongly associate the prisoner's name with the policeman in the background. 2) Since the user writing the caption with a strong context, they can correctly annotate some extremely hard cases, *e.g.* a face with heavy makeup. However, forcing the model to learn from such noisy cases may impair the performance.

More applications of MovieFace The collected MovieFace derives a new research problem, namely, caption supervised face recognition. As a dataset of rich annotations, the MovieFace can also facilitate the research in other areas. As shown in sec. 5.3, both MIL-based methods and SSL-based methods are far from satisfactory. Existing methods for MIL-based methods and SSL-based methods usually rely on some specific assumptions, the MovieFace poses a challenge for applying these methods in a more practical setting. Besides, with the time stamp of each photo, it provides a good source for age-invariant face recognition. The rapid growth of such data also provides a good source for continuous learning.

6 Conclusion

In this paper, we address a meaningful research topic named caption-supervised face recognition. It aims to train a face recognizer with the millions of web images with captions, which are free and continuously growing. We build a large-scale dataset named MovieFace, containing more than $6.3M$ faces from $305K$ identities, to support this research topic. With the proposed dataset, we demonstrate that we can train a state-of-the-art face model without any manual annotation by a simple approach, which shows the immeasurable potential of this topic. Also, extensive experiments and analyses are executed to promote further researches on caption-supervised face recognition.

Acknowledgment This work is partially supported by the SenseTime Collaborative Grant on Large-scale Multi-modality Analysis (CUHK Agreement No. TS1610626 & No. TS1712093), the General Research Fund (GRF) of Hong Kong (No. 14203518 & No. 14205719), and Innovation and Technology Support Program (ITSP) Tier 2, ITS/431/18F.

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence (2013) 5
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research (2003) 3
3. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. CoRR (2011) 4
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015) 7
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018) 1, 2, 7

6. Chen, B., Deng, W.: Weakly-supervised deep self-learning for face recognition. In: IEEE International Conference on Multimedia and Expo, ICME (2016) 4
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 9
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 7, 10
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence (1997) 5
10. Gallo, I., Nawaz, S., Calefati, A., Piccoli, G.: A pipeline to improve face recognition datasets and applications. In: International Conference on Image and Vision Computing New Zealand, IVCNZ (2018) 4
11. Gao, Y., Ma, J., Yuille, A.L.: Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. IEEE Transactions on Image Processing (2017) 3
12. Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D.: Curriculumnet: Weakly supervised learning from large-scale web images. Lecture Notes in Computer Science (2018) 4
13. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016) 1, 2, 3, 7, 9
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3, 10
15. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008) 3, 7, 9
16. Huang, H., Zhang, Y., Huang, Q., Guo, Z., Liu, Z., Lin, D.: Placepedia: Comprehensive place understanding with multi-faceted annotations. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7
17. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 7
18. Huang, Q., Xiong, Y., Lin, D.: Unifying identification and context learning for person recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 7
19. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7
20. Huang, Q., Xiong, Y., Xiong, Y., Zhang, Y., Lin, D.: From trailers to storylines: An efficient way to learn from movies. arXiv preprint arXiv:1806.05341 (2018) 7
21. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (2018) 5
22. Jiang, L., Zhou, Z., Leung, T., Li, L., Fei-Fei, L.: Mentornet: Regularizing very deep neural networks on corrupted labels. CoRR (2017) 4
23. Jin, C., Jin, R., Chen, K., Dou, Y.: A community detection approach to cleaning extremely large face database. Comp. Int. and Neurosc. (2018) 4
24. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1, 2, 7, 9, 10

25. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 9, 10
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012) 7
27. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory (1982) 11
28. Loy, C.C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., Zhou, D., Xia, W., Li, Q., Luo, P., et al.: Wider face and pedestrian challenge 2018: Methods and results. arXiv preprint arXiv:1902.06854 (2019) 7
29. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: ICIP (2014) 9
30. Ng, H., Winkler, S.: A data-driven approach to cleaning large face datasets. In: IEEE International Conference on Image Processing, ICIP (2014) 4
31. Patrini, G., Rozza, A., Menon, A.K., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2017) 4
32. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3, 5
33. Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7
34. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 7
35. Roli, F., Marcialis, G.L.: Semi-supervised pca-based face recognition using self-training. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR) (2006) 3
36. Rolnick, D., Veit, A., Belongie, S.J., Shavit, N.: Deep learning is robust to massive label noise. CoRR (2017) 4
37. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 7
38. Sukhbaatar, S., Fergus, R.: Learning from noisy labels with deep neural networks. In: International Conference on Learning Representations (ICLR) Workshop (2015) 4
39. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2014) 7
40. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Loy, C.C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 2, 4, 7, 9
41. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 1

42. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition (2018) 3, 5, 11
43. Wang, Z., Zheng, L., Li, Y., Wang, S.: Linkage based face clustering via graph convolution network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 11
44. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016) 1
45. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3, 5
46. Xia, J., Rao, A., Xu, L., Huang, Q., Wen, J., Lin, D.: Online multi-modal person search in videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7
47. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2015) 4
48. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: The IEEE International Conference on Computer Vision (ICCV) (2019) 7
49. Yang, L., Chen, D., Zhan, X., Zhao, R., Loy, C.C., Lin, D.: Learning to cluster faces via confidence and connectivity estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 1
50. Yang, L., Huang, Q., Huang, H., Xu, L., Lin, D.: Learn to propagate reliably on noisy affinity graphs. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3
51. Yang, L., Zhan, X., Chen, D., Yan, J., Loy, C.C., Lin, D.: Learning to cluster faces on an affinity graph. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 11
52. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014) 7, 9
53. Zhan, X., Liu, Z., Yan, J., Lin, D., Change Loy, C.: Consensus-driven propagation in massive unlabeled data for face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 1, 3, 4, 11
54. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters (2016) 8
55. Zhao, X., Evans, N., Dugelay, J.L.: Semi-supervised face recognition with lda self-training. In: IEEE International Conference on Image Processing (2011) 3