Face Representation Learning using Composite Mini-Batches

Evgeny Smirnov
Speech Technology Center
smirnov-e@speechpro.com

Andrei Oleinik
Speech Technology Center
oleynik@speechpro.com

Aleksandr Lavrentev
Speech Technology Center
lavrentyev@speechpro.com

Elizaveta Shulga
Speech Technology Center
shulga-e@speechpro.com

Vasiliy Galyuk
Speech Technology Center
galyuk@speechpro.com

Nikita Garaev
ITMO University
garaev@speechpro.com

Margarita Zakuanova
Speech Technology Center
zakuanova@speechpro.com

Aleksandr Melnikov
ITMO University
melnikov-a@speechpro.com

Abstract

Mini-batch construction strategy is an important part of the deep representation learning. Different strategies have their advantages and limitations. Usually only one of them is selected to create mini-batches for training. However, in many cases their combination can be more efficient than using only one of them. In this paper, we propose Composite Mini-Batches - a technique to combine several mini-batch sampling strategies in one training process. The main idea is to compose mini-batches from several parts, and use different sampling strategy for each part. With this kind of mini-batch construction, we combine the advantages and reduce the limitations of the individual mini-batch sampling strategies. We also propose Interpolated Embeddings and Priority Class Sampling as complementary methods to improve the training of face representations. Our experiments on a challenging task of disguised face recognition confirm the advantages of the proposed methods.

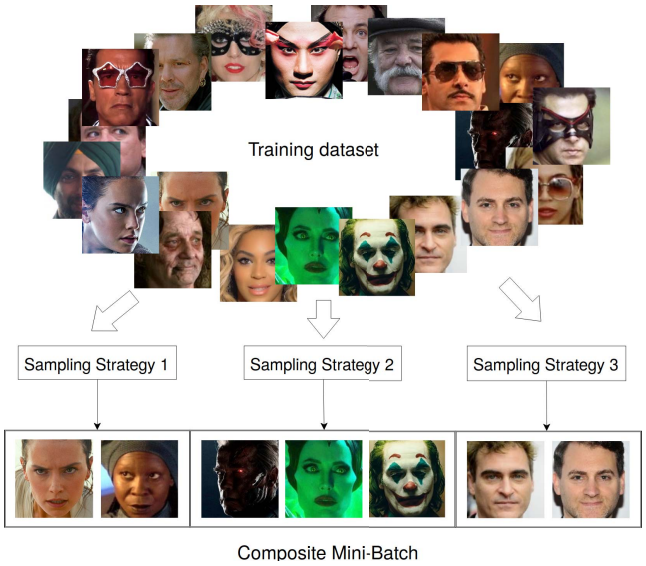

Figure 1. Example of Composite Mini-Batch with three sampling strategies

1. Introduction

Training discriminative representations (embeddings) with deep neural networks is an established method in different areas like face [48, 37, 41, 42] and speaker [30, 32, 31, 29] recognition, image retrieval [56, 28, 50], person [57, 10] and vehicle [36, 3] re-identification, landmark recognition [7], fine-grained recognition [16] and few-shot learning [44, 34, 60].

Training of such representations is usually performed on large datasets in a mini-batch regime: at each training iteration a small portion of data (mini-batch) is used to calculate the gradients and adjust the weights of the neural network. Mini-batches can be constructed in many different ways: items could be sequentially selected from the whole dataset, or sampled using some kind of importance measure [6, 20], or selected based on their classes [45, 41]. Some kind of hard example mining [22, 25, 46, 41, 42] can also be used. Each of these sampling strategies has its advantages and limitations.

Usually only one sampling strategy is used for the mini-batch construction. In this paper, we present the way to simultaneously utilize several strategies using Composite Mini-Batches (Figure 1). The main idea is to use different sampling strategies for different parts of the mini-

batch. With this kind of mini-batch construction, the advantages of several sampling strategies could be combined, and their limitations could be diminished. We evaluate the proposed Composite Mini-Batches on the task of disguised face recognition and confirm the advantages of this mini-batch construction method.

We also propose Interpolated Embeddings and Priority Class Sampling as a complimentary ways to improve the training of deep representations with Composite Mini-Batches.

2. Related work

2.1. Face representation learning

Learning face representations with deep neural networks is an established method to perform face recognition [35]. Training of these representations (embeddings) requires large and rich face datasets [17, 4, 8], appropriate neural network architectures [18, 58, 49, 9], loss functions [53, 12, 14] and mini-batch sampling strategies [45, 41, 42, 25].

2.2. Loss functions

Face representations supposed to be discriminative [55]. This is usually achieved with specialized loss functions, encouraging intra-class compactness and inter-class separability.

One type of these functions is based on training a classifier with a version of normalized Softmax function [52], employing some kind of margin [27, 26, 53, 12, 25, 59, 54]. Training with this type of loss functions encourages exemplar embeddings of each class to be close in the embedding space to the prototypes of their corresponding classes and far from the prototypes of other classes. These two objectives are usually attempted to be achieved jointly, but they could be "dissected" and accomplished separately [19]. For face datasets with large number of identities Softmax-based loss functions could become too computationally demanding (very large number of classes in the classifier), but there are ways to reduce computation in this case [62, 19].

Another way to train discriminative representations is to use exemplar-based loss functions, which operate on distances between examples in the embedding space. These functions use pairs [11, 41], triplets [37] or other groups of examples [33], where distances between examples of the same class are stimulated to be small, and distances between examples of different classes are stimulated to be large, usually larger than some margin [37, 56]. Since examples are used in pairs, and the training is performed in mini-batches, these functions heavily rely on hard example mining and smart mini-batch construction methods [37, 22, 41, 42], which ensure that there is enough useful pairs (or triplets) of examples in every mini-batch.

2.3. Mini-batch construction

There are several main ways to constuct mini-batches for training. The simplest way is to iterate over the whole dataset and use all the images sequentially [55, 12]. When the end of the dataset is reached (training for one epoch is performed), images in the dataset are shuffled and iterated again. The process is repeated until convergence. We refer this mini-batch sampling strategy as "*Iterate-Shuffle*" (see Figure 3 (a)). This strategy achieves good results when the dataset is balanced and no exemplar-based loss is used. With this strategy each image in the dataset has the same probability to be sampled in the mini-batch, and also it is guaranteed to be used for training if at least one epoch is performed. However, for the cases with severely unbalanced datasets (which is very common in face recognition, see Figure 2 for the example), with this kind of strategy the major part of the training process will be spent on the few classes with large number of images. This kind of training can be helpful to teach the network to be robust to the pose and appearance variations of faces (because there are many different faces of the same person used for this type of training), but it will more likely fail to distinguish between faces of similarly-looking people in large datasets (because there are likely not enough similarly-looking people in the few image-rich classes). Also with this kind of training using exemplar-based loss functions becomes less efficient, because there is no hard example mining involved, and so there are very small possibility of finding useful pairs of examples in the mini-batches.

The second type of mini-batch sampling strategy can be referred as "*Classes-Then-Images*" (see Figure 3 (b)). It consists of two phases. At the first phase, some classes are selected to be included into the current mini-batch. This can be done randomly or using some kind of hard class mining method [45, 41]. At the second phase, for each selected class several images are chosen to be included into the mini-batch. It also can be done randomly or with some kind of hard example mining method [42]. With this kind of sampling we can ensure that each class will appear in the training approximately same number of times. Also it is well suited for exemplar-based loss functions because of hard class and example mining. This strategy also works best when the dataset is balanced and has difficulties in case of unbalanced datasets. However, these difficulties are of the different kind than for the first strategy: if the classes are sampled randomly, then the images from image-rich classes are less likely to be sampled into the mini-batch than images of other, smaller classes. For some cases there is even a possibility, that some images may never be used in training at all. For the "long-tail" datasets with large number of classes (like in Figure 2) this sampling strategy (especially with hard class mining methods like Doppelganger Mining [41]) can be well-trained to distinguish between similarly-

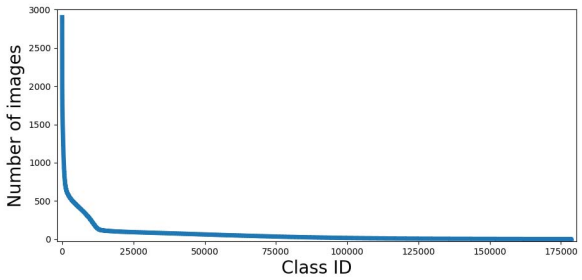

Figure 2. Training dataset with long-tail distribution

looking faces of different people, but fail in cases of large appearance and pose variations of the faces of same person (because it will be trained less on the intra-person face image variations of image-rich classes).

There are other possible ways of mini-batch construction like using of importance sampling [20], stratified sampling [24], unequal training [63] etc. They all have their advantages and limitations.

3. Proposed methods

3.1. Composite Mini-Batches

As a way to combine the advantages of several different sampling strategies, we propose to use "*Composite Mini-Batches*" (see Figure 3 (c)). This mini-batch construction strategy presumes the composition of a mini-batch from the several parts (sub-mini-batches), which can be of same or different size, and using different sampling strategies for each part. With this kind of mini-batch construction, it can benefit from each sampling strategy simultaneously. For example, by using "*Iterate-Shuffle*" and "*Classes-Then-Images*" strategies for two different parts of a Composite Mini-Batch, we ensure, that each training image from the dataset will be sampled into the mini-batch regularly (because of the first strategy), and that each class will regularly appear in the mini-batch (because of the second strategy). Also there will be enough hard example pairs for exemplar-based loss (if we include hard example mining method in the second strategy). This way we combine the advantages and reduce the limitations of the individual sampling strategies mentioned above.

One possible downside of the proposed mini-batch construction method is that for some cases not advantages, but disadvantages of several sampling strategies could be combined together. To prevent this, one should select sampling strategies carefully, considering their collaborative abilities.

3.2. Interpolated embeddings

In real-world datasets with the long-tail distribution there is a problem with image-poor classes: there are not enough training examples to fill the "holes" in the embedding space (see Figure 4 (a)). Gradients, computed in these cases, could point in wrong directions, thus reducing the efficiency of the exemplar-based training methods.

To alleviate these problems, we propose Interpolated Embeddings (see Figure 4 (b)). These embeddings are calculated at the face representation layer of the neural network. They are computed using an interpolation of embeddings, which are presented for the selected class in current mini-batch.

To create one Interpolated Embedding and add it to the current mini-batch, we first randomly select a class, which has at least 2 items in the current mini-batch. Then we randomly select a subset of current mini-batch items, belonging to this class. For each of them, we randomly generate a weight in range $(0, 1)$. At the face representation layer, we take an embedding for each selected item, multiply it by generated weight, sum them together and L_2-normalize the resulting embedding vector and add it to the current mini-batch.

Exemplar-based loss function uses Interpolated Embeddings to select better hard example pairs (positive and negative) and fill the "holes" in the embedding space.

3.3. Priority Class Sampling

For some specialized problems like disguised face recognition, the amount of available training data is too small to train from scratch. Using this training data only for fine-tuning could make the network to overfit on it and later fail on more general problems. We can add this specialized training data to the full training dataset and train the network on it to ensure that it will work on both tasks. However, if the ratio of specialized data is too small compared to the full dataset, it will be sampled too infrequently. To alleviate this problem, we propose Priority Class Sampling strategy. The main idea is to create the priority class list (which includes classes and images from specialized training data), and use it as a training dataset for one of the sampling strategies of the Composite Mini-Batch. This way, a small number of specialized training data is sampled in each mini-batch, ensuring that the network is training for the specialized problem, while other parts of mini-batch keeping it suitable for more general problems.

4. Experiments

In this section, we evaluate proposed Composite Mini-Batch sampling strategy, Interpolated Embeddings and Priority Class Sampling on the challenging task of disguised face recognition (DFW2018 [23] and DFW2019 [39]).

4.1. Disguised face recognition

Modern face recognition algorithms demonstrate high accuracy in controlled conditions as well as in the wild

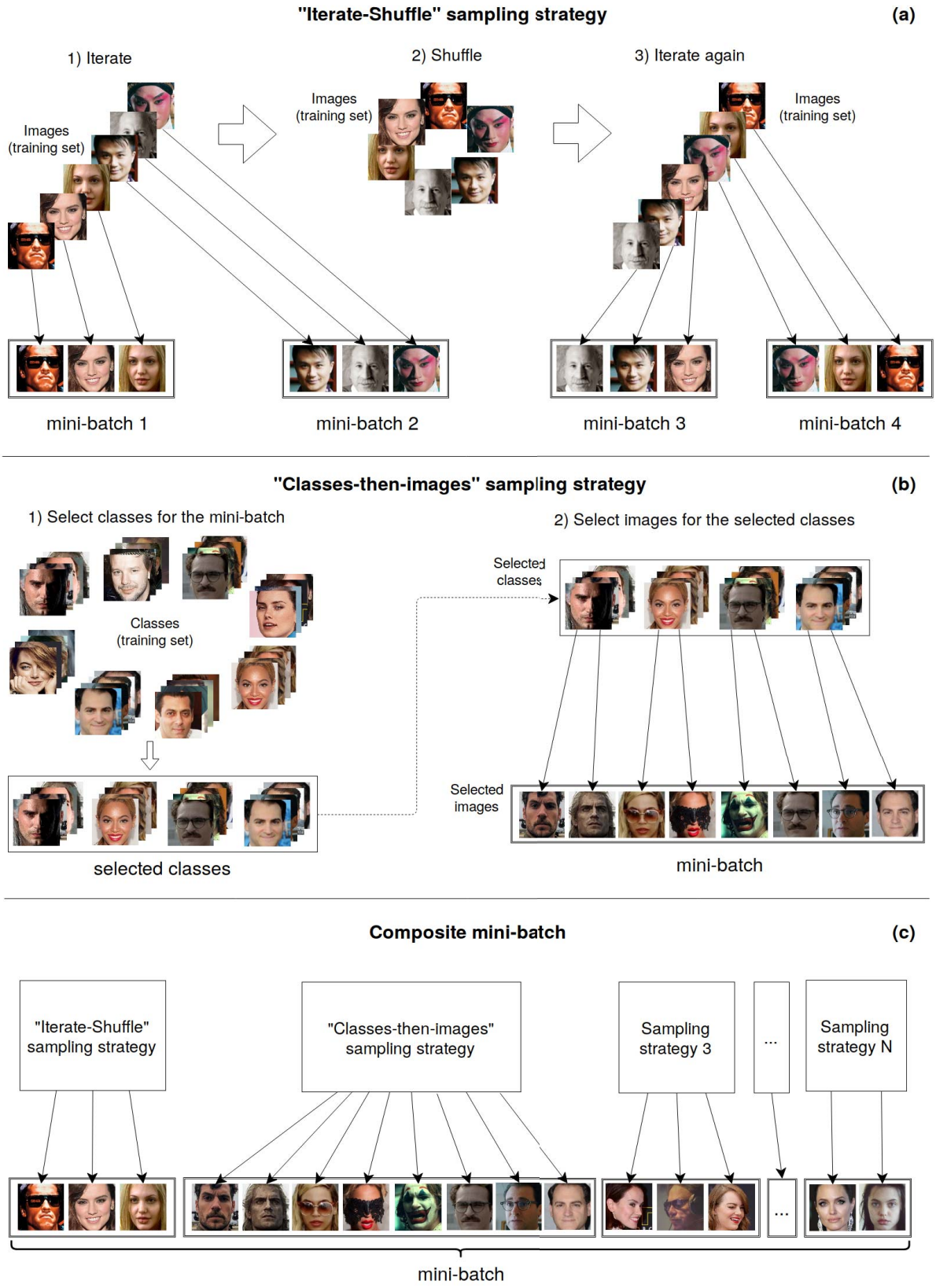

Figure 3. a) ”Iterate-Shuffle” sampling strategy. b) ”Classes-then-images” sampling strategy. c) Composite Mini-Batch.

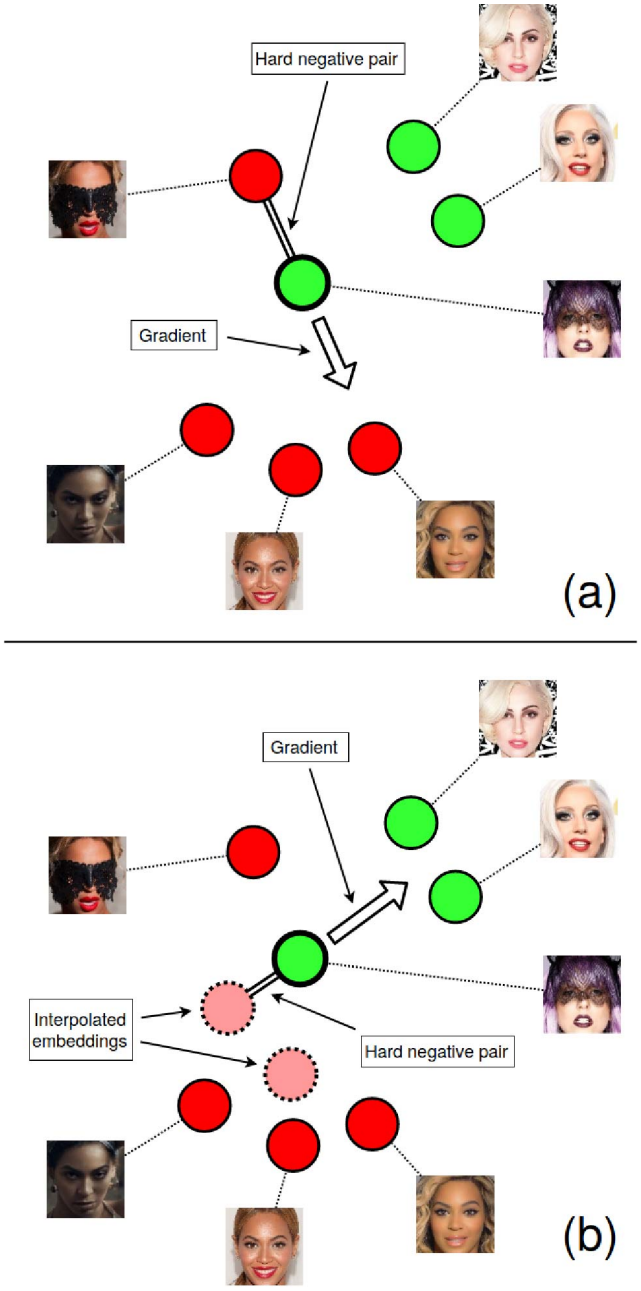

Figure 4. Examples of hard negative pairs a) without Interpolated Embeddings, b) with Interpolated Embeddings

scenario. One of the major challenges nowadays is significant appearance variations, which include heavy make-up, masks, sunglasses, beard (fake or natural), etc. In the worst case, these changes are made intentionally to hide ones identity or imitate the appearance of another person [15, 38, 47]. This issue was addressed in the Disguised Faces in the Wild (DFW) competition, which was held in 2018 [23, 40] and in 2019 [39]. In order to deal with this problem, several solutions have been proposed

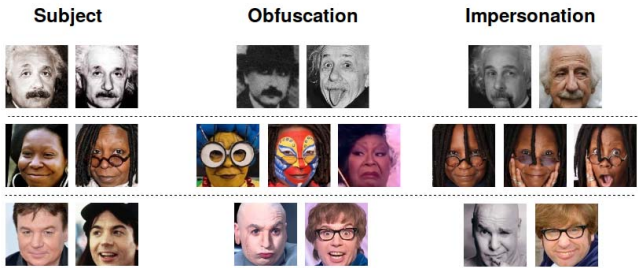

Figure 5. Examples of the DFW faces. Subject images are at the left, obfuscated images in the middle and impersonator images at the right

[5, 21, 51, 61, 42].

For each person, the DFW dataset contains images of the following types:

- "*Subject*": a normal face image of a person.
- "*Obfuscation*" (also referred as "*disguised*"): a disguised face image of the same person. The disguise may be achieved by means of beard, sunglasses, make-up, etc.
- "*Impersonator*" is a face image of another person that bears a significant resemblance to the subject image.

Figure 5 shows the examples of DFW face images. In our experiments, $3,386$ images of 400 people from the DFW2018 dataset were included into the training set, $7,771$ test images from the DFW2018 dataset were used to perform an independent assessment of the performance of the proposed sampling strategies, and $3,840$ images from the new DFW2019 test set were used to perform comparison with existing state-of-the-art algorithms of disguised face recognition. Compared to DFW2018, new DFW2019 test set also contains faces, modified by plastic surgery.

4.2. Implementation details

4.2.1 Preprocessing

We used RetinaFace detector [13] for face detection. All faces were aligned and cropped to the size of 112x112. Pixel values were normalized to $[-1, 1]$. We used random horizontal flipping for the training phase and mirror trick [41] at the testing phase.

4.2.2 Neural network architecture and training

In the experiments, we used SE-LResNet50E-IR [12] neural network architecture, which takes images of size 112x112 and outputs L_2-normalized embedding of size 512. We used SV-Arc-Softmax [54] with $m = 0.5$, $s = 64$ and $t = 1.2$. We also added COPRA margin [59] $m_2 = 0.01$. In the experiments containing Doppelganger Mining [41] we also

used Combined Margin Pairwise Loss [2] with $m_1 = 1.0$, $m_2 = 0.5$, $m_3 = 0.0$ and loss weight of 2.0.

For training we used mini-batches of size 300 and Stochastic Gradient Descent with momentum of 0.9 and weight decay of 0.0005. OneCycle learning rate policy [43] with base value of 0.01, maximum value of 0.1 and minimum value of 0.0001 was used for training: first $150,000$ iterations learning rate linearly increased from the base value to the maximum value, then for another $150,000$ iterations learning rate linearly decreased back to base value, and then for $300,000$ more iterations it was linearly decreased to minimum value.

For DFW2019 submission, we trained an ensemble of SE-LResNet101E-IR [12] networks.

4.2.3 Training dataset

Training is performed on the combined dataset, which consists of public face datasets: MS-Celeb-1M [17], VGGFace2 [8] and TrillionPairs-Asians [1]. We also included training subset of the DFW2018 dataset [23].

To combine datasets properly (as they can contain same people, which should be merged in the same classes), we used the following strategy: using a pre-trained network, we retrieved L_2-normalized face embeddings for every image in all datasets. Then, we calculated average embedding for each person and used pairwise cosine distances to find similar pairs of people. After that, we:

- merged people with cosine distances less than 0.3,
- treated pairs of people with cosine distances larger than 0.5 as different people,
- dropped people with smaller number of images in pairs, where cosine distance is between 0.3 and 0.5.

Cosine distance values 0.3 and 0.5 were chosen using a manually examined subset of training data pairs. The resulting dataset contains $178,688$ people and $11,121,926$ images, ranging from 1 to $2,901$ images per person. This dataset has severe long-tail distribution as shown on Figure 2.

4.2.4 Sampling strategies

We have trained seven neural networks with these sampling strategies:

- "*I-S*": This network is trained with Iterate-Shuffle (I-S) strategy, mini-batch size is 300.
- "*DM-AE*": This network is trained with Doppelganger Mining with Auxiliary Embeddings (DM-AE) strategy. Mini-batch of size 300 is sampled using a variant of "Class-Then-Images" sampling strategy, with Doppelganger Mining [41] used for class selection (two classes are selected randomly, others - using Doppelganger List) and Hard Example Mining with Auxiliary Embeddings [42] used to select images for each class (we used 4 as a number of images per class in the mini-batch, 0.3 as a probability of hard positive and hard negative pairs, and 100 as a maximum number of candidates). Auxiliary embeddings for the training dataset are calculated with the final layer of the neural network, after $300,000$ iterations of training. Before that all images for selected classes are chosen randomly.
- "*I-S + DM-AE (2:1)*": This network is trained with Composite Mini-Batch, where 200 images are sampled with I-S strategy and 100 images are sampled with DM-AE strategy.
- "*I-S + DM-AE (1:2)*": This network is trained with Composite Mini-Batch, where 100 images are sampled with I-S strategy and 200 images are sampled with DM-AE strategy.
- "*I-S + DM-AE (1:2), with IE*": This network is trained with Composite Mini-Batch, where 100 images are sampled with I-S strategy, 200 images are sampled with DM-AE strategy, and 150 Interpolated Embeddings were added to the mini-batch at the embedding layer.
- "*I-S + DM-AE (1:2), with PCS*": This network is trained with Composite Mini-Batch, where 100 images are sampled with I-S strategy, 192 images are sampled with DM-AE strategy and 8 more images (2 persons, 4 images per person) are sampled with PCS strategy. For the Priority Class List, we used a list of classes, which are added to the combined dataset from the DFW2019 training set. These classes are known to have a large number of disguise variations, obfuscated faces, impersonators and so on.
- "*I-S + DM-AE (1:2), with IE and PCS*": This network is trained with Composite Mini-Batch, where 100 images are sampled with I-S strategy, 192 images are sampled with DM-AE strategy, 8 more images (2 persons, 4 images per person) are sampled with PCS strategy, and 150 Interpolated Embeddings were added to the mini-batch at the embedding layer.

4.3. Results

The results on the DFW2018 test set are presented in Table 1 and Figure 6. For the DFW2019 challenge submission we trained an ensemble of three larger networks.

Method	GAR@1%	GAR@0.1%
I-S	91.13%	86.95%
DM-AE	90.82%	86.38%
I-S + DM-AE (2:1)	91.68%	87.66%
I-S + DM-AE (1:2)	91.55%	87.61%
I-S + DM-AE (1:2), IE	91.78%	87.83%
I-S + DM-AE (1:2), PCS	91.80%	87.86%
I-S + DM-AE (1:2), IE, PCS	**91.95%**	**87.96%**
DFW2019 submission	*91.92%*	*88.45%*
Ensemble	***92.40%***	***88.80%***

Table 1. Results on the Disguised Faces in the Wild 2018 test set. Evaluation metric is genuine accept rate (GAR) at false accept rates (FAR) of 1% and 0.1%

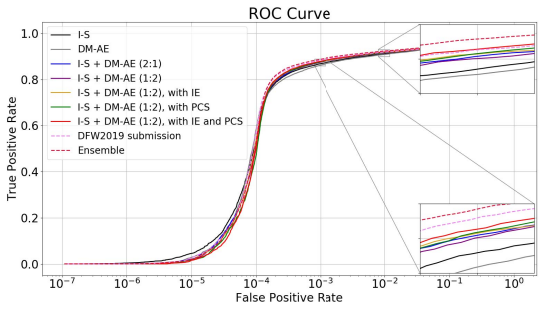

Figure 6. Results on the Disguised Faces in the Wild 2018 test set

Method	GAR@0.1%	GAR@0.01%
Our submission	**93.2%**	**52.0%**
ResNet-50 [39]	47.60%	38.40%
LightCNN-29v2 [39]	74.40%	51.20%

Table 2. Results on the Protocol-1 (Impersonation) of DFW2019 test set. Evaluation metric is genuine accept rate (GAR) at false accept rates (FAR) of 0.1% and 0.01%

We tested this ensemble on DFW2018 ("*DFW2019 submission*" in Table 1) and DFW2019 ("*Our submission*" in Tables 2-5) test sets. We also report the results, achieved after the challenge deadline by the ensemble of multiple networks, including networks from DFW2019 submission and from experiments with sampling strategies ("*Ensemble*" in Table 1).

The results on DFW2019 test set for Protocol-1 (Impersonation), Protocol-2 (Obfuscation), Protocol-3 (Plastic Surgery) and Protocol-4 (Overall) are presented in Tables 2, 3, 4, 5. Baseline results are taken from [39].

Composite Mini-Batch achieved improvement over the individual sampling strategies. Interpolated Embeddings and Priority Class Sampling improved results even more. The best single network result on the DFW2018 test set

Method	GAR@0.1%	GAR@0.01%
Our submission	**92.3%**	**87.7%**
ResNet-50 [39]	35.38%	16.42%
LightCNN-29v2 [39]	55.56%	36.90%

Table 3. Results on the Protocol-2 (Obfuscation) of DFW2019 test set. Evaluation metric is genuine accept rate (GAR) at false accept rates (FAR) of 0.1% and 0.01%

Method	GAR@0.1%	GAR@0.01%
Our submission	**95.6%**	**92.0%**
ResNet-50 [39]	46.40%	22.40%
LightCNN-29v2 [39]	69.20%	47.20%

Table 4. Results on the Protocol-3 (Plastic Surgery) of DFW2019 test set. Evaluation metric is genuine accept rate (GAR) at false accept rates (FAR) of 0.1% and 0.01%

Method	GAR@0.1%	GAR@0.01%
Our submission	**92.1%**	**83.1%**
ResNet-50 [39]	35.96%	16.89%
LightCNN-29v2 [39]	55.74%	36.50%

Table 5. Results on the Protocol-4 (Overall) of DFW2019 test set. Evaluation metric is genuine accept rate (GAR) at false accept rates (FAR) of 0.1% and 0.01%

(91.95% GAR@1%FAR and 87.96% GAR@0.1%FAR) was achieved by a network using Composite Mini-Batch with a combination of all proposed sampling strategies.

Our DFW2019 challenge submission achieved the results of 91.92% GAR@1%FAR and 88.45% GAR@0.1%FAR on the DFW2018 test set and performed better than baseline on the DFW2019 test set for all four protocols. We also report the results of multi-network ensemble, which achieved 92.40% GAR@1%FAR and 88.80% GAR@0.1%FAR on the DFW2018 test set.

5. Conclusions

In this paper, we have presented a novel method of Composite Mini-Batch construction to improve the training of deep neural networks. The main idea of this method is to use different sampling strategies for different parts of the mini-batch. With this kind of mini-batch sampling, the advantages of several sampling strategies are utilized simultaneously.

Experimental results on the challenging task of disguised face recognition confirmed the advantages of Composite Mini-Batches over individual sampling strategies.

We also presented Interpolated Embeddings and Priority Class Sampling as a complementary ways to improve the training of face representations. We used them to get improvements on DFW2018 and DFW2019 datasets.

Acknowledgement

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08).

References

[1] http://trillionpairs.deepglint.com. 6

[2] Authors. Working memory prototype loss for deep representation learning. *Manuscript in preparation*. 6

[3] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan. Group sensitive triplet embedding for vehicle re-identification. *IEEE Transactions on Multimedia*, 2018. 1

[4] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *IJCB*, 2017. 2

[5] A. Bansal, R. Ranjan, C. D. Castillo, and R. Chellappa. Deep features for recognizing disguised faces in the wild. In *CVPR Workshops*, 2018. 5

[6] Y. Bengio, J.-S. Senécal, et al. Quick training of probabilistic neural nets by importance sampling. In *AISTATS*, 2003. 1

[7] A. Boiarov and E. Tyantov. Large scale landmark recognition via deep metric learning. In *CIKM*, 2019. 1

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 2, 6

[9] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, 2018. 2

[10] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 1

[11] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 2

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 5, 6

[13] J. Deng, J. Guo, Z. Yuxiang, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*, 2019. 5

[14] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *CVPR Workshops*, 2017. 2

[15] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PloS one*, 9(7):e99212, 2014. 5

[16] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan. Incorporating intra-class variance to fine-grained visual recognition. In *ICME*, 2017. 1

[17] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 2, 6

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[19] L. He, Z. Wang, Y. Li, and S. Wang. Softmax dissection: Towards understanding intra-and inter-clas objective for embedding learning. *arXiv preprint arXiv:1908.01281*, 2019. 2

[20] A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *ICML*, 2018. 1, 3

[21] N. Kohli, D. Yadav, and A. Noore. Face verification with disguise variations via deep disguise recognizer. In *CVPR Workshops*, 2018. 5

[22] V. B. Kumar, B. Harwood, G. Carneiro, I. Reid, and T. Drummond. Smart mining for deep metric learning. In *ICCV*, 2017. 1, 2

[23] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *CVPR Workshops*, 2018. 3, 5, 6

[24] E. Liberty, K. Lang, and K. Shmakov. Stratified sampling meets machine learning. In *ICML*, 2016. 3

[25] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *CVPR*, 2019. 1, 2

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 2

[27] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016. 2

[28] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. No fuss distance metric learning using proxies. 2017. 1

[29] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov. Stc speaker recognition systems for the voices from a distance challenge. *arXiv preprint arXiv:1904.06093*, 2019. 1

[30] S. Novoselov, O. Kudashev, V. Schemelinin, I. Kremnev, and G. Lavrentyeva. Deep cnn based feature extractor for text-prompted speaker recognition. In *ICASSP*, 2018. 1

[31] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, and I. Kremnev. Triplet loss based cosine similarity metric learning for text-independent speaker recognition. In *Interspeech*, 2018. 1

[32] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin. On deep speaker embeddings for text-independent speaker recognition. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018. 1

[33] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 2

[34] B. Oreshkin, P. R. López, and A. Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NIPS*, 2018. 1

[35] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, 2015. 2

[36] A. Sanakoyeu, V. Tschernezki, U. Buchler, and B. Ommer. Divide and conquer the embedding space for metric learning. In *CVPR*, 2019. 1

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 2

[38] A. Singh, D. Patil, G. M. Reddy, and S. Omkar. Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network. In *ICCV Workshops*, 2017. 5

[39] M. Singh, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa. Disguised faces in the wild 2019. *Technical Report, IIIT Delhi*, 2019. 3, 5, 7

[40] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019. 5

[41] E. Smirnov, A. Melnikov, S. Novoselov, E. Luckyanets, and G. Lavrentyeva. Doppelganger mining for face representation learning. In *ICCV Workshops*, 2017. 1, 2, 5, 6

[42] E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, and E. Luckyanets. Hard example mining with auxiliary embeddings. In *CVPR Workshops*, 2018. 1, 2, 5, 6

[43] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. 6

[44] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 1

[45] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 1, 2

[46] Y. Suh, B. Han, W. Kim, and K. M. Lee. Stochastic class-based hard example mining for deep metric learning. In *CVPR*, 2019. 1

[47] A. Suri, M. Vatsa, and R. Singh. A-link: Recognizing disguised faces via active learning based inter-domain knowledge. In *BTAS*, 2019. 5

[48] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[49] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 2

[50] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *NIPS*, 2016. 1

[51] S. Vishwanath Peri and A. Dhall. Disguisenet: A contrastive approach for disguised face verification in the wild. In *CVPR Workshops*, 2018. 5

[52] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: l_2 hypersphere embedding for face verification. In *ACM Multimedia*, 2017. 2

[53] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 2

[54] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei. Support vector guided softmax loss for face recognition. *arXiv preprint arXiv:1812.11317*, 2018. 2, 5

[55] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 2

[56] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 1, 2

[57] Q. Xiao, H. Luo, and C. Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017. 1

[58] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 2

[59] D. Xu and Q. Zhao. Contrapositive margin softmax loss for face verification. In *ICRCA*, 2018. 2, 5

[60] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019. 1

[61] K. Zhang, Y.-L. Chang, and W. Hsu. Deep disguised faces recognition. In *CVPR Workshops*, 2018. 5

[62] X. Zhang, L. Yang, J. Yan, and D. Lin. Accelerated training for massive classification via dynamic class selection. In *AAAI*, 2018. 2

[63] Y. Zhong, W. Deng, M. Wang, J. Hu, J. Peng, X. Tao, and Y. Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *CVPR*, 2019. 3