SCAF: Skip-Connections in Auto-encoder for Face Alignment with Few Annotated Data

Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, Bertrand Coüasnon

SCAF: Skip-Connections in Auto-encoder for Face alignment with few annotated data

Martin Dornier[1,2], Philippe-Henri Gosselin[1], Christian Raymond[2],
Yann Ricquebourg[2], and Bertrand Coüasnon[2]

[1] InterDigital
{martin.dornier,philippehenri.gosselin}@interdigital.com
[2] Univ Rennes, CNRS, IRISA, France
{christian.raymond,yann.ricquebourg,bertrand.couasnon}@irisa.fr

Abstract. Supervised face alignment methods need large amounts of training data to achieve good performance in terms of accuracy and generalization. However face alignment datasets rarely exceed a few thousand samples making these methods prone to overfitting on the specific training dataset. Semi-supervised methods like TS^3 or 3FabRec have emerged to alleviate this issue by using labeled and unlabeled data during the training. In this paper we propose Skip-Connections in Auto-encoder for Face alignment (SCAF), we build on 3FabRec by adding skip-connections between the encoder and the decoder. These skip-connections lead to better landmark predictions, especially on challenging examples. We also apply for the first time active learning to the face alignment task and introduce a new acquisition function, the Negative Neighborhood Magnitude, specially designed to assess the quality of heatmaps. These two proposals show their effectiveness on several face alignment datasets when training with limited data.

Keywords: Face alignment · Semi-supervised training · Active learning.

1 Introduction

Face alignment (also called facial landmark detection) aims to localize a set of pre-defined facial anatomical keypoints such as the corners of the mouth, the boundaries of the eyes or the tip of the nose [25, 28, 17]. Many applications rely on this task, for example, facial expression recognition or face swapping.

Although the rise of deep learning methods significantly improved the performances, the algorithms are still limited by the amount of labeled data available for training. Semi-supervised methods [1, 33, 23, 13, 8, 24, 12, 9] have emerged in the field of face alignment to alleviate the lack of labeled data. In this work, we follow this principle, we try to train face alignment models with as little labeled data as possible. To do so, we build on 3Fabrec [1], this semi-supervised method achieves impressing performance in face alignment even with very limited training data. However, because of its relatively simple architecture, its performance degrades significantly on challenging datasets such as WFLW [28]. We try in this work to alleviate this issue. Our contribution is twofold:

- We enhance 3FabRec architecture with skip-connections between its encoder and decoder during the supervised training. This addition significantly improves its performances on both 300-W [25] and WFLW [28] datasets.
- We successfully apply active learning to face alignment and introduce a new acquisition function, the Negative Neighborhood Magnitude, improving even further the performance of our method when training with limited data.

The rest of the paper is organized as follows. Section 2 sums up the existing work on face alignment, in particular with limited data, and introduces the active learning procedure. In Section 3, we present our proposed methods to address face alignment with limited data. The results of these methods are shown in Section 4. Finally, we conclude this paper in Section 5.

2 Related work

In our context, face alignment methods can be divided into two families: supervised and semi-supervised methods.

2.1 Supervised face alignment

Before the deep learning development in the computer vision domain, face alignment algorithms usually relied on parametric models, such as active shape model [6] or active appearance model [20], or on cascade regression [5, 30, 29]. Nowadays, almost all methods are based on artificial neural networks. Among recent approaches, while some methods still try to regress directly the landmark coordinates [10], most of them are now based on heatmap regression [22, 3, 28, 27, 18, 7]. In this latter case, the network outputs a probabilistic heatmap for each landmark and the landmark coordinates are computed from it, usually with the best local maximum. Wu et al. [28] use facial boundaries heatmaps instead of landmark heatmaps making the algorithm more robust to large poses and occlusions. To take into account occlusions Kumar et al. [18] model the uncertainty and visibility of landmarks as a mixture of random variables while Zhu et al. [32] add in the model weights based on occlusion probability.

2.2 Semi-supervised face alignment

Annotating facial landmarks is time-consuming and can be difficult on faces with large pose or occlusions. For this reason, face alignment datasets rarely exceed a few thousand annotated faces. Semi-supervised methods try to alleviate the lack of annotated training data by incorporating non annotated, or weakly annotated, data into the learning process. Zhu et al. [33] augment the training dataset with synthetics faces generated from a 3D face model. Similarly, Qian et al. [23] generate images with different styles from an input pose image. Honari et al. [13] impose the equivariance of landmark predictions over multiple transformations of a face image. To deal with the large variance of different images styles Dong et al. [8] transforms images into style-aggregated images, and Robinson et al. [24] generate fake landmark heatmaps from unlabeled images using a Generative Adversarial Network [12]. Dong et al. [9] train a teacher to assess the quality of student predicted landmarks, the best samples are added, along with real data, to the next training set for retraining the student detectors. Finally, Browatzki et al. [1] propose 3FabRec that we will detail in the next section.

2.3 3FabRec

In 3FabRec [1], first, an auto-encoder is trained on a large number of unlabeled face images. During this unsupervised training, the hidden representation of the auto-encoder learns implicit knowledge about face features (shape, skin color, gender...) in order to reconstruct the image. The massive amounts of images used for training make this representation robust to a large diversity of faces.

After the unsupervised training, the auto-encoder is modified to perform face alignment, the decoder (also called generator) weights are frozen and convolutional layers called Interleaved Transfer Layers (ITLs) are added between its layers to take advantage of its generative power to also generate landmark heatmaps. The model is then trained on labeled face alignment datasets.

This method achieves impressive results even with few labeled data.

2.4 Active learning

In the academic field, authors usually sample randomly labeled data from the full annotated training set to demonstrate the effectiveness of their method with limited training data. However, in real-world applications, at first, no labeled data is available and one must decide which samples to annotate.

Active learning aims to select the best samples to annotate to get the best possible model. It is particularly useful when annotation is time-consuming such as facial landmark annotations. It follows an iterative procedure: from an unlabeled dataset $\mathcal{U}_N$, an initial set $\mathcal{L}^0$ is annotated, the model is trained on this labeled set and all the remaining unlabeled samples are ranked using an *acquisition function*, the K best samples are annotated and added to $\mathcal{L}^0$ giving a new labeled set $\mathcal{L}^1$. The model is then trained from scratch on this new labeled set and this procedure is repeated until the annotation budget has been exhausted.

The acquisition functions can be divided into two approaches even though some combine both [16]. The first one is based on "uncertainty sampling" [11, 31], meaning the acquisition function will try to select the samples where the model is the least confident, the acquisition function acts as a proxy of the training loss which is not available. The second approach is based on "diversity sampling" [26], the acquisition function tries to find samples that represent the diversity existing in the unlabeled dataset, it is particularly suited for classification tasks where having a class-balanced training dataset is crucial. To the best of our knowledge, before this work active learning had never been applied to face alignment.

3 Methods

3.1 SCAF: adding skip-connections to 3FabRec

To detect precisely a facial landmark, spatial information must be kept. However, in common convolutional networks such as the auto-encoder of 3FabRec

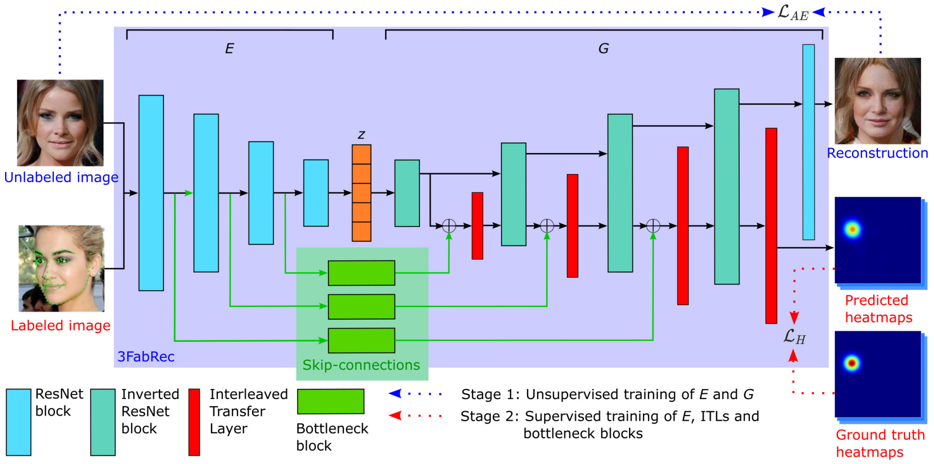

Fig. 1. Our network architecture and the 2-stage training pipeline. We add skip-connections between the encoder and the generator of 3FabRec.

[1], the spatial dimensions of the features maps are progressively reduced in the encoder as global information emerges leading to a compact representation of the face image. This representation contains strong semantic information useful to reconstruct the whole face but may lack local details crucial to detect precisely the landmarks. To address this issue, many recent architectures for face alignment [22, 3, 28, 27, 9] use the Hourglass architecture where "skip-connections" are added between the encoder and decoder. These skip-connections between the two parts of the network preserve spatial information at multiple resolutions, the decoder can combine these different resolutions to generate better heatmaps.

Following this principle, we propose SCAF which stands for Skip-Connections in Auto-encoder for Face alignment, we enhance the 3FabRec architecture with skip-connections between the encoder and the ITLs. Thus, the input of an ITL is the element-wise sum of the output of the previous ResNet layer of the generator and the output of the corresponding encoder layer (the one with the same spatial dimensions). Before the sum, the output of the encoder layer is transformed by a set of convolutions called "bottleneck block" as it is done in Hourglass architectures. The full architecture can be seen in Fig. 1.

We also noticed that splitting the supervised training into two steps: (1) Training only the ITLs, (2) Finetuning the ITLs and the encoder, is not necessary and we obtained better results training directly both. As in 3FabRec, we use as loss the $L2$ distance between the predicted ($\tilde{H}$) and ground truth heatmaps H.

3.2 Active learning for face alignment

When training with very few examples, we optionally use active learning to select the best examples. We introduce a *new* acquisition function called Negative Neighborhood Magnitude (NNM) based on uncertainty sampling applied to the landmark heatmaps. When the model is not confident about its predictions, we noticed that the magnitude of the heatmaps near the predicted landmark is lower than when it is confident (see Fig 2). Thus, to compute the NNM, for each predicted heatmap $\tilde{H}$, we compute the sum of the heatmap pixels in a square window W_i of size s around the predicted landmark position $\tilde{l}_i$, then we sum all heatmaps and take the negative so that the NNM behaves the same way as entropy, the less confident the model is, the greater NNM is.

$$NNM(\tilde{H}) = -\sum_{i=1}^{L} \sum_{u,v \in W_i} \tilde{H}_i(u,v) \tag{1}$$

After each model training, we rank the unlabeled samples. Some datasets contain very hard images where the most face is occluded which, thus, have a large NNM but are not useful for annotation. So, when we select images to label, we discard a percentage of the images with the largest NNM (see Section 4.6).

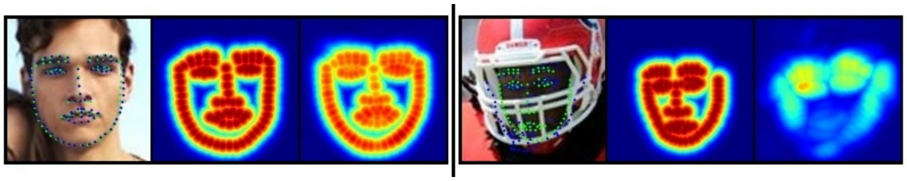

Fig. 2. Original image with ground truth (green dots) and predicted (blue dots) landmarks, ground truth heatmaps and predicted heatmaps for two images from WFLW.

4 Experiments

4.1 Datasets

Unsupervised training datasets We used a combination of two datasets for the unsupervised training:

AffectNet [21]: dataset created to capture a wide range of facial emotions. It contains 748K images.

CelebA [19]: dataset with 202K images of celebrities.

Our final dataset for unsupervised training contains about 950K images. The authors of 3FabRec [1] used as dataset a combination of 228K images from AffectNet [21] and 1.8M images from the VGGFace2 dataset [4] yielding a total of 2.1M images. However, due to copyright issues, VGGFace2 [4] is no longer available, so we could not use it for our experiments.

Supervised training datasets We trained and evaluated our supervised models on two facial landmark datasets.

300-W [25]: combination of several facial landmark datasets re-annotated with 68 landmarks. Following the usual splits [1, 9], our training set contains 3148 images, the *full* test set contains 689 images and is split into a *common* test set of 554 images and a *challenging* test set of 135 images.

WFLW [28]: dataset containing 7500 training images and 2500 testing images annotated with 98 landmarks. Many faces are heavily occluded or blurred making it a challenging dataset.

4.2 Experimental settings

Unsupervised training Apart from the training datasets (see Section 4.1), we follow the same procedure as in 3FabRec [1], we use the same network (without ITLs and skip-connections) and the same hyperparameters.

Supervised training. Using the ground truth bounding boxes, we crop and resize the images to 256 x 256 pixels. To generate the ground truth heatmaps, we use Gaussian kernels with $\sigma = 7$. Each ITL layer is a 3x3 convolutional layer and for the bottleneck blocks we use the hierarchical, parallel, and multi-scale block of [2]. The modified auto-encoder generates landmark heatmaps of size 128 x 128 pixels by skipping the last generator layer (the authors of 3FabRec [1] showed that the higher generator layers contain mostly decorrelated local appearance information). We use the same data augmentations as in 3FabRec [1], we also use Adam [15] to optimize the ITL and bottleneck layers, their learning rate is set to 0.001 while the encoder's one remains to 2×10^{-5}, the Adam's β_1 is reset to the default value of 0.9. Unlike 3FabRec [1], we train directly the three modules in parallel without any ITL-only-training stage before.

Active learning The initial labeled $\mathcal{L}^0$ set always contains 10 random samples from the unlabeled set $\mathcal{U}_N$, the number K of added samples after each training depends on the final training set size. For 300-W we used K=60, 30, 10 for a final training size of 315 (10% of dataset), 158 (5%), 50 respectively. For WFLW, we used K=100, 75, 40, 20, 10 for a final training size of 750 (10% of dataset), 375 (5%) 200, 100, 50 respectively.

Evaluation To evaluate our models, we use the Normalized Mean Error (NME) with the distance between outer eye-corners as "inter-ocular" normalization.

4.3 Unsupervised training results

We trained the auto-encoder with the same training parameters as in 3FabRec [1] except for the datasets (see section 4.1) but because we had less than half of the number of images used in 3FabRec [1] for our unsupervised training, we obtained worse results on the supervised training. Fortunately, the authors of 3FabRec [1] provide the source code and pre-trained weights for the auto-encoder at https://github.com/browatbn2/3FabRec, so we decided to use these weights to focus on the supervised training and get fair comparisons with their results.

4.4 Qualitative results

During the supervised training of SCAF, the reconstruction error increases because details non-necessary for landmark detection such as gender or skin color fade away. Only the shape of the face remains but sometimes some reconstructed facial parts do not even match with the predicted landmarks. For example (see Fig. 3), the mouth is always reconstructed as close even if it is open in the original face, but the landmark predictions align with the original mouth.

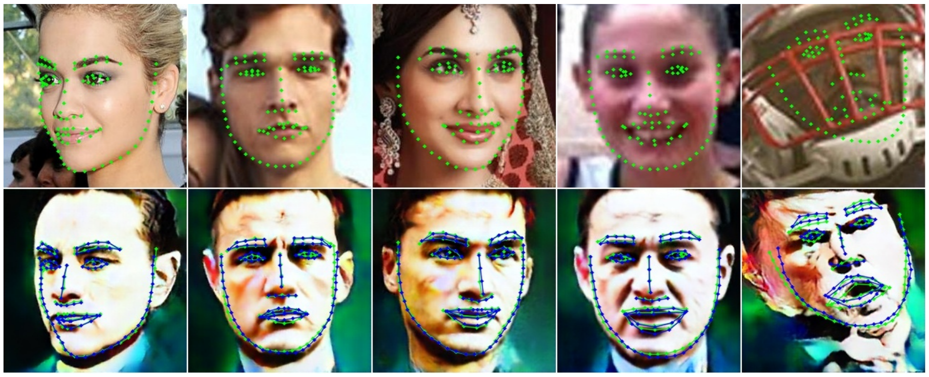

Fig. 3. Comparison of the reconstructions and landmark predictions. Top row shows some original images from the WFLW Full test set and their ground truth landmarks (green dots). Bottom row shows the reconstructed images with SCAF. Predicted landmarks are displayed in blue along with the ground truth landmarks in green.

4.5 Comparison with state-of-the-art

Comparison with fully supervised methods Table 1 compares our methods with fully supervised methods on 300-W [25] and WFLW [28] when training on the full training set. We re-trained 3FabRec [1] from the provided unsupervised weights available at https://github.com/browatbn2/3FabRec to get a fair comparison between the original network and our modified architecture. When training on 300-W, our implementation of 3FabRec gets worse results than the ones reported in the paper of 3FabRec [1] however SCAF improves our results, especially on the Challenging test set. For WFLW, this time, our implementation of 3Fabrec obtain NME results slightly better than the ones reported in the 3FabRec paper [1]. The addition of the skip-connections improves again the NME from 5.58 to 5.50. Recent fully supervised methods beat our approach when trained on the full training set of WFLW or 300-W but the point of our semi-supervised method is to keep good performance even when training with limited data as we will see in the next paragraph.

Training with limited data For 300-W, our implementation of 3FabRec gets better results than the ones reported in 3FabRec paper [1] when training on reduced training size. Apart from the training size of 50, SCAF outperforms our implementation of 3FabRec for any training dataset size. If we also apply active learning, the NME is reduced on the Challenging test but increased on the Common test set meaning that the model is more robust to challenging examples but a bit less precise for common examples. Results are reported in Table 2.

Table 1. Normalized mean error (%) on 300-W on the Common, Challenging and Full test sets and on WFLW Full test set

300-W			
Method	Com.	Chall.	Full
SDM [29]	5.57	15.40	7.52
SAN [8]	3.34	6.60	3.98
LAB [28]	2.98	5.19	3.49
ODN [32]	3.56	6.67	4.17
SA [23]	3.21	6.49	3.86
TS^3 [9]	2.91	5.90	3.49
AWing [27]	2.72	4.52	3.07
LUVLi [18]	2.76	5.16	3.23
3FabRec [1]	3.36	5.74	3.82
3FabRec (Our impl.)	3.54	5.93	4.01
SCAF (Ours)	3.48	5.83	3.95

WFLW	
Method	Full
SDM [29]	10.29
SAN [8]	5.22
LAB [28]	5.27
SA [23]	4.39
AWing [27]	4.36
LUVLi [18]	4.37
3FabRec [1]	5.62
3FabRec (Our impl.)	5.58
SCAF (Ours)	5.50

Table 2. Normalized mean error (%) with reduced training sets on 300-W on the Common, Challenging and Full test sets (first, second and third columns respectively for each training set size). AL stands for active learning.

300-W dataset															
Method	Training set size														
	100%			20%			10%			5%			50(1.5%)		
RCN+ [14]	3.00	**4.98**	**3.46**	-	**6.12**	**4.15**	-	6.63	4.47	-	9.95	5.11	-	-	-
SA [23]	3.21	6.49	3.86	3.85	-	-	4.27	-	-	6.32	-	-	-	-	-
TS^3 [9]	**2.91**	5.90	3.49	4.31	7.97	5.03	4.67	9.26	5.64	-	-	-	-	-	-
3FabRec [1]	3.36	5.74	3.82	3.76	6.53	4.31	3.88	6.88	4.47	4.22	6.95	4.75	4.55	7.39	5.10
3FabRec (Our impl.)	3.54	5.93	4.01	3.79	6.33	4.29	3.93	6.70	4.47	4.10	6.86	4.64	**4.27**	7.23	4.85
SCAF (Ours)	3.48	5.89	3.95	**3.66**	6.23	4.17	**3.87**	6.60	**4.40**	**3.93**	6.84	**4.50**	4.33	7.60	4.97
SCAF+AL (Ours)	-	-	-	-	-	-	3.99	**6.49**	4.48	4.19	**6.78**	4.70	4.29	**6.93**	**4.81**

Table 3 compares our models to other semi-supervised methods on WFLW. Firstly, when training 3FabRec network on WFLW, skipping the ITL-only training step significantly improves the NME, especially when training with very limited data. Apart from the training size of 50, SCAF consistently outperforms 3FabRec when training on full or limited training data. When combined with Active learning, its performance is improved even further, especially with the training size of 50 where it beats our implementation of 3FabRec this time.

4.6 Ablation studies

Comparison of acquisition functions We tried three different acquisition functions on the 3FabRec [1] and SCAF architectures. Two of them are based on uncertainty sampling: our proposed Negative Neighborhood Magnitude (NNM) and the mean of the spatial entropy of the heatmaps. The last function is based on diversity sampling, it is the K-center-greedy algorithm used in [26].

Table 3. Normalized mean error (%) with reduced training sets on WFLW Full test set. AL stands for active learning.

WFLW dataset					
Method	Training set size				
	100%	20%	10%	5%	50
SA [23]	**4.39**	**6.00**	7.20	-	-
3FabRec [1]	5.62	6.51	6.73	7.68	8.39
3FabRec (Our impl.)	5.58	6.23	6.42	6.84	7.74
SCAF (Ours)	5.50	6.07	6.28	6.72	8.06
SCAF+AL (Ours)	-	-	**6.24**	**6.59**	**7.60**

Table 4. Normalized mean error (%) on WFLW Full test set for different active learning methods and different training set sizes (5% = 375 examples and 10% = 750 examples), for our implementation of 3FabRec and SCAF.

WFLW dataset										
Method	3FabRec (Our impl.)					SCAF				
	Final training set size					Final training set size				
	50	100	200	5%	10%	50	100	200	5%	10%
Random	7.74	7.44	7.04	6.84	6.42	8.06	7.40	6.88	6.72	6.28
NNM	8.27	7.57	7.15	6.77	6.36	8.04	7.44	7.01	6.63	6.22
Entropy	8.17	7.53	7.06	6.71	6.32	7.95	7.44	7.02	6.61	6.22
NNM$_{10\%}$	**7.63**	7.20	6.82	**6.62**	**6.31**	7.60	6.99	**6.72**	**6.59**	6.24
Entropy$_{10\%}$	7.71	**7.12**	**6.83**	**6.62**	6.34	**7.53**	**6.96**	6.73	6.62	**6.22**
K-center-greedy	7.85	7.36	6.95	6.65	6.32	7.74	7.18	6.82	6.61	6.28

Table 4 reports the NME on WFLW for these acquisition functions. Apart from the final training size of 50 on the 3FabRec network, the K-center-greedy function improves consistently the results over random sampling. When using NNM or Entropy to select the samples among *all* the unlabeled samples, when the final training size is small (≤200), the results are worse (or barely superior) than sampling at random. However, if we discard the top 10% ranked samples when selecting the samples, then we improve the NME and strongly outperform the random and K-center-greedy samplings. These methods are referred as NNM$_{10\%}$ and Entropy$_{10\%}$ in Table 4. This shows that very hard samples in the WFLW training dataset should not be added to the training set because they are outliers and won't help the model to generalize to unseen data. However, as the final training set size increases, the benefit of discarding the worst examples tends to disappear; because more samples are added, the proportion of outliers decreases, and "normal" challenging samples are added to the training set.

Fig. 4 shows the top-5 ranked samples according to the NNM after training SCAF on 10 random samples for the WFLW training set. The top row displays the top-5 samples among *all* unlabeled samples, the five images are clearly outliers: blue color, distorted face for the second image, non-human face for the last image and won't help much the model to generalize to unseen data after training. The bottom row displays the top-5 images after discarding the top-10% images from the unlabeled dataset. These five images are still challenging (low resolution, occlusion, baby face) but closer to "normal" images, adding them to the training set should improve the model predictions.

Entropy and NNM have close results in terms of NME. However, in the case of Entropy, the whole heatmaps must be normalized before computing the entropy on the whole heatmaps too, whereas the computation of the NNM only requires summing heatmaps values of small windows. In our experiments, with an Intel Core i7-9850H CPU, computing the Entropy took on average 0.042 seconds whereas computing the NNM only took 0.012 seconds. Thus, the NNM is 3.5 times faster to compute than the Entropy while achieving comparable results.

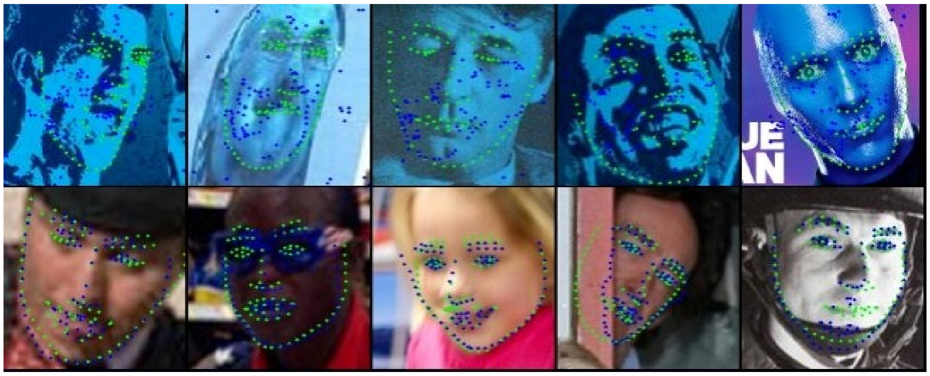

Fig. 4. Top-5 ranked images for NNM after training the model on 10 random samples. Ground truth landmarks are displayed with green dots while blue ones are the predicted landmarks. Top row shows the top-5 ranked images among all the unlabeled samples while bottom row displays the top-5 ranked images after removing the top-10% images.

5 Conclusion

In this paper, we addressed the problem of training face alignment models with limited labeled data. To achieve this goal, we improved 3FabRec [1] architecture by adding skip-connections between the encoder and decoder during the supervised training. This makes the network predict more accurately the facial landmarks heatmaps, especially for challenging examples where the hidden representation fails to capture all the specificities of the image. We also applied active learning to the face alignment task to improve even further the performance with limited training data and showed its effectiveness by introducing a new acquisition function for heatmaps called Negative Neighborhood Magnitude. This function achieves similar performance to spatial entropy in terms of NME while being much faster to compute.

Acknowledgement This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012376 made by GENCI.

References

1. Browatzki, B., Wallraven, C.: 3fabrec: Fast few-shot face alignment by reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6110–6120 (2020)
2. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3706–3714 (2017)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age (2018)
5. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International journal of computer vision **107**(2), 177–190 (2014)
6. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61**(1), 38–59 (1995)
7. Dapogny, A., Bailly, K., Cord, M.: Decafa: Deep convolutional cascade for face alignment in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6893–6901 (2019)
8. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 379–388 (2018)
9. Dong, X., Yang, Y.: Teacher supervises students how to learn from partially labeled images for facial landmark detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 783–792 (2019)
10. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2245 (2018)
11. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International Conference on Machine Learning. pp. 1183–1192. PMLR (2017)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
13. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018)
14. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018)

15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kirsch, A., Van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems **32**, 7026–7037 (2019)
17. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2144–2151. IEEE (2011)
18. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8236–8246 (2020)
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
20. Matthews, I., Baker, S.: Active appearance models revisited. International journal of computer vision **60**(2), 135–164 (2004)
21. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
22. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
23. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10153–10163 (2019)
24. Robinson, J.P., Li, Y., Zhang, N., Fu, Y., Tulyakov, S.: Laplace landmark localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10103–10112 (2019)
25. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 397–403 (2013)
26. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
27. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6971–6981 (2019)
28. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2129–2138 (2018)
29. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 532–539 (2013)
30. Yan, J., Lei, Z., Yi, D., Li, S.: Learn to combine multiple hypotheses for accurate face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 392–396 (2013)
31. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 93–102 (2019)

32. Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3486–3496 (2019)
33. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)