Unsupervised Learning of Landmarks based on Inter-Intra Subject Consistencies

Weijian Li*, Haofu Liao*, Shun Miao†, Le Lu† and Jiebo Luo*
*Department of Computer Science, University of Rochester, Rochester, NY, USA
†PAII, Inc., Bethesda, MD, USA
Email:*{wli69, hliao6, jluo@cs.rochester.edu}, †{shwinmiao, tiger.lelu@gmail.com}

Abstract—We present a novel unsupervised learning approach to image landmark discovery by incorporating the inter-subject landmark consistencies on facial images. This is achieved via an inter-subject mapping module that transforms original subject landmarks based on an auxiliary subject-related structure. To recover from the transformed images back to the original subject, the landmark detector is forced to learn spatial locations that contain the consistent semantic meanings both for the paired intra-subject images and between the paired inter-subject images. Our proposed method is extensively evaluated on two public facial image datasets (MAFL, AFLW) with various settings. Experimental results indicate that our method can extract the consistent landmarks for both datasets and achieve better performances compared to the previous state-of-the-art methods quantitatively and qualitatively.

I. INTRODUCTION

Image landmark localization aims to detect a set of semantic keypoints on the given objects from images, such as the eyes, nose, and ears of human faces. It has been an essential process to assist many high-level computer vision tasks [1], [2]. Traditional fully supervised approach relies on a set of annotated landmark locations that are labeled by human experts. These landmarks are subsequently used to train a supervised model before it can be applied to unseen images. Although many efforts have been made in this direction and promising results have been achieved [3], [4], [5], [6], the challenge of supervised models remains that a large amount of human labeling efforts are required to have desirable performance, which is expensive and the annotation processing is subjective.

Another recent approach follows the unsupervised learning strategy to extract keypoints with self-supervision [7], [8], [9], [10]. Many of the existing methods propose to apply a group of random transformations, such as rotations and translations, on the original image to generate the transformed and paired images. Machine learning models are trained to predict landmark locations based on the fact and constraint that the paired landmarks should follow the same transformation.

Despite the popularity and success, training landmark detectors with only paired images from the same subject images may be insufficient to discover the inter-subject consistency among different subjects. The trained detector may be biased to learn landmark locations that are meaningful for the transformation within the same-subject pairs, but make different predictions on the same landmark across different subjects.

To this end, we propose a novel unsupervised learning method for image landmark discovery via exploring and integrating on the inter-subject consistency. Our method follows the standard equivariance approach by using image reconstruction as supervision cues, added with injecting a subject mapping module between the image encoder and decoder to ensure the inter-subject landmark semantics. Specifically, (1) our model first extracts the feature maps from the input image, then computes a landmark heatmap from an auxiliary subject image as the structural guidance. (2) We implement a subject mapping module to perform structural transformation on the input image according to the structure defined by the extracted landmark heatmap of the auxiliary image. (3) The transformed image is then sent into a second transformation guided by the landmark heatmap of a paired image of the input subject and the final generated image is output. In this manner, we adopt a cycle-like design to complete the transformation cycle between the paired intra-subject images in both directions.

By modeling an intermediate landmark based inter-subject transformation, the landmark detector is enforced to extract semantically-consistent facial landmark locations across different subjects to produce accurate landmark based image generation. The cycle-like intra-subject translation enables additional supervision that encourages our network to learn consistent referential keypoints for both forward and backward image translations. These two factors together help our network to not only extract discriminative landmark locations for each subject in accordance with the provided transformation, but also simultaneously retain landmark semantics across different subjects.

In summary, our main contributions are as follows:
- We propose an unsupervised learning method for image landmark discovery by focusing on both inter and intra landmark consistencies.
- We construct the inter-subject consistency directly through landmark representations with the use of auxiliary images.
- We model the intra-subject transformation as a cycle and build a two-path end-to-end trainable structure to improve the intra-subject landmark consistency.
- Comprehensive quantitative and qualitative evaluations on two public facial image datasets demonstrate that the consistent superior landmark localization performances using our method are observed.
II. Related Work

During recent years, several studies have been conducted towards object landmark discovery with unsupervised learning [8], [9], [7], [11], [12], [13], [14], [15]. The equivariance [16] constraint is widely adopted as a supervision signal to learn meaningful landmark locations. For example, Thewlis et al. [7] propose to extract landmarks compatible with the input image deformations by regressing the probabilistic maps; Suwajanakorn et al. [12] detect 3D object keypoints by predicting the known rigid transformations between paired input objects; Zhang et al. [11] introduce a generative framework with learnable landmark locations under a set of transformation constraints. Meanwhile, Jakab et al. [8] also adopt image generation as supervision signal to discover landmarks. They propose to construct a heatmap bottleneck for landmarks by applying the Gaussian-like function centered on the highest responses on the extracted feature map. Then the built Gaussian-like heatmaps act as driving signal to deform input image to the target image. Sanchez et al. [9] study the effect of domain adaptation in unsupervised landmark detection as well as measurement of detection stability by introducing a new evaluation metric.

While these methods have made great success in different perspectives, inter-subject supervision is usually missing. The paired deformed images are able to help the network locate geometry positions from the same subject but may fail to ensure the position consistency across different subjects. Zhang et al. [11] recognize this issue as Cross-object correspondence but they rely on network’s implicit learning without supervision. Lately, Thewlis et al. [10] address this issue by proposing a vector exchange process. During this process, each original image’s features at each pixel is first replaced by a weighted aggregation of all pixel features from the auxiliary images. Then maximizing the pixel-wise feature similarities over spatial locations between the exchanged image and the deformed paired image is used as supervision for the learning process. Though our method shares the same idea in the sense to include auxiliary subject images to enable inter-subject information exchange, Thewlis et al. [10] et al. mainly focus on learning general feature representations and extract landmarks as a separate follow up step. Instead, we directly encode auxiliary images into landmark representations in the same form as the source subject images, and include them as a driving signal to enforce consistent position exchanges that valid for both intra and inter subject relationships.

III. Method

Given an image \( x \in X \subseteq \mathbb{R}^{H \times W \times C} \) and its deformed version \( x' \in X' \), our goal is to learn a function \( \Phi(x) = y \in Y \) that extract \( K \) structural representations as landmarks without any annotations. Following previous works [8], [9], we address this problem through conditional image generation. The overall framework can be seen in Figure 1 which mainly contains three parts: 1) landmark detectors, 2) inter and intra image generators, and 3) a backward cycle path. This design aims to explore the landmark consistency across the inter- and intra-subject image pairs generated by geometry transformations. In the following sections, we will describe the proposed method in detail.

A. Landmark Detector

A landmark detector takes an image \( x \) as input and outputs \( K \) sets of landmark representations where each corresponding to a landmark location. We adopt a similar structure as proposed in [8], [9]. In particular, the input image \( x \) is first encoded by a standard convolutional neural network to extract visual feature maps \( S \in \mathbb{R}^{H \times W \times K} \). Spatial coordinates for the \( K \) landmarks are then obtained from the feature maps and remain differentiable with a Sotfmax [17] operation.

Specifically, the predicted \( k \)-th landmark location \( u_k \) is the weighted average of the spatial locations \( i \), where the weights are computed by the softmax of the \( k \)-th feature map \( \hat{S}_k \), i.e.,

\[
    u_k = \sum_i \exp(\beta S_k(i)) \cdot \frac{1}{\sum_i \exp(\beta S_k(i))},
\]

where \( \beta \) is a hyperparameter for the smoothness. Each prediction is then mapped back to a Gaussian-like probabilistic heatmap centered at \( u_k \): 

\[
    \Phi_H(x; k) = \exp\left(-\frac{1}{2\beta^2} \| u - u_k \|^2 \right). 
\]

\( u_k \) will be the final landmark detection results and \( \Phi_H(x; k) \) will be used by later modules as a driving signal to complete the image generation task achieving a self-supervision for \( u_k \).

B. Inter-Intra Image Generator

Previous studies have shown success in unsupervised learning of landmark locations given pairs of images with different geometries. However, since both the two images of a pair come from the same subject, the method does not consider the inter-subject landmark consistency and fails to learn the inter-subject semantics. To this end, we propose to include an auxiliary image which comes from a different subject, and incorporate it as an intermediate transformation as shown in Figure 1.

Specifically, we denote the geometrically deformed image pairs as \( x \) and \( x' \), and an auxiliary image as \( x^a \). An image encoder \( \Phi_E \) is first applied to the source image \( x \) to extract a visual feature map \( F_s = \Phi_E(x) \in \mathbb{R}^{H \times W \times D} \). At the first stage, we transform the object structure from source image \( x \) into auxiliary image \( x^a \) based on the landmark representation \( \Phi_H(x^a) \) of the auxiliary image, where \( \Phi_H \) is the landmark detector we described in the previous section. This is achieved via an image generation function \( \Psi \) which takes the concatenation of the visual feature map and the landmark heatmap as inputs, and outputs the generated image:

\[
    I_a = \Psi(F_s, \Phi_H(x^a)) = \Psi(\Phi_E(x), \Phi_H(x^a)).
\]

Next, in the second stage, image \( I_a \) is further transformed by the landmark representations \( \Phi_H(x') \) extracted from the
by constructing training target images on images from are generated. However, as long as we construct deformed trivial and can be removed as more deformed image pairs results themselves. One may argue that this modification is lead to the generation instead of focusing on the generated that we focus on learning the landmarks (the conditions) that source image and our goal is to reconstruct .

also consider a reversed-order scenario where target image to be generated. Similar to Zhu et al. [18], we consider a reversed-order scenario where $x'$ is used as the source image and its groundtruth image which focuses on generation details:

$$
\mathcal{L}_R(I, I_{gt}) = \|I - I_{gt}\|^2.
$$

2) perceptual loss [19]: an MSE loss on the layer outputs of an ImageNet [20] pretrained VGG-16 [21] network with the generated target image and its groundtruth image as inputs respectively. It focuses on high level feature representations:

$$
\mathcal{L}_P(I, I_{gt}) = \sum_l \|VGG^l(I) - VGG^l(I_{gt})\|^2.
$$

The overall loss is thus a combination of these two losses on both directions of the cycle:

$$
\mathcal{L} = \mathcal{L}_R(I_{ct}, x) + \mathcal{L}_R(I_{tx}, x') + \mathcal{L}_P(I_{ct}, x) + \mathcal{L}_P(I_{tx}, x')
$$

Our model is trained end-to-end with the overall loss $\mathcal{L}$.

IV. EXPERIMENTS

A. Implementation Details

**Landmark Detector**: We follow the previous work [9] to adopt a Hourglass [27] based network as our landmark detector which is experimented to be effective in keypoint
The network is first pretrained on a Human Pose Estimation heatmap using Equation 2. To keep a fair comparison with [9], the coordinates are further mapped back to a Gaussian-like distribution to reconstruct the target image.

The landmarks are further linearly regressed into \( C \) landmarks where \( C \) is set to be 10, 30 or 50. These landmarks are further linearly regressed into \( K \) landmarks by training a linear regressor on the AFLW and MAFL training set with all the other parameters of the network fixed. The obtained results are considered the final detection for AFLW and MAFL datasets. To generate geometrically deformed image \( x_t \), a combination of scaling, rotation and translation is applied on the original image \( x \). The auxiliary image \( x^a \) for each deformed pair \( x \) and \( x_t \) is randomly selected from the original images \( X \).

In our implementation, it is randomly drawn from the other \( x \) images in the same batch.

We set parameter \( \beta \) in Equation 1 to 10, parameter \( \sigma \) in Equation 2 equal to 0.5. The VGG-16 layers we use for the perceptual loss of Equation 6 are \( \{relu_1, relu_2, relu_3, relu_4\} \). Batch size is set to 32. Adam optimizer is adopted with initial learning rate 0.001 with a 0.1 decay rate every 30 epochs. The proposed model is implemented in PyTorch and is experimented on a single NVIDIA GeForce GTX 1080Ti GPU.

### B. Quantitative Evaluation

Following previous works [8], [7], we evaluate the proposed method based on a point-to-point MSE metric normalized by the inter-ocular distance on the detected landmarks on the test sets. A baseline method is constructed without the proposed inter-mapping layer and the cycle backward path. As shown in Table I, integrating the inter-subject mapping module brings improvements comparing to the baseline method indicating the importance of introducing the auxiliary images. Integrating the cycle backward module also improves our model’s performance. By combining the two proposed modules together, our complete model Ours-All further achieves 3.08% and 6.20% error rates when detecting \( K = 10 \) intermediate landmarks.

| Method        | K | MAFL   | AFLW  |
|---------------|---|--------|-------|
| **Supervised**|   |        |       |
| TCDCN [22]    | 7.95 | 7.65   |       |
| RAR [23]      | -  | 7.23   |       |
| MTCNN [24]    | 5.39 | 6.90   |       |
| **Unsupervised** | |        |       |
| Thewlis [25]  | 5.83 | 8.80   |       |
| Shu [15]      | 5.45 | -      |       |
| Sahasrabudhe [26] | 6.01 | -      |       |
| Wiles [14]    | 3.44 | -      |       |
| Thewlis [7]   | 7.95 | -      |       |
| Sanchez [9]   | 3.99 | 6.69   |       |
| Zhang [11]    | 3.46 | 7.01   |       |
| Jakab [8]     | 3.19 | 6.86   |       |
| Jakab [8]     | 2.58 | 6.31   |       |
| Jakab [8]     | 2.54 | 6.33   |       |
| Thewlis [10]  | 2.86 | 6.54   |       |
| **Baseline**  | | 3.41  | 6.59  |
| w. Inter-Subject | | 3.10 | 6.24  |
| w. Cycle      | 3.12 | 6.28   |       |
| Ours-All      | 3.08 | 6.20   |       |
| Ours-All      | 2.89 | 6.08   |       |
| Ours-All      | 2.85 | **6.04** |       |

### TABLE I: Normalized MSE evaluations on the public MAFL and AFLW dataset. Baseline*: our re-implementation of [9].

### TABLE II: Normalized MSE evaluations on the MAFL test set for varying number (n) of supervised samples from MAFL training set used for learning the regressor. We use \( K = 10 \) intermediate landmarks.

| n supervised | Thewlis [7] | Sanchez [9] | Ours |
|--------------|------------|-------------|------|
| 1            | 10.82      | 18.70       | **9.03** |
| 5            | 9.25       | 8.77        | **7.50** |
| 10           | 8.49       | 7.13        | **7.09** |
| 100          | -          | 4.53        | **3.71** |
| 500          | -          | 4.13        | **3.23** |
| 1000         | -          | 4.16        | **3.17** |
| 5000         | -          | 4.05        | **3.09** |
| All          | 7.15       | 3.99        | **3.08** |

Localization tasks such as Human Pose Estimation, Facial Landmark Detection, etc. It takes \( \mathbb{R}^{128 \times 128 \times 3} \) RGB images as input and outputs \( \mathbb{R}^{32 \times 32 \times K} \) feature maps. Each heatmap is then transformed to be \( \mathbb{R}^{K \times 2} \) landmark coordinates \( u \) with the Softargmax operation on each \( k \)th channel of the feature maps. The coordinates are further mapped back to a Gaussian-like heatmap using Equation 2. To keep a fair comparison with [9], the network is first pretrained on a Human Pose Estimation dataset MPII [28] to detect \( K = 19 \) landmarks. Then all the trained network parameters are fixed. A set of linear projection matrices are applied on the weights of the convolutional layers and are trained for the new detection tasks. We also tried training all the parameters from scratch. The details can be found in the Ablation Studies section.

**Inter-Intra Image Generator:** The image encoder \( \Phi_E \) takes \( \mathbb{R}^{128 \times 128 \times 3} \) RGB images as input and spatially down-sampled the image into a \( \mathbb{R}^{32 \times 32 \times 256} \) feature map. It is then concatenated with the obtained landmark heatmaps from the landmark detector \( \Phi_H \) along the channel dimension. The concatenated result is sent into the generator network \( \Psi \) which contains 6 residual blocks and two spatial upsampling blocks to reconstruct the target image.

**Learning Facial Landmarks:** To examine the effectiveness of the proposed method, we follow previous works [8], [9], [7], [10] to adopt the CelebA [29] dataset which contains over 200k training images from the celebrities faces excluding 1,000 common images from the MAFL [24] dataset; the AFLW [30] and MAFL [24] datasets which contains 10,122/2,991 and 18,997/1,000 training/testing images respectively. During training, the network is first trained on CelebA dataset outputing \( K = C \) landmarks as intermediate detections where \( C \) is set to be 10, 30 or 50. These landmarks are further linearly regressed into \( K = 5 \) landmarks by training a linear regressor on the AFLW and MAFL training set with all the other parameters of the network fixed. The obtained results are considered the final detection for AFLW and MAFL datasets. To generate geometrically deformed image \( x_t \), a combination of scaling, rotation and translation is applied on the original image \( x \). The auxiliary image \( x^a \) for each deformed pair \( x \) and \( x_t \) is randomly selected from the original images \( X \).
Fig. 2: Visualization results of the detected landmarks on CelebA. **Top two rows:** 16 detection results without the proposed modules (Baseline model). **Bottom two rows:** 16 detection results from the same face subjects as in the top rows but with the proposed modules (Ours full model). Each colored dot corresponding to a detected landmark. Landmarks with the same color within the top or bottom two rows are the same, but are not necessary the same across the top and bottom two rows.

compositional effect. Though our method predicts lesser 10 intermediate landmarks, we notice an even better performance on the AFLW dataset comparing to [10] which predicts 50 intermediate landmarks. For $K = 30$ and $K = 50$, our model also produce competitive results with a better performance on AFLW when $K$ is set to 50. We believe the reason is that our model is able to locate semantically more meaningful and consistent landmarks for effective inference. This can be further illustrated by examining the visualization results by comparing the detection stability in Figure 2.

To check our model’s capability with different dataset scales, we conduct another evaluation by varying the number of training images when training the final linear regressor on the MAFL dataset following previous works [7], [9]. As can be seen in Table II, our method achieves better performance across all the experimental settings comparing to the previous state-of-the-arts. Notice that a smaller standard deviation is also achieved by our method compared with [9]. The results tend to saturate when 100 or more images are used for training, and are almost the same best performance when using 5,000 images and all. It indicates a desirable capability of our method for datasets with less training data.

### C. Qualitative Evaluation

To qualitatively examine the detection results and verify the proposed method, we visualize the detected landmarks on the images from CelebA dataset with and without the proposed modules in Figure 2. It is clear to see that most of the landmarks predicted by our method are meaningful, e.g., eye corners, nose, mouse corners, cheek, while some of the landmarks predicted by the baseline method are located outside the facial regions, for example, hair strains, neck or collar. We believe these regions should not be considered as valid landmarks since they may not even exist in all the images. Moreover, looking at each detected landmark, we notice that our model is able to extract more consistent locations. For example, comparing the pink dot in the top rows and the blue dot in the bottom rows, we find both of them tend to focus on the forehead region. While some of the pink dots drift to other places such as the chin, background or hair, the blue dots, on the contrary, are apparently more stable across different face subjects. However, we also notice some problems predicted by our model. Although each landmark focuses on the same region in general, local variance still exists when occlusion or pose changes occur, such as the red landmark in the bottom rows that seems to find the

| Method     | K  | MAFL | AFLW |
|------------|----|------|------|
| VGG-19     | 10 | 3.23 | 6.42 |
| $\mathcal{L}_R$ only | 10 | 7.87 | 14.98 |
| $\mathcal{L}_P$ only | 10 | 3.25 | 6.28 |
| Proposed-All | 10 | 3.08 | 6.20 |
right cheek region, but sometime may drift upper or lower marginally; the blue landmark shifts to the open area when the forehead is covered by hair where the visual appearance is more semantically consistent but geometrically not. Therefore, we consider that integrating landmark spatial constraints will be beneficial for better performance.

D. Ablation Study

We examine variations of the modules and understand their effects to our model including: 1) different network structure for computing Perceptual Loss; 2) different choice of loss function. As shown in Table III, the VGG-19 model cannot perform as well as the VGG-16 model. Adopting the reconstruction loss alone is not sufficient to work well for the overall task. When both reconstruction and perceptual losses are adopted, we achieve the best performance. This indicates the importance of the perceptual loss for extracting semantic similarities as well as the benefit from the detailed context information.

V. CONCLUSION

In this work, we introduce an image generation based landmark discovery model with unsupervised learning. Our model extracts inter- and intra-subject consistent landmarks by including an inter-subject mapping module as intermediate translation as well as a backward cycle path for additional intra-subject supervision. The superior performance on two public facial image datasets under varies evaluation settings validate the effectiveness of the proposed model.

REFERENCES

[1] A. Bulat and G. Tzimiropoulos, “Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 109–117.

[2] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2377–2386.

[3] X. Xiong and F. De la Torre, “Global supervised descent method,” in CVPR, 2015, pp. 2664–2673.

[4] X. Wang, L. Bo, and L. Fuxin, “Adaptive wing loss for robust face alignment via heatmap regression,” in CVPR, 2019, pp. 6971–6981.

[5] Z. H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in CVPR, 2018, pp. 2235–2245.

[6] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, “Look at boundary: A boundary-aware face alignment algorithm,” in CVPR, 2018, pp. 2129–2138.

[7] J. Thewlis, H. Bilen, and A. Vedaldi, “Unsupervised learning of object landmarks by factorized spatial embeddings,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5916–5925.

[8] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, “Unsupervised learning of object landmarks through conditional image generation,” in Advances in Neural Information Processing Systems, 2018, pp. 4016–4027.

[9] E. Sanchez and G. Tzimiropoulos, “Object landmark discovery through unsupervised adaptation,” in Advances in Neural Information Processing Systems, 2019, pp. 13498–13509.

[10] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, “Unsupervised learning of landmarks by descriptor vector exchange,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6361–6371.

[11] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, “Unsupervised discovery of object landmarks as structural representations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2694–2703.

[12] S. Suwajanakorn, N. Snavely, J. J. Tompson, and M. Norouzi, “Discovery of latent 3d keypoints via end-to-end geometric reasoning,” in Advances in Neural Information Processing Systems, 2018, pp. 2059–2070.

[13] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, “Unsupervised learning of object keypoints for perception and control,” in Advances in Neural Information Processing Systems, 2019, pp. 10723–10733.

[14] O. Wiles, A. Koepke, and A. Zisserman, “Self-supervised learning of a facial attribute embedding from video,” arXiv preprint arXiv:1808.06882, 2018.

[15] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos, “Deforming autoencoders: Unsupervised disentangling of shape and appearance,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 650–665.

[16] K. Lenc and A. Vedaldi, “Learning covariant feature detectors,” in European conference on computer vision. Springer, 2016, pp. 100–117.

[17] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “Lift: Learned invariant feature transform,” in European Conference on Computer Vision. Springer, 2016, pp. 467–483.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[19] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European conference on computer vision. Springer, 2016, pp. 694–711.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[21] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Learning deep representation for face alignment with auxiliary attributes,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 5, pp. 918–930, 2015.

[23] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3476–3483.

[24] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European conference on computer vision. Springer, 2014, pp. 94–108.

[25] J. Thewlis, H. Bilen, and A. Vedaldi, “Unsupervised learning of object frames by dense equivariant image labelling,” in Advances in neural information processing systems, 2017, pp. 844–855.

[26] M. Sahasrabudhe, Z. Shu, E. Bartrum, R. Alp Guler, D. Samaras, and I. Kokkinos, “Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[27] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in European conference on computer vision. Springer, 2016, pp. 483–499.

[28] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738.

[30] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in 2014 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011, pp. 2144–2151.