IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 February 2023, accepted 9 March 2023, date of publication 14 March 2023, date of current version 20 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257180

RESEARCH ARTICLE

HybridMatch: Semi-Supervised Facial Landmark Detection via Hybrid Heatmap Representations

SEOUNGYOON KANG[1], (Graduate Student Member, IEEE), MINHYUN LEE[1], MINJAE KIM[2], AND HYUNJUNG SHIM[3]

[1]School of Integrated Technology, Yonsei University, Yeonsu-gu, Incheon 21983, Republic of Korea
[2]Vision AI Laboratory, AI Center, NCSOFT, Bundang-gu, Seongnam-si, Gyeonggi-do 13494, Republic of Korea
[3]Kim Jaechul Graduate School of Artificial Intelligence, KAIST, Dongdaemun-gu, Seoul 02455, Republic of Korea

Corresponding author: Hyunjung Shim (kateshim@kaist.ac.kr)

This work was supported in part by NCSOFT, Seongnam, South Korea; the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by MSIP under Grant NRF-2022R1A2C3011154; IITP grant funded by the Korea government (MSIT) and KEIT grant funded by the Korea government (MOTIE) (No. 2022-0-00680); and IITP grant funded by the Korea government (MSIT) and KEIT grant funded by the Korea government (MOTIE) (No. 2022-0-01045).

ABSTRACT Facial landmark detection is an essential task in face-processing techniques. Traditional methods, however, require expensive pixel-level labels. Semi-supervised facial landmark detection has been explored as an alternative, but previous approaches only focus on training-oriented issues (e.g., noisy pseudo-labels in semi-supervised learning), neglecting task-oriented issues (i.e., the quantization error in landmark detection). We argue that semi-supervised landmark detectors should resolve the two technical issues simultaneously. Through a simple experiment, we found that task- and training-oriented solutions may negatively influence each other, thus eliminating their negative interactions is important. To this end, we devise a new heatmap regression framework via hybrid representation, namely HybridMatch. We utilize both 1-D and 2-D heatmap representations. Here, the 1-D and 2-D heatmaps help alleviate the task-oriented and training-oriented issues, respectively. To exploit the advantages of our hybrid representation, we introduce curriculum learning; relying more on the 2-D heatmap at the early training stage and gradually increasing the effects of the 1-D heatmap. By resolving the two issues simultaneously, we can capture more precise landmark points than existing methods with only a few annotated data. Extensive experiments show that HybridMatch achieves state-of-the-art performance on three benchmark datasets, especially showing 26.3% NME improvement over the existing method in the 300-W *full* set at 5% data ratio. Surprisingly, our method records a comparable performance, 5.04 (*challenging* set in the 300-W) to the fully-supervised facial landmark detector 5.03. The remarkable performance of HybridMatch shows its potential as a practical alternative to the fully-supervised model.

INDEX TERMS Facial landmark detection, facial key-points, landmark detection, semi-supervised facial landmark detection, heatmap-based landmark detection.

I. INTRODUCTION

Facial landmark detection aims to identify predefined key points of the face image, including eyes, nose, mouth, and facial contour. It has been widely utilized in various applications, such as face morphing, tracking, expression analysis, and face identification. Existing landmark detection methods can be divided into the coordinate-based or heatmap-based approach, depending on the representation of a landmark point set. Since the coordinate-based approach does not fully utilize spatial and contextual information of landmark points, it shows relatively lower performance than the heatmap-based approach. For this reason, recent studies tend to develop heatmap-based methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

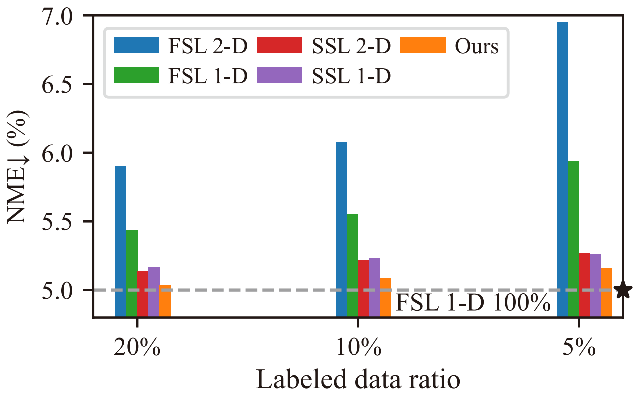

FIGURE 1. Effects of 1-D/2-D heatmap in the fully- (FSL) and semi- (SSL) landmark detection. FSL uses only *r*% of the training dataset, while SSL uses *r*% as supervised learning and the rest as unsupervised learning. All results are based on HRNet.

The heatmap-based methods mostly utilize 2-D heatmap as the landmark representation. Since a 2-D heatmap requires $\mathcal{O}(N^2)$ memory complexity to express $N \times N$ heatmap resolution, N cannot be large in practical applications. However, various studies [1], [2], [3], [4] have pointed out that a small N yields the large quantization error of the heatmap and thus is a performance bottleneck of facial landmark detection. To alleviate this issue, recent models have been actively studied to restore the residual parts (fractional components after quantization). However, restoring residual parts is sensitive to the size of the original 2-D heatmap; a poor performance with a considerably small N. Recently, Yin et al. [5] addresses this issue by representing the coordinate (x, y) of a heatmap with only $\mathcal{O}(2N)$ memory complexity. That is, they replace a 2-D heatmap with two 1-D heatmaps by assuming the separability of the 2-D heatmap. However, a 2-D heatmap is often not separable and the correlation between x and y coordinate is informative for landmark regression. To compensate for the loss of correlation information, they introduce a co-attention module between two 1-D heatmaps. Overall, the method by Yin et al. [5] effectively reduces the quantization error by increasing N. Finally, it significantly improves the accuracy of fully-supervised facial landmark detection over the 2-D heatmap-based methods (17% NME improvement in 300-W *common* dataset).

Despite the significant performance advantages of the fully-supervised methods, these models heavily rely on a large number of clean annotations. In particular, the landmark detection scenario requires pixel-level labels, which involves an expensive annotation cost. Besides, it easily suffers from noisy labels as creating precise pixel-level labels is challenging even for human annotators. To reduce the labeling budget and the sensitivity to data quality, recent studies have investigated the semi-supervised learning regime for facial landmark detection. The semi-supervised models utilize a mixture of a small amount of labeled data and a large amount of unlabeled data for model training, which is a reasonable setting for practical applications.

To this end, existing semi-supervised landmark detection methods focus only on training-oriented issues, such as effectively handling unlabeled data. They implicitly learn facial shapes via unsupervised training [6], developing a selective pseudo-labeling scheme by assessing pseudo-label quality [7], formulating multi-task learning [8], or utilizing style transfer to increase training dataset [9]. However, we argue that semi-supervised landmark detection should resolve *two technical challenges at the same time*; (i) the *task-oriented issue* such as quantization errors caused by low-resolution heatmap representation, and (ii) the *training-oriented issue* such as noisy pseudo-labels caused by the semi-supervised learning scenario.

Assuming that the task-oriented solution and training-oriented solution independently affect the performance, it is natural to combine the state-of-the-art of each side and then constitute the framework. Therefore, we attempt to combine the 1-D heatmap-based method by Yin et al. [5] for handling the quantization error and the high-performance semi-supervised framework (FixMatch [10]). Interestingly, through a simple experiment, we found that a 1-D heatmap is no longer more effective than a 2-D heatmap in the semi-supervised setting. Figure 1 compares facial landmark detection performances using a 1-D and 2-D heatmap, respectively. As expected, 1-D heatmap (FSL 1-D [5]) outperforms 2-D heatmap (FSL 2-D [11]) in the fully-supervised setting as reported in [5]. Counter-intuitively, under the semi-supervised scenario, we found that using the 1-D heatmap (SSL 1-D) and 2-D heatmap (SSL 2-D) reported similar NME values. It means that the semi-supervised training is facilitated better with a 2-D heatmap than a 1-D heatmap when observing the performance gain of each method over its fully-supervised counterpart; 24.2% and 11.4% NME improvements at 5% data ratio for 2-D and 1-D representation, respectively. From these results, we conclude that task-oriented and training-oriented solutions may negatively influence each other, thus eliminating their negative interactions is an important issue to bridging the performance gap.

To understand what causes negative feedback, we further investigate the learning profiles using 1-D and 2-D heatmaps and confirm the positive role of the 2-D heatmap in the semi-supervised setting. The estimated 2-D heatmap is more accurate than the estimated 1-D heatmap at the early training stage. Then, we propose a new training strategy that uses 1-D and 2-D heatmap representations simultaneously to enjoy the advantages of both sides, namely HybridMatch. The proposed model is built upon FixMatch [10]; it learns unlabeled data via consistency regularization between weakly-augmented and strongly-augmented samples and labeled data via conventional cross-entropy loss. Then, it utilizes the high-resolution 1-D representation to reduce the quantization errors in the heatmap. The low-resolution 2-D representation plays a central role in facilitating semi-supervised learning. To enjoy the advantages of both 1-D and 2-D representation, we employ the curriculum learning strategy. It focuses only on the feedback from the 2-D heatmap at the beginning of training, and gradually increases the feedback from the 1-D heatmap.

The motivation behind our training scheme is two-fold. First, it increases the quality of the pseudo-label at the early training stage with 2-D heatmaps. Second, it exploits higher performance with a high-resolution 1-D heatmap when it is available (i.e., at the end of training). Through this coarse-to-fine approach, our HybridMatch significantly outperforms the existing semi-supervised models on three datasets (26.3% NME improvement in 300-W *full* set at 5% data ratio), even comparable with those of the fully-supervised models.

II. RELATED WORKS

A. SUPERVISED FACIAL LANDMARK DETECTION

Supervised facial landmark detection techniques can be categorized into two groups; (i) coordinate regression methods [12], [13] and (ii) heatmap regression methods [2], [14], [15], [16], [17], [18], [28]. The coordinate regression predicts a normalized landmark coordinate, and the heatmap regression estimates a heatmap per landmark coordinate. Owing to the performance advantages, recent methods adopt heatmap regression for facial landmark detection. Several heatmap-based methods utilize additional geometric constraints to improve performance; Merget et al. [19] using a PCA-based 2-D shape model, LAB [18] exploiting face boundary information, and ODN [20] learning additional weighting from occlusion probabilities. Awing [21] resolves the imbalance between foreground and background on the heatmap and significantly improves the performance. HRNet [11] develops a new model architecture by exploiting high-resolution representation. SLPT [22] proposes a sparse local patch transformer for learning the inherent relation between facial landmarks.

However, heatmap-based methods commonly suffer from large memory complexity because they estimate a 2-D heatmap per landmark coordinate value. Consequently, the heatmap resolution is usually lower than the resolution of the input image in practice. Then, the fractional part of the landmark coordinates is neglected, resulting in severe quantization errors. HIH [23] observes that NME caused by quantization error is even larger than 1/3 of the state-of-the-art item. To tackle the quantization error, Sun et al. [2] exploits the probabilities of all landmarks to estimate the landmark coordinates with fractional parts. DSNT [1] converts discrete 2-D heatmaps into continuous coordinates by adding a differential layer. FHR [3] estimates fractional parts of landmarks by fitting the 2-D Gaussian distribution from samples of a 2-D heatmap. DARK [4] estimates landmarks by approximating the distribution using Taylor expansion. However, when the resolution of a 2-D heatmap is considerably low, it no longer carries the informative spatial distribution. Then, various strategies for estimating the fractional part are often not effective. Recently, Yin et al. [5] introduces the idea of representing a 2-D heatmap via two 1-D heatmaps with a co-attention mechanism and shows that significant performance gain can be achieved by effectively handling the quantization error.

B. SEMI-SUPERVISED FACIAL LANDMARK DETECTION

RCN [8] proposes a multi-task framework, performing both attribute classification and landmark detection. SA [9] employs a data augmentation method by generating style-translated examples to secure more training data. TS^3 [7] utilizes a teacher-students framework. Here, the teacher criticizes the quality of the student-generated samples, and the students are re-trained with the refined pseudo samples via quality filtering. 3FabRec [6] shows that unsupervised generative training captures implicit facial shape information. Then, it sufficiently trains a facial landmark detector with supervised follow-up training, only using small supervised samples.

While our HybridMatch uses pseudo-labels like TS^3, we do not require multiple trainable networks with multi-stage training. More importantly, existing semi-supervised facial landmark detectors focus only on the training-oriented issue (i.e., how to use unlabeled data). That is, they do not consider task-oriented issues such as quantization errors in a semi-supervised setting, leading to sub-optimal performance. To the best of our knowledge, we are the first to argue that semi-supervised landmark detection should resolve the task-oriented issue as well as the training-oriented issue. In this work, we propose an integrated 1-D and 2-D heatmap representation and an effective training strategy for semi-supervised facial landmark detection, which effectively utilizes unlabeled data and tackles quantization errors at the same time.

III. METHODS

A. PRELIMINARY: FixMatch

According to the semi-supervised learning scenario, the training dataset can be partitioned into a labeled set $\{x_s, y_s\} \sim \mathcal{D}_s$ and an unlabeled set $\{x_u\} \sim \mathcal{D}_u$. Here, x_s and x_u are image data and y_s is a corresponding label of x_s. Following the convention, the model learns $\mathcal{D}_s$ in the same way as a fully-supervised model. For unlabeled data $\mathcal{D}_u$, one of the common approaches is to extract the pseudo-label $\hat{y}_u = PL(x_u)$ via the pseudo-label extractor $PL(\cdot)$ and then guide the model training using $\{x_u, \hat{y}_u\}$. Recently, FixMatch [10] shows impressive performance improvement in the semi-supervised image classification task. The method selectively uses pseudo-labels and exploits consistency regularization to handle unlabeled data. Specifically, FixMatch generates high-confidence one-hot pseudo-labels from weakly-augmented unlabeled data $\hat{y}_u = PL(T_w(x_u))$. Then, it trains the model $f_\theta(\cdot)$ parameterized by θ by mapping strongly-augmented unlabeled data $T_s(x_u)$ to $\hat{y}_u$, where $T_w(\cdot)$ and $T_s(\cdot)$ are weak and strong data augmentation polices, respectively. Finally, the supervised data loss $\mathcal{L}_s$ and the unsupervised data loss $\mathcal{L}_u$

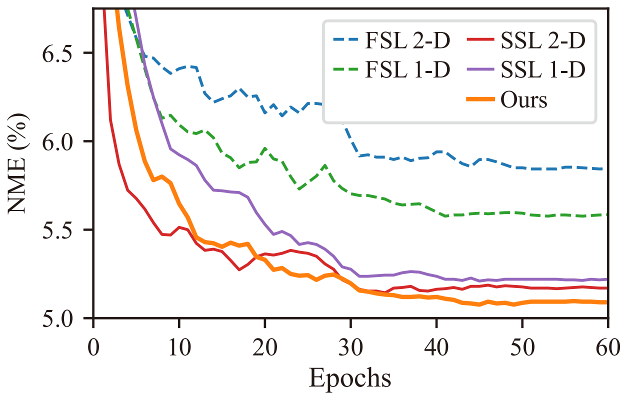

FIGURE 2. Convergence trends of fully- and semi-supervised scenarios using NME across training epochs. Interestingly, SSL 2-D shows faster convergence than SSL 1-D. Here, FSL uses 10% of labeled data, while SSL uses 10% of labeled data with 90% of unlabeled data. All results are based on HRNet.

of FixMatch are expressed as follows:

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}_s|} \sum_{\{x_s, y_s\} \sim \mathcal{D}_s} d\left(f_\theta(x_s), y_s\right), \tag{1}$$

$$\mathcal{L}_u = \frac{1}{|\mathcal{D}_u|} \sum_{x_u \sim \mathcal{D}_u} d\left(f_\theta(T_s(x_u)), PL(T_w(x_u))\right), \tag{2}$$

where $d(\cdot)$ is an error metric. The final objective $\mathcal{L}$ is,

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u, \tag{3}$$

where $\lambda_u(\geq 0)$ is a weighting parameter of the unsupervised loss. In the following, we develop the proposed model based on the semi-supervised learning framework of FixMatch for handling unlabeled data. (For a fair comparison, the same FixMatch framework is employed for semi-supervised competitors.)

B. MOTIVATION

Semi-supervised facial landmark detection inherits two performance bottlenecks; (i) limited representation power due to low-resolution heatmap (i.e., task-oriented bottleneck), and (ii) noisy pseudo-labels in semi-supervised learning [7], [24] (i.e., training-oriented bottleneck). Presuming the solution for each issue does not affect the other, one may introduce a 1-D heatmap representation into the FixMatch framework, collectively choosing the state-of-the-art methods for each side. As shown in Figure 1, we observe that the 1-D heatmap-based method in the semi-supervised setting has *less performance gain than the 2-D heatmap-based method, unlike the fully-supervised setting*.

From the counter-intuitive observation, we further investigate why the 2-D heatmap-based method is more suitable for the semi-supervised setting. For that, we compare the convergence trends under different fully-supervised settings based on HRNet [11], and semi-supervised settings based on FixMatch [10] using the same feature extractor [11]. Figure 2 shows that the 2-D heatmap-based method converges faster than the 1-D-based method. We conjecture that the different convergences are induced by the different levels of representation power. By definition, the 2-D heatmap inherently encodes the relationship between x-y coordinates and thus can reveal the relationship naturally in the heatmap output. On the other hand, the 1-D heatmap-based method ignores the dependency along x-y coordinates to simplify the representation. Although a co-attention module is used to restore their relationships, gaps in convergence speed are inevitable. That is, the 1-D heatmap-based method shows a slower convergence speed than that of the 2-D heatmap-based method.

Slow convergence is critically negative in a semi-supervised setting. Note that, at the early stage of the semi-supervised training, the 2-D heatmap-based method provides more accurate pseudo-labels than those of the 1-D heatmap-based method. This large gap in the early training stage affects even the final performance, especially in the semi-supervised training [25]. The importance of the early training stage is also discussed in Liu et al. [24]. Based on our experiments on convergence trends, we confirm that the state-of-the-art method for reducing the quantization error can provide negative feedback to the state-of-the-art semi-supervised framework. Our method eliminates negative feedback by enjoying fast convergence by 2-D representation and accurate performance by 1-D representation simultaneously. Details of the method will be discussed in the next section.

C. HybridMatch

We propose HybridMatch, utilizing both high-resolution 1-D and low-resolution 2-D heatmap representations. Our key motivation is to eliminate the negative effects between task- and training-oriented solutions in semi-supervised landmark detection, thus enjoying the advantages of both sides. Specifically, the high-resolution 1-D heatmap enables the model to reduce quantization errors. Meanwhile, the low-resolution 2-D heatmap provides more accurate pseudo-labels at the early stage of semi-supervised training. Figure 3 depicts the overall architecture for training unlabeled data. Our model is built upon the HRNet architecture. (i) It first regresses the 2-D heatmap supervised by 2-D pseudo-labels like HRNet. (ii) It estimates the 1-D heatmap via a 1-D heatmap regressor using the estimated 2-D heatmap. For the parameter updates, both the 1-D and 2-D pseudo-labels are used to update all parameters except for the 1-D heatmap regressor. Here, the parameters for the 1-D regressor are updated only with 1-D pseudo-labels. In this way, we enjoy the advantages of both 1-D and 2-D heatmaps because both heatmaps participate in the training loss.

Motivated by our analysis of convergence trends in Section III-B, we rely more on feedback from the 2-D heatmap at the early training stage. For that, we have gradually increased the importance of 1-D heatmap feedback by controlling λ_u^{1D} in Eqn. 5 as the training evolves. Finally,

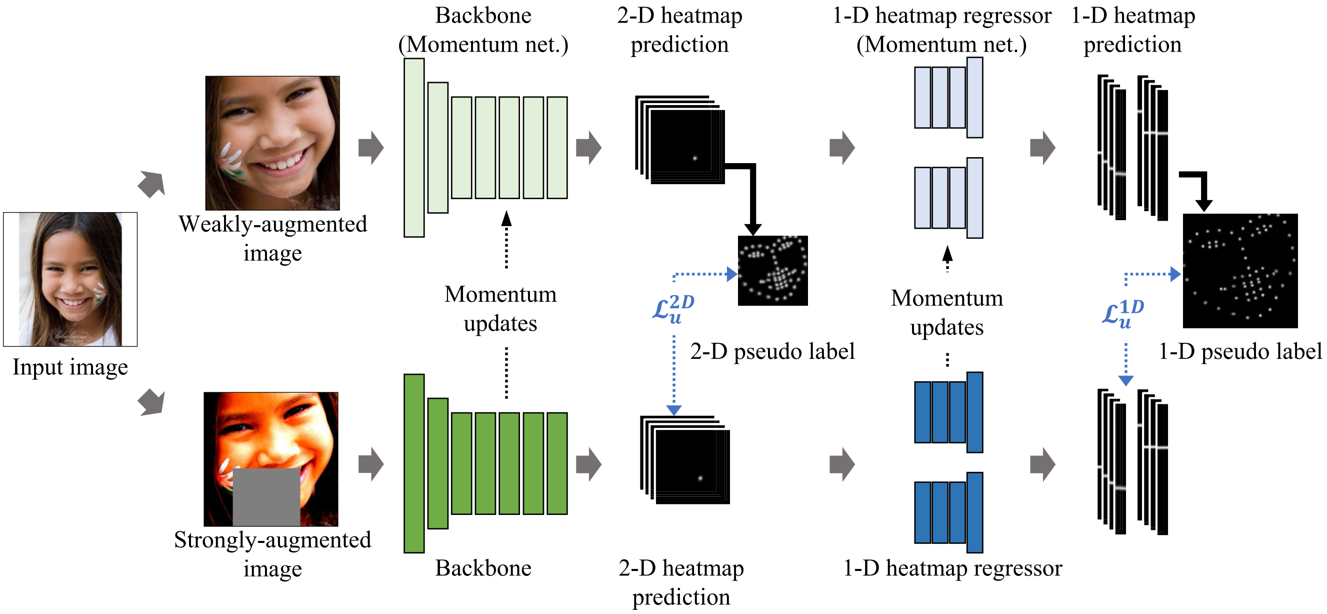

FIGURE 3. Overall framework of HybridMatch for training unlabeled data. $\mathcal{L}_u^{2D}$ and $\mathcal{L}_u^{1D}$ are pseudo-labeling-based 2-D and 1-D heatmap regression feedback, respectively (see Eqn. 5).

the total loss for the labeled data $\mathcal{L}_s$ and the unlabeled data $\mathcal{L}_u$ are written as follows:

$$\mathcal{L}_s = \mathcal{L}_s^{2D} + \lambda_s^{1D} \cdot \mathcal{L}_s^{1D}, \tag{4}$$

$$\mathcal{L}_u = \mathcal{L}_u^{2D} + \lambda_u^{1D} \cdot \mathcal{L}_u^{1D}, \tag{5}$$

where λ_s^{1D} is a constant weighting factor for $\mathcal{L}_s^{1D}$ and λ_u^{1D} is a varying weighting factor for $\mathcal{L}_u^{1D}$. We use linear scheduling for λ_u^{1D} in Eqn. 5, which is defined as follows:

$$\lambda_u^{1D} = \begin{cases} i/K \cdot \lambda^{1D} & \text{if } i < K. \\ \lambda^{1D} & \text{if } K \leq i \leq I, \end{cases} \tag{6}$$

where i is the index of the current training iteration, I is the total iterations, and K is the end iteration when the linear ramp-up ends. With our adaptive training strategy, HybridMatch can receive high-quality feedback at the beginning of the training from the 2-D heatmap representation. Then, it finally has the advantage of high-resolution 1-D heatmap representation that helps reduce quantization errors.

In the following, we describe three training skills to further improve our model. They are data augmentation, a mean teacher framework, and confidence-regularized hard pseudo-labeling.

1) DATA AUGMENTATION

We adopt a data augmentation strategy from Sohn et al. [10] and make several modifications for facial landmark detection. Firstly, to ensure the pixel alignment between pseudo-labels from the weakly and strongly augmented image, we set $T_s(\cdot) = T_s'(T_w(\cdot))$, where T_s' only includes photometric transformation. By doing so, the geometric alignment between $T_w(\cdot)$ and $T_s(\cdot)$ is guaranteed. Specifically, we choose AutoAugment [26] followed by Cutout [27] except for rotation, shear, and translation for $T_s'(\cdot)$.

2) MEAN TEACHER FRAMEWORK

We employ a mean teacher framework [28], widely used in the semi-supervised scenario to improve the training stability and prediction quality. Instead of generating the pseudo-label $\hat{y}_u = PL(T_w(x_u))$ using the model $PL(\cdot) = f_\theta(\cdot)$, we use the model $PL(\cdot) = f_\phi(\cdot)$, where ϕ is an exponential moving average of the previous values in θ throughout the optimization. The mean teacher framework is regarded as a temporal ensemble and leads to stable prediction without an expensive computing cost. Following the convention, the model f_ϕ is used to obtain the pseudo-label of $T_w(x_u)$. The model f_θ provides the predicted heatmap for $T_s(x_u)$. The parameters θ are updated using the loss between the predicted heatmaps and the pseudo-labels.

3) PSEUDO-LABELING

How to use the prediction $f_\phi(T_w(x_u))$ as the pseudo-label (i.e., whether to use the soft or hard label) remains an open question in semi-supervised learning. Many semi-supervised methods use a hard label with confidence-based thresholding for entropy minimization [10], [29]. Confidence-based thresholding can also be applied to facial landmark detection. It selects a point of the highest intensity on the heatmap as the confidence score for the heatmap and then uses it if the confidence is greater than the threshold. However, it is non-trivial to choose an appropriate threshold value for different datasets. Furthermore, we observe that the optimal threshold value should change upon each landmark point. To bypass the reliability issue of thresholding, we use a hard label with confidence regularization as follows:

$$\mathcal{L}_u^m = \frac{1}{|\mathcal{D}_u|} \sum_{x_u \sim \mathcal{D}_u} \mathcal{A}_{conf}^m \cdot d\left(f_\theta^m(T_s(x_u)), \hat{y}_u^m\right),$$
$$\mathcal{L}_u = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_u^m, \tag{7}$$

where $f_\theta^m(\cdot)$, $\mathcal{A}_{conf}^m$ and $\hat{y}_u^m$ are the m^{th} heatmap prediction, its highest intensity value, and its pseudo-label. Here, x_u is the unlabeled input and M is the number of landmarks. Since $\mathcal{A}_{conf} \in [0, 1]$ is the highest value in the heatmap, it naturally penalizes the low-confidence heatmap (i.e., generally having low intensities) but promotes the high-confidence heatmap. In this way, we can focus on the high-confidence heatmap without additional parameters for thresholding.

By conducting experiments on various benchmark datasets, we confirmed that both soft labels and hard labels with confidence regularization are successful on datasets with small variations, such as human faces. Meanwhile, hard labels with confidence regularization are more effective on datasets with large variations, such as caricature faces (see Section IV-F2). To achieve the generalized performances on various datasets, we use hard labels with confidence regularization and then generate pseudo-labels from the prediction $f_\phi(T_w(x_u))$ in all experiments. To render hard labels, we specifically apply the `argmax` operation on the heatmap to obtain coordinate information. Then, we transform the coordinate information to the heatmap by fitting the Gaussian distribution; it is a common protocol, thus identical to all other methods of producing the heatmap label from the coordinate label. We use the generated pseudo-labels as guidance for unlabeled data.

IV. EXPERIMENTAL RESULTS

A. DATASET

300-W is a semi-automatically annotated facial landmark dataset with 68 landmark points, including LFPW [30], AFW [31], HELEN [32], XM2VTS [33], and additional data [34]. We use the same data split as Ren et al. [35], which composes 3,148 training and 689 testing images (*full*). The test split consists of 135 images for a *challenging* subset and 554 images for a *common* subset. For the experiments, we report performances on *challenging*, *common*, and *full* testing sets.

AFLW is a large-scale collection of annotated face images from Flickr, exhibiting a large variety of appearances (e.g., pose, expression, ethnicity, age, and gender). AFLW [36] contains 24,386 images, and we use splits of 20,000 images for training and 4,386 images for testing (*full*). The test split includes 1,165 images for a *frontal* subset. Following the convention as in [37], we use only 19 out of 21 annotated landmarks.

WFLW [18] is a manually annotated facial landmark dataset with 98 landmark points, whose images are sourced from the WIDER FACE dataset [38]. WFLW contains 10,000 faces with 7,500 training images and 2,500 test images. The test split consists of several different test subsets, where each subset varies in pose, illumination, expression, occlusion, make-up, or blur.

WebCari [39] is a large photograph-caricature dataset with 252 identities collected from the web. We composed the WebCari dataset for our work with only caricature images, which vary in artistic styles with 17 landmark points. The WebCari dataset includes a total of 6,042 caricature images with 246 identities and we divide them into 3,942 training images and 2,100 test images. Unlike a real face dataset, the structural information and style information of each image vary significantly in this dataset.

B. EXPERIMENTAL SETUP

1) NETWORK ARCHITECTURE

Our model is based on HRNetV2-W18 [11], which performs 2-D heatmap regression on the input. We follow the same model configuration as in Wang et al. [11]. In order to perform 1-D heatmap regression, we add 1-D heatmap regressor as suggested in [5] at the end of the last layer.

2) IMPLEMENTATION DETAILS

We follow the configuration from HRNetV2-W18 [11], which is widely used in facial landmark detection. All images are cropped and resized to 256 × 256. We choose random horizontal flipping ($p = 0.5$), random rotation (±30°), and random scaling (±25%) for weak data augmentation $T_w(\cdot)$. The total training epoch is 60. We use an Adam optimizer with a linear warmup. The learning rate is 0.0001, where the rate decreases by 0.1 times in the 30th and 50th epochs. The output resolution is 64 × 64 and 256 × 2 for the 2-D and 1-D heatmap, respectively. We randomly sample $r\%$ of data using a fixed seed value and report the best performance out of 3 runs in all experiments. Our experiments are carried out on NVIDIA Titan Xp GPUs.

Curriculum learning. In Eqn. 6, we use linear scheduling for λ_u^{1D}. For the hyper-parameters in the equation, we use $K = 0.1I$ and $\lambda^{1D} = \lambda_s^{1D} = 0.05$ for the rest of the paper.

C. EVALUATION METRIC

1) NME

For facial landmark detection, the mean squared error has a significant limitation in that it neglects the scale of the face. Thus, the normalized mean squared error (NME) is widely adopted as an evaluation metric in the literature. NME includes a normalization factor of L, which is often defined as the distance between eyes and is defined as follows:

$$\text{NME} = \frac{1}{M} \sum_{m=1}^{M} \frac{||p_1^m - p_2^m||_2}{L}, \tag{8}$$

where p_1^m and p_2^m are landmark coordinate points and M is the number of landmarks. We use *Inter-ocular* norm for 300-W and WFLW datasets, which defines L as the outer-eye-corner

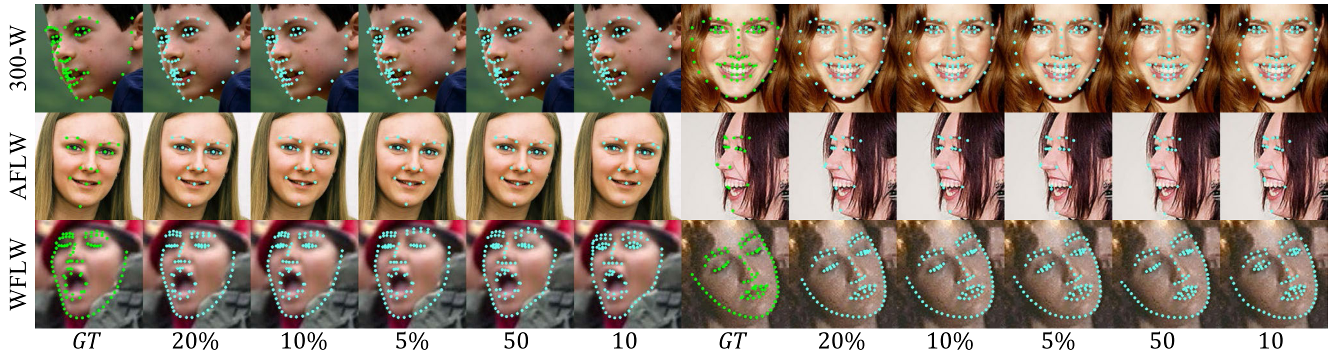

FIGURE 4. Qualitative results on 300-W, AFLW, and WFLW datasets. The second to sixth columns depict predicted landmark points overlaid input images along with the labeled image ratio (%) or the number of images. *GT* indicates ground-truth landmark points overlaid images. Better viewed when zoomed in.

TABLE 1. Upper-bound performance on 300-W, AFLW, and WFLW datasets. "†" and "‡" denote semi-supervised method and fully-supervised method, respectively. "-" denotes unavailable, {100%, 20%} indicate the labeled data ratio, and the results are NMEb↓ (%). We highlight that our HybridMatch achieves comparable performance to the HRNet using a fully labeled dataset.

Method	Setting	300-W			AFLW		WFLW						
		Com.	Chall.	Full	Full	Frontal	Full	Pose	Exp.	Ill.	Mk. Up	Occ.	Blur
SA †	FSL (100%)	3.21	6.49	3.86	-	-	4.39	8.24	4.68	4.24	4.27	5.60	4.86
TS^3 †		3.17	6.41	3.78	-	-	-	-	-	-	-	-	-
3FabRec†		3.36	5.74	3.82	1.84	1.59	5.62	10.23	6.09	5.55	5.68	6.92	6.38
HRNet ‡		2.87	5.15	3.32	1.57	1.46	4.60	7.94	4.85	4.55	4.29	5.44	5.42
SAAT ‡		2.87	5.03	3.29	-	-	-	-	-	-	-	-	-
HybridMatch †	SSL (20%)	2.99	5.04	3.40	1.61	1.50	4.73	7.86	5.01	4.59	4.33	5.46	5.44

distance for quantitative evaluations as following [6], [7]. For AFLW, we define L as the width of the square bounding box following Zhu et al. [40].

2) FAILURE RATE AND AREA UNDER CURVE

Failure Rate (FR_r) indicates the ratio of failed predictions out of the given images. We consider an image that has NME larger than the threshold r as a failed prediction. We use $r = 10\%$ as a threshold.

Area Under Curve (AUC) computes the area of the cumulative error distribution curve ($\mathrm{CED}(x) = 1 - \mathrm{FR}_x$). A larger AUC means higher accuracy and lower sensitivity to the threshold. We evaluate $\mathrm{CED}(x)$ for $x \in [0\%, 10\%]$.

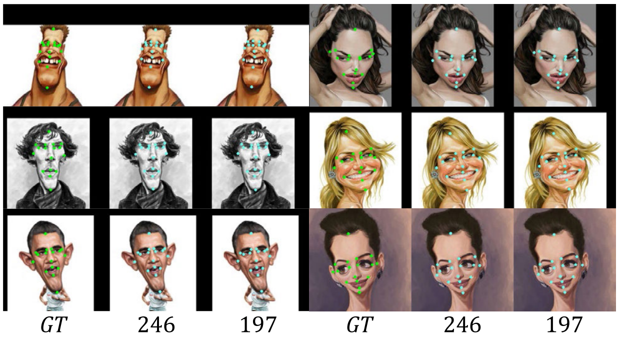

FIGURE 5. Qualitative results on the WebCari dataset. The second and third columns depict the predicted landmarks using 246 and 197 (5%) training images, respectively. For 246 images, each image is sampled from 246 identities of the WebCari dataset. Better viewed when zoomed in.

D. COMPARISON WITH STATE-OF-THE-ART

Table 1 compares the upper-bound performances of existing semi-supervised methods and HRNet. The upper-bound performances of SSL methods are computed with the model trained with the entire training data (100% supervision). From these comparisons, HRNet shows comparable or superior upper-bound performances over existing semi-supervised models, thus our HybridMatch is implemented based on HRNetV2-W18. HRNet is denoted as FSL 2-D throughout this paper, and Figure 1 clearly shows that HybridMatch significantly improves the HRNet baseline.

In Table 2, we compare our method with state-of-the-art semi-supervised models on the 300-W dataset. Our HybridMatch outperforms the existing methods for all data ratios. Semi-supervised models are evaluated under 20%, 10%, 5%, and even extreme ratios such as 1.4% (50 samples) and 0.3% (10 samples). Our method is more robust against data ratio variation than the other methods; the effects of HybridMatch are more pronounced under harsh conditions. For example, 3FabRec shows a performance degradation of 0.93%p (*full* testing set) when it only uses 5% labeled data ratio, compared to the same model using full supervision. On the other hand, HybridMatch shows a performance degradation of 0.18%p under the same setting. It indicates that our method is less sensitive to the size of the training set. Notably, given only 50 labeled samples, HybridMatch outperforms 3FabRec with 100% training data.

TABLE 2. Quantitative results of semi-supervised methods on the 300-W dataset. Each column represents {*common, challenging, full*} testing set. The results are NMEb↓ (%).

Method	20%			10%			5%			50 (1.4%)			10 (0.3%)		
RCN	-	6.12	4.15	-	6.63	4.47	-	9.95	5.11	-	-	-	-	-	-
SA	3.85	-	-	4.27	-	-	6.32	-	-	-	-	-	-	-	-
TS^3	4.31	7.97	5.03	4.67	9.26	5.64	-	-	-	-	-	-	-	-	-
3FabRec	3.76	6.53	4.31	3.88	6.88	4.47	4.22	6.95	4.75	4.55	7.39	5.10	4.96	8.29	5.61
HybridMatch	**2.99**	**5.04**	**3.40**	**3.03**	**5.09**	**3.44**	**3.10**	**5.16**	**3.50**	**3.30**	**5.64**	**3.76**	**3.70**	**6.37**	**4.22**

TABLE 3. Quantitative results of semi-supervised methods on AFLW dataset. Each column represents {*full, frontal*} testing set, respectively. The results are NMEb↓ (%).

Method	20%		10%		5%		1%		50 (0.25%)		10 (0.05%)	
RCN	-	-	-	-	2.17	-	2.88	-	-	-	-	-
TS^3	1.99	1.86	2.14	1.94	2.19	2.03	-	-	-	-	-	-
3FabRec	1.96	1.74	2.03	1.74	2.13	1.86	2.38	2.03	2.74	2.23	3.05	2.56
HybridMatch	**1.61**	**1.50**	**1.66**	**1.56**	**1.68**	**1.57**	**1.77**	**1.62**	**1.92**	**1.73**	**2.29**	**2.18**

TABLE 4. Quantitative results of semi-supervised methods on WFLW *full* testing set. The results are NMEb↓ (%).

Method	20%	10%	5%	50	10
SA	6.00	7.20	-	-	-
3FabRec	6.51	6.73	7.68	8.39	9.66
HybridMatch	**4.73**	**4.85**	**5.04**	**5.47**	**7.11**

TABLE 5. Area Under Curve (AUC) and Failure Rate (FR$_{10\%}$) on the 300-W test set.

Method	Setting	AUC	FR
M^{3}CSR[41]	FSL (100%)	47.52	5.50
CFSS [40]		49.87	5.05
DenseReg+MDM [42]		52.19	3.67
JMFA [43]		54.85	1.00
LAB [18]		58.85	0.83
3FabRec[6]		54.61	**0.17**
HybridMatch (ours)	SSL (20%)	**60.56**	**0.17**

Table 3 summarizes NME scores on the AFLW dataset. HybridMatch records the robust performances across varying data ratios on the AFLW dataset as it is on the 300-W. More importantly, even with only 1% annotation labels, we achieve 1.77/1.62 (*full* and *frontal* testing set), which is comparable to the accuracy of 3FabRec using full supervision (1.84/1.59).

Table 4 shows NME scores on the WFLW, which is considered the most challenging dataset. Our HybridMatch still outperforms all existing methods with large gaps. In particular, SA reports outstanding performance on the WFLW *full* testing set, even outperforming HRNetV2-W18 in the fully-supervised setting of Table 1. However, in the semi-supervised setting, our method performs remarkably better than SA with a 2.35%p gain on 10% labeled data ratio.

To investigate the effectiveness of our HybridMatch, we compare the proposed method with other fully-supervised models using FR and AUC metrics. We evaluate the methods on the 300-W test set. Table 5 shows that our HybridMatch with only 20% of labeled data achieves the best results in both measurements. Our FR indicates that only one image out of the full 300-W test set has a larger NME than the threshold. Furthermore, our HybridMatch outperforms 3FabRec in AUC at a high margin. This result indicates that our model shows accurate results with low deviation. Note that our HybridMatch depicts $\text{FR}_{10\%} = 0.67$ and $\text{AUC} = 56.33$ only with 50 labeled images training.

E. QUALITATIVE RESULTS

Figure 4 visualizes our landmark prediction results on the 300-W, AFLW, and WFLW datasets. Although NME increases by reducing the labeled data, we observe that our predicted landmark points are sufficiently close to ground-truth landmark points. Besides, the predictions are generally robust against facial orientation, expression, and occlusion. In the second row on the left image in the WFLW dataset, we find that the predicted facial contour is imprecise when only 10 samples are labeled. This is because the labeled samples are extremely few, thus the model is incapable of learning challenging cases, such as occluded or rotated faces. In general, our model achieves compelling quality at a 20% labeled data ratio, which is on par with the fully supervised model. Considering the impressive results and the computational efficiency, HybridMatch can serve as a good alternative to fully-supervised facial landmark detection.

Figure 5 shows landmark prediction results on the WebCari dataset. Although WebCari is a challenging dataset with high variation (i.e., diverse shapes and textures), the proposed model successfully provides accurate predictions. In the

TABLE 6. Effects of the heatmap representation. All results are based on HybridMatch architecture. The results are NMEb↓ (%) on the 300-W dataset. FSL and SSL indicate fully- and semi-supervised learning, respectively. FSL uses only 10% of the training dataset.

Setting	Ratio	Com.	Chall.	Full
FSL 2-D	10%	3.21	6.07	3.77
FSL 1-D	10%	3.13	5.55	3.60
SSL 2-D	10%	3.12	5.22	3.53
SSL 1-D	10%	3.07	5.23	3.50
HybridMatch	10%	**3.03**	**5.09**	**3.44**

TABLE 7. Effects of pseudo-label type. All results are based on HybridMatch architecture. The results are NMEb↓ (%) on 300-W (*full*) and WebCari datasets.

	300-W	WebCari
Ratio	10%	246 (6.2%)
Hard label	**3.41**	**6.43**
Soft label	3.44	6.69

image of the first row on the left, we observe that the predicted forehead point is inaccurate. However, since the corresponding subject is bald, the human also suffers from pinpointing the accurate forehead point. Considering the difficulty of the task, we conclude that our model is fairly robust against the highly varying dataset.

F. ABLATION STUDY

1) EFFECTS OF HEATMAP REPRESENTATION

Figure 1 and Table 6 exhibit the effects of the heatmap representations on the 300-W dataset. In a fully-supervised setting based on [11], the 1-D heatmap representation [5] (FSL 1-D) outperforms the 2-D heatmap representation [11] (FSL 2-D). This is expected because the high-resolution 1-D heatmap helps reduce quantization errors. However, in the semi-supervised setting based on [10], the 2-D heatmap provides more accurate pseudo-labels from the early training stage. Thus, the performance gap between the 2-D heatmap-based method (SSL 2-D) and the 1-D heatmap-based method (SSL 1-D) is significantly reduced. On the other hand, our method utilizes 1-D and 2-D heatmap representations simultaneously to have the advantages of both sides, improving the final performance.

2) EFFECTS OF PSEUDO-LABELING METHOD

Table 7 compares the accuracy of the pseudo-labeling methods. We observe that both the hard labeling and soft labeling methods as $PL(\cdot)$ show no significant difference on the dataset with low geometric variation such as 300-W. However, we observe a considerable performance gap in high geometric variation datasets such as WebCari. Based on our observation, we conjecture that the soft pseudo-label under the dataset with high geometric variation tends to produce a low-confidence heatmap, which is blurry and distorted. Then, the key point no longer obeys the Gaussian distribution. This misleads the feedback from unlabeled data, thus resulting in performance degradation.

G. MEMORY USAGE AND MODEL PARAMETERS

Since HybridMatch uses both 1-D and 2-D heatmaps, one might consider the increase in model parameters as a negative side effect. However, compared to the 1-D based model [5], our model only adds two convolutional layers with batch normalization (for regressing a 2-D heatmap), thus the increase in parameters is negligible. Furthermore, as our model provides feedback from the 2-D representation, we reduce the network size of the 1-D heatmap branch without much performance drop. As a result, the total weight of our model is 11.23M while the 1-D-based model [5] is 16.44M. Our memory cost is much lower than that of the state-of-the-art semi-supervised model [6] (25.97M). Another computational factor is the memory capacity during training. Since the 2-D heatmap requires much more capacity, increasing the 2-D resolution can incur out-of-memory issues. Although our hybrid representation inherits the same memory capacity issue, adding the 1-D heatmap is marginal in terms of memory capacity. Overall, our hybrid representation does not consume many model parameters over the 1-D heatmap model [5] and memory capacity compared to the 2-D heatmap model [11].

V. CONCLUSION

We propose an effective semi-supervised facial landmark detection framework via hybrid representation, namely HybridMatch. This paper first identifies that we should consider both task-oriented (i.e., quantization error) and training-oriented (i.e., noisy pseudo-labels) issues simultaneously when tackling a semi-supervised landmark detection problem. To this end, we propose HybridMatch for simultaneously mitigating the performance bottlenecks caused by quantization error and noisy pseudo-labels. Specifically, our HybridMatch utilizes the high-resolution 1-D heatmap representation for reducing quantization error and the low-resolution 2-D heatmap for facilitating the fast convergence of semi-supervised learning. Extensive evaluations demonstrate the effectiveness of our HybridMatch and the outstanding performances, the new state-of-the-art accuracies for semi-supervised facial landmark detection on 300-W, AFLW, and WFLW datasets. Concretely, our method achieves 26.3% NME improvement over the existing method in 300-W *full* set at 5% data ratio. Our HybridMatch can capture more precise facial landmark points than existing methods with only a few annotated data (e.g., even when training with only 10 annotation labels). More importantly, HybridMatch achieves comparable performance, 2.99/5.04/3.40 (*common*, *challenging* and *full* testing set in 300-W) to the fully-supervised facial landmark detector (2.87/5.03/3.29) even using 20% labeled data.

ACKNOWLEDGMENT

Seoungyoon Kang and Minhyun Lee contributed equally to this work.

REFERENCES

[1] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," 2018, *arXiv:1801.07372*.

[2] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 529–545.

[3] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen, "Towards highly accurate and stable face alignment for high-resolution videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8893–8900.

[4] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7093–7102.

[5] S. Yin, S. Wang, X. Chen, E. Chen, and C. Liang, "Attentive one-dimensional heatmap regression for facial landmark detection and tracking," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 538–546.

[6] B. Browatzki and C. Wallraven, "3FabRec: Fast few-shot face alignment by reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6110–6120.

[7] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 783–792.

[8] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1546–1555.

[9] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10153–10163.

[10] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 596–608.

[11] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, and Y. Zhao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[12] H. Liu, J. Lu, J. Feng, and J. Zhou, "Two-stream transformer networks for video-based face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2546–2554, Nov. 2018.

[13] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[14] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.

[15] L. Chen, H. Su, and Q. Ji, "Deep structured prediction for facial landmark detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2450–2460.

[16] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1831–1840.

[17] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[18] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2129–2138.

[19] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 781–790.

[20] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3486–3496.

[21] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6971–6981.

[22] J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, "Sparse local patch transformer for robust face alignment and landmarks inherent relation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4052–4061.

[23] X. Lan, Q. Hu, and J. Cheng, "Revisting quantization error in face alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1521–1530.

[24] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 20331–20342.

[25] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," 2020, *arXiv:2001.06001*.

[26] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 113–123.

[27] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[28] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–7.

[29] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," 2017, *arXiv:1706.05208*.

[30] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

[31] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 679–692.

[32] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[33] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, vol. 964, 1999, pp. 965–966.

[34] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.

[35] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.

[36] 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 *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

[37] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3317–3326.

[38] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1281–1290.

[39] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "Webcaricature: A benchmark for caricature recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.

[40] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4998–5006.

[41] J. Deng, Q. Liu, J. Yang, and D. Tao, "M^3 CSR: Multi-view, multi-scale and multi-component cascade shape regression," *Image Vis. Comput.*, vol. 47, pp. 19–26, Mar. 2016.

[42] R. A. Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "DenseReg: Fully convolutional dense shape regression in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6799–6808.

[43] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, "Joint multi-view face alignment in the wild," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3636–3648, Jul. 2019.

SEOUNGYOON KANG (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree with the School of Integrated Technology. He is currently a Contract Researcher with Kim Jaechul Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology (KAIST), South Korea. His research interests include computer vision and machine learning. In particular, he is interested in manipulating images using generative models, such as generative adversarial networks.

MINHYUN LEE received the B.S. degree in computer science from Yonsei University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree with the School of Integrated Technology. His research interests include computer vision and machine learning. In particular, he is interested in weakly-supervised semantic segmentation and active learning.

MINJAE KIM received the B.S. and Ph.D. degrees in electrical engineering from Korea University, South Korea, in 2007 and 2015, respectively. He was a Senior Researcher of autonomous driving technology with LG Electronics, from 2015 to 2017. He joined NCSOFT, in 2017, and leading the Vision AI Laboratory to research on large-scale multimodal models and generative models for contents creation technology. His research interests include generative models (GAN, diffusion) and vision-language tasks based on a large language model.

HYUNJUNG SHIM received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2002, and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2004 and 2008, respectively. From 2008 to 2013, she was with the Samsung Advanced Institute of Technology, Samsung Electronics Company Ltd., Suwon, South Korea. She was an Assistant/Associate Professor with the School of Integrated Technology, Yonsei University, from 2013 to 2022. She is currently an Associate Professor with Kim Jaechul Graduate School of Artificial Intelligence, Korea Advanced Institute of Science and Technology. Her research interests include encompass generative models, deep neural networks, classification/recognition algorithms, 3D vision, inverse rendering, face modeling, and medical image analysis.