IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 22, 2020, accepted January 5, 2021, date of publication January 18, 2021, date of current version February 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052460

Multistage Probabilistic Approach for the Localization of Cephalometric Landmarks

HYUK JIN KWON[1], (Graduate Student Member, IEEE), HYUNG IL KOO[2], (Member, IEEE), JAEWOO PARK[1], (Student Member, IEEE), AND NAM IK CHO[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, INMC, Seoul National University, Seoul 08826, South Korea
[2]Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea

Corresponding author: Hyung Il Koo (hikoo@ajou.ac.kr)

This work was supported by the Ministry of Trade, Industry, and Energy (MOTIE), South Korea, through the ''Regional Innovation Cluster Development Program'' (Research and Development) supervised by the Korea Institute for Advancement of Technology (KIAT) under Grant P0002072.

ABSTRACT The accurate and reproducible localization of cephalometric landmarks is an important procedure for treatment planning and clinical practice in orthodontics and maxillofacial surgery. In this paper, we propose a new multistage cephalometric landmark localization method that exploits local appearances and global characteristics simultaneously. To be precise, a convolutional neural network(CNN) is trained by minimizing the sum of all landmark errors. Since landmarks are considered simultaneously, global hard/soft tissue characteristics, as well as landmark relations, can be reflected in this stage. Then, we exploit local appearances by using high-resolution cropped images. In this second stage, we train CNNs for individual landmarks, respectively. Finally, we improve the localization performance of cephalometric landmarks of the mandible with linear estimators. Experiments on ISBI2015 dataset have shown that the proposed method outperforms conventional methods. Also, the proposed method allows us to evaluate confidence (e.g., standard deviational ellipses) due to its probabilistic formulation.

INDEX TERMS Cephalometric landmark detection, cephalometry, dental radiography.

I. INTRODUCTION

Quantitative cephalometry is an essential step in orthodontics and orthognathic surgery. The accurate localization of cephalometric landmarks in lateral cephalograms allows us to classify anatomic facial types and growth patterns, which is crucial to the treatment planning and clinical practice. However, manual placements of landmarks are time-consuming and often suffer from intra-examiner and inter-examiner differences [1], [2]. Therefore, the demand for reliable and reproducible automatic detection of cephalometric landmarks has been increasing [3].

To this end, Wang *et al.* [4] presented a cephalometric landmark detection benchmark (ISBI2015 dataset). In this dataset [4], there are two kinds of landmarks: (a) *anatomic landmarks* and (b) *derived landmarks*. The former landmarks correspond to anatomic structures and the latter are defined from neighboring anatomic structures. To localize both types of landmarks, neighboring regions and characteristics of related landmarks should be considered [5]. Cephalometric landmarks can also be classified into bilateral and unilateral landmarks depending on whether they appear on both sides or not [5]. Due to the bilateral nature and asymmetric growth of mandible, it is common that the bilateral landmarks of the mandible do not coincide in the lateral cephalograms [6]. Clinically, positions of bilateral landmarks are defined as the mid-point of landmarks on both sides, however, their estimation is difficult due to high inter-examiner and intra-examiner variations [7]. Automated methods also suffer from inaccurate localization [4]. Actually, the difficulty in the localization of *gonion* (landmark 10, one of the bilateral landmarks of the mandible) has been reported in [3], [8]. This is usually caused by asymmetry of the mandible, and we need to consider related anatomic structures for the accurate localization.

Like other cases of image-based tasks, deep learning methods are providing ever-increasing performance on the above mentioned benchmark. The CNN-based methods usually formulate the detection of cephalometric landmarks as a regression problem that estimates pre-defined heatmaps [9]–[13]. To be precise, CNNs are trained to predict heatmaps rather than landmarks (i.e., heatmap regression), since non-differentiable argmax operators are employed to get landmarks from heatmaps [14]. In this approach, the ground truth heatmap is determined in ad-hoc manners

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

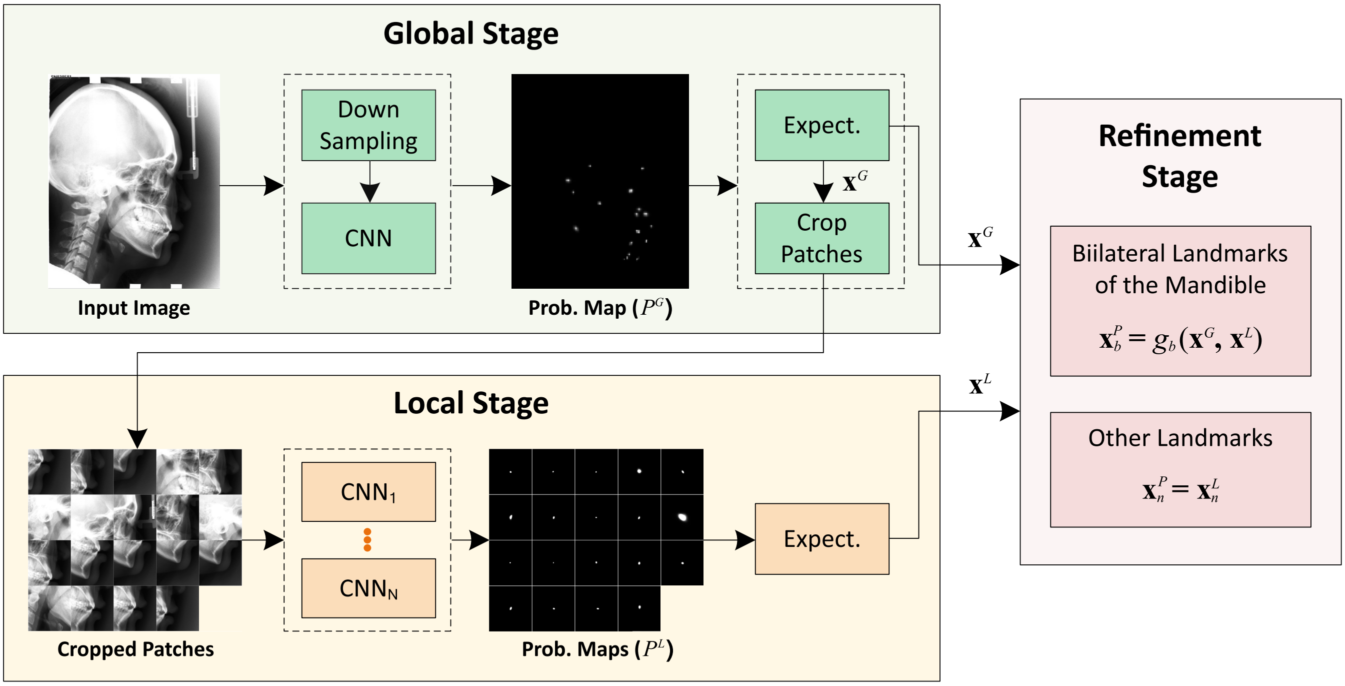


FIGURE 1. Overview of the proposed multistage probabilistic method. The global and local stages estimate probability density functions P^G and P^L of landmark positions using global hard/soft tissue characteristics and local details of anatomical structures respectively. The landmarks positions from the global stage ($\mathbf{x}^G$) and the local stage ($\mathbf{x}^L$) are given by the expectations (Expect.) of P^G and P^L. The refinement stage improves the positions of bilateral landmarks of the mandible ($\mathbf{x}_b^P$) with linear estimators ($g_b(\cdot,\cdot)$) using $\mathbf{x}^G$ and $\mathbf{x}^L$.

and the probabilistic interpretation of predicted heatmap and objective functions is not clear [8].

To alleviate these problems, we develop a new multistage probabilistic approach. The proposed method trains the positions of cephalometric landmarks by using the distance measure between ground truth landmarks and estimated ones, which is achieved by formulating neural network outputs as the probability density functions of landmark locations as in [14]. As illustrated in Fig. 1, we first estimate landmark positions by training a CNN that yields the locations of all landmarks from down-sampled input images [14], [15]. Since we can consider whole images in this stage, the network can learn global characteristics and landmark relations. Then, we predict individual landmark positions using local but high-resolution images (i.e., images are cropped based on the estimation results of the first stage). Since we focus on local appearances in this stage, we can train CNNs for individual landmarks independently.

In the final stage, we address the challenges on bilateral landmarks of the mandible. We can estimate the growth pattern of cranium and mandible from cephalometric landmarks (as in the clinical measurement methods for classifications of anatomical facial types) [6], [16], [17]. Therefore, we believe that other landmarks can be used to improve the localization performance of bilateral landmarks of the mandible, i.e., *gonion* (landmark 10) and *articulare* (landmark 19), and develop a linear filter to exploit the information.

In the experiment, we show that our method achieves state-of-the-art performance on the ISBI2015 benchmark [4]. Specifically, the proposed method achieves mean radial errors of 1.12 mm on *Test1* and 1.41 mm on *Test2*. Also, it shows the best successful detection rate 77.16% and 84.74% for 2.0 mm and 2.5 mm thresholds in the *Test2*, respectively. Also, we apply our detection results to clinical measurement methods for the anatomic facial type classification, which classifies anatomic facial types into several kinds, and obtain the comparable results to existing methods. Since our method is a probabilistic approach, we can have confidence regions for each prediction (e.g., standard deviational ellipses) [18].

II. RELATED WORK

Numerous methods have been proposed to localize cephalometric landmarks and it is beyond the scope of this paper to review all these techniques. Rather, we focus on methods that were published after the release of ISBI2015 dataset [4].

A. RANDOM FOREST-BASED APPROACH

The method of Lindner and Cootes [19] was based on the random forest regression voting and constrained local models [20]–[23]. They first trained statistical shape models by re-sampling target images in a standardized reference frame. After the standardization of images, Haar features were extracted with random displacements from annotated landmark positions and random forest regressors were trained to predict the most likely positions of landmarks. In the test, trained regressors generate initial positions of landmarks and multiple predictions are followed to vote for the final landmark positions.

Ibragimov *et al.* [24] trained random forest regressors to get posterior probabilities of landmark positions. The candidates for each landmark are given by the local maxima of estimated posterior probabilities. They used a game-theoretic framework by considering landmarks as players, candidate points as player strategies, and likelihoods as player payoffs. The positions of landmarks are determined so as to maximize the total payoff of all players. The whole estimation process can be iterated for further improvement [25].

B. DEEP LEARNING-BASED APPROACH

The majority of deep learning based approaches formulated the detection of the cephalometric landmarks as the heatmap regression, where heatmaps are usually shaped with Gaussian or Laplace distribution functions. From regressed heatmaps, the estimation of landmark positions is obtained by applying argmax operations.

Zhong *et al.* [26] used a two-stage approach. In the first stage, they calculated coarse candidate positions in down-sampled input images by using U-Net. In the second stage, a set of U-Nets compute refined heatmaps of landmarks using high-resolution patches around coarse estimations. Lee *et al.* [8] attempted to use confidence maps in this two stage approach. The first stage gives the rough estimates of landmark positions using down-sampled input images. Then, a set of Bayesian CNNs [27] estimate uncertainties of points in the high-resolution patches. They proposed Score Weighting Method that computes confidence scores of each point and estimated landmark positions with a confidence-weighted average. Multiple stages were used in the work of Gilmour and Ray [28]. They constructed multi-scale features by stacking the outputs of a trained CNN for multiple scales. Then, multilayer perceptrons (MLPs) are employed to predict landmark positions from multi-scale features respectively. They iterated the same procedure for 10 times by using predictions from MLPs as new initial positions.

The work of Qian *et al.* [11], called CephaNN, is based on the multi-attention mechanism. The first part of CephaNN computes two sets of features with a multi-headed structure. Then, the attention weights are computed from features, and the weighted features are concatenated to estimate heatmaps for landmarks. CephaNN used bottleneck blocks [29] and final landmark positions obtained by applying argmax operators to heatmaps. Oh *et al.* [13] attempted to improve the attention mechanism by considering anatomic structures. They argued that angles and distances between landmarks reflect anatomical properties, and defined anatomical context weight (ACW) as the sum of differences between predicted landmarks and ground-truth. Then, an attentive U-Net [30] is trained to minimize a weighted L_2 loss between predicted heatmaps and ground-truth heatmaps by using ACW as weight factors.

III. PROPOSED METHOD

As illustrated in Fig. 1, we first detect all landmarks with a single network and refine individual landmarks using cropped images in the second stage. This two-step approach allows us to consider local details as well as their relations, and works well for most landmarks as a result. However, there are challenges in localizing bilateral landmarks of the mandible. We address this problem by re-estimating these landmarks using all current landmark estimates (obtained in the global and local stages) [6], [31]. In this estimation, we use a simple linear filter to avoid overfitting (due to the limited number of training samples). Unlike conventional methods, our estimation method is based on the probabilistic modeling of landmarks, which naturally enables the probabilistic interpretation of the results.

A. GLOBAL STAGE

We estimate the landmark positions by training a CNN that yields heatmaps (probability density functions) for all landmarks. This stage is designed such that we can consider global hard and soft tissue characteristics. To be precise, we train this CNN to produce $P^G(\cdot)$ from a down-scaled input image, so that the estimated locations of landmarks are computed with an expectation:

$$\mathbf{x}_k^G = \int \mathbf{x}\, P_k^G(\mathbf{x}) d\mathbf{x} \in \Re^2, \tag{1}$$

where k is a landmark (channel) index, and we consider $\mathbf{x}$ as a two-dimensional location vector [14], [15]. To learn the global statistical relationship between landmarks, we train the CNN by using a loss function considering all landmark errors:

$$L^G = \sum_{k=1}^{N} \|\mathbf{x}_k^G - \mathbf{x}_k\|_2^2, \tag{2}$$

where $\mathbf{x}_k$ is the ground-truth position of the k-th landmark and $N = 19$ is the number of landmarks. The whole process is differentiable and we can make the CNN produce all landmark locations at the same time.

B. LOCAL STAGE

From the previous step, we have estimations of landmark positions. In the local stage, we refine them using the local details of anatomical structures with a set of CNNs: Each CNN yields one landmark using a high-resolution cropped input image. To train a CNN that focuses on the local details around each landmark, we generate training data by augmenting original training samples. Given the ground-truth annotation of the k-th landmark, we build these training samples by cropping square patches with random perturbations (Details will be presented in the experimental section).

Similar to (1), the estimated position ($\mathbf{x}_k^L$) is given by an expectation:

$$\mathbf{x}_k^L = \int \mathbf{x}\, P_k^L(\mathbf{x}) d\mathbf{x} \in \Re^2. \tag{3}$$

and we use L_2 loss in the training,

$$L_k^L = \|\mathbf{x}_k^L - \mathbf{x}_k\|_2^2, \tag{4}$$

Here, we focus on local details and each CNN is trained independently. During the inference, a square patch whose center is $\mathbf{x}_k^G$ and side length is s is cropped and fed into a corresponding CNN, and the final landmark position is estimated by (3). These results are used as final estimates except for bilateral landmarks of the mandible.

TABLE 1. Definitions and criteria of 8 clinical measurement methods for the anatomic facial type classification used in successful classification rate (SCR). We use following abbreviations: MP = mandibular plane, PP = palatal plane, FH = Frankfort horizontal plane and FP = facial plane.

Method	Definition
ANB	The angle between point A, nasion and point B.
SNB	The angle between sella, nasion and point B.
SNA	The angle between sella, nasion and point A.
ODI	The arithmetic sum of the angle between the AB plane to MP and the angle of the PP to FH.
APDI	The arithmetic sum of the angle between FH and FP, the angle between FP and AB plane and the angle between FH and PP.
FHI	The ratio of the Posterior Face Height to the Anterior Face Height.
FMA	The angle between the line from sella to nasion and the line from gonion to gnathion.
MW	The distance between upper incisal incision and lower incisal incision.

Method	Type 1	Type 2	Type 3
ANB	Class I (Normal): 3.2° ∼ 5.7°	Class II: > 5.7°	Class III: < 3.2°
SNB	Normal mandible: 74.6° ∼ 78.7°	Retrognathic madible: < 74.6°	Prognathic mandible: > 78.7°
SNA	Normal mandible: 79.4° ∼ 83.2°	Prognathic maxilla: > 83.2°	Retrognathic maxilla: < 79.4°
ODI	Normal mandible: 74.5° ± 6.07°	Deep bite tendency: > 80.5°	Open bite tendency: < 68.4°
APDI	Normal: 81.4° ± 3.8°	Class II tendency: < 77.6°	Class III tendency: > 85.2°
FHI	Normal: 0.65 ∼ 0.75	Short face tendency: > 0.75	Long face tendency: < 0.65
FMA	Normal: 26.8° ∼ 31.4°	Mandible high angle tendency: > 31.4°	Mandible lower angle tendency: < 26.8°
MW	Normal: 2 mm ∼ 4.5 mm	Edge to edge: 0 mm, Anterior cross bite: <0 mm	Large over jet: > 4.5 mm

C. REFINEMENT STAGE

We further refine the positions of bilateral landmarks of the mandible by applying linear filters to $\mathbf{x}^G$ and $\mathbf{x}^L$. We believe these points have information to estimate asymmetric growth of the mandible and compensate its effect [16], [31]. One might think that using deep neural networks (DNNs) is a desirable choice, because DNNs have more expressive powers than simple linear filters. However, linear filters have intrinsic interpretability [33] and we have found that they show the best validation errors, probably due to the small number of training samples and the lack of valid augmentation methods for this problem.

When we estimate the position of the b-th bilateral landmark of the mandible ($\mathbf{x}_b^P$), we train a linear estimator $g_b(\cdot, \cdot)$:

$$\mathbf{x}_b^P = g_b(\mathbf{x}^G, \mathbf{x}^L) = [\boldsymbol{\theta}^x, \boldsymbol{\theta}^y]^\top \mathbf{x}^V \tag{5}$$

where $\boldsymbol{\theta}^x, \boldsymbol{\theta}^y \in \Re^{4N}$ are trainable parameters and $\mathbf{x}^V \in \Re^{4N}$ is a vectorized representation of $\mathbf{x}^G$ and $\mathbf{x}^L$. To represent bilateral landmarks of the mandible as the internal divisions of landmark estimations $\mathbf{x}^G$ and $\mathbf{x}^L$, we train $g_b(\cdot, \cdot)$ by minimizing

$$L_b^P = \|\mathbf{x}_b^P - \mathbf{x}_b\|_2^2 \tag{6}$$

under the constraints

$$\sum_{k=1}^{N} \boldsymbol{\theta}_k^i = 1, \quad \boldsymbol{\theta}_k^i \geq 0 \tag{7}$$

for $i \in \{x, y\}$. Intuitively, (7) can be considered a regularization term.

IV. EXPERIMENTS

We evaluate the performance of our method on the ISBI2015 dataset [4]. ISBI2015 dataset has 400 lateral cephalograms: *Training* (150 images), *Test1* (150 images), and *Test2* (100 images). We have used *Test1* as the validation set, *Test2* as the test set as in [13]. In the dataset, each image has a size of 2400 × 1935 with a spatial resolution of 0.1 mm in both directions. There are two sets of annotations of $N = 19$ cephalometric landmarks: Annotations are made by two experienced orthodontists. In evaluations, the average points of annotations are used as ground-truth positions [10], [11], [13].

A. EVALUATION METRICS

As presented in [4], we evaluate performance in terms of mean radial error (MRE), successful detection rate (SDR), and successful classification rate (SCR).

MRE and SDR for a set of size M are defined as

$$\mathrm{MRE} = \frac{1}{NM} \sum_{i=1}^{M} \sum_{k=1}^{N} R(i, k), \tag{8}$$

$$\mathrm{SDR} = \frac{1}{NM} \sum_{i=1}^{M} \sum_{k=1}^{N} \mathbb{1}\{R(i, k) < \tau\}, \tag{9}$$

respectively. Here $R(i, k)$ is a radial error for the k-th landmark in the i-th image, $\mathbb{1}\{\cdot\}$ is an indicator function, and τ is a reference threshold, whose typical values are 20 (2.0 mm), 25 (2.5 mm), 30 (3.5 mm) and 40 (4.0 mm). SCR is defined as the average of diagonal entries of a confusion matrix, where the confusion matrix is obtained by performing type classification using (1) one of clinical measurement methods in Table 1, and (2) landmark positions provided by the algorithm (e.g., our algorithm). Since 8 clinical measurement methods for the anatomic facial type classification (i.e., ANB, SNB, SNA, ODI, ADPI, FHI, FMA, and MW) are used for the evaluation as summarized in Table 1, each landmark localization method has 8 SCR values [4].

B. IMPLEMENTATION DETAILS

We have implemented the proposed method with PyTorch[1] [37]. The architectures of neural networks are based on DeepLabv3 [36].

[1] https://github.com/hjkwonispl/mpa

TABLE 2. Comparison of mean radial error (MRE) and successful detection rate (SDR).

Model	Test1 (Validation Set)					Test2 (Test Set)				
	MRE (mm)	SDR (%)				MRE (mm)	SDR (%)			
		2.0mm	2.5mm	3.0mm	4.0mm		2.0mm	2.5mm	3.0mm	4.0mm
Ibragimov et al. [24]	1.84	71.72	77.40	81.93	88.04	2.01	62.74	70.47	76.53	85.11
Lindner and Cootes [19]	1.67	73.68	80.21	85.19	91.47	1.92	66.11	72.00	77.63	87.43
Arik et al. [12]	-	75.37	80.92	84.32	88.25	-	67.68	74.16	79.11	84.63
Qian et al. [10]	-	82.50	86.40	89.30	90.60	-	72.40	76.15	79.65	85.90
Chen et al. [9]	1.17	86.67	92.67	95.54	98.35	1.48	75.05	82.84	88.53	95.05
Zhong et al. [26]	**1.12**	86.91	91.82	94.88	97.90	1.42	76.00	82.90	88.74	94.32
Oh et al. [13]	1.18	86.20	91.20	94.40	97.70	1.45	75.89	83.36	**89.26**	**95.73**
Qian et al. [11]	1.15	**87.61**	**93.16**	**96.35**	**98.74**	1.43	76.32	82.95	87.95	94.63
Zeng et al. [32]	1.34	81.37	89.09	93.79	97.86	1.64	70.58	79.53	86.05	93.32
Ours	**1.12**	86.91	91.44	94.21	97.68	**1.41**	**77.16**	**84.79**	89.21	94.95

TABLE 3. Comparison of successful classification rate (SCR). Numbers are given in percents (%).

Methods	Test1 (Validation Set)								Test2 (Test Set)							
	ANB	SNB	SNA	ODI	ADPI	FHI	FMA	MW	ANB	SNB	SNA	ODI	ADPI	FHI	FMA	MW
Ibragimov et al. [24]	59.42	71.09	59.00	78.04	80.16	58.97	77.03	83.94	76.64	75.24	70.24	63.71	79.93	86.74	78.90	77.53
Lindner and Cootes [19]	64.99	**84.52**	68.45	84.64	82.14	67.92	75.54	82.19	75.83	81.92	**77.97**	71.26	87.25	90.90	80.66	82.11
Arik et al. [12]	61.47	70.11	63.57	75.04	82.38	65.92	73.90	81.31	77.31	69.81	66.72	72.28	87.18	69.16	78.01	77.45
Oh et al. [13]	78.80	83.92	66.33	83.34	84.01	74.30	79.62	**91.10**	84.05	87.20	72.96	72.52	**89.37**	**94.75**	83.03	**82.67**
Zeng et al. [32]	**78.84**	81.46	71.08	84.88	86.22	90.32	85.11	84.19	82.06	**89.69**	64.75	71.47	88.90	71.86	83.50	81.90
Ours	78.04	84.12	**73.93**	**89.85**	**86.40**	**91.52**	**86.65**	91.02	**87.38**	84.18	71.38	**78.37**	89.08	78.71	**84.39**	82.22

1) GLOBAL STAGE

The last convolution layer of a DeepLabv3 network is modified to have 19 channels. Input images are resized to 720×580 with bilinear interpolation. Images are augmented by random rotations in $[-25°, 25°]$, scaling in $[0.9, 1.2]$ and random variations of brightness, hue, contrast and saturation in $[0, 0.25]$. We have used Adam optimizer with a learning rate of 10^{-4} for $1, 500$ epochs.

2) LOCAL STAGE

Since we train one network per each landmark, we modify the last convolution layer of DeepLabv3 to have one channel output. In the training, patches having a side length of 512 are cropped from perturbed ground-truth positions. Augmentation schemes used in the global stage are also applied. All networks are trained with Adam optimizer of a learning rate of 10^{-4} for 50 epochs.

3) REFINEMENT STAGE

We trained the linear filter with Adam optimizer of a learning rate 10^{-6} for 10 epochs.

C. COMPARISON WITH EXISTING METHODS

We compare the proposed method with conventional methods. Table 2 shows MRE and SDR values. As shown, our method achieves the best MRE in *Test1* and *Test2*. In terms of SDR, the proposed method takes the first place when τ is 2.0 mm and 2.5 mm, the second place when τ is 3.0 mm and the third place when τ is 4.0 mm in the *Test2*. For *Test1*, the method in [11] shows the best SDR for all thresholds, however, our method follows their results when $\tau = 2.0$ mm, which is a clinically acceptable error value in cephalometric analysis [8]. Also note that *Test1* is used as a validation set in the experimental settings.

TABLE 4. Comparison of mean radial error (MRE) and successful decision rate (SDR) for bilateral landmarks of the mandible on *Test2* (test set).

Landmark	Methods	MRE (mm)	SDR (%)			
			2.0mm	2.5mm	3.0mm	4.0mm
Gonion (Landmark 10)	Qian et al. [11]	1.38	81.00	87.00	94.00	98.00
	Oh et al. [13]	1.41	83.00	89.00	92.00	96.00
	Ours	**1.20**	**85.00**	**92.00**	**99.00**	**99.00**
Articulare (Landmark 19)	Qian et al. [11]	1.22	83.00	93.00	94.00	**100.00**
	Oh et al. [13]	**1.19**	**87.00**	**97.00**	**98.00**	98.00
	Ours	1.21	83.00	91.00	94.00	99.00

SCR results are summarized in Table 3. The proposed method shows the best and the second-best accuracies for 5 and 2 methods in *Test1*, respectively. For *Test2*, our method achieves the best accuracy for 3, and the second-best result for 2 methods. Since each measurement method has its own anatomical meaning and clinical importance, there is no a single comparison metric. However, results on SCR show that the proposed method can be adopted in more measurement methods for the anatomic facial types classification than others. To evaluate the localization performance of bilateral landmarks of the mandible, we also compare MRE and SDR of *gonion* (landmark 10) and *articulare* (landmark 19) with other methods. Table 4 shows that our method shows improved performance to localize *gonion*. For the *articulare*, our method shows the second-best result in MRE and SDR when $\tau = 2.0$ mm.

Another advantage of our probabilistic approach is that we can provide the confidence regions of predictions. This property is important in treatment planning since clinicians need to review and correct estimated landmark positions

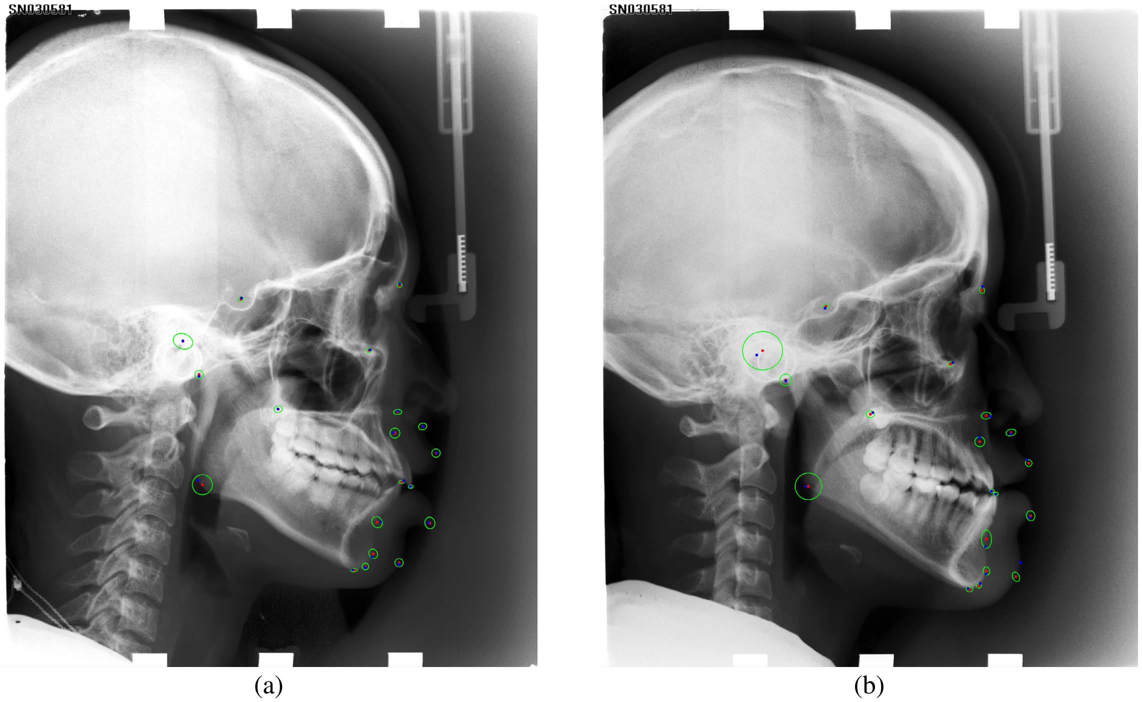


FIGURE 2. Visualizations of estimated landmarks (red dot), their confidence regions (green ellipse), and ground-truth annotations (blue dot) of samples from (a) *Test1* and (b) *Test2*.

TABLE 5. Comparison of mean radial error (MRE) and successful decision rate (SDR) of the global stage with different networks.

Model	Test1 (Validation Set)					Test2 (Test Set)				
	MRE (mm)	SDR (%)				MRE (mm)	SDR (%)			
		2.0mm	2.5mm	3.0mm	4.0mm		2.0mm	2.5mm	3.0mm	4.0mm
U-Net[34]	1.73	67.93	78.42	85.37	93.23	2.27	53.95	64.84	73.16	86.63
FCN[35]	1.16	86.28	**92.07**	94.77	**98.04**	1.56	72.00	80.11	86.42	**94.21**
DeepLabv3[36]	**1.15**	**86.53**	91.68	**95.30**	98.00	**1.54**	**72.47**	**80.32**	**87.21**	93.79

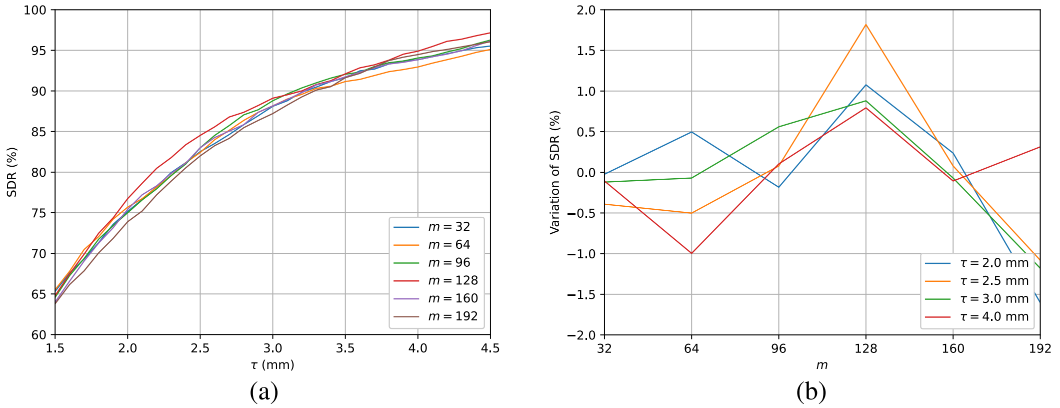
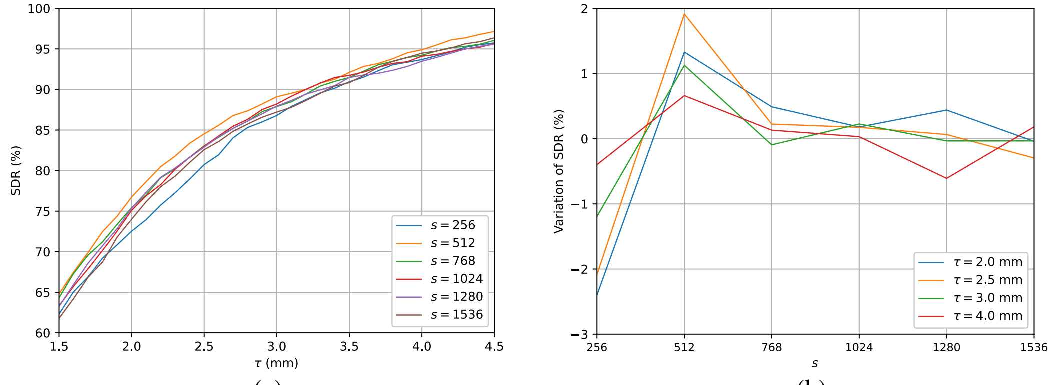


FIGURE 3. Ablations of the boundary (*m*) of uniform distribution on [−*m*, *m*] with *Test2* when the patch size (*s*) is 512: (a) Comparison of successful decision rate (SDR) for different values of *m* when swing τ from 1.5 mm to 4.5 mm, (b) Variations of SDR from averages for *m* is in [32, 192] when τ is 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm.

with confidence regions [8]. In Fig 2, we compute statistical values from $P^G(\cdot)$ and $P^L(\cdot)$ for each landmark and visualize confidence regions (standard deviational ellipses of 3σ) [18]. As shown, all ground-truth positions are located around standard deviational ellipses except *soft tissue pogonion* (landmark 16) of *Test2*. However, it is reported that the *soft tissue pogonion* is annotated in a different way from training set and *Test1* [13].



(a) (b)

FIGURE 4. Ablations of patch size (*s*) on *Test2* when the boundary (*m*) of uniform distribution on [−*m*, *m*] is 128: (a) Comparison of successful decision rate (SDR) for different values of *s* when swing τ from 1.5 mm to 4.5 mm, (b) Variations of SDR from averages for *s* is in [256, 1536] when τ is 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm.

TABLE 6. Comparison of training loss (TL) and mean radial error (MRE) of *Test1* (validation set) and *Test2* (test set) for the refinement stage for linear filters (Linear), MLPs with two layers (MLP2) and three layers (MLP3).

Model	Gonion (Landmark 10)			Articulate (Landmark 19)		
	TL	MRE (mm)		TL	MRE (mm)	
		Test1	Test2		Test1	Test2
MLP2	**0.10**	2.20	2.60	5.24	1.80	1.59
MLP3	0.50	2.19	1.94	**4.24**	1.74	1.72
Linear	12.30	**1.73**	**1.20**	14.71	**1.61**	**1.21**

D. ABLATION STUDY

We also evaluate the contribution of each block. First, we compare the performance of the global stage using other backbone networks (FCN [35], UNet [34], and DeepLabv3 [36]). Table 5 shows that the DeepLabv3 achieves the best performance in terms of MRE and SDR at $\tau = 2.0$ mm.

For the local networks, we evaluate the effects of data augmentation in terms of SDR. Let us denote the patch size s and crop box dislocations as (Δ_x, Δ_y). When the ground truth landmark is (p_x, p_y), we crop the patch whose center is $(p_x + \Delta_x, p_y + \Delta_y)$ and side length is s. First, we sample Δ_x and Δ_y from a uniform distribution in $[-m, m]$, and compare results for a range of m. Fig. 3 shows that local networks achieve the best performance when $m = 128$ for $s = 512$. We also evaluate SDR for $s \in [256, 1536]$ when $m = 128$. Fig. 4 shows that the best SDR is achieved when $s = 512$. Note that a larger value of s does not always improve the performance. We believe that this result justifies the global-to-local approach in terms of performance besides memory usages (compared with using very large inputs). CNNs can work well when they are provided only necessary information, i.e., proper local neighborhood areas, especially when training samples are limited.

To validate the use of linear filters for the refinement stage, we compare the training loss and MRE on *Test1* and *Test2* using MLP of two and three layers. Each MLP has 256 hidden units for all layers, and we trained them with the same inputs and loss functions used for linear filters. Table 6 shows that linear filters give the best MRE on *Test1* and *Test2*. Both MLPs seem to be overfitted when compared to the results of linear filter. This result shows that the use of linear filters is a reasonable choice for our case (150 training data with $4N = 76$ input dimension). Finally, we compare SDR of bilateral landmarks of the mandible for three stages. Fig. 5 shows the proposed refinement method gives improved SDR for all τ.

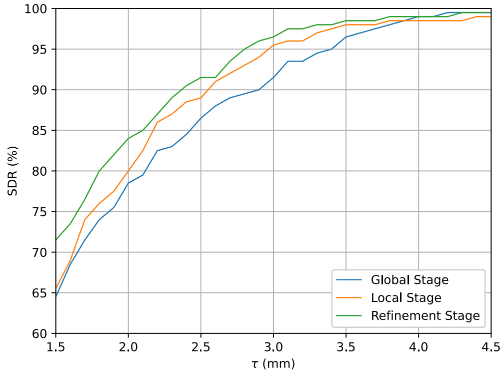


FIGURE 5. Successful decision rate (SDR) comparison of bilateral landmarks of the mandible. We plot SDR of global, local and refinement stages on the *Test2* for several τ values.

V. CONCLUSION

In this paper, we have presented a cephalometric landmark localization method based on the multistage probabilistic approach. The multistage method allows us to estimate the positions of cephalometric landmarks using (1) global hard/soft tissue characteristics and (2) local details of anatomical structures. Also, this probabilistic framework provides confidence regions for estimated landmarks. To handle the asymmetries of anatomic structures in the mandible, we have also developed linear estimators that use all landmark estimations. Experiments on ISBI2015 dataset shows that our method achieved state-of-the-art performances in terms of mean radial error, successful detection rate, and anatomic facial type classification accuracy.

REFERENCES

[1] J.-H. Park, H.-W. Hwang, J.-H. Moon, Y. Yu, H. Kim, S.-B. Her, G. Srinivasan, M. N. A. Aljanabi, R. E. Donatelli, and S.-J. Lee, "Automated identification of cephalometric landmarks: Part 1—Comparisons between the latest deep-learning methods YOLOV3 and SSD," *Angle Orthodontist*, vol. 89, no. 6, pp. 903–909, Nov. 2019.

[2] H.-W. Hwang, J.-H. Park, J.-H. Moon, Y. Yu, H. Kim, S.-B. Her, G. Srinivasan, M. N. A. Aljanabi, R. E. Donatelli, and S.-J. Lee, "Automated identification of cephalometric landmarks: Part 2-might it be better than human?" *Angle Orthodontist*, vol. 90, no. 1, pp. 69–76, Jan. 2020.

[3] C. Lindner, C.-W. Wang, C.-T. Huang, C.-H. Li, S.-W. Chang, and T. F. Cootes, "Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms," *Sci. Rep.*, vol. 6, no. 1, p. 33581, Sep. 2016.

[4] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, P. Fischer, T. F. Cootes, and C. Lindner, "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, Jul. 2016.

[5] B. Phulari, *An Atlas on Cephalometric Landmarks*. New Delhi, India: JP Medical Ltd, 2013.

[6] E. Whaites and N. Drage, *Essentials of Dental Radiography and Radiology*. Amsterdam, The Netherlands: Elsevier, 2013.

[7] S. Malkoc, Z. Sari, S. Usumez, and A. E. Koyuturk, "The effect of head rotation on cephalometric radiographs," *Eur. J. Orthodontics*, vol. 27, no. 3, pp. 315–321, Jun. 2005.

[8] J.-H. Lee, H.-J. Yu, M.-J. Kim, J.-W. Kim, and J. Choi, "Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks," *BMC Oral Health*, vol. 20, no. 1, pp. 1–10, Dec. 2020.

[9] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, "Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 873–881.

[10] J. Qian, M. Cheng, Y. Tao, J. Lin, and H. Lin, "CephaNet: An improved faster R-CNN for cephalometric landmark detection," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 868–871.

[11] J. Qian, W. Luo, M. Cheng, Y. Tao, J. Lin, and H. Lin, "CephaNN: A multi-head attention network for cephalometric landmark detection," *IEEE Access*, vol. 8, pp. 112633–112641, 2020.

[12] S. Ö. Arik, B. Ibragimov, and L. Xing, "Fully automated quantitative cephalometry using convolutional neural networks," *J. Med. Imag.*, vol. 4, no. 1, Jan. 2017, Art. no. 014501.

[13] K. Oh, I.-S. Oh, T. V. N. Le, and D.-W. Lee, "Deep anatomical context feature learning for cephalometric landmark detection," *IEEE J. Biomed. Health Inform.*, early access, Jun. 15, 2020, doi: 10.1109/JBHI.2020.3002582.

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

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

[16] L. R. Leslie, T. E. Southard, K. A. Southard, J. S. Casko, J. R. Jakobsen, E. A. Tolley, S. L. Hillis, C. Carolan, and M. Logue, "Prediction of mandibular growth rotation: Assessment of the Skieller, Björk, and Linde-Hansen method," *Amer. J. Orthod. Dentofacial Orthop.*, vol. 114, no. 6, pp. 659–667, 1998.

[17] Y.-J. Yoon, K.-S. Kim, M.-S. Hwang, H.-J. Kim, E.-H. Choi, and K.-W. Kim, "Effect of head rotation on lateral cephalometric radiographs," *Angle Orthodontist*, vol. 71, no. 5, pp. 396–403, 2001.

[18] B. Wang, W. Shi, and Z. Miao, "Confidence analysis of standard deviational ellipse and its extension into higher dimensional Euclidean space," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118537.

[19] C. Lindner and T. F. Cootes, "Fully automatic cephalometric evaluation using random forest regression-voting," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, 2015.

[20] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862–1874, Sep. 2015.

[21] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, T. F. Cootes, arcOGEN Consortium, "Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Berlin, Germany: Springer, 2013, pp. 181–189.

[22] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2012, pp. 278–291.

[23] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, vol. 41, no. 10, pp. 3054–3067, Oct. 2008.

[24] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "Computerized cephalometry by game theory with shape- and appearance-based landmark refinement," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, Dec. 2015.

[25] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "A game-theoretic framework for landmark-based image segmentation," *IEEE Trans. Med. Imag.*, vol. 31, no. 9, pp. 1761–1776, Sep. 2012.

[26] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, and X. Gao, "An attention-guided deep regression model for landmark detection in cephalograms," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 540–548.

[27] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.

[28] L. Gilmour and N. Ray, "Locating cephalometric X-ray landmarks with foveated pyramid attention," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2020, pp. 262–276.

[29] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[30] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, and B. Glocker, "Attention U-Net: Learning where to look for the pancreas," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2018, pp. 1–10.

[31] T. Graber, R. Vanarsdall, K. Vig, and G. Huang, *Orthodontics: Current Principles and Techniques*. Maryland Heights, MO, USA: Mosby, Inc, 2000.

[32] M. Zeng, Z. Yan, S. Liu, Y. Zhou, and L. Qiu, "Cascaded convolutional networks for automatic cephalometric landmark detection," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101904.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. Assoc. Comput. Mach. Special Interest Group Knowl. Discovery Data (ACM SIGKDD)*, 2016, pp. 1135–1144.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[37] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

HYUK JIN KWON (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and the D.D.S. degree from Seoul National University, Seoul, South Korea, in 2012 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include dental imaging, computer vision, and machine learning.

HYUNG IL KOO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, Seoul National University, Seoul, South Korea, in 2002, 2004, and 2010, respectively. From 2010 to 2012, he was a Research Engineer with the Qualcomm Research Korea. He joined the Department of Electrical and Computer Engineering, Ajou University, in 2012, where he is currently an Associate Professor. His research interests include computer vision and machine learning.

JAEWOO PARK (Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include image processing, computer vision, and machine learning.

NAM IK CHO (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1986, 1988, and 1992, respectively. From 1991 to 1993, he was a Research Associate with the Engineering Research Center for Advanced Control and Instrumentation, Seoul National University. From 1994 to 1998, he was an Assistant Professor of Electrical Engineering with the University of Seoul. In 1999, he joined the Department of Electrical and Computer Engineering, Seoul National University, where he is currently a Professor. His research interests include image processing, adaptive filtering, digital filter design, and computer vision.