IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 June 2022, accepted 26 July 2022, date of publication 10 August 2022, date of current version 16 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3197749

RESEARCH ARTICLE

Implicit Semantic Data Augmentation for Hand Pose Estimation

KYEONGEUN SEO[1], HYEONJOONG CHO[2], DAEWOONG CHOI[2], AND JU-DERK PARK[3]
[1]Information Media Research Center, Korea Electronics Technology Institute, Seoul 03924, South Korea
[2]Department of Computer Convergence Software, Korea University, Sejong 30019, South Korea
[3]Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea

Corresponding author: Hyeonjoong Cho (raycho@korea.ac.kr)

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) Grant through the Ministry of Land, Infrastructure and Transport under Grant 22AMDP-C160501-02.

ABSTRACT Data augmentation is a well-known technique used for improving the generalization performance of modern neural networks. After the success of several traditional random data augmentation for images (including flipping, translation, or rotation), a recent surge of interest in implicit data augmentation techniques occurs to complement random data augmentation techniques. Implicit data augmentation augments training samples in feature space, rather than in pixel space, resulting in the generation of semantically meaningful data. Several techniques on implicit data augmentation have been introduced for classification tasks. However, few approaches have been introduced for regression tasks with continuous/structured labels, such as object pose estimation. Hence, we are motivated to propose a method for implicit semantic data augmentation for hand pose estimation. By considering semantic distances of hand poses, the proposed method implicitly generates extra training samples in feature space. We propose two additional techniques to improve the performance of this augmentation: metric learning and hand-dependent augmentation. Metric learning aims to learn feature representations to reflect the semantic distance of data. For hand pose estimation, the distribution of augmented hand poses can be regulated by managing the distribution of feature representations. Meanwhile, hand-dependent augmentation is specifically designed for hand pose estimation to prevent semantically meaningless hand poses from being generated (e.g., hands generated by simple interpolation between both hands). Further, we demonstrate the effectiveness of the proposed technique using two well-known hand pose datasets: STB and RHD.

INDEX TERMS Hand pose estimation, data augmentation, semantic learning, feature learning.

I. INTRODUCTION

Data augmentation is an effective technique for alleviating the problem of overfitting in training deep neural networks. For hand pose estimation [1], [2], [3], [4], [5], data augmentation usually involves applying several random transformations to existing images (e.g., cropping, rotation, and translation). These simple transformation methods help neural networks avoid overfitting the training data. Unfortunately, they cannot always generate semantically transformed data, such as changing hand poses or their backgrounds [15]. Recently, researchers have studied generative adversarial networks (*GAN*s). After training, GAN generators can synthesize an infinite number of semantically meaningful training samples [6]. However, a separate GAN model should be trained to use GANs for data augmentation, in addition to the target model. Moreover, according to [9], GANs require nontrivial training that involves intensive computation.

For efficient semantic data augmentation, implicit data augmentation [7], [8], [9], [47] has emerged as the new paradigm that can address the weakness of generative model-based approaches. It augments training data by selecting feature representations of a certain layer in neural networks and by generating new feature representations via interpolation or translation using selected features. Using the newly augmented features, the method trains models to minimize predefined loss. Although such methods provide efficient data augmentation, most approaches have focused on tasks

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

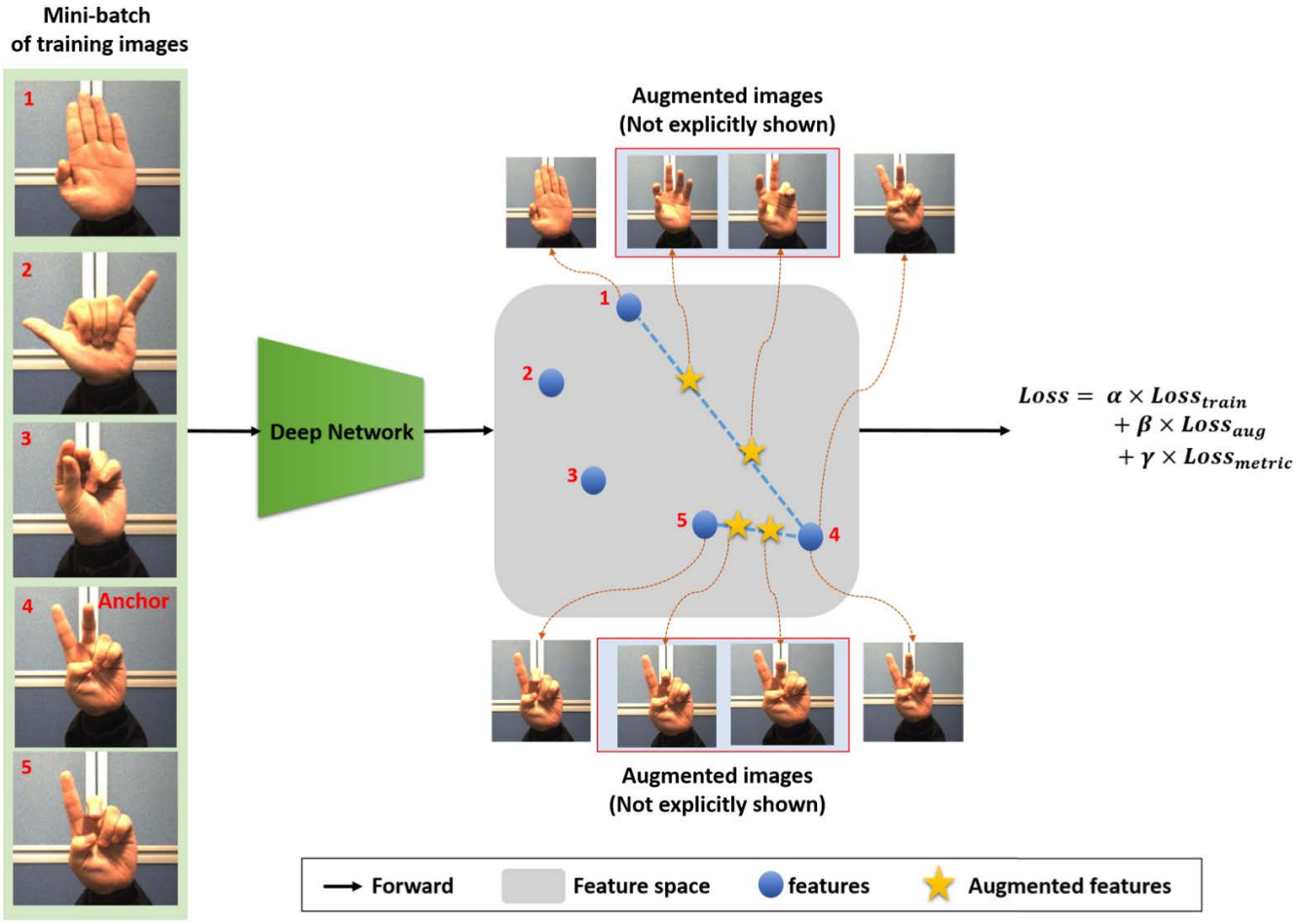

FIGURE 1. Overview of the proposed implicit semantic data augmentation for hand pose estimation. The proposed method generates semantically meaningful samples through interpolation with feature representations selected based on the semantic distance of hand poses. Deliberate selection makes the distribution of augmented samples more balanced compared to that of randomly selected samples. Detailed procedure is described in Section III.

with discrete labels (e.g., image classification). Whereas, few approaches consider tasks with continuous and structured labels.

We propose an implicit semantic data augmentation method to train deep neural models for hand pose estimation. The proposed method is motivated by intriguing observations made in a recent study [10]. The study demonstrates the importance of understanding the semantic similarity between images in numerous areas of computer vision (e.g., human pose estimation, face identification, image retrieval, and video object tracking). For hand pose estimation, learning the semantic similarity of hand poses allows neural networks to be sensitive to joint positions and invariant to illumination, background, clutter, and occlusions.

Current methods on implicit data augmentation [8], [9] randomly select two minibatches and then interpolate with a randomly associated pair of features, leading to a mixed minibatch. Regarding hand pose estimation, the random selection of features without considering the semantic distance of hand poses may cause a biased distribution of augmented data. To address this drawback, we consider features' distance that reflects the semantic distance of hand poses to synthesize new data. By deliberately selecting features based on the semantic distance of hand poses and by applying interpolation between them, we can augment new hand poses with a more balanced distribution than those of existing methods. Figure 1 illustrates the conceptual procedure of the proposed method.

To improve the effectiveness of our method, we propose to apply two additional components: metric learning and hand-dependent augmentation. Metric learning learns feature representations to reflect the semantic distance of hand poses, allowing the regulation of the distribution of augmented hand pose data by managing the distribution of their feature representations. Meanwhile, hand-dependent augmentation imposes an additional constraint dedicated to hand pose estimation. It prevents the generation of semantically meaningless data, for example, hand poses synthesized via simple interpolation between the left and right hands.

The remainder of this paper is organized as follows. In Section II, previous studies on data augmentation,

metric learning, and hand pose estimation were discussed. Section III describes in detail the proposed technique. In Section IV, we report the experiments we performed to demonstrate the effectiveness of the proposed method using two different datasets. Finally, Section V concludes the paper.

II. RELATED WORK

A. DATA AUGMENTATION

Artificial data created by using data augmentation methods have been commonly used to alleviate overfitting in training deep neural networks. Approaches for initial data augmentation for images mainly include label-preserving transformations, such as random flipping, scaling, rotation, and translation [1], [2], [3], [4], [5], [13], [14]. Furthermore, hue, saturation, brightness, and darkness changes in images were used to increase the diversity of the generated images [15], [16]. These methods typically augment images by retaining semantic information.

Beyond label-preserving transformations, several studies that generate synthetic images in pixel space have been proposed. *Mixup* [32], [33] linearly combined two samples in pixel space and their labels using a weighted linear combination. The images generated by Mixup were extremely different from the original data and unreal to human perception when compared to those generated by label-preserving data augmentation. Recently, GANs have been used to generate semantically transformed training images [6], [17]. After training GANs with training data, the generators of GANs can synthesize an infinite number of samples.

Beyond data augmentation in pixel space, recent studies [7], [8], [9] augmented data in feature space, called as implicit data augmentation. This generates features that correspond to the augmented images in pixel space and uses them to train the models. Wang *et al.* [7] augmented training data semantically while preserving the labels by translating their features along with the feature distributions of each class. Li *et al.* [9] proposed implicit augmentation of data by simultaneously interpolating two training inputs in both the feature and label spaces. To increase stability, Verma *et al.* [8] used the moments (mean and standard deviation) of features, by exchanging one training image with another and interpolating the features and labels. Swapping the moments between two features facilitates the generation of different samples with asymmetric proportions to the mean and standard deviation. These methods were proven to be effective in terms of classification tasks. However, directly using them in regression tasks, such as hand pose estimation, is challenging because of the intrinsic difference in data labels (i.e., discrete and continuous).

B. METRIC LEARNING

Metric learning plays an essential role in numerous areas of computer vision, including pose estimation, because it helps estimate the similarity between images, which is a basic component of human reasoning [18]. For deep metric learning, contrastive loss and triplet loss are standard loss functions [19], [20], [21]. For instance, given a triplet of anchor, positive, and negative images, the triplet loss forces the distance between the anchor and positive images to be smaller than that between the anchor and negative images. To learn the similarities between continuous labels, some studies [34], [35], [36] applied metric learning after mapping continuous labels into discrete labels. This approach quantized continuous similarities into binary levels through distance thresholding or nearest-neighbor search. Unfortunately, both strategies are unusual for continuous metric learning.

Recently, Kim *et al.* [10] proposed a new loss called *log-ratio loss* for metric learning of retrieval tasks with continuous labels, including human poses, room layouts, and caption-aware images. Considering a triplet of an anchor and two neighbors, the model can approximate the ratio of label distances using the ratio of feature distances. Inspired by their work, we used log-ratio loss for hand pose estimation in this study.

C. HAND POSE ESTIMATION

Reference [22] classified 3D hand pose estimation methods into two categories: generative [23], [24], [25] and discriminative [1], [2], [3], [4], [5], [26], [27], [28], [40], [41], [42], [43], [44], [45], [46]. Generative approaches generate hypothetical hand poses and compare them to observed data to find a solution that minimizes the objective function defined as a discrepancy between hypothetical and observed hand poses. Most generative approaches rely on local searches after initialization and, therefore, are susceptible to local optima.

Conversely, discriminative approaches learn direct mapping from observations of hand poses [1], [2], [3], [4], [5], [26], [27], [28], [40], [41], [42], [43], [44], [45], [46]. With the advent of convolutional neural networks (*CNN*s) and large-scale datasets, discriminative methods have exhibited promising performance and have proven to be suitable alternatives to generative approaches. Recently, Moon *et al.* [5] introduced a CNN-based model, *InterNet*, achieving significantly improved performance in terms of hand pose estimation. InterNet can estimate both hand poses and four components concurrently: right- and left-hand pose handedness and right hand-relative left-hand depth.

Discriminative methods [1], [2], [3], [4], [5], [43], [44], [45], [46] generally use a traditional random data augmentation comprising various transformations (e.g., rotation, translation, flip, scale, hue, saturation, jitter) to improve the estimation performance. Because random data augmentation does not generate semantically meaningful hand poses, it restricts the performance improvement of hand pose estimation methods. To the best of our knowledge, our study is the first attempt to augment semantically transformed samples for hand pose estimation through implicit data augmentation.

III. PROPOSED METHOD

The proposed implicit data augmentation method was designed for hand pose estimation tasks, which intrinsically

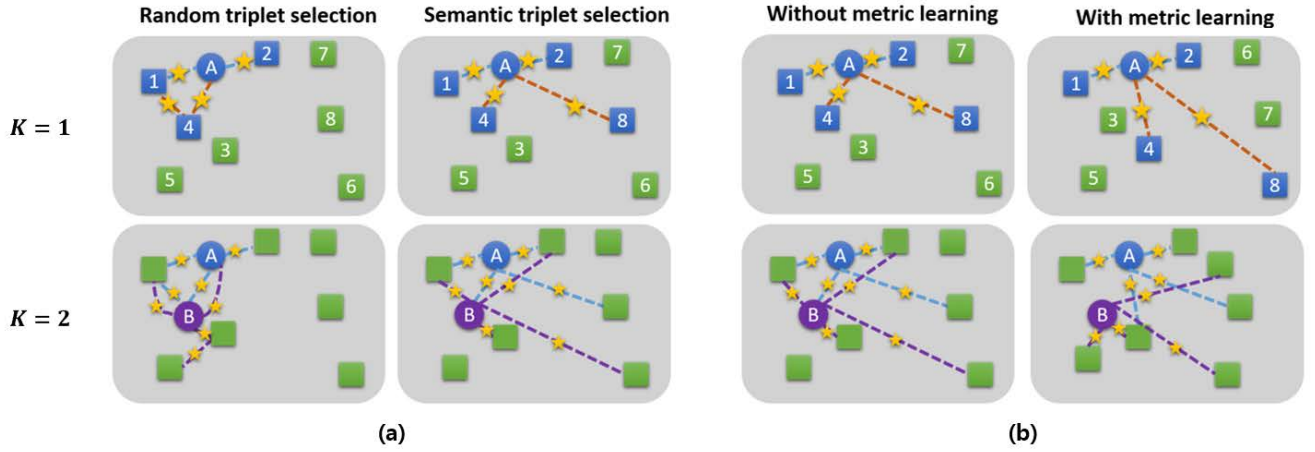

FIGURE 2. Implicit semantic data augmentation and metric learning. Here, K is the number of selected anchors. Circles in the figures are anchor features, squares are other selected features, and stars are augmented features. The number on a feature denotes the rank of the hand pose similarity in the pixel space between an anchor and a feature. (a) The semantic triplet selection considers hand pose similarities. Thus, the distribution of augmented samples becomes more balanced than that of random triplet selection. (b) Metric learning makes the features learn to reflect the semantic distance of hand poses. Consequently, the features are forced to be well-arranged in the feature space according to the hand pose similarity, resulting in a more balanced distribution of augmented samples.

Algorithm 1: Procedure of the Proposed Method

Input: a minibatch $\boldsymbol{B} = \{(x_i, y_i)\}_{i=1}^{N}$
Input: the number of selected anchors, K
Input: learnable parameters $\boldsymbol{\theta}$
Output: updated parameters $\boldsymbol{\theta}$

if *random data augmentation is permitted* **then**
  $\boldsymbol{B}$ = Apply random data augmentation to $\boldsymbol{B}$
for $(x_i, y_i) \in \boldsymbol{B}$ **do**
  Compute $f_i = Net(x_i, \boldsymbol{\theta})$ and add to $\boldsymbol{F}^{train}$
Sample anchors $\boldsymbol{F}^{A} = \{(f^a, y^a)_k\}_{k=1}^{K}$ from $\boldsymbol{F}^{train}$;
for $(f_k^a, y_k^a) \in \boldsymbol{F}^{A}$ **do**
  for $(f_i, y_i) \in \boldsymbol{F}^{train}$ **do**
    Compute $D(y_k^a, y_i)$ using Eq. (1)
  Construct a triplet T_k^c and add to $\boldsymbol{T^c}$;
  Construct a triplet T_k^d and add to $\boldsymbol{T^d}$
for $T_k \in \boldsymbol{T} = \boldsymbol{T^c} \cup \boldsymbol{T^d}$ **do**
  Generate (f_k^{aug}, y_k^{aug}) using Eq.(2) and add to $\boldsymbol{F}^{aug}$
for $(f_k, y_k) \in \boldsymbol{F}^{aug} \cup \boldsymbol{F}^{train}$ **do**
  Compute $\hat{y}_k = Net^{+}(f_k, \boldsymbol{\theta}^{+})$
Construct $\boldsymbol{F}^{aug'}$ by selecting the samples from $\boldsymbol{F}^{aug}$;
Compute *Loss* using Eq. (6);
Update $\boldsymbol{\theta}$ using backpropagation

involve continuous and structural labels. We added two additional components, namely, metric learning and hand-dependent augmentation, to further enhance the performance of our implicit data augmentation. The detailed description of each component is presented in the following subsections. The procedure of the proposed method is also presented in Algorithm 1.

A. IMPLICIT SEMANTIC DATA AUGMENTATION

Consider training a neural network model $Net(\cdot)$ parameterized by its weights $\boldsymbol{\theta}$ with a minibatch $\boldsymbol{B} = \{(x_i, y_i)\}_{i=1}^{N}$, where y_i is a set of positions of J hand joints for ith sample x_i and N is the size of the minibatch. We assume that the ith feature vector $f_i = Net(x_i, \boldsymbol{\theta})$ is obtained from $Net(\cdot)$ and define a feature minibatch $\boldsymbol{F}^{train} = \{(f_i, y_i)\}_{i=1}^{N}$. Note that the entire model is defined as $\hat{y}_i = Net^{+}(Net(x_i, \boldsymbol{\theta}), \boldsymbol{\theta}^{+})$.

We constructed a set of triplets $\boldsymbol{T} = \{T_k\}_{k=1}^{2K}$ to obtain augmented samples, where $T_k = ((f^a, y^a), (f^m, y^m), (f^n, y^n))$. f^a and y^a are an anchor feature and its label, respectively. Further, (f^m, y^m) and (f^n, y^n) are two other pairs carefully selected. In particular, (f^m, y^m) and (f^n, y^n) should be carefully selected by considering the discrepancy of hand poses to make the distribution of augmented data as balanced as possible. The degree of discrepancy of hand poses provides the information regarding how much different two hand poses are. We computed the discrepancy between two hand poses, y and y', as follows:

$$D(y_i, y_i') = \sum_{j=1}^{J} \|y_{i,j} - y_{i,j}'\|_2^2, \tag{1}$$

where $y_{i,j}$ is the joint position of a hand pose y_i.

To construct a balanced set of triplets, we first randomly selected anchor pairs, $\boldsymbol{F^A} = \{(f_k^a, y_k^a)\}_{k=1}^{K}$ from $\boldsymbol{F}^{train}$. Then, we constructed two types of triplet sets: $\boldsymbol{T^c}$ and $\boldsymbol{T^d}$. Note that $\boldsymbol{T} = \boldsymbol{T^c} \cup \boldsymbol{T^d}$. We constructed $\boldsymbol{T^c} = \{T_k^c\}_{k=1}^{K}$, where T_k contains an anchor (f^a, y^a), a feature having the closest distance from the anchor (f^m, y^m), and a feature having second closest distance (f^n, y^n). We constructed $\boldsymbol{T^d} = \{T_k^d\}_{k=1}^{K}$, where T_k contains an anchor feature (f^a, y^a), a feature having middle distance (f^m, y^m), and a feature having the farthest distance

(f^n, y^n). The strategy of deliberate triplet selection leads to a balanced distribution of augmented data, as illustrated in Figure 2.

After constructing the set of triplets, $\boldsymbol{T}$, we generated new samples by interpolating the anchor and selected feature as performed in [9] as follows:

$$f^{aug}(f^a, f') = \lambda \times f^a + (1-\lambda) \times f' \text{and}$$
$$y^{aug}(y^a, y') = \lambda \times y^a + (1-\lambda) \times y', \quad (2)$$

where $\lambda \in [0, 1]$ is the interpolation constant. Whenever the augmented data are generated, the interpolation constant is randomly selected. A single interpolation was performed per triplet; consequently, a new augmented minibatch $\boldsymbol{F}^{aug} = \{(f_k^{aug}, y_k^{aug})\}_{k=1}^{4K}$ was obtained from $\boldsymbol{T}$.

B. HAND-DEPENDENT AUGMENTATION

After performing data augmentation, we formed a sample set $\boldsymbol{F}^{aug'}$ by including meaningful samples only from the augmented minibatch $\boldsymbol{F}^{aug}$. $\boldsymbol{F}^{aug'}$ excludes the samples generated by combining the features of both hands. For training, we consequently used the augmented image generated by features that include same side hands, but do not include results generated by left and right hands. Using $\boldsymbol{F}^{aug'}$, we obtained $Loss_{aug}$ as

$$Loss_{aug} = \sum_{\forall y_k^{aug} \in \boldsymbol{F}^{aug'}} D(y_k^{aug}, \hat{y}_k^{aug}), \quad (3)$$

where $\hat{y}_k^{aug}$ is an estimated hand pose. Because most open hand datasets provide the information about what hand types are included, the loss can be easily achieved.

C. METRIC LEARNING

Metric learning aims to learn feature representations capable of effectively representing the semantic similarity between hand poses. As shown in Figure 2, metric learning regulates the distance between features in feature space and eventually arrange the features by reflecting their pose similarities.

Given triplet (f^a, f^n, f^m), we used the log-ratio loss [10] for metric learning. Log-ratio loss was designed to make the neural model approximate the ratio of label distances in the feature space. It is defined as

$$Loss_{metric} = \sum_{\forall triplet} \left\{ log \frac{d(f^a, f^n)}{d(f^a, f^m)} - log \frac{d(y^a, y^n)}{d(y^a, y^m)} \right\}^2, \quad (4)$$

where $d(\cdot)$ denotes the Euclidean distance.

Finally, $Loss_{train}$ is defined for training original data as follows:

$$Loss_{train} = \sum_{i=1}^{N} D(y_i, \hat{y}_i) \quad (5)$$

where y is the ground-truth hand pose and $\hat{y}$ is the hand pose estimated by the model.

TABLE 1. Ablation studies with the STB and RHD datasets. (I: Implicit data augmentation, H: Hand dependent augmentation, M: Metric learning, R: Random triplet selection, S: Semantic triplet selection.)

Method	STB		RHD	
	EPE	MPJPE	EPE	MPJPE
Baseline	7.87	7.91	16.70	19.59
Baseline + I (R)	7.92	7.93	16.61	19.62
Baseline + I (S)	7.75	7.79	16.54	19.62
Baseline + M (R)	8.05	8.08	16.40	19.11
Baseline + M (S)	7.71	7.88	16.07	18.90
Baseline + I (S)	7.75	7.79	16.54	19.62
Baseline + I (S) + H	7.69	7.71	15.94	18.75
Baseline + I (S) + H + M (S)	**7.64**	**7.68**	**15.88**	**18.50**

Considering all discussed aspects, we defined the entire loss function as

$$Loss = \alpha \times Loss_{train} + \beta \times Loss_{aug} + \gamma \times Loss_{metric}. \quad (6)$$

For the experiments, we set all components of the loss function to have the same importance ($\alpha = 1, \beta = 1\ \gamma = 1$).

In addition, we normalized the two intermediate minibatches with positional normalization (*PONO*) [29], which is beneficial for the convergence of neural networks.

IV. EXPERIMENTS

A. ABLATION STUDIES

We conducted ablation studies to determine the effect of each component of the proposed method for hand pose estimation. We set a deep model trained with random data augmentation to our baseline model. Further, we added various combinations of the proposed components (i.e., implicit semantic data augmentation, metric learning, and hand-dependent augmentation for comparison).

1) EXPERIMENTAL SETUPS

a: DATASETS

We used two well-known hand pose datasets, namely, *STB* and *RHD*, which have different contexts. STB [30] is a common dataset used to evaluate the performance of hand pose estimation techniques. It includes stereo video sequences of diverse poses of a single person with different backgrounds, captured in third-perspective views. Meanwhile, RHD [26] is a large, synthesized image dataset with 20 subjects performing 31 different actions in random backgrounds, without hand object interaction.

b: EVALUATION METRICS

For the evaluation, we used two metrics, end-point error (*EPE*) and mean per joint position error (*MPJPE*). EPE is defined as the mean Euclidean distance between the estimated hand pose and its ground truth, whereas MPJPE [5] is used to measure the aligned hand pose error, which is defined as the Euclidean distance between the estimated hand pose and its ground truth after root joint alignment.

c: RANDOM DATA AUGMENTATION

We used a random data augmentation consisting of five transformations (i.e., translation (15%), rotation (45%), scaling (25%), horizontal flip, and color jittering (20%)), which is denoted by RA(5), as a part of the baseline model. The five transformations were reasonably selected because they are frequently used in hand images.

d: DEEP MODELS AND IMPLEMENTATION

As a deep neural model for hand pose estimation, we used InterNet [5], which exhibits state-of-the-art performance. We followed its original learning settings, except for the original data augmentation. The architecture of InterNet has two parts: feature extraction and output estimation. We used ResNet50 as the backbone network.

The backbone network was initialized using the parameters of ResNet pretrained using the Image dataset. The weights were updated using the ADAM optimizer [31] with a mini-batch containing 16 samples. To crop the hand region from the input image, we used a ground-truth bounding box in both training and testing stages. The cropped hand image was resized to 256×256 pixels. The initial learning rate was set to 10^{-4} and reduced by a factor of 10 at the $45th$ and $47th$ epochs. Further, we trained our model for 50 epochs on an NVIDIA TitanX GPU.

For implicit data augmentation, we set K to 4. According to the setting, we randomly chose four features as anchor features from 16 features and then generated 16 features as augmented samples per minibatch.

2) EXPERIMENTAL RESULTS

Table 1 presents the results of the experiments. I(R) and I(S) denote implicit data augmentation with random triplet selection and semantic triplet selection, respectively. M(R) implies metric learning with a randomly constructed triplet; M(S), metric learning with semantic triplet selection; and H, hand-dependent data augmentation. The results show that each component of our method helps improve the performance of hand pose estimation using both datasets. The effect of each component is presented as follows.

TABLE 2. $Loss_{metric}$ **analysis on STB and RHD.**

Method	STB		RHD	
	Mean	Standard deviation	Mean	Standard deviation
Baseline	0.60	0.91	0.54	0.84
Ours	0.51	0.79	0.50	0.79

a: RANDOM VS SEMANTIC TRIPLET SELECTION

When training, the random selection approach may focus on a limited set of poses because it does not consider the distribution of features. It can degrade the performance due to biased learning, as shown in the results of Baseline+I(R) against those of Baseline+I(S). Conversely, the semantic triplet selection using the semantic distances of hand poses improves the performance of data augmentation. This is because a semantically balanced distribution of synthetic hand poses was generated. It therefore improves the generalization performance of deep neural models for hand pose estimation.

b: WITHOUT VS WITH METRIC LEARNING

The correlation between the distance of feature representations and the semantic distance of hand poses is reinforced through metric learning. The distribution of synthetic hand poses becomes increasingly manageable because our implicit semantic data augmentation generates a new hand pose through interpolation with two existing hand poses. Unfortunately, previous triplet mining methods that used log ratio loss [10] was not directly applicable to hand pose estimation as it was designed for task retrieval. Therefore, we proposed a method for constructing a triplet and experimentally proved the superiority of our method, as shown in the results of Baseline+M(S) and Baseline+M(R).

To determine the effect of metric learning, we computed the mean and standard deviation of $Loss_{metric}$. We randomly selected three samples per minibatch and computed the mean and standard deviation of $Loss_{metric}$ for the entire training dataset. Table 2 shows that applying metric learning reduces the mean and standard deviation of $Loss_{metric}$ for both datasets. Therefore, metric learning enhances the proposed method in terms of proportionally mapping the distance of feature representations to the distance of hand poses. This eventually helped our method deliberately select feature samples from a balanced distribution of hand poses.

c: WITHOUT VS WITH HAND-DEPENDENT AUGMENTATION

Our hand-dependent augmentation achieved a more accurate performance, as demonstrated in the results of Baseline+I(S) and Base- line+I(S)+H on both datasets. This emphasizes the importance of the proposed simple technique that excludes the feature representation of semantically meaningless hand poses.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

We evaluated our method in the terms of hand pose estimation with two datasets: STB and RHD. The proposed method was compared with state-of-the-art data augmentation methods: *Manifold Mixup* [9] and *MoEx* [8]. Note that both Manifold Mixup and MoEx were originally proposed for classification tasks. The two counterparts and our method generate labels for synthetic samples based on Equation 2.

We considered two different random data augmentations, denoted by RA(5) and RA(12), as part of our baseline model. RA(5) is the data augmentation that was used in the ablation study, whereas RA(12) involves 12 widely used image transformations [5], [15] (i.e., identity, auto-contrast, equalization, rotation, solarization, color jittering, posterization, contrast, brightness, sharpness, translation, and shear). To preserve the hand shapes, we excluded shear transformation. For training, we used the same setups as used in ablation studies.

TABLE 3. Comparison with state-of-the-art implicit data augmentation methods.

Method	STB			RHD		
	EPE	MPJPE	Time hour/epoch	EPE	MPJPE	Time hour/epoch
RA(5)	7.87	7.91	0.18	16.70	19.59	0.24
RA(5)+Mixup [9]	7.91	8.02	0.28	15.94	18.75	0.38
RA(5)+MoEx [8]	7.80	7.88	0.28	16.21	18.78	0.38
RA(5)+Ours	**7.64**	**7.68**	0.30	**15.88**	**18.50**	0.40
RA(12)	7.55	7.43	0.18	16.56	19.17	0.24
RA(12)+Mixup [9]	7.76	7.78	0.28	16.64	19.17	0.38
RA(12)+MoEx [8]	7.54	7.57	0.28	16.63	19.23	0.38
RA(12)+Ours	**7.49**	**7.31**	0.30	**16.43**	**18.95**	0.40

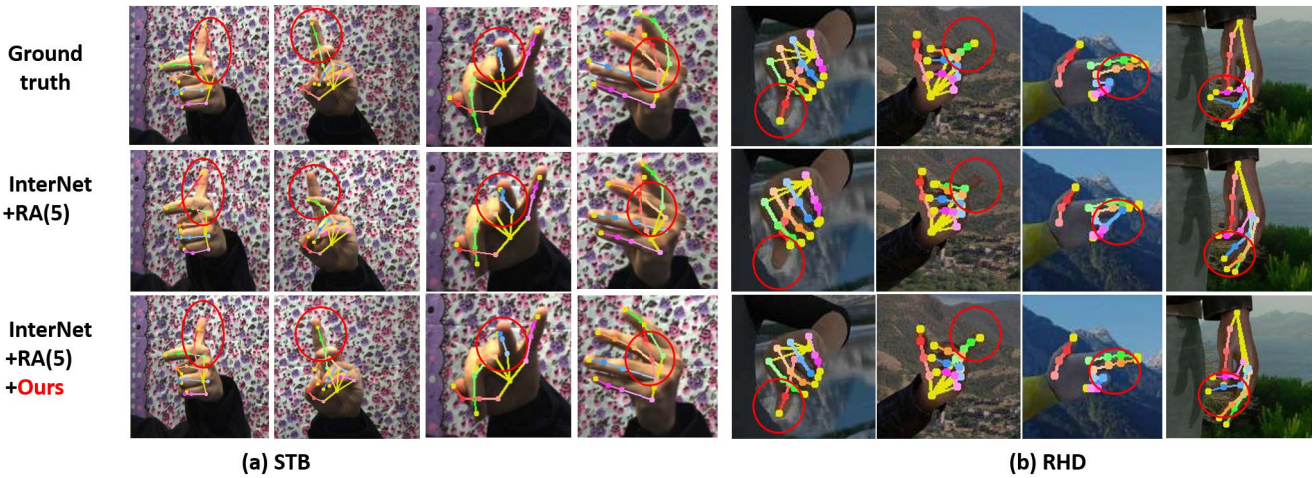

FIGURE 3. Qualitative comparison between InterNet trained using a random data augmentation technique and the proposed method.

TABLE 4. Comparison with state-of-the-art hand pose estimation methods using the STB dataset.

Method	STB (EPE)
Zimm *et al.* [26]	8.68
Yang *et al.* [28]	8.66
Spurr *et al.* [3]	8.56
Wu *et al.* [41]	8.38
Moon *et al.* [5]	7.95
Seo *et al.* [40]	7.92
InterNet [5] + Ours	**7.49** (RA(12))

TABLE 5. Comparison with state-of-the-art hand pose estimation methods using the RHD dataset.

Method	RHD (EPE)
Zimm *et al.* [26]	30.42
Yang *et al.* [28]	19.95
Spurr *et al.* [3]	19.73
Moon *et al.* [5]	20.80
Liu *et al.* [42]	19.30
InterNet [5] + Ours	**15.88** (RA(5))

Table 3 shows the results of the experimental comparison. Evidently, in some cases, either Mixup or MoEx degrades the performance of the baseline models. The possible reason for the occasional performance degradation is because both methods were originally developed for classification tasks and were not designed for hand pose estimation tasks. Conversely, the proposed method enhances the performance of hand pose estimation in every case.

We also found that complicated random data augmentations, RA(12), can improve the performance on STB. However, it can unnecessarily increase learning complexity and provide degraded performance on RHD. On the other hand, our method improved the performance of both datasets. This implies that our approach would be helpful in alleviating this phenomenon in future.

Table 4 and 5 show the results of the comparison with current state-of-the-art methods, which were outperformed by the proposed approach in terms of EPEs on the two datasets. Figure 3 shows several examples of qualitative comparisons. These examples show that InterNet trained using the proposed method more accurately estimates hand poses than when trained with only RA(5). The results show that the proposed method maintains a high accuracy in terms of hand pose estimation, even with cluttered backgrounds and hand occlusion.

V. CONCLUSION

In this paper, we presented a novel implicit semantic data augmentation method to complement existing data augmentation techniques for hand pose estimation. Unlike most existing

methods on implicit data augmentation that focus on classification tasks with discrete labels, the proposed approach, designed to address the hand pose estimation problem, focused on regression tasks with continuous and structural labels. By considering the semantic distances of hand poses, the proposed method implicitly generates extra training samples in feature space. We proposed two additional techniques: metric learning, allowing us to regulate the distribution of augmented hand poses by reflecting the semantic distance of the data to feature space, and hand-dependent augmentation, preventing semantically meaningless hand poses from being augmented. Using two well-known datasets, we empirically showed that the proposed method improves the performance of hand pose estimation compared with several state-of-the-art techniques. In the future, we will explore the possibility of applying our method to different regression tasks with continuous/structured labels for their performance improvement.

REFERENCES

[1] J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5079–5088.

[2] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. J. Argus, and T. Brox, "FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–10.

[3] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 89–98.

[4] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D human pose estimation in RGBD images for robotic task learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–7.

[5] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. Mu Lee, "InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 548–564.

[6] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "GANerated hands for real-time 3D hand tracking from monocular RGB," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 49–59.

[7] Y. Wang, X. Pan, S. Song, H. Zhang, C. Wu, and G. Huang, "Implicit semantic data augmentation for deep networks," in *Proc. Neural Inf. Proccess. Syst. (NIPS)*, 2019, pp. 1–10.

[8] B. Li, F. Wu, S.-N. Lim, S. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12383–12392.

[9] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, A. Courville, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 6438–6447.

[10] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2288–2297.

[11] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jun. 2015.

[12] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[13] L. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Proccess. Syst. (NIPS)*, 2012, pp. 1–9.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[15] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. Neural Inf. Proccess. Syst. (NIPS)*, 2020, pp. 1–10.

[16] I. Sato, H. Nishimura, and K. Yokoi, "APAC: Augmented pattern classification with neural networks," 2015, *arXiv:1505.03229*.

[17] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernandez, J. Wardlaw, and D. Rueckert, "GAN augmentation: Augmenting training data using generative adversarial networks," 2018, *arXiv:1810.10863*.

[18] W. Van Orman Quine, *Ontological Relativity, and Other Essays*. New York, NY, USA: Columbia Univ. Press, 1969.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[20] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[21] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3D pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–10.

[22] R. Li, Z. Liu, and J. Tan, "A survey on 3D hand pose estimation: Cameras, methods, and datasets," *Pattern Recognit.*, vol. 93, pp. 251–272, Sep. 2019.

[23] P. Panteleris and A. Argyros, "Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1–10.

[24] N. Kyriazis and A. Argyros, "Scalable 3D tracking of multiple interacting objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1–8.

[25] I. Oikonomidis, M. I. A. Lourakis, and A. A. Argyros, "Evolutionary quasi-random search for hand articulations tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3422–3429.

[26] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 4903–4911.

[27] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, "SO-HandNet: Self-organizing network for 3D hand pose estimation with semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6961–6970.

[28] L. Yang and A. Yao, "Disentangling latent hands for image synthesis and pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9877–9886.

[29] B. Li, F. Wu, K. Q. Weinberger, and S. Belongie, "Positional normalization," in *Proc. Neural Inf. Proccess. Syst. (NIPS)*, 2019, pp. 1–13.

[30] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, "3D hand pose tracking and estimation using stereo matching," 2016, *arXiv:1610.07214*.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.

[32] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5486–5494.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Rep. (ICLR)*, 2017, pp. 1–13.

[34] A. Gordo and D. Larlus, "Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6589–6598.

[35] G. Mori, C. Pantofaru, N. Kothari, T. Leung, G. Toderici, A. Toshev, and W. Yang, "Pose embeddings: A deep architecture for learning to match human poses," 2015, *arXiv:1507.00302*.

[36] O. Sumer, T. Dencker, and B. Ommer, "Self-supervised learning of pose embeddings from spatiotemporal relations in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4298–4307.

[37] M. Li, Y. Gao, and N. Sang, "Exploiting learnable joint groups for hand pose estimation," in *Proc. Assoc. Adavn. Arti. Intel. (AAAI)*, 2021, pp. 1–9.

[38] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[39] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[40] K. Seo, H. Cho, D. Choi, and T. Heo, "Stereo feature learning based on attention and geometry for absolute hand pose estimation in egocentric stereo views," *IEEE Access*, vol. 9, pp. 116083–116093, 2021.

[41] L. Wu, Z. Yu, Y. Liu, and Q. Liu, "Limb pose aware networks for monocular 3D pose estimation," *IEEE Trans. Image Process.*, vol. 31, pp. 906–917, 2021.
[42] Y. Liu, J. Jiang, J. Sun, and X. Wang, "Internet+: A light network for hand pose estimation," *Sensors*, vol. 21, no. 20, p. 6747, Oct. 2021.
[43] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2761–2770.
[44] T. Ohkawa, Y. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato, "Domain adaptive hand keypoint and pixel localization in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–26.
[45] J. Cheng, Y. Wan, D. Zuo, C. Ma, J. Gu, P. Tan, H. Wang, X. Deng, and Y. Zhang, "Efficient virtual view selection for 3D hand pose estimation," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, 2022, pp. 1–8.
[46] Y. Ye, A. Gupta, and S. Tulsiani, "What's in your hands? 3D reconstruction of generic objects in hands," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3895–3905.
[47] N. Nida, M. H. Yousaf, A. Irtaza, and S. A. Velastin, "Video augmentation technique for human action recognition using genetic algorithm," *ETRI J.*, vol. 44, no. 2, pp. 327–338, Jan. 2022.

KYEONGEUN SEO received the B.S., M.S., and Ph.D. degrees in computer and information science from Korea University, South Korea, in 2013, 2015, and 2022, respectively. She is currently a Senior Researcher with the Korea Electronics Technology Institute, South Korea. Her research interests include deep-learning-based hand pose estimation, gesture-based user interface, and human–computer interaction.

HYEONJOONG CHO received the B.S. degree in electronic engineering from Kyungpook National University, South Korea, in 1996, the M.S. degree in electronic and electrical engineering from the Pohang University of Science and Technology, in 1998, and the Ph.D. degree in computer engineering from Virginia Polytechnic Institute and State University (Virginia Tech), in 2006. He was a Senior Researcher at the Electronics and Telecommunications Research Institute, South Korea. He was a Senior Software Engineer at Samsung Electronics, South Korea. He is currently a Professor with the Department of Computer and Information Science, Korea University, South Korea. His research interests include real-time systems, cyber-physical systems, and machine learning.

DAEWOONG CHOI received the B.S. and M.S. degrees in computer and information science from Korea University, South Korea, in 2016, where he is currently pursuing the Ph.D. degree in computer and information science. His research interests include text entry, ten finger typing on virtual keyboards, human–computer interaction, and machine learning.

JU-DERK PARK received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Chungbuk National University, Cheongju, South Korea, in 1995, 1997, and 2011, respectively. He has been working with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, since 2000. His major research interests include analysis of EM fields and communications for the Internet of Things.