Self-Supervised RGB-NIR Fusion Video Vision Transformer Framework for rPPG Estimation

Soyeon Park*, Bo-Kyeong Kim*, and Suh-Yeon Dong*, Member, IEEE

Abstract—Remote photoplethysmography (rPPG) is a technology that can estimate noncontact heart rate (HR) using facial videos. Estimating rPPG signals requires low cost, and thus, it is widely used for noncontact health monitoring. Recent HR estimation studies based on rPPG heavily rely on the supervised feature learning on normal RGB videos. However, the RGB-only methods are significantly affected by head movements and various illumination conditions, and it is difficult to obtain large-scale labeled data for rPPG to determine the performance of supervised learning methods. To address these problems, we present the first of its kind self-supervised transformer-based fusion learning framework for rPPG estimation. In our study, we propose an end-to-end fusion video vision transformer (Fusion ViViT) network that can extract long-range local and global spatiotemporal features from videos and convert them into video sequences to enhance the rPPG representation. In addition, the self-attention of the transformer integrates the spatiotemporal representations of complementary RGB and near-infrared (NIR), which, in turn, enable robust HR estimation even under complex conditions. We use contrastive learning as a self-supervised learning (SSL) scheme. We evaluate our framework on public datasets containing both RGB, NIR videos and physiological signals. The result of near-instant HR (approximately 6 s) estimation on the large-scale rPPG dataset with various scenarios was 14.86 of root mean squared error (RMSE), which was competitive with the state-of-the-art accuracy of average HR (approximately 30 s). Furthermore, transfer learning results on the driving rPPG dataset showed a stable HR estimation performance with 16.94 of RMSE, demonstrating that our framework can be utilized in the real world.

Index Terms—Near-infrared (NIR), remote heart rate (HR) measurement, remote photoplethysmography (rPPG), RGB, self-supervised learning (SSL), video vision transformer (ViViT).

I. INTRODUCTION

Hearth rate (HR) is a physiological signal that reflects a person’s physical and mental state. Traditionally, electrocardiogram (ECG) or photoplethysmography (PPG) is used to monitor this physiological signal. However, these methods require direct human contact, making it difficult for use in a daily life or for some cases (e.g., newborns and patients with burn [1], or people who are driving [2]). Therefore, the remote PPG (rPPG) method, which is a noncontact method, was proposed to predict cardiac activities based on subtle changes in lights that are reflected from the skin in facial videos [3], [4].

In recent years, there have been several end-to-end deep learning methods that estimate rPPG or HR from facial frames of normal RGB videos [5], [6]. However, in general, these models were trained only on static RGB videos, and the estimation quality is not adequate in complex situations such as head movements and illumination changes. As a result, researchers introduced models that estimated the rPPG or HR using near-infrared (NIR) videos [7], [8]. The NIR cameras integrate a subset of wavelengths in the infrared band, and thus, they can yield clear images even under poor lightings. A recent study demonstrated the improved performance of NIR cameras under low-light conditions, and subsequently verified the feasibility of the devices under complex conditions [9]. Despite the benefits of NIR cameras, HR accuracy using only NIR videos is relatively lower than that using RGB videos under bright conditions mainly because of the light absorption characteristics of blood [10], [11]. As a result, a majority of the recent research is directed toward utilizing both RGB and NIR videos for robust HR estimation. Recent studies have shown that RGB-NIR fusion methods can contribute to an accurate HR estimation by automatically picking suitable face patches in both domains [12] or adaptively exploiting RGB and NIR signals based on the correlations between the lights from the background and the facial areas [13]. On the other hand, these methods have the drawbacks that they require strict preprocessing procedure and domain expertise, contrary to end-to-end deep learning strategies. In addition, there is a limit to obtaining generalized results on the fusion method of rPPG estimation by performing the evaluation on a small-scale self-collection dataset. In this study, we propose a new end-to-end deep learning-based learning framework that aims to design a method to effectively capture temporal information while fusing the rPPG features of the two RGB and NIR modalities. This simplifies the process by treating video frames as input and predicting rPPG signals directly. It also leads to generalized results with an efficient approach for robust rPPG estimation even in complex environments.

Recent methods [14], [15] utilize convolutional neural networks (CNNs) to form the mapping between spatiotemporal physiological features and their corresponding ground
truths. However, such CNN-based methods may be difficult to extract long-term dependencies from sequence data because of inherent architectural constraints. Transformer has been used successfully in natural language processing applications with its outstanding long-range attentional modeling abilities to address sequence-to-sequence issues [16], [17]. Subsequently, it was applied to various computer vision tasks such as classification [18], [19] and object detection [20]. In addition, the self-attention mechanism of the transformer was used in sensor fusion methods for object detection [21], and in multimodal representation learning [22] by taking multimodal tokens as input. A recent method for autonomous driving [23] combines the global context of RGB and LiDAR scenes by applying the self-attention on convoluted features to handle multimodal data well. However, in the field of computer vision, convolutional architectures are still dominant, so applications of the transformer are limited [24]. When the self-attention mechanism is used in the field of vision, it is mainly used in combination with the CNN-like architectures, or replaced only a part of the convolutions. Recently, the vision transformer (ViT) [24] that works well in image classification was developed using only a pure transformer without CNN architecture injecting image-specific inductive biases into the model. Subsequently, the video ViT (ViViT) [25], a transformer for video classification, was recently introduced, inspired by ViT. This model yielded the state-of-the-art results with regard to video classification benchmarks.

The video sequence in rPPG measurements can be treated as a signal sequence problem utilizing the long-range contextual clues [17]. ViViT can be used as an architecture to model a long-range spatiotemporal context relationship for processing video sequence problem of extracting 1-D time-series signals from 3-D videos. In addition, the self-attention mechanism of the transformer has the advantage of effectively fusing features of various modalities by automatically highlighting important parts, such as widely used in sensor fusion and multimodal representation learning [21], [22], [23]. Therefore, we consider that transformer-based ViViT could not only combine the spatiotemporal information of different RGB and NIR modalities well, but also be suitable for video sequence problems by exploiting a long-range contextual clues. Unlike previous studies that combine the convolution structure and the attention-based structure [23], [26], we propose an end-to-end approach that can extract rPPG features from raw videos of RGB and NIR using a ViViT as a convolution-free transformer.

Recent RGB-based rPPG or HR studies based on supervised deep learning were dependent on the availability of large-scale labeled data. However, large-scale biometric data are particularly scarce and are expensive to annotate because they require domain expertise. In addition, transformers have a fatal drawback that it does not generalize well when training with inadequate data, because it lacks some inductive biases inherent to CNNs, such as locality [24]. An approach to resolve this issue is to transfer knowledge using a pretrained model in different domain (i.e., ImageNet) and fine-tune it with a relatively small amount of rPPG- or HR-labeled facial videos [27]. Another approach is to overcome the insufficient amount of training data by synthesizing physiological signals [28], [29]. These approaches were applied in recent RGB-based rPPG or HR models. However, it is difficult to apply these approaches to fusion models that include NIR videos. First of all, it is difficult to find a pretrained model for large-scale NIR videos. Moreover, the partial information transfer from existing RGB-based models is also challenging. Not only can the networks between fusion and single models differ, but most of the RGB-based pretrained models are targeted at classification not requiring temporal information from videos. Thus, weight transfer from such models is not sufficient to directly apply to a model reconstructing temporal information like a rPPG signal. In addition, the second approach has a risk that a domain shift between original and synthetic data may occur [30], [31]. Therefore, we hypothesized that self-supervised learning (SSL) could be a feasible alternative for learning representations from unlabeled data.

SSL can be broadly divided into pretext task and contrastive learning methods. The pretext task performs a simple task with SSL to learn visual representations, with the aim of applying the learned representation or model weights to downstream (real) task. However, the pretext task can be significantly different from the downstream task, and is computationally expensive as training progresses on each augmented data. Also, it is difficult to generalize because pretext tasks are heuristically applied to each downstream task. Therefore, to overcome these limitations and learn richer feature representations of data, contrastive learning proposes augmenting the data itself to different versions during training, and contrasting a model’s representations of the data in ways that encourage learning features with invariance or equivariance to particular augmentations [30], [32]. This augmentation process for contrastive learning acts as specific constraints, such as an inductive bias, to extract representative features and effectively regulate and generalize the domain information. Usual standard augmentation techniques such as rotation, colorization, or optical flow, were processed for each spatial and temporal domain. Unlike these, in this study, both spatiotemporal domains are considered to estimate the rPPG signal by introducing spatiotemporal augmentation for contrastive learning. Contrastive learning is concerned with bringing representation of different augmented views of the same image closer and spreading representations of views from distinct images apart [33]. Recent studies revealed that the SSL techniques that employed contrastive learning show a similar or even better performance in rPPG estimation compared to traditional supervised approaches [30], [31].

Therefore, we propose an end-to-end self-supervised fusion learning framework for rPPG estimation. The aim of this work is to automatically learn fused representations of both the RGB and NIR in unsupervised manner for robust rPPG estimation under complex real-world scenarios. Our contribution includes as follows.

1) The proposed fusion video vision transformer (Fusion ViViT) effectively combines complementary spatiotemporal features from both RGB and NIR modalities and extracts them into video sequence. To the best of our knowledge, this is the first end-to-end deep
learning-based model that estimates rPPG using the RGB-NIR fusion method.

2) We adopt contrastive learning for SSL of transformer networks to extract better representations from insufficient data without overfitting issues. Further, this approach can provide rich transferable representations for rPPG estimation when the weights of the trained model are applied to other complex rPPG datasets.

3) We experimentally validate the feasibility of our approach under various realistic scenarios using public datasets containing both the RGB and NIR videos. We further demonstrate the scalability of the proposed approach in diverse applications.

II. METHODOLOGY

A. RGB-NIR Fusion Learning Framework for rPPG Estimation

Fig. 1 shows the overview of our proposed framework. A video encoder takes RGB and NIR facial video frames as input and produces an RGB-NIR fused rPPG representation vector. The video encoder called Fusion ViViT is described in detail in Section II-B. In the rPPG expression training process, the extracted vector enters the projection head to perform contrastive learning. In the testing process for HR estimation, the extracted vector is then passed through an rPPG estimator, and the resultant rPPG signals are utilized to calculate HR. We use a simple MLP as the rPPG estimator to extract 1-D features from high-level representations that summarize both temporal and spatial contexts.

B. Fusion Video Vision Transformer

The proposed Fusion ViViT utilizes the self-attention mechanism of transformers to effectively unify the global context of both modalities. For the self-attention computation process of this architecture, model 3 of ViViT\(^1\) [25] is used to model the interactions between all the space-time tokens in the videos. Fig. 2 shows the architecture of the fusion network. This network is adopted as a video encoder in our proposed fusion learning framework for rPPG estimation.

The spatiotemporal augmented input videos from each of two modalities, \(V^\text{RGB} \in \mathbb{R}^{T \times H \times W \times C}\) and \(V^\text{NIR} \in \mathbb{R}^{2T \times H \times W \times C}\), are stacked in the temporal dimensions, resulting in the combined video \(\widetilde{V} \in \mathbb{R}^{2T \times H \times W \times C}\), where \(T\) is the number of frames, \((H, W)\) is the resolution of the frame, and \(C\) is the number of channels. We use tubulet embedding [25] for mapping the combined video \(\widetilde{V}\) to a sequence of tokens \(\widetilde{s} \in \mathbb{R}^{2n_t \times n_h \times n_w \times d}\). The video \(\widetilde{V}\) is converted into a sequence of tubes \(x = [x_1, \ldots, x_N] \in \mathbb{R}^{N \times t \times h \times w \times C}\), where \((t \times h \times w)\) is the tubulet tube size, \(N = 2n_t n_h n_w\), \(n_t = \lfloor T/t \rfloor\), \(n_h = \lfloor (H/h) \rfloor\), and \(n_w = \lfloor (W/w) \rfloor\) is the number of tubes. Each spatiotemporal tube is rasterized into 1D token and then linearly projected to \(\mathbb{R}^d\). Finally, the positional embedding \(P \in \mathbb{R}^{N \times d}\), which incorporates positional inductive biases, is added to obtain a sequence of token embeddings \(\overline{s} \in \mathbb{R}^{N \times d}\)

\[
\overline{s} = [\text{Ex}_1, \text{Ex}_2, \ldots, \text{Ex}_N] + P
\]

where \(E \in \mathbb{R}^{d \times (\text{few}C)}\) denotes a 3-D convolution. Since \(\overline{s} \in \mathbb{R}^{2n_t \times n_h \times n_w \times d}\) is the tokenized result of \(\widetilde{V} \in \mathbb{R}^{2T \times H \times W \times C}\) stacking with two modalities on the time axis, \(\overline{s}\) preserves the order of each modality for sequence of tokens. Therefore, it facilitates the separation the resulting token after self-attention operation into each modality \(\overline{s}^\text{RGB}\) and \(\overline{s}^\text{NIR}\) in turn.

Model 3 of ViViT [25] factorizes computations by dividing them temporally and spatially, instead of computing multi-headed self-attention (MSA) across all pairs of tokens, \(\overline{s}\), at layer \(l\). For computational efficiency in each divided temporal and spatial block, the computation proceeds after the reshaping process. Therefore, self-attention is first computed only spatially by reshaping input tokens \(\overline{s}\) from \(R^{1 \times 2n_t \times n_h \times n_w \times d}\) to \(R^{2n_t \times n_h \times n_w \times d}\) (denoted by \(s_i\)) (2), and then computed temporally by reshaping \(y_i\) from \(R^{2n_t \times n_h \times n_w \times d}\) to \(R^{n_t \times n_h \times n_w \times d}\) (3). Each transformer layer \(l\), in which the factorized self-attention operation for time and space is computed, consists of two multi MSA [34] blocks, with layer normalization (LN) [35] and residual connections

\[
y'_i = \text{MSA}([\text{LN}(s'_i)] + s'_i) \tag{2}
\]

\[
y'_i = \text{MSA}([\text{LN}(y'_i)] + y'_i) \tag{3}
\]

The output is separated into two feature maps representing each of the individual modality sequentially along the time axis, \(\overline{s}^\text{RGB}\) and \(\overline{s}^\text{NIR}\) (Fig. 2). This fusion process, in which the input token combined with the two modalities, \(\overline{s}\), undergoes self-attention and then divides into \(\overline{s}^\text{RGB}\) and \(\overline{s}^\text{NIR}\), was applied twice, the number of transformer feature extractors. After performing the dense fusion through multiple transformers, we apply average pooling and flattening operations to obtain modality feature vectors, \(h^\text{RGB}\) and \(h^\text{NIR}\). These vectors are further combined using element-wise summation to form a final fused representation, \(f\), that effectively represents the spatiotemporal information of both the RGB and NIR videos.

C. Self-Supervised Learning

To overcome the limitations of supervised learning on insufficient labeled data, we pretrain the fusion network in a self-supervised manner, as shown in Fig. 3. For this purpose, we adopt the data augmentation technique and contrastive learning method for the remote HR measurement proposed in [31]. Spatiotemporal augmentation, including sparsity-based
temporal augmentation [31] and random horizontal flip augmentation, is applied in the same way to each video clip sampled from RGB and NIR raw video datasets. For each modality, two augmented clips $c_{\text{modality}1}$, $c_{\text{modality}2}$ are generated from the same video $v_{\text{modality}i}$, and a total of four augmented clips from both modalities are input into the video encoder. This produces the corresponding fused features $f_i^1$, $f_i^2$. They are fed into an MLP projection head, resulting in a projection vectors $z_i^1$, $z_i^2$. We use the following contrastive loss [32] for a pair of positive samples $(i, j)$ in the feature learning stage of SSL:

$$L_{i,j} = -\log \frac{\exp(\langle z_i, z_j \rangle/\tau)}{\sum_{k=0}^{2M} I_{k,ji} \exp(\langle z_i, z_j \rangle/\tau)}$$
where \( I_{k,gi} \in \{0, 1\} \) is an indicator function that means 1 iff \( k \neq i \), \( \tau \) is the temperature hyperparameter, and \( \text{sim}(u, v) \) computes the cosine similarity between \( u \) and \( v \). This loss enforces positive pairs (i.e., augmented images from the same input) to be close in the feature space and pushes negative pairs far away from each other. On the other hand, at the testing stage, the projection head is removed, and the fused pairs far away from each other. On the other hand, at the input) to be close in the feature space and pushes negative pairs (i.e., augmented images from the same \( i \neq k \) is an indicator function that means 1 iff \( i \neq k \).

III. RESULTS

This section describes how the proposed approach was implemented and compares it to other state-of-the-art methods to evaluate its effectiveness. First, two publicly available datasets are introduced, which consists of physiological signals and both RGB and NIR facial videos: VIPL-HR [36], and MR-NIR-Car [37]. Then, the evaluation metrics and the experimental setup are described. Finally, the experimental results are reported and analyzed, including a comparison with other noncontact HR measurement approaches, and the effectiveness of the proposed Fusion ViViT network and the learning method of the proposed framework is validated.

A. Datasets

VIPL-HR [36] dataset is a large-scale HR estimation benchmark, which consists of 3130 videos recorded in various scenarios, including head movements and illumination variations, from 107 subjects. There are nine scenarios in the database, each captured by three different devices (Logitech C310 webcam, the front camera of HUAWEI P9 smartphone, and RealSense F200 camera). Most of the existing methods [38], [39] are designed for well-controlled scenarios and their ability to generalize into less-constrained scenarios are unknown.

On the other hand, VIPL-HR, a large-scale dataset including various changes including NIR videos, can confirm the generalization ability of the fusion model. We only used a subset of RGB and NIR datasets, which were measured simultaneously on the RealSense F200 camera for RGB and NIR matching. The frame rates of the videos in VIPL-HR database range from 25 fps to 30 fps, and the resolution is 960 \( \times \) 720 and 1920 \( \times \) 1080. The ground-truth HR and PPG signals were obtained using a pulse oximeter (CONTEC CMS-60C BVP sensor).

MR-NIR-Car [37] dataset is the first video dataset that contains the ground-truth pulse signals captured under driving situations. This dataset contains 190 videos of 19 subjects captured during driving and inside parked cars. The MR-NIR-Car dataset can confirm whether the rPPG-based HR estimated from the fusion model can be applied in a new driving situation, not in a laboratory environment. The frame rate of the videos is 30 frames/s, and the resolution is 640 \( \times \) 640. The ground-truth PPG signals were extracted from a CMS 50D+ finger pulse oximeter recording at 60 frames/s, which was then downsampled to 30 frames/s and synchronized with the video recording. Following [37], [40], we considered only RGB and NIR driving recordings for the “minimal head motion” condition.

B. Evaluation Metrics

To evaluate the average HR, we use the standard deviation (SD), mean absolute error (MAE), root mean square error (RMSE), and Pearson’s correlation coefficient (R), which are all extensively adopted metrics in noncontact HR measurement [41], [42]. The evaluation metrics are calculated based on the HR error [41] representing the difference between the predicted HR and the ground truth

\[
H_e(i) = H_p(i) - H_{gt}(i)
\]  

where \( H_e(i) \) is the measurement error, \( H_p(i) \) is the predicted HR, and \( H_{gt}(i) \) is the ground truth.

1) Standard Deviation: It is used to quantify how much variance in data values. The SD is utilized to evaluate the dispersion of \( H_e \) in our experiments. Lower value indicates that the data points of \( H_e \) tend to be closer to the mean value

\[
SD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (H_e(i) - \bar{H}_e)^2}
\]  

where \( N \) is the number of measurement.

2) Mean Absolute Error: It is used to measure the differences between the predicted HR by an estimator and the ground truth by averaging the absolute values of the HR error

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |H_e(i)|.
\]

3) Root Mean Squared Error: It is used to measure the differences between the predicted HR by an estimator and the ground truth by taking the root of the mean of the HR error sum of squares. It is sensitive to outliers; lower value indicates fewer outliers in \( H_e \)

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} H_e(i)^2}.
\]

4) Pearson’s Correlation Coefficient: It measures the linear correlation between the predicted HR and the ground truth. It is defined as the covariance of two variables divided by the product of the SD

\[
\rho = \frac{\sum_{i=1}^{N} (H_{gt}(i) - \bar{H}_{gt})(H_p(i) - \bar{H}_p)}{\sqrt{\sum_{i=1}^{N} (H_{gt}(i) - \bar{H}_{gt})^2} \sqrt{\sum_{i=1}^{N} (H_p(i) - \bar{H}_p)^2}}
\]  

where \( \rho \) is \( R \), \( \bar{H}_{gt} \) is the average value of ground truths, and \( \bar{H}_p \) is the average value of predicted HRs. The value range is between +1 and −1, and a positive linear correlation close to 1 indicates good performance.

C. Experimental Settings

In the preprocessing stage before training, referring to previous studies [30], [43], the Viola and Jones [44] face detector for VIPL-HR and the S3FD [45] face detector for MR-NIRP-Car were used to crop the region with faces in the video clips. If face was not detected, the current frame was
TABLE I

SUPERVISED AND LINEAR EVALUATION RESULTS ON THE VIPL-HR DATASET

| Method            | Architecture | Modalities (source) | SD   | MAE  | RMSE  | R    |
|-------------------|--------------|---------------------|------|------|-------|------|
| Supervised        | PhysNet      | RGB-only            | 11.83| 13.81| 16.31 | 0.102|
|                   | PhysNet-late fusion | RGB-NIR           | 13.16| 11.24| 14.44 | 0.027|
|                   | Fusion ViViT (ours) | RGB-NIR           | 18.27| 25.72| 30.26 | -0.11|
| Self-supervised   | PhysNet      | RGB-only            | 19.87| 20.88| 26.43 | 0.005|
|                   | PhysNet-late fusion | RGB-NIR           | 14.15| 10.75| 14.20 | -0.039|
|                   | Fusion ViViT (ours) | RGB-NIR           | 13.58| 11.70| 14.86 | -0.085|

cropped after applying facial region detected in the previous frame. The resultant images were then resized to 112 × 112. Both videos were down-sampled to 25 frames/s for efficiency, and the length of each clip was limited to about 6.5 s (total of 160 frames) to predict a near instantaneous HR, unlike previous studies that estimated an average HR of 30 s [27], [42]. Moreover, the PPG signals were subsampled and synchronized with the video frame rate due to the discrepancy between the PPG sampling frequency and the video frame rate. The linear evaluation and ablation studies were conducted by randomly dividing the VIPL-HR [36] dataset into five subsets, selecting fold-1 as the test set, and selecting the rest as the training set. The MR-NIRP-Car [37] dataset for transfer learning is subject-separated by a ratio of 7:3. Therefore, during the evaluation process for fair comparisons, we reduced the individual dependencies of deep learning models by adopting subject-exclusive tests, meaning that the training and test sets have different subjects.

In the training stage, for contrastive learning, we applied spatiotemporal data augmentation, which consists of a sparsity-based temporal augmentation [31] followed by a random horizontal flip augmentation process. First, for temporal augmentation, sparsity-based temporal augmentation augments the video for each modality with [160 × 112 × 112] (T × H × W) size into two clips with [40 × 112 × 112] size using two randomly sampled strides from the list {1, 2, 3, 4}. Both modalities are augmented with the same sampled strides, resulting in a total of four clips. The length of each augmented clip was limited to 40 frames. The longest clip (i.e., stride of 4) contains about 6.5 s of information, and the shortest clip (i.e., stride of 1) contains about 1.5 s of information. Existing techniques widely used for spatial data augmentation (i.e., color jitter, Gaussian blur) were not designed to capture subtle color changes of pixel values from face videos, and this had a negative impact on the rPPG estimation performance [31]. Therefore, we only applied random horizontal flips to the sparsity-based temporal augmented images for spatial augmentation. In addition, we implemented the projection head as a linear layer that consists of 512 output nodes that represent the spatiotemporal video information for rPPG estimation. We also compare one RGB-only model (PhysNet [6]) and two RGB-NIR fusion models (physNet-late fusion, Fusion ViViT). In the case of physNet-late fusion, the features using the existing physNet architecture for each modality are fused at the end.

For the supervised method, when comparing PhysNet and PhysNet-late fusion model, which use the same architecture, the RGB-NIR fusion model, PhysNet-late fusion (14.44 of RMSE) performs similarly or better than the RGB-only model, PhysNet (16.31 of RMSE). This result showed the potential for performance improvement when using multiple modalities rather than a single modality for rPPG estimation. However, in the case of our Fusion ViViT, which is also an RGB-NIR fusion model, the HR estimation performance was the worst as a result of supervised learning. On the other hand, applying SSL to Fusion ViViT (14.86 of RMSE) resulted in significant performance improvement, highlighting the strengths of SSL, especially in the absence of sufficient pretraining on transformer-based models. PhysNet-late fusion also showed

D. Results and Analysis

1) Linear Evaluation: First, experiments were conducted to compare models and learning methods for rPPG estimation. Linear classification protocol [32], [46], which fixes the weights of self-supervised video encoder layers and trains only a following FC layer of the video encoder, is a common protocol to assess the quality of extracted rPPG representations. Table I shows the supervised learning and linear evaluation results for the SSL approach using VIPL-HR [36]. We also compare one RGB-only model (PhysNet [6]) and two RGB-NIR fusion models (physNet-late fusion, Fusion ViViT).
Table II
TRANSFER LEARNING RESULTS OF THE MR-NIRP-CAR DATASET FOR THE LINEAR EVALUATION SETTING

| Method       | Architecture          | Modalities (source) | SD  | MAE | RMSE | R   |
|--------------|------------------------|---------------------|-----|-----|------|-----|
| Supervised Pre-Training | PhysNet                | RGB-only            | 15.81 | 42.57 | 45.42 | 0.178 |
|               | PhysNet:late fusion    | Fusion ViViT (ours) | 26.44 | 53.80 | 59.95 | -0.42 |
|               | Fusion ViViT (ours)    |                     | 15.26 | 24.31 | 28.69 | 0.152 |
| SSL Pre-Training | PhysNet                | RGB-only            | 23.85 | 36.03 | 43.21 | -0.583 |
|               | PhysNet:late fusion    | Fusion ViViT (ours) | 29.25 | 33.85 | 42.40 | -0.84 |
|               | Fusion ViViT (ours)    |                     | 10.97 | 12.90 | 16.94 | 0.511 |

Table III
TRANSFER LEARNING RESULTS OF THE MR-NIRP-CAR DATASET FOR THE FINE-TUNED SETTING

| Method       | Architecture          | Modalities (source) | SD  | MAE | RMSE | R   |
|--------------|------------------------|---------------------|-----|-----|------|-----|
| Supervised Pre-Training | PhysNet                | RGB-only            | 20.76 | 29.32 | 33.77 | 0.294 |
|               | PhysNet:late fusion    | Fusion ViViT (ours) | 18.58 | 37.81 | 42.13 | -0.103 |
|               | Fusion ViViT (ours)    |                     | 17.73 | 19.13 | 24.92 | 0.14 |
| SSL Pre-Training | PhysNet                | RGB-only            | 32.02 | 44.17 | 51.49 | -0.696 |
|               | PhysNet:late fusion    | Fusion ViViT (ours) | 13.20 | 19.28 | 23.36 | 0.129 |
|               | Fusion ViViT (ours)    |                     | 10.81 | 22.15 | 24.64 | 0.432 |

performance improvement when SSL was applied. Therefore, through the comparison results of supervised learning and SSL method in Table I, we confirmed that RGB-NIR fusion models, especially our Fusion ViViT, are effective when SSL is applied. This demonstrates the generalizability of the rPPG estimation of the fusion model in various scenarios. However, when comparing the two fusion models in SSL, there was little difference in performance between the two RGB-NIR fusion models, and even the performance of our Fusion ViViT (14.86 of RMSE) based on transformer was slightly worse than that of physNet-late fusion (14.20 of RMSE) based on 3-DCNN. There is still a limitation that the fusion effect of the transformer-based model for rPPG is not clear.

2) Transfer Learning: Next, we evaluated the transfer learning performance on the MR-NIRP-Car [37] in both linear evaluation (fixed feature extractor) and fine-tuned settings. Transfer learning was first pretrained on the VIPL-HR datasets without labels. After that, the MR-NIRP-Car dataset with labels was used to fine-tune only the weights of a FC layer except the video encoder in the linear evaluation setting, and fine-tune all in the fine-tuned setting. We compared the transfer ability of the extracted rPPG features according to three VIPL-HR pretrained models.

For linear evaluation results in Table II, it was confirmed that the SSL pretraining method performed better than the supervised pretraining method for all three models. While the performance of the RGB-only model showed little change according to the learning method, the performance of the two RGB-NIR fusion models greatly improved in SSL, as the RMSE reduced by more than 10. This indicates that the fused features are well-learned from RGB and NIR videos through the SSL approach. In addition, when comparing the two fusion models, SSL results in Table I showed little difference between the two models, whereas in the case of the SSL results in Table II, the transformer-based Fusion ViViT (16.94 of RMSE) performed much better than the 3-DCNN-based physNet-late fusion (42.40 of RMSE). The learning framework of the SSL-based Fusion ViViT architecture can better learn the fused features, enabling faster adaptation to new datasets. In addition, our Fusion ViViT performed best in both supervised and SSL pretraining methods, demonstrating that RGB and NIR fusion through self-attention of the transformer enables robust rPPG estimation in real-world complex environments. This highlights the need for transformer-based structures rather than CNN structures, and the strength of SSL, especially in transformer-based models, when estimating rPPG through a fusion method for robust HR.

For the fine-tuned setting in Table III, similar to the linear evaluation results in transfer learning, the SSL pretraining method performed better than the supervised pretraining in both fusion models, indicating the stable HR estimation. When comparing SSL-based fusion models in both settings of transfer learning, our fusion ViViT can learn useful representations by simply training a linear FC layer on top of the learned features (16.94 of RMSE in Table II). However, physNet-late fusion did not perform well in the linear setting (42.40 of RMSE in Table II), and only showed stable performance in the fine-tuned setting (23.36 of RMSE in Table III). These results show that applying our proposed framework to pretrain using the large-scale VIPL-HR dataset allowed for faster rPPG feature extraction adaptation without deep fine-tuning. As a result, comparison with the results in Table II showed that our transformer-based Fusion ViViT framework can quickly and effectively learn fused representations for robust rPPG estimation in nonlaboratory environments such as driving situations, without having to fine-tune the entire model.

In all the results, when the SSL method was applied, the performance of the fusion models improved, whereas
TABLE IV
ABLATION ON THE NUMBER OF TRANSFORMERS AND THE TUBE TOKENIZATION IN THE RGB-NIR FUSION ViViT

| # Transformer | Tube size [time x height x width] | RMSE (bpm) |
|---------------|----------------------------------|------------|
| 1             | [4 x 8 x 8]                      | 20.16      |
| 2             | [4 x 8 x 8] [2 x 2 x 2]          | 14.86      |
| 3             | [2 x 8 x 8] [2 x 2 x 2] [2 x 7 x 7] | 15.41     |

In complex real-world environments, such as less-constrained laboratory or driving environment, beyond a well-controlled laboratory environment. The results using VIPL-HR dataset suggest the generalization ability of the fusion model, especially Fusion ViViT, through the performance improvement when SSL is applied to the fusion models. The results of transfer learning with the MR-NIR-Car dataset are the first case showing that the fusion model, pretrained with VIPL-HR, can be effectively used for rPPG-based HR measurement in a new driving situation. Also, unlike previous studies comparing the average HR accuracy over 30 s, measuring the near-instantaneous HR corresponding to about 6.5 s can quickly adapt to sudden changes such as rapid HR fluctuations. Furthermore, these performance levels are competitive even when compared to the 30 s average HR accuracy of the previous RGB-only end-to-end deep learning-based SOTA models [42] (i.e., I3D [48]: 15.9 of RMSE, PhysNet [6]: 14.8 of RMSE, DeepPhys [5]: 13.8 of RMSE). In particular, the MR-NIR-Car [37] dataset contains artifacts such as changes in ambient lightning, sudden stopping, accelerating, and turning, making it difficult to estimate rPPG signals. However, fast and stable performance of shorter rPPG-based HR measurement in the proposed SSL-based transformer-based framework increases usability in realistic situations. Fig. 4 shows stable rPPG estimation over a short period of time when transfer learning on the MR-NIR-Car dataset. Also, the proposed Fusion ViViT has 56.22 M parameters and 26.5 GFLOPs. Compared to the 3-D CNN-late fusion model with 770.22 K parameters and 107.52 GFLOPs, the number of parameters is large, but the computational complexity is low, which is advantageous for real-time inference.

However, there are several limitations. We adopted the holdout method rather than \( k \)-fold cross-validation, due to the large size of the video dataset, limited time, and resources. Although previous self-supervised studies [32] used large batch sizes to provide more negative samples, we experimented with small batch sizes because a large number of frames are required as input in limited resources for sequence estimation. However, it is significant in that it is the first attempt to verify experimentally the possibility of the spatiotemporal feature fusion using RGB and NIR based on end-to-end deep learning for robust rPPG estimation in complex situations. In addition, the experimental conditions to explain the cause of performance variation for specific variables were not varied, and the rPPG-based HR measurement performance of the newly proposed framework was simply compared with other methods. For example, we anticipate that parametric test such

IV. DISCUSSION
We verify whether the rPPG-based HR measured through the proposed fusion framework is at a level that can be utilized the performance of the RGB-only model hardly changed or deteriorated. In the case of the fusion models, domain discrepancy exists because fused features from RGB and NIR videos are obtained by different sensors with different imaging mechanisms. As reported in a previous image fusion study [47], the fusion model requires a self-supervised method to extract features through an adaptive process while retaining important information by reconstructing the input of sources. On the other hand, in the case of the RGB-only model, it is considered that the SSL method may be limited in improving the rPPG estimation performance. In addition, as a result of transfer learning, the HR estimation performance of the RGB-only model significantly deteriorated due to the large environmental difference between the data. Most of the driving dataset, including the MR-NIRP-Car, was in a dark low-light environment. This emphasized that using the RGB-NIR fusion model rather than the RGB-only model can provide more stable performance for estimating rPPG in various nonlaboratory environments.

3) Ablation Study: In our experiments, we implemented the self-supervised fusion architecture by partially modifying ViViT [25]. We demonstrated the impact of varying the number of transformers and tokenization of tubes, as shown in Table IV. The results according to the number of transformers were the worst when using a single transformer and the best when using two transformers. In particular, we observed a significant performance improvement in RMSE (about \(-5 \) b/min) when using two or more transformers. Similar to ResNet where the CNN layer learns different types of features for effective representation learning, our model learns different types of spatiotemporal representations through tokenization with different tube size of each transformer. Therefore, the multiscale fusion can effectively integrate features from different modalities. We further investigated the impact of the spatial and temporal domains on tube tokenization when the number of transformers was fixed to two. Based on the tube size (e.g., [4 x 8 x 8]) used in one transformer, we selected a tube (e.g., [2 x 8 x 8] [2 x 2 x 2] that only additionally modulates the spatial domain, and the other that modulates both space and time (e.g., [4 x 8 x 8] [2 x 2 x 2]). As a result, it was confirmed that the performance improved slightly when both the spatial and temporal domains were adjusted. Therefore, we used two transformers with [4 x 8 x 8], and [2 x 2 x 2] tube size in each embedding, consisting of 384, 768 hidden dimensions, six layers, and 12 heads in each layer.

IV. DISCUSSION
We verify whether the rPPG-based HR measured through the proposed fusion framework is at a level that can be utilized

![Fig. 4. Estimated rPPG signal of approximately 6.5 s (40 frames) compared to the ground truth in the transfer learning on the MR-NIRP-Car dataset.](image-url)
as a ANOVA test are expected to be useful in future work to analyze the effect of skin tone on the performance of various rPPG estimation methods. We plan to gradually develop it in the field of verification or application in future work.

V. CONCLUSION

In this work, we present a novel self-supervised RGB-NIR fusion learning framework for robust rPPG estimation. The proposed fusion network was built based on the self-attention mechanism of transformers to effectively combine spatiotemporal rPPG features of RGB and NIR videos and extract them into sequences. The contrastive loss was utilized for self-supervised representation learning to alleviate problems of scarce labeled biometric data. The results showed that our approach outperformed supervised baselines in HR accuracy under complex scenarios including head movements and lighting changes. Furthermore, we showed that the Fusion ViViT network, pretrained with SSL, can learn transferable features which can be applied to other relevant tasks. We believe that our proposed framework can be an effective pretraining strategy for remote PPG or HR estimation under real-world scenarios such as driving. These factors indicate the potential for use in remote physiological measurement.

REFERENCES

[1] L. A. Aarts et al., “Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—A pilot study,” *Early Hum. Develop.*, vol. 89, no. 12, pp. 943–948, 2013.

[2] Z. Guo, Z. J. Wang, and Z. Shen, “Physiological parameter monitoring of drivers based on video data and independent vector analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4374–4378.

[3] Y. Sun and N. Thakor, “Photoplethysmography revisited: From contact to noncontact, from point to imaging,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 463–477, Mar. 2016.

[4] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, “Video-based heart rate measurement: Recent advances and future prospects,” *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.

[5] W. Chen and D. McDuff, “DeepPhys: Video-based physiological measurement using convolutional attention networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 349–365.

[6] Z. Yu, X. Li, and G. Zhao, “Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks,” *2019, arXiv:1905.02419*.

[7] N. Martinez, M. Bertran, G. Sapiro, and H.-T. Wu, “Non-contact photoplethysmogram and instantaneous heart rate estimation from infrared face video,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2020–2024.

[8] M. V. Gastel, S. Stuijk, and G. D. Haan, “Motion robust remote-PPG in infrared,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1425–1433, May 2015.

[9] M. Hu et al., “Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep condition,” *PloS ONE*, vol. 13, no. 1, 2018, Art. no. e0190466.

[10] L. F. C. Martinez, G. Paez, and M. Strojnik, “Optimal wavelength selection for noncontact reflection photoplethysmography,” in *Proc. SPIE*, vol. 8011, pp. 2388–2394, Nov. 2011.

[11] W. Wang, A. C. den Brinker, and G. de Haan, “Discriminative signatures for remote-PPG,” *IEEE Trans. Biomed. Eng.*, vol. 67, no. 5, pp. 1462–1473, May 2020.

[12] S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, “Remote heart rate measurement from RGB-NIR video based on spatial and spectral face patch selection,” in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 5676–5680.

[13] K. Kurihara, D. Sugimura, and T. Hamamoto, “Adaptive fusion of RGB/NIR signals based on facebackground cross-spectral analysis for heart rate estimation,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4534–4538.

[14] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, “Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7411–7421, Oct. 2020.

[15] M. Hu, F. Qian, D. Guo, X. Wang, L. He, and F. Ren, “ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement,” *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 2506212.

[16] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, and G. Zhao, “PhysFormer: Facial video-based physiological measurement with temporal difference transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4186–4196.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[18] F. Wang et al., “Residual attention network for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[19] T. Ye, J. Zhang, Y. Li, X. Zhang, Z. Zhao, and Z. Li, “CT-Net: An efficient network for low-altitude object detection based on convolution and transformer,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[20] M. Yasuda, Y. Ohishi, C. Sun, M. Lučić, and C. Schmid, “ViViF: A video vision transformer,” *2021, arXiv:2103.15691*.

[21] X. Niu et al., “Robust remote heart rate estimation from face utilizing spatial-temporal attention,” in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG’ 2019)*, May 2019, pp. 1–8.

[22] X. Niu, S. Shan, H. Han, and X. Chen, “RhythmNet: End-to-end remote heart rate estimation from face via spatial-temporal representation,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 29, Oct. 2019, pp. 2404–2409.

[23] X. Niu, H. Han, S. Shan, and X. Chen, “SynRhythm: Learning a deep heart rate estimator from general to specific,” in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3580–3585.

[24] F. Condrea, V.-A. Ivan, and M. Leordeanu, “In search of life: Learning representations from videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 298–299.

[25] J. Gideon and S. Stent, “The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled videos,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3995–4004.

[26] H. Wang, E. Ahn, and J. Kim, “Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss,” 2021, arXiv:2107.07695.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. ICML*, Jul. 2020, pp. 1597–1607.

[28] S. Aitio, M. Aawais, and J. Kittler, “SiT: Self-supervised vision transformer,” *2021, arXiv:2104.03602*.

[29] A. Vaswani et al., “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.

[30] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, arXiv:1607.06450.

[31] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, arXiv:1607.06450.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. ICML*, Jul. 2020, pp. 1597–1607.

[33] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, “Near-infrared imaging photoplethysmography during driving,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3589–3600, Apr. 2022.
[38] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 42–55, Jan. 2012.

[39] R. Stricker, S. Müller, and H.-M. Gross, “Non-contact video-based pulse rate measurement on a mobile service robot,” in Proc. 23rd IEEE Int. Symp. Robot Human Interact. Commun., Aug. 2014, pp. 1056–1062.

[40] A. Comas, T. K. Marks, H. Mansour, S. Lohit, Y. Ma, and X. Liu, “Turnip: Time-series U-Net with recurrence for NIR imaging PPG,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2021, pp. 309–313.

[41] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A. El Saddik, “EVM-CNN: Real-time contactless heart rate estimation from facial video,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1778–1787, Jul. 2019.

[42] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, “AutoHR: A strong end-to-end baseline for remote heart rate measurement with neural searching,” IEEE Signal Process. Lett., vol. 27, pp. 1245–1249, 2020.

[43] X. Liu, X. Yang, Z. Meng, Y. Wang, J. Zhang, and A. Wong, “MANet: A motion-driven attention network for detecting the pulse from a facial video with drastic motions,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2021, pp. 2385–2390.

[44] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Dec. 2001, pp. 1–I.

[45] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3FD: Single shot scale-invariant face detector,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 192–201.

[46] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9729–9738.

[47] F. Zhao, W. Zhao, L. Yao, and Y. Liu, “Self-supervised feature adaption for infrared and visible image fusion,” Inf. Fusion, vol. 76, pp. 189–203, Dec. 2021.

[48] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.

Soyeon Park received the integrated B.S./M.S. degree from the Department of Information Technology Engineering (double B.S. major at the Department of Statistics), Sookmyung Women’s University, Seoul, South Korea. Her research interests include human–computer interaction, machine learning, and signal processing.

Bo-Kyeong Kim received the B.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2020, respectively. She is currently a Machine Learning Researcher at Nota Inc., Seoul, South Korea. Her research interests include generative models, network compression, and machine learning.

Suh-Yeon Dong (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2010, 2011, and 2016, respectively. She is currently an Assistant Professor at the Department of Information Technology Engineering, Sookmyung Women’s University, Seoul, South Korea. Her research interests include machine-learning-based biosignal processing and cognitive neuroscience.