Hindawi
Journal of Sensors
Volume 2022, Article ID 4302179, 14 pages
https://doi.org/10.1155/2022/4302179

Research Article

Visual Interaction Force Estimation Based on Time-Sensitive Dual-Resolution Learning Network

Feilu Wang,[1] **Yanan Jiang**,[1] **Yang Song**,[1,2] **Shanna Lv**,[1] **Mingkun Li**,[1] **and Rungen Ye**[1]

[1]*School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China*
[2]*Key Laboratory of Building Information Acquisition and Measurement Control Technology, Anhui Jianzhu University, Hefei 230601, China*

Correspondence should be addressed to Yang Song; esunny@ahjzu.edu.cn

Received 13 December 2021; Revised 5 January 2022; Accepted 26 January 2022; Published 8 February 2022

Academic Editor: Bin Gao

Haptic force feedback is an important perception method for humans to understand the surrounding environment. It can estimate tactile force in real time and provide appropriate feedback. It has important research value for robot-assisted minimally invasive surgery, interactive tactile robots, and other application fields. However, most of the existing noncontact visual power estimation methods are implemented using traditional machine learning or 2D/3D CNN combined with LSTM. Such methods are difficult to fully extract the contextual spatiotemporal interaction semantic information of consecutive multiple frames of images, and their performance is limited. To this end, this paper proposes a time-sensitive dual-resolution learning network-based force estimation model to achieve accurate noncontact visual force prediction. First, we perform continuous frame normalization processing on the robot running the video captured by the camera and use the hybrid data augmentation to improve the data diversity; secondly, a deep semantic interaction model is constructed based on the time-sensitive dual-resolution learning network, which is used to automatically extract the deep spatiotemporal semantic interaction information of continuous multiframe images; finally, we construct a simplified prediction model to realize the efficient estimation of interaction force. The results based on the large-scale robot hand interaction dataset show that our method can estimate the interaction force of the robot hand more accurately and faster. The average prediction MSE reaches 0.0009 N, $R2$ reaches 0.9833, and the average inference time for a single image is 6.5532 ms; in addition, our method has good prediction generalization performance under different environments and parameter settings.

1. Introduction

With the rapid development of artificial intelligence and sensor technology, as well as the urgent demand for robots in medical, intelligent services, and other fields, research on intelligent robots has important value and significance [1, 2]. Among them, sensing and estimating the force information between the robot and the object are a key step to realize the humanoid robot [3–5].

Haptics is one of the five perception methods that humans use to perceive the external environment. Humans use the tactile information obtained in the process of contact with the outside world to determine their next behavior. Similar to human intelligence, we hope that robots can freely interact with the outside world like humans to obtain dynamic or static tactile information from the outside world and then intelligently judge the current state and execute the next action based on the real-time state [6, 7].

How to make a robot able to grasp and release an object stably, the accurate estimation and feedback of the force exerted on the object by the manipulator is a crucial step. Traditional force estimation techniques are mainly based on hardware design. Researchers designed a clever touch sensor and embedded it in the hands of the robot and used hardware circuits combined with digital signal processing algorithms to measure the interaction force between the

manipulator and the object, thereby improving the accuracy of manipulator operation [8].

However, such methods have many problems such as low precision of the tactile sensor hardware, difficulty in overcoming biocompatibility, and excessive size that affects the flexibility of manipulator operation, which limits the application and promotion of such methods. For example, in the process of robot-assisted minimally invasive surgery, the tactile force sensor embedded in the robot hand needs to follow the scalpel into the human tissue in real time to continuously estimate and feedback the interaction force. There will be biological phenomena in this process. Problems such as capacitance and excessive sensor size may cause surgical risks and the flexibility and accuracy of minimally invasive surgical robots.

Therefore, in recent years, researchers have turned to the use of noncontact visual information to evaluate and feedback the interaction force, in order to make up for the shortcomings of traditional contact force estimation [4]. This type of method only uses cameras installed on the robotic arm or peripheral locations to capture the operation process of the robotic hand in real time and combines intelligent algorithms to estimate and feedback the interaction force between the robotic hand and the object. A large number of studies have shown that this type of method discards redundant tactile sensor devices and uses visual information for estimation, which can well overcome the difficulties of traditional contact force estimation methods [5, 9–21]. Therefore, research on interactive force estimation based on noncontact visual information has gradually become the mainstream technical direction and has great application value and prospects in the fields of robot-assisted minimally invasive surgery and interactive tactile robots [5, 9–23].

The current research methods for force estimation using noncontact visual information can be roughly divided into two categories: methods based on recurrent neural networks and methods based on CNN+RNN/LSTM. The first type of method based on recurrent neural network is the method mainly used in the early stage [5, 12, 13]. This method uses RNN or LSTM as the main body of the model and uses continuous multiframe images to estimate the force; the second type is based on CNN+RNN/LSTM. The research idea of the method is first use CNN (2D or 3D) to extract the deep semantic information of a single frame image or multiple frames of continuous images and then use RNN/LSTM to build a depth model to achieve continuous force estimation [15, 16, 18–21, 24].

However, the above two types of traditional methods have the following shortcomings:

(a) Force estimation method based on RNN/LSTM: this type of method uses the constructed deep RNN/LSTM to predict continuous force. However, due to the loss of information in the sequence processing process, LSTM can only obtain high-level significant visual information at the top level, but fails to obtain key low-level visual information, which makes the extracted spatiotemporal feature information insufficient and limits the force estimation performance. In addition, when the deep RNN/LSTM network is transmitted in the reverse direction, the training is quite time-consuming and difficult due to multiframe spreading, and it is difficult to converge

(b) A force estimation method based on CNN+RNN/LSTM: this type of method adds 2D/3D CNN to extract the visual salient features of the image and combines RNN/LSTM to achieve force estimation. However, the existing 2D/3D CNN+LSTM architecture method focuses on using 2D/3D CNN to extract the important visual information of static images with the same spatiotemporal resolution, instead of extracting interactive information of different spatiotemporal resolutions, which will cause the loss of dynamic visual features. Therefore, the loss of spatiotemporal information at different resolutions will result in the inability to describe dynamically changing interactive actions, which will affect the performance of force estimation. In addition, due to the subsequent deep LSTM architecture, this will also lead to time-consuming training and difficult to converge

To this end, in response to the above-mentioned insufficient extraction of spatiotemporal information and time-consuming training, inspired by the design of slowfast network [25, 26], we propose a time-sensitive dual-resolution learning network-based force estimation model, referred to as TDL network, to achieve interaction force prediction. TDL network is to construct a time-efficient dual-resolution learning structure to extract the time and space depth interactive semantic information of multiple consecutive frame images in parallel and then use the prediction module to achieve accurate estimation of continuous force. The main contributions of this paper are as follows:

(1) In order to fully extract the depth spatiotemporal semantic feature information of consecutive multiframe images, we constructed a dual-resolution learning network. The network is designed with two parallel 3DCNN branches for feature extraction and interaction with different spatiotemporal resolutions, so as to obtain the depth spatiotemporal semantic features of continuous multiframe images, which can overcome the lack of spatiotemporal information in the traditional CNN architecture

(2) In order to improve the timeliness of interaction force prediction, we streamlined the TDL architecture. First, this paper introduces three hyperparameters and reduces the number of 3D convolution channels to compress the capacity and computational complexity of parallel dual-resolution 3DCNN. Further, we also simplify the force prediction module. Two parallel global average pooling layers and a concatenation layer are designed to fuse the feature information of the two resolution branches; and to improve the generalization performance, dropout is introduced to obtain sparse

semantic interaction information. Finally, the regression module is directly used to realize the efficient prediction of the interaction force

(3) In order to improve the robustness of the force estimation model constructed in this paper, we design a hybrid data augmentation method to increase data diversity for continuous multiframe interactive data

2. Related Work

Research on force estimation using computer vision methods has become the mainstream [5, 9–21]. Through detailed investigations, we roughly divide the current research methods for force estimation into two categories: methods based on recurrent neural networks [5, 12, 13] and methods based on 2D/3D CNN+RNN/LSTM [15, 16, 18–21, 24].

The first type of method based on recurrent neural networks is mainly used for early-stage research ideas [12, 13]. This method uses RNN or LSTM as the main architecture of the model and uses continuous multiframe images to estimate the force. The research team built a behavior prediction model based on RNN and predicted the recorded hand motion video data through RNN, including scoop, stir, wash, wipe, and other actions. The results on the two datasets show that the proposed RNN model can accurately predict behavior in real time [12]. The research team proposed an LSTM-based interaction force prediction method. This method is to construct multiple parallel LSTM networks and extract the continuous interactive image features of each time stamp; the extracted features are averaged to obtain the fused interactive features; then, a regression is used to predict the interaction force. Based on the dataset constructed by the team, the effectiveness of the proposed method is evaluated [13].

The second method based on 2D/3D CNN+RNN/LSTM is the main technical method currently used. The research idea is as follows. First, a CNN (2D or 3D) network is constructed to extract the deep semantic information of a single frame image or multiple frames of continuous images; then, RNN/LSTM was introduced to build a depth model to achieve continuous force estimation.

The research team built a force estimation model based on 2DCNN+LSTM, which is composed of a 2DCNN and multiple LSTM networks, and its CNN architecture consists of three convolutional layers and pooling layers. The input of the model is a single-frame gray image of 128 multiplied by 128, and the output is the predicted interaction force. Experimental results show that the force prediction model using this model has better force prediction performance [18]. The researchers constructed a large-scale robot hand interaction scene dataset, which includes interaction scenes images under different interaction conditions. Based on this, the researchers proposed a force estimation model that combines deep learning and attention mechanism. The model uses 3DCNN as the backbone network and introduces an attention mechanism to improve feature extraction capabilities; furthermore, multiple LSTMs are used to predict the interaction force of multiple continuous images [19]. The research team designed a new method based on visual measurement to estimate the contact force of the machine, and at the same time, it was able to identify the grasped material. In this work, neuromorphic camera technology and tactile sensors are introduced to collect data on interactive scenes. On this basis, a joint discrimination method based on time-delay deep neural network (TDNN) and Gaussian process (GP) is proposed, which realizes the prediction of contact force and the identification of grasped materials. The experimental results show that the mean square error of using TDNN and GP methods to predict the contact force is 0.16 N and 0.17 N, respectively, and the average recognition accuracy of materials is 79.17% [20]. Furthermore, the team proposed a deep model fused with convolutional neural network and long- and short-term memory network. Experimental results show that the contact force can be predicted by the depth model to obtain a mean square error of 0.1 N, and it can be estimated every 10 milliseconds [21]. In addition, the research team built an optical flow fully connected deep neural network to predict the contact force in the three-dimensional direction [27].

3. Methods

The technical framework that is proposed in this paper is illustrated in Figure 1. Our method includes three main steps: performing continuous frame normalization processing on the robot running video collected by the camera, constructing a deep semantic interaction model based on the TDL network and automatically extract deep fusion spatiotemporal semantic information, and building a streamlined predictive model to achieve efficient estimation of the interaction force of the robot. In the following sections, we will describe these three steps in detail.

3.1. Definition and Enhancement of Perception Data. The purpose of this paper is to use the continuous multiframe images captured by the camera to estimate the interaction force of the sensor from the perspective of computer vision. Therefore, we use a large-scale sensor interaction force dataset as the research object [19].

The dataset is a data acquisition system composed of an electric probe system [19]. During the interaction between the probe and the object, the interaction image is recorded and captured by a high-resolution camera, and the interaction force is recorded at the same time. The data acquisition system uses an RC servo motor connected to the translation stage to control the movement of the probe. The rod-type tool installed on the translation stage will automatically move up and down to exert interactive force on the object. The research team measured the interaction force between the tool tip and the interactive object through a load cell (BCL-1L, CAS), and the force recorded by this sensor was used as the true value of the interaction force. At the same time, a high-resolution camera (Chameleon3, CM3) was used to simultaneously collect interactive images between the probe and the object. Further, a corresponding relationship between the image and the interaction force is

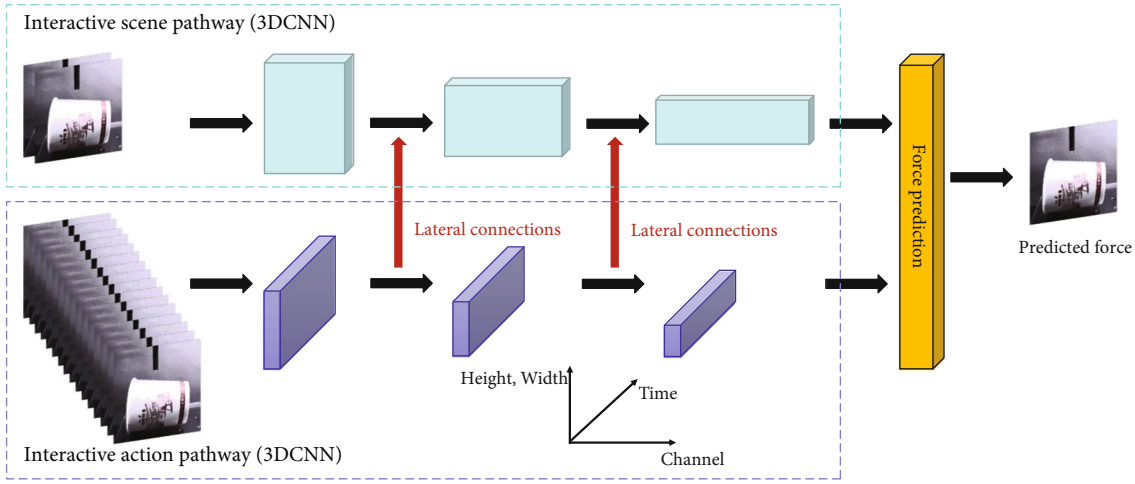

Figure 1: Schematic diagram of our proposed method, which includes three processing steps: (a) performing continuous frame normalization processing on the robot running video collected by the camera; (b) constructing a deep semantic interaction model based on the TDL network and automatically extract deep fusion spatiotemporal semantic information; (c) building a streamlined predictive model to achieve efficient estimation of the interaction force of the robot.

established, that is, each frame of image corresponds to a standard ground truth of the interaction force.

In addition, in order to increase the richness of the dataset, four objects composed of materials with different stiffness were selected, namely, sponge, paper cup, tube, and stapler, as shown in Figures 2(a)–2(d). Further, in order to increase the performance diversity of interactive images under different environmental conditions, four different angles of pressure must be applied to each type of object, which are 0°, 10°, 20°, and 30°, respectively. In addition, three different levels of light intensity are applied to the image scene of each type of object, which are, respectively, 350, 550, and 750 lux. The interactive images under different environmental conditions are shown in Figures 2(a)–2(d).

Therefore, for 4 types of objects, the camera collects interactive images of each type of object under different environmental conditions (4 types of angles, 3 types of light intensity) at a frame rate of 120 fps. A total of 17 groups of interactive images are collected, each with approximately 500 consecutive images. Finally, the dataset collected a total of $17 \times 500 \times 4 \times 3 \approx 400000$ continuous interactive images [19].

In order to effectively characterize continuous multiframe data, we first define the data source to be processed in this article. To this end, we construct a four-dimensional coordinate system, namely, x, y, z, and t, where x, y, and z represent the pixel values of the R, G, and B channels of the collected interactive image and t represents the time stamp. Therefore, each image can be represented as follows:

$$X = \left(x(t_1), x(t_2), \cdots, x\left(t_{N_X}\right)\right)^T, \tag{1}$$

where N_X represents the size of a certain group of continuous datasets, that is, the time step. The value of N_X in each group of datasets may not have the same length, which is about 500. This paper includes $4 \times 17 \times 4 \times 3 = 816$ continuous datasets. Then, the data of each time stamp t is the image data of rgb, expressed as $x(t)$, as shown in Equation (1).

Taking into account the difference in pixel fluctuations of different images collected and in order to facilitate the fusion of multiple consecutive images, we use standard deviation normalization to preprocess the dataset X and normalize X to the standard distribution range. Then, we normalize the data at each time step t as $n(t)$, which is defined as follows:

$$X_n = \left(n(t_1), n(t_2), \cdots, n\left(t_{N_X}\right)\right)^T, \tag{2}$$

where $n(t)$ in Equation (2) is each rgb image after normalization. Further, in order to extract contextual features between interactive images and reduce the negative impact of noise, we use time window Δt to window the original time series image data, and the windowed data is expressed as

$$X_n(\Delta t) = (n(\Delta t), n(\Delta t + \lambda), \cdots, n(\Delta t + k\lambda))^T, \tag{3}$$

where Δt represents the length of the window and λ represents the moving step length of the sliding window, that is, the sliding time length of each time window. k is the time multiple of the moving step, that is, represents the k-th time window data. Therefore, this article uses the above-mentioned windowed data as the basic unit for feature extraction and interaction force prediction.

In addition, in order to improve the robustness of the force estimation model, we adopt a hybrid data augmentation method. Specifically, it includes the standardization of input data (mean is 0, input data divided by standard deviation), the random rotation angle of the image (set to 0.15 in this paper), the horizontal offset ratio of the image (set to

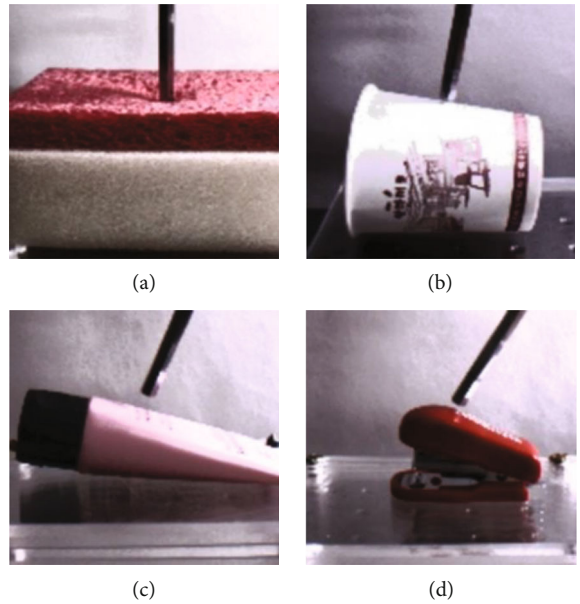

Figure 2: Image examples of the interactive force image dataset used in this paper: (a) sponge (0°, 350 lux); (b) paper cup (10°, 350 lux); (c) tube (20°, 550 lux); (d) stapler (30°, 750 lux), where 30° represents the pressure angle and 750 lux represents the light intensity.

0.125), and the vertical image offset amplitude ratio (set to 0.125). Therefore, for the windowed data $X_n(\Delta t)$, the aforementioned hybrid data augmentation method is used to diversify the input data, and the enhanced data is used to train the deep network to improve the robustness.

3.2. Deep Spatiotemporal Semantic Feature Extraction Based on TDL Network. Based on the windowed time series image data created in the previous section, we need to construct an effective feature extractor to extract the semantic information of continuous image sequence data and then realize the prediction of interaction force.

The overall framework of the proposed method is shown in Figure 1, which mainly includes the following three parts: a normalized processing module for windowing perceptual data (as described in Section 3.1), which performs windowing and data augmentation on the original perceptual video data; depth spatiotemporal semantic feature extraction module based on dual-resolution channel, which is used to extract significant semantic feature information; and time sensitivity prediction module for fast and accurate interaction force prediction.

Therefore, in this section, we will introduce the TDL network constructed in this paper in detail. The core part is the deep spatiotemporal semantic feature extraction module based on dual-resolution channels, which is used to extract deep interactive semantic features. The content described in this section is also the main contribution of this article.

3.2.1. Dual-Resolution Learning Mechanism. Inspired by the inherent mechanism of primate visual system and related research in video recognition field [25, 26], this project innovatively introduces dual-resolution path mechanism into force estimation field.

The biological study of retinal ganglion cells in the visual system of early primates found that 15-20% of retinal ganglion cells were small cells, called magnocellular (M-cells), and 80% were are smaller cells, called parvocellular (P-cells). M-cells run at a relatively high time frequency and can respond to rapid time changes, but are not sensitive to spatial details or color. In contrast, P-cells can capture the basic invariable spatial details and colors, but have low temporal resolution and slow response to stimuli. Based on the difference of response of M-cells and P-cells to different characteristics, the research team in the field of video recognition proposed the slowfast network, which uses two channels with high and low resolution to simulate the biological mechanism of M-cells and P-cells and constructs a deep learning model to realize the classification of video clips [25, 26].

Therefore, this paper draws on the above-mentioned high- and low-resolution ideas and constructs an efficient time-sensitive dual-resolution learning (TDL) network for the dynamic prediction problem faced by the cutting-edge interaction force of the robotic hand. The network designs two processing channels with different resolutions and uses them, respectively. The 3DCNN network performs deep semantic feature extraction and feature interaction between the two channels in the process, as shown in Figure 1.

For the continuous video data captured by the camera, its essence is the image data of continuous framing, as shown in Equation (1). In order to be able to obtain better context information, we have carried out windowing processing to obtain the original continuous windowed image data to be processed, as shown in Equation (3).

Through a large number of early experimental studies, we found that it is difficult to obtain better prediction results if the windowed interactive image data is processed as a whole object by purely using machine learning or deep

learning methods. The main reason is that when the continuous interactive image data is predicted as a whole, it is a symmetrical processing method, that is, the dynamic time, the width, and height of the image are treated equally, and the change process of the image characteristics over time is not described. In short, it does not perform differentiated processing and analysis of time and space characteristics, so it does not achieve better prediction results.

However, force prediction is a regression problem based on three-dimensional coordinate system, that is, as time changes, the spatiotemporal characteristics will also change. From a more essential point of view, with the time migration, there are significant differences in the class semantics and action change degree in the continuous sequence images. For example, as shown in Figure 1, in a windowed multiframe continuous image sequence, with the change of time, the class attribute of the image has not changed greatly, and it has always been the interaction scene between a certain type of medium and the cutting edge of robotic hands. In contrast, the interaction action between the cutting edge of the robotic hand and the medium is continuously and rapidly changing. Therefore, different processing methods need to be adopted for interactive scenes and interactive actions in order to improve the information capture capability.

For this reason, different from the traditional force prediction of continuous interactive image data as a whole, this paper adopts dual-resolution channel sampling for interactive scenes (semantic class attributes) and rapidly changing interactive actions, as shown in the leftmost side of Figure 1. The sampling channel for the interactive scene is called the interactive scene channel, that is, interactive scene pathway. The windowed data is sampled with a lower time sampling frequency, and the interactive images with a large time interval and low frame rate are extracted as the description data of the interactive scene.

To this end, we introduce the hyperparameter α, which is used to control the sampling frequency of the interactive scene channel, that is, an image is extracted every α time interval, which is defined as follows:

$$X_{\text{scene}}\left(\frac{\Delta t}{\alpha}\right) = \left(X_n(1), X_n(1+\alpha), \cdots, X_n(1+\Delta t-\alpha)\right)^T, \quad (4)$$

where Δt is the original windowed data; we use the above formula to complete the equal interval low-frequency sampling of the original windowed data, and the length of each interactive scene data sample is $\Delta t/\alpha$.

The sampling channel for fast changing interactive actions is called the interactive action channel, i.e., interactive action pathway. The windowed data is sampled at a higher time sampling frequency to extract high frame rate interactive action images with small time interval as the description data of fast changing interactive actions.

To this end, we introduce the hyperparameter β, which is used to control the sampling frequency of the interactive action channel, that is, to sample an image every α/β time interval. Different from the previously set hyperparameter α, this parameter is a proportional parameter, which can ensure a dynamic proportional relationship between the interactive scene and the interactive action channel. The definition of hyperparameter β is as follows:

$$X_{\text{action}}\left(\Delta t \cdot \frac{\beta}{\alpha}\right) = \left(X_n(1), X_n\left(1+\frac{\alpha}{\beta}\right), \cdots, X_n\left(1+\Delta t-\frac{\alpha}{\beta}\right)\right)^T, \quad (5)$$

where Δt is the original windowed data. We use the above formula to complete the equal interval high-frequency sampling of the original windowed data, and the length of each interactive action data sample is $\Delta t/(\alpha/\beta) = \Delta t \cdot \beta/\alpha$. The action continuity of the interactive action can be preserved, and the amount of data can be reduced, which can be extracted in this way.

Through the differential resolution sampling of the two channels, the personalized description of the original windowed data can be obtained, which is different from the traditional force prediction method.

3.2.2. Spatiotemporal Interaction Feature Extraction Based on Time-Sensitive Dual-Resolution Learning with 3DCNN. Based on the dual-resolution sampling channels constructed in the previous section, namely, the interactive scene channel and the interactive action channel, we design two deformable 3DCNN to extract the deep semantic features of the two resolution channels. In addition, in the forward extraction process, a horizontal connection strategy is added to share and interact with the feature information of the two channels, thereby obtaining richer deep semantic interaction features, as shown in Figure 1.

First, we construct two symmetrical 3DCNN networks with the same depth and extract the channel features with resnet3D-50 as the backbone network [2, 28, 29]. The resnet3D-50 model has been restructured and streamlined. In order to reduce the channel capacity and increase the forward inference speed, we have modified the number of channels of each convolution bottleneck of the resnet3D network, that is, the number of filters, and we have reduced it by a factor of 2. Therefore, the paper includes two models; one is based on the original resnet3D-50 as the backbone model, and the other is a simplified model of resnet3D-50. In the subsequent experimental sections, we also compared the performance of the simplified model with the original model. Furthermore, we adjust the channel level of the backbone network. Because of the large number of frames in the interactive action channel, if the number of channels designed by the backbone network is too large, the speed of feature extraction will be too slow and it is difficult to obtain real-time requirements. Therefore, we need to simplify the feature channel dimension of interactive action channel to improve the efficiency of feature extraction.

To this end, we introduce a hyperparameter γ to control the channel ratio of interactive scene channels and interactive action channels in the backbone network. Through the introduction of this hyperparameter γ, the number of interactive action channels can be dynamically adjusted, greatly

reducing the amount of calculation of interactive action channels and increasing the inference speed. For example, when we set $\gamma = 1/8$ and the number of feature channels in a certain layer of the interactive scene channel is 64, the number of feature channels in the corresponding interactive action channel is $64 \cdot \gamma = 8$. Moreover, the introduction of hyperparameter γ can adaptively adjust the channel structure of the model instead of manual adjustment, which has good generalization performance.

$$C^l_{X_{\text{action}}} = \gamma \cdot C^l_{X_{\text{scene}}}, \tag{6}$$

where C^l represents the number of feature channels corresponding to the lth layer, that is, the number of convolution kernels.

Finally, we introduce the lateral connections strategy, that is, in the middle layer of the backbone network of the upper and lower branches, we introduce a feature interaction module, that is, the features of interactive action channels are convoluted and integrated into the features of interactive scene channels for feature interaction, so as to improve the feature correlation between different channels. The specific process of the lateral connection strategy is as follows. We use 3D convolution to extract features of interactive action channels, the size of the convolution kernel is set to $5 \times 1 \times 1$, and the number of channels is set to $2\gamma C^l_{X\text{scene}}$, stride $= \beta$. After 3D convolution feature extraction, the feature map of the interactive action channel is integrated into the feature map of the interactive scene channel by using the concatenation operation, so as to realize the feature interaction of the two parallel channels, as shown in Figure 1.

After the double branch feature extraction channel is completed, we use the global average pooling layer to fuse the extracted features at the end of each branch to obtain the global semantic feature information.

3.3. Interaction Force Estimation Based on Time-Sensitive Prediction Module. Based on the extracted deep semantic information, we constructed a lightweight interactive prediction module. First, we design a concatenation layer to fuse the features extracted from the two-channel 3DCNN network to obtain global semantic features with dual resolution; furthermore, a dropout layer is added to sparse the above features to improve the generalization performance of the model; finally, a fully connected regression layer is connected to predict the interaction force, where the activation function is set to the sigmoid function. Among them, the input of fully connected regression layer is the final sparse feature, its dimension is concatenation feature dimension, and the output is prediction interaction force and its dimension is 1. The activation function is set in this way because the normalized range of the interaction force is $(0, 1)$, and the range of sigmoid is also $(0, 1)$, which is more suitable. It should be noted that this paper inputs continuous multiframe images as a sample into the constructed network, and the output is the interaction force corresponding to the last image in the continuous multiframe images.

4. Experimental Results

In this section, we implement a series of experiments to evaluate the effectiveness of our proposed time-sensitive dual-resolution learning network-based force estimation method. The hardware environment of all the verification experiments is a desktop workstation running Windows 10 64-bit, equipped with Intel (R) Core (TM) i5-7500 CPU @ 3.40 GHz, GeForce RTX2080 Ti 11 G GPU, 48 G RAM.

4.1. Datasets and Hyperparameter Configuration. This paper uses a large-scale force estimation dataset to conduct a series of verification experiments. As mentioned in Section 3.1, this paper uses a large-scale open source robotic hand interaction dataset to verify the performance of our method. There are four different contact media in this dataset, namely, sponge, paper cup, tube, and stapler. In addition, in order to increase the diversity of data, the interactive scene between the robot hand and the medium was collected and built a dataset from 4 different angles and 3 different light conditions. To this end, as described in Section 3.1, we obtain 400000 continuous interactive images through the cleaning of the dataset, in which there are 100,000 images for each contact medium, which is the original robotic interaction image dataset created in this paper.

Further, based on the above-mentioned original interactive image dataset, we use the windowing method described in Section 3.1 to obtain continuous windowed data. Therefore, a windowed interactive image dataset is established, which is used to verify the performance of our method. Each sample in the dataset can be expressed as shown in Equation (3), where Δt is the size of the window, that is, the number of frames contained in a sample.

In order to evaluate the performance of our method in this paper more scientifically, a training-validation-test strategy is used, that is, the training set is used for iterative training of the model, and the relevant parameters are updated; the verification set is used to verify the performance of the model online so that the relevant model parameters can be adjusted in time to make the training more effective; finally, the test set is used to evaluate the overall performance of the trained model, including evaluation from a variety of indicators. Therefore, we randomly divide the original windowed interaction dataset into 10 groups: 8 groups are randomly selected as the training set, 1 group is used as the validation set, and the remaining 1 group is used as the test set.

The verification experiment in this paper has the following unified settings: we scale the original image input size to $112 \times 112 \times 3$; the output label dimension is set to 1; the window length Δt of each sample is set to 16, and we also compared the model performance when $\Delta t = 32$; the backbone architecture of 3DCNN is set to resnet3D-50; the loss function of the model is set to mean square error; the optimizer is set to adam; the training epochs are set to 200; the learning rate is set to 0.001; the batch size is set to 32. In addition, α, β, and γ are three important hyperparameters. We have carried out different settings, and the comparison of experimental results can be seen in detail below. In the end, these three hyperparameters are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$.

4.2. Evaluation Measures. In order to effectively evaluate and compare the force estimation performance of different methods, this paper uses the following mainstream indicators to evaluate the predictive performance of all force estimation methods, namely, RMSE, MAE, MSE, and $R2$. The calculation method is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{pre}_m - \mathrm{gt}_m)^2}, \tag{7}$$

$$\mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M}|\mathrm{pre}_m - \mathrm{gt}_m|, \tag{8}$$

$$\mathrm{MSE} = \frac{1}{M}\sum_{m=1}^{M}(\mathrm{pre}_m - \mathrm{gt}_m)^2, \tag{9}$$

$$R2 = 1 - \frac{\mathrm{MSE}(\mathrm{pre}, \mathrm{gt})}{\mathrm{Var}(\mathrm{gt})}, \tag{10}$$

where pre_m represents the predicted interaction force of the mth sample and gt represents the ground truth of the mth sample, which is the label. M is the total number of samples in the test set, and var $(\bullet)$ represents the variance of the sample set.

These indicators are used to describe the performance of regression problems, but there are some differences in these indicators, which can describe the performance of the model from different angles. Therefore, through the statistical results of the above indicators, the force prediction performance of the model can be described more comprehensively. In addition, we calculate the average inference time of the statistical model, which is the average time for the model to predict the interaction force of each sample image.

4.3. Experimental Results of Force Estimation Based on TDL. In this section, we evaluate the performance of our proposed force estimation method. The loss in Figure 3 is the MSE value between the predictive force and the ground truth. It can be seen from Figure 3 that with the increase of iteration epochs, the MSE loss of the model on the training set has been showing a downward trend, and the loss value on the verification set also shows the same trend, which indicates that the model can converge on the data and is effective for the prediction of interaction force. In Figure 3, there is an obvious fluctuation around 130 epochs, which may be caused by bad data and has no effect on the training of the whole model.

4.4. Comparison with State-of-the-Art Force Estimation Methods. In order to evaluate the effectiveness of the proposed method, we verify the proposed method on a complete dataset and compare it with the state-of-the-art force estimation methods [2, 18, 19, 24, 30]. Table 1 shows the comparison results of force prediction performance between the proposed method and the state-of-the-art force estimation methods. The main results are as follows:

(1) First, we compare the performance of the method in this paper with the existing state-of-the-art force estimation methods and observe (a) to (e) and (g). The time window length of various methods in the experiment is $\Delta t = 32$, and the total amount of data is 400000 images. The ratio of training, verification, and testing is 8 : 1 : 1. The three hyperparameters of our method are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$, and training epochs are 200. It should be noted that the methods (a) to (d) are reproduced according to the methods in the references and the experimental results obtained on our dataset using these methods. Regarding the results of method (e), we directly quoted the best results in the references and did not reproduce. The $R2$ index was not calculated in the references. Therefore, we did not analyze the R 2 index of method (e). The results show that our method obtains the best prediction accuracy performance, with RMSE, MSE, MAE, and $R2$ indicators reaching 0.0397, 0.0243, 0.0016, and 0.9725, respectively. Compared with the best force estimation method (e) (3DCNN+Attention+LSTM) [19], our method reduces RMSE, MSE, and MAE by 0.0558, 0.0075, and 0.0069, respectively

In addition, we compared the average inference time of each sample, and our method (g) is faster than method (e), which is nearly 35 times faster. Compared with the traditional VGG19, Resnet, VGG+LSTM, and Resnet+LSTM methods, our method does not increase too much computational consumption. Therefore, combining the above results shows that our method has better force estimation performance.

(2) Further, we compared the effects of different sample sizes on the performance of our method and observed (f) and (g). We changed the total amount of data and randomly selected 40000 images. The ratio of training, validation, and testing was 8 : 1 : 1. The three hyperparameters are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$, and training epochs are 200. Comparing the results, it can be found that after the data volume is increased by 10 times, our method reduces the RMSE, MSE, and MAE indicators by 0.0084, 0.0087, and 0.0007, and $R2$ increases by 0.0007. This result shows that the increase in the amount of data is positive for the improvement of model performance. Therefore, this paper uses 400000 images as the original sample set for subsequent experimental evaluation

(3) Finally, we compared the simplified model and the original resnet3D-50 as the backbone of the model effect; observe (g) and (h). The length of the time window is $\Delta t = 32$, the total amount of data is set to 400000, and the ratio of training, verification, and testing is 8 : 1 : 1. The three hyperparameters are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$, and training epochs are 200. By comparing (g) and (h), it can be

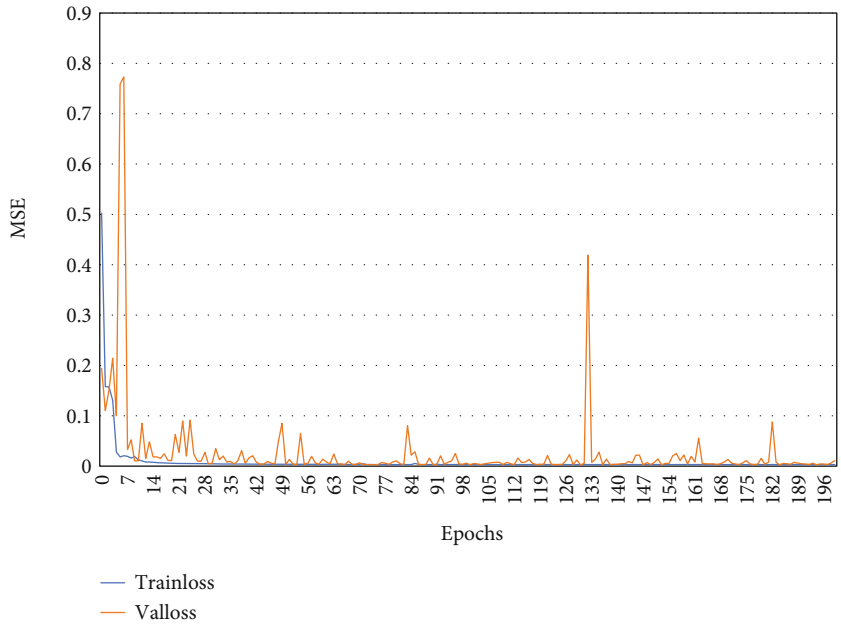

Figure 3: MSE loss curves of the proposed method on the training set and the validation set versus the number of epochs.

Table 1: Comparison results with state-of-the-art force estimation methods.

ID	Method	RMSE	MAE	MSE	$R2$	Inference time (s)
(a)	VGG19 ($\Delta t = 32$, 400000 images)	0.2201	0.1969	0.0484	-0.0279	$2.0052e-3$
(b)	Resnet ($\Delta t = 32$, 400000 images)	0.2197	0.1914	0.0483	-0.0215	$0.9593e-3$
(c)	VGG+LSTM ($\Delta t = 32$, 400000 images)	0.2161	0.1825	0.0467	-0.0026	$2.0421e-3$
(d)	Resnet+LSTM ($\Delta t = 32$, 400000 images)	0.2182	0.1840	0.0476	-0.1165	$1.3578e-3$
(e)	3DCNN+Attention+LSTM	0.0955	0.0312	0.0091	—	$3.4204e-1$
(f)	Our method ($\Delta t = 32$, 40000 images)	0.0480	0.0329	0.0023	0.9718	$9.8222e-3$
(g)	Our method ($\Delta t = 32$, 400000 images)	0.0397	0.0243	0.0015	0.9725	$9.8482e-3$
(h)	Simplified version of our method ($\Delta t = 32$, 400000 images)	0.0313	0.0183	0.0009	0.9833	$6.5532e-3$
(i)	Simplified version of our method ($\Delta t = 16$, 400000 images)	0.0295	0.0185	0.0008	0.9850	$4.5861e-3$

found that the simplified model's prediction accuracy has been improved. The three indicators of RMSE, MSE, and MAE are reduced by 0.00840, 0.00590, and 0.0006, respectively, and $R2$ is increased by 0.0108. Moreover, the average inference speed for each sample is increased by 0.003 seconds. This result shows that the streamlined network we designed can also learn the semantic features of the interactive scene well, the prediction is more accurate, and it also has a faster prediction speed

(4) In addition, we compared the impact of shortening the time window Δt on the performance of the model. The results are shown in (h) and (i). It can be found that when the length of the time window is reduced doubled, that is, $\Delta t = 16$, the prediction performance of our method is further improved, with improvements in MSE, MSE, MAE, and $R2$ indicators, and the inference speed is accelerated by 0.002 s. This result shows that in order to better implement real-world applications in the future, we can further streamline the length of the time window, increase the speed of inference, and have the same prediction accuracy

4.5. Comparison with State-of-the-Art Spatiotemporal Methods. The prediction of the interaction force between the robotic hand and the contact medium is the core problem to be solved in this paper. In essence, this problem also belongs to the video behavior analysis, that is, we perform spatiotemporal analysis for continuous multiframe images to realize the prediction of interaction force. Therefore, we investigated the latest spatiotemporal analysis methods in the field of video behavior analysis and applied these methods to the force prediction problem studied in this paper [25, 28, 29, 31, 32]. It should be noted that these

Table 2: Comparison results with state-of-the-art spatiotemporal methods.

Method	RMSE	MAE	MSE	R2	Inference time (s)
C3D-Desnet3D ($\Delta t = 32$, 400000 images)	0.1541	0.0929	0.0237	0.5797	$5.3426e-2$
T3D-EfficientnetB0 ($\Delta t = 32$, 400000 images)	0.1652	0.0768	0.0273	0.4491	$1.6993e-2$
T3D-Resnet50 ($\Delta t = 32$, 400000 images)	0.1225	0.0652	0.0150	0.7926	$1.5352e-2$
Our method ($\Delta t = 32$, 400000 images)	0.0397	0.0243	0.0015	0.9725	$9.8482e-3$
Simplified version of our method ($\Delta t = 32$, 400000 images)	0.0313	0.0183	0.0009	0.9833	$6.5532e-3$

current spatiotemporal analysis methods are mainly used to classify the types of video content, which belong to the classification problem, and the interaction force prediction that this article focuses on belongs to the regression problem. Therefore, we adjusted the final densely connected layer of these comparison methods to predict the interaction force.

The length of the time window is $\Delta t = 32$, the total amount of data is set to 400000, and the ratio of training, verification, and testing is 8 : 1 : 1. The three hyperparameters are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$, and training epochs are 200. The comparison results between the method in this paper and the state-of-the-art spatiotemporal analysis methods are shown in Table 2. C3D-Desnet3D uses Desnet3D as the backbone network. T3D-EfficientnetB0 and T3D-Resnet50 indicate that the backbone network is different. Resnet50 and EfficientnetB0 are currently the best-performing feature extraction networks, which are usually used as backbone networks to concentrate in other task frameworks. In the process of experimental comparison, we introduce it into the force estimation problem. In the process of experimental comparison, we introduce it into the force estimation problem.

Compared with C3D-Desnet3D, T3D-EfficientnetB0, and T3D-Resnet50 methods, our method reduced the RMSE, MAE, and MSE indicators by an average of 0.1076, 0.0540, and 0.0205, respectively, and the $R2$ indicator increased by 0.3654 on average. The inference time is also shorter, with an average acceleration of 0.0187 seconds. In addition, compared with the simplified version of our method, the RMSE, MAE, and MSE indicators are reduced by an average of 0.1159, 0.0600, and 0.0211, respectively, and the $R2$ indicator is increased by an average of 0.3762. The inference time is also shorter, with an average acceleration of 0.0220 seconds. Therefore, the results show that our method is more suitable for the prediction of the interaction force, with better prediction accuracy and faster inference speed than the most advanced spatiotemporal analysis methods.

4.6. The Influence of Different Contact Media on the Performance of Our Method. Different contact media have different materials, which will cause different deformations after the tip of the robotic hand touches. From a visual point of view, the predicted force will also be different, which also increases the difficulty of interactive force prediction. To this end, we evaluated the interaction force prediction performance of the method proposed in this article under different media conditions. The length of the time window is $\Delta t = 32$.

Table 3: The influence of different contact media on the performance of our method.

Contact media	RMSE	MAE	MSE	$R2$
Paper cup	0.0381	0.0208	0.0014	0.9821
Sponge	0.0334	0.0205	0.0011	0.9798
Stapler	0.0332	0.0186	0.0011	0.9803
Tube	0.0342	0.0222	0.0012	0.9837

The three hyperparameters are set to $\alpha = 16$, $\beta = 8$, and $\gamma = 1/8$, and training epochs are 200.

The statistical results are shown in Table 3. For the four different materials of paper cup, sponge, stapler, and tube, the average values of RMSE, MAE, MSE, and $R2$ of the interaction force predicted by our method are 0.0347, 0.0205, 0.0012, and 0.9814, respectively. The results show that the method proposed in this paper has stable applicability to different media and shows good generalization performance.

4.7. The Influence of Different Hyperparameters on the Performance of Our Method. The three parameters α, β, and γ are the key hyperparameters that control our method. For this reason, we compare the robustness of the model under different hyperparameter conditions. We set multiple sets of conventional α, β, and γ and retrained and evaluated the models we built. From the results in Table 4, it can be seen that under different hyperparameter settings, the model performs well and stable on the four indicators, without much change and fluctuation. The results show that the method proposed in this paper is robust and can adapt to different hyperparameters.

In addition, during the experiment, we found that when the three parameters of α, β, and γ are 8, 8, and 1/8, respectively, the prediction effect is the best, but at the same time, it also consumes inference time. Therefore, in consideration of the balance of accuracy and speed, this paper sets the three parameters of α, β, and γ to 16, 8, and 1/8, respectively, for the experimental verification in this paper.

4.8. Visualization Results of Predicted Interaction Force. In order to show the force prediction effect of our method more intuitively, we visualize the force prediction results of 100 consecutive frames of the interactive force image, as shown in Figure 4. In Figure 4, the horizontal axis is the number of each predicted frame, the range is 0 to 100, the vertical axis is the normalized predicted interaction force, the range is 0 to 1, and the unit is N. From the change trend in Figure 4, it can be found that the results of the interaction

Table 4: The influence of different hyperparameters on the performance of our method.

Model	RMSE	MAE	MSE	*R*2
Alpha = 4, beta = 1/8, tau = 16	0.0472	0.0288	0.0022	0.9733
Alpha = 8, beta = 1/8, tau = 16 (baseline)	0.0523	0.0332	0.0027	0.9650
Alpha = 2, beta = 1/8, tau = 16, batch size = 32	0.0704	0.0446	0.0049	0.9446
Alpha = 16, beta = 1/8, tau = 16, batch size = 32	0.0596	0.0404	0.0035	0.9567
Alpha = 8, beta = 1/4, tau = 16, batch size = 32	0.0455	0.0284	0.0021	0.9739
Alpha = 8, beta = 1/2, tau = 16, batch size = 32	0.0568	0.0346	0.0032	0.9639
Alpha = 8, beta = 1/16, tau = 16, batch size = 32	0.0401	0.0272	0.0016	0.9794
Alpha = 8, beta = 1/8, tau = 8, batch size = 16	0.0359	0.0228	0.0013	0.9841
Alpha = 8, beta = 1/8, tau = 32, batch size = 32	0.0503	0.0366	0.0025	0.9701

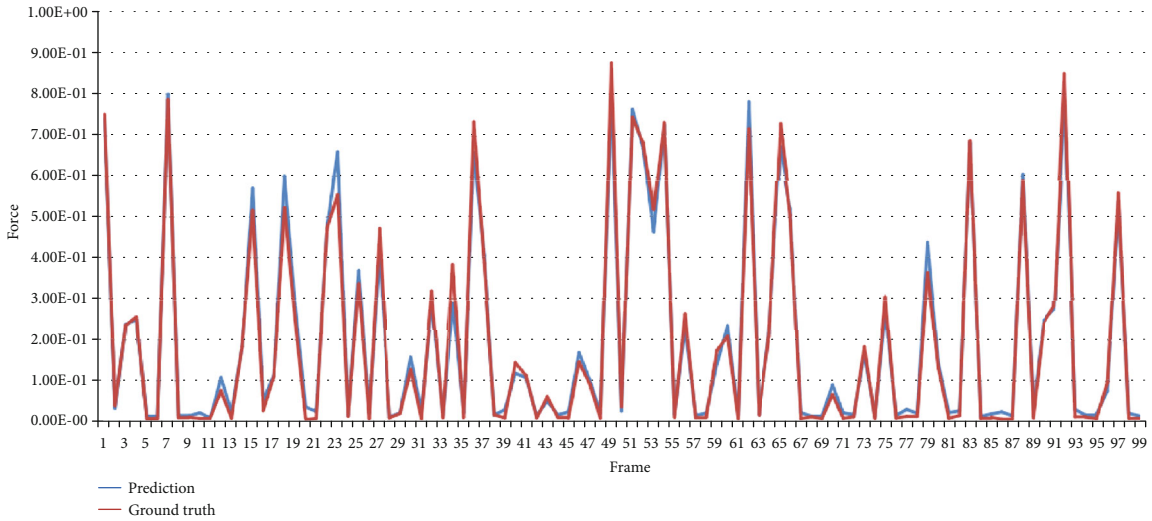

Figure 4: Visualized results of predicted interaction forces.

force predicted by the method here are basically consistent with the ground truth, and the difference is small, indicating that the prediction performance of the method is better and can well capture the change of the interaction force.

4.9. Ablation Experiments. In this section, we completed a series of ablation experiments to verify the effectiveness of our proposed method. The main innovation of our method is the construction of two resolution parallel processing pathways (interactive scene pathway and interactive action pathway), which is also the main difference from the traditional single-link CNN architecture method. In addition, we have carried out a lightweight design of the network to improve the timeliness; and in the network training stage, we designed a hybrid data augmentation to improve the robustness of our method.

Therefore, in order to verify the effectiveness of our method, we conducted a detailed ablation verification of each part of our innovative strategy. The ablation experiments include a CNN network with only interactive scene pathway, a CNN network with only interactive action pathway, a lightweight design strategy for the network, and a hybrid data augmentation strategy. Our training dataset and test dataset division and hyperparameter settings all adopt the unified standard described in Section 4.1. In terms of evaluation indicators, we selected a more balanced *R*2 indicator that reflects the accuracy of the interaction force prediction and the average inference time of a single image that reflects the prediction speed of the interaction force.

The results of the ablation experiments are shown in Table 5. Among them, method (1) represents a CNN network with only interactive scene pathway, which is the upper part of our method shown in Figure 1; method (2) represents a CNN network with only interactive action pathway, which is the lower part of our method shown in Figure 1; method (3) represents the CNN architecture fused with dual-resolution pathways, which is an integral part of our method shown in Figure 1; method (4) means adding a hybrid data augmentation strategy to the CNN architecture fused with dual-resolution pathways; method (5) indicates that the CNN architecture fused with dual-resolution pathways is added with a hybrid data augmentation strategy, and a lightweight design is carried out. This is the final simplified model TDL proposed in this paper. The detailed analysis results are as follows:

Table 5: Comparison results of ablation experiments.

Method	Interactive scene pathway	Interactive action pathway	Lightweight network design	Hybrid data augmentation	$R2$	Inference time (s)
(1)	√				0.9268	$8.2443e-3$
(2)		√			0.8224	$6.9621e-3$
(3)	√	√			0.9613	$9.8482e-3$
(4)	√	√		√	0.9725	$9.8482e-3$
(5)	√	√	√	√	0.9833	$6.5532e-3$

(a) The comparison results of methods (1), (2), and (3) show that when we use (1) or (2) alone, the accuracy of $R2$ is significantly lower than that of method (3). This result verifies the effectiveness of our dual-resolution processing pathways

(b) Comparing the results of methods (3) and (4), it can be seen that in the training phase, adding the hybrid data augmentation strategy we designed, that is, method (4), its $R2$ increased by 0.0112. This result shows that the hybrid data augmentation strategy can effectively improve the accuracy of interaction force prediction

(c) Comparing the results of methods (4) and (5), it can be seen that when we introduce a lightweight design strategy, that is, method (5), which is our proposed TDL network, its $R2$ has increased by 0.0108. In addition, the inference time of our method is shortened by 0.0033 s compared with method (4). This result shows that the TDL network proposed in this paper not only has the highest interaction force prediction accuracy but also has a faster interaction force prediction speed

In summary, the results of a series of ablation experiments show that the various innovative designs of the TDL network proposed in this paper are effective in terms of interaction force prediction.

5. Discussion and Conclusions

Real-time estimation of tactile force and appropriate feedback has important research value for robot-assisted minimally invasive surgery, interactive tactile robots, and other application fields. Compared with traditional contact tactile sensors, which are restricted by biocompatibility or excessive sensor size, force estimation and feedback through noncontact visual information has become a mainstream solution. However, the existing noncontact visual ability estimation methods are all implemented using traditional machine learning or 2D/3D CNN combined with LSTM. These methods are difficult to fully mine the contextual temporal and spatial interaction semantic information of multiple consecutive image frames, and their performance is limited.

Therefore, this paper proposes a noncontact visual force estimation method based on a time-sensitive dual-resolution learning network (TDL) to achieve accurate and rapid prediction of the interaction force. First, the continuous robot hand interactive video collected by the running camera is framed, windowed, and normalized. Furthermore, this paper constructs a deep semantic interaction model based on a time-sensitive dual-resolution learning network and automatically extracts the deep fusion spatiotemporal semantic information of consecutive multiple frames of images. Finally, we design a simplified interaction force prediction module to achieve efficient prediction of interaction force.

According to the experimental results on the large-scale robot hand interaction dataset, our method can estimate the interaction force of the robot hand more accurately than the traditional interaction force prediction method or advanced spatiotemporal analysis method. At the same time, it did not bring more time consumption, and the inference time was shorter. The average prediction MSE reaches 0.0009 N, $R2$ reaches 0.9833, and the average inference time for a single image is 6.5532 ms. In addition, through experiments under different hyperparameter conditions, experiments under different contact media conditions, and ablation experiments, the above results show that our method can still capture the interactive features well, has stable predictive performance, and shows good generalization performance. In the future, we will consider conducting verification experiments on our proposed algorithm in more real interactive scenarios, including some interactive scenarios in microsurgery. In addition, automatic deep learning is widely used in many fields [33–35], and in the future, it can be considered to be introduced into our model optimization.

Data Availability

The data used in the experimental verification of this paper is a public standard dataset, which is only supported for scientific research. The data used to support the study is available upon request to the corresponding author.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Nature Science Foundation of Anhui Province (1908085MF204), the Key Outstanding Young Talents Support Project for Universities in Anhui Province (gxyqZD2018057 and gxyq2019056), the Science and Technology Project for Urban and Rural Housing Construction of Anhui Province (2021-YF24), the Quality Engineering Project for Universities in Anhui Province (2019zyrc081, 2020szsfkc0294, and 2021jyxm0374), and the Quality Engineering Project of Anhui Jianzhu University (2020szkc03 and 2021jy70).

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[3] K. Kumagai and K. Shimonomura, "Event-based tactile image sensor for detecting spatio-temporal fast phenomena in contacts," in *2019 IEEE World Haptics Conference (WHC)*, pp. 343–348, Tokyo, Japan, 2019.

[4] R. Amin, B. N. Fariborz, M. Dimitrios, and Z. Yahya, "A novel event-based incipient slip detection using dynamic active-pixel vision sensor (DAVIS)," *Sensors*, vol. 18, no. 2, p. 333, 2018.

[5] H. Cho, H. Kim, D. K. Ko, S. C. Lim, and W. Hwang, "Which LSTM type is better for interaction force estimation?," in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, pp. 61–66, Daejeon, Republic of Korea, 2019.

[6] J. Tegin and J. Wikander, "Tactile sensing in intelligent robotic manipulation - a review," *Industrial Robot-an International Journal*, vol. 32, no. 1, pp. 64–70, 2005.

[7] G. Westling and R. S. Johansson, "Factors influencing the force control during precision grip," *Experimental Brain Research*, vol. 53, no. 2, pp. 277–284, 1984.

[8] Y. Wan, Y. Wang, and C. F. Guo, "Recent progresses on flexible tactile sensors," *materials today physics*, vol. 1, pp. 61–73, 2017.

[9] A. I. Aviles, S. M. Alsaleh, J. K. Hahn, and A. Casals, "Towards retrieving force feedback in robotic-assisted surgery: a supervised neuro-recurrent-vision approach," *IEEE Transactions on Haptics*, vol. 10, no. 3, pp. 431–443, 2017.

[10] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S. C. Zhu, "Inferring forces and learning human utilities from videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3823–3833, Las Vegas, NV, USA, 2016.

[11] T. H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros, "Towards force sensing from vision: observing hand-object interactions to infer manipulation forces," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2810–2819, Boston, MA, USA, 2015.

[12] C. Fermüller, F. Wang, Y. Yang et al., "Prediction of manipulation actions," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 358–374, 2018.

[13] H. Wonjun and L. Soo-Chul, "Inferring interaction force from visual information without using physical force sensors," *Sensors*, vol. 17, no. 11, p. 2455, 2017.

[14] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force distribution estimation using GelSlim and inverse FEM," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7, Montreal, Canada, 2019.

[15] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 951–958, Singapore, 2017.

[16] Y. Wenzhen, D. Siyuan, and A. Edward, "GelSight: high-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[17] K. Kamiyama, K. Vlack, T. Mizota, H. Kajimoto, K. Kawakami, and S. Tachi, "Vision-based sensor for real-time measuring of surface traction fields," *IEEE Computer Graphics and Applications*, vol. 25, no. 1, pp. 68–75, 2005.

[18] D. H. Lee, W. Hwang, and S. C. Lim, "Interaction force estimation using camera and electrical current without force/torque sensor," *IEEE Sensors Journal*, vol. 18, no. 21, pp. 8863–8872, 2018.

[19] S. Hochul, C. Hyeon, K. Dongyi, K. Dae-kwan, L. Soo-Chul, and H. Wonjun, "Sequential image-based attention network for inferring force estimation without haptic sensor," *IEEE Access*, vol. 7, pp. 150237–150246, 2019.

[20] F. B. Naeini, A. M. AlAli, R. Al-Husari et al., "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 1881–1893, 2020.

[21] F. B. Naeini, D. Makris, D. Gan, and Y. H. Zweiri, "Dynamic-vision-based force measurements using convolutional recurrent neural networks," *Sensors*, vol. 20, no. 16, pp. 4469–4484, 2020.

[22] F. Wang, R. Ye, S. Yang et al., "Modeling analysis and 3D force prediction of a novel piezoelectric tactile sensor," *Journal of Sensors*, vol. 2021, Article ID 3667833,, 15 pages, 2021.

[23] F. Wang and S. Yang, "Three-dimensional force prediction of a flexible tactile sensor based on radial basis function neural networks," *Journal of Sensors*, vol. 2021, Article ID 8825019, 12 pages, 2021.

[24] D. Kim, H. Cho, H. Shin, S. C. Lim, and W. Hwang, "An efficient three-dimensional convolutional neural network for inferring physical interaction force from video," *Sensors*, vol. 19, no. 16, p. 3579, 2019.

[25] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *International Conference on Computer Vision*, pp. 6202–6211, Seoul, Korea, 2019.

[26] J. Liang, L. Cao, X. Xiong, T. Yu, and A. G. Hauptmann, "Spatial-temporal alignment network for action recognition and detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

[27] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: a finite element approach," *IEEE Access*, vol. 7, pp. 173438–173449, 2019.

[28] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, Salt Lake City, UT, USA, 2018.

[29] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "ConvNet architecture search for spatiotemporal feature learning," 2017, https://arxiv.org/abs/1708.05038.

[30] J. Donahue, L. A. Hendricks, M. Rohrbach et al., "Long-term recurrent convolutional networks for visual recognition and description," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 677–691, Honolulu, HI, USA, 2017.

[31] A. Diba, M. Fayyaz, V. Sharma et al., "Temporal 3D ConvNets: new architecture and transfer learning for video classification," 2017, https://arxiv.org/abs/1711.08200.

[32] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, Long Beach, CA, USA, 2019.

[33] J. Chen, Y. N. Jiang, Z. X. Huang et al., "Fine-grained detection of driver distraction based on neural architecture search," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5783–5801, 2021.

[34] T. Elsken, J. H. Metzen, and F. Hutter, "Correction to: Neural architecture search," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1997–2017, 2019.

[35] B. Zoph and V. L. Quoc, "Neural architecture search with reinforcement learning," 2016, https://arxiv.org/abs/1611.01578.