applied sciences

MDPI

Article

A Multiorder Attentional Spatial Interactive Convolutional Neural Network (MoAS-CNN) for Low-Resolution Haptic Recognition

Kailin Wen [1,2], Jie Chu [1,*], Yu Chen [1,3], Dong Liang [1,3], Chengkai Zhang [2] and Jueping Cai [1,*]

[1] School of Microelectronics, Xidian University, Xi'an 710071, China
[2] Suzhou Honghu Qiji Electronic Technology Co., Ltd., Suzhou 215008, China
[3] The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China
* Correspondence: chujie@xidian.edu.cn (J.C.); jpcai@mail.xidian.edu.cn (J.C.)

Abstract: In haptic recognition, pressure information is usually represented as an image, and then used for feature extraction and classification. Deep learning that processes haptic information in end-to-end manner has attracted attention. This study proposes a multiorder attentional spatial interactive convolutional neural network (MoAS-CNN) for haptic recognition. The asymmetric dual-stream all convolutional neural network with integrated channel attention module is applied for automatic first-order feature extraction. Later on, the spatial interactive features based on the overall feature map are computed to improve the second-order description capability. Finally, the multiorder features are summed to improve the feature utilization efficiency. To validate the MoAS-CNN, we construct a haptic acquisition platform based on three-scale pressure arrays and collect haptic letter-shape (A–Z) datasets with complex contours. The recognition accuracies are 95.73% for 16 × 16, 98.37% for 20 × 20 and 98.65% for 32 × 32, which significantly exceeds the traditional first- and second-order CNNs and local SIFT feature.

Keywords: haptic recognition; convolutional neural network; channel attention; spatial interactive second-order feature; multiorder feature

check for updates

Citation: Wen, K.; Chu, J.; Chen, Y.; Liang, D.; Zhang, C.; Cai, J. A Multiorder Attentional Spatial Interactive Convolutional Neural Network (MoAS-CNN) for Low-Resolution Haptic Recognition. *Appl. Sci.* **2022**, *12*, 12715. https://doi.org/10.3390/app122412715

Academic Editor: Rocco Furferi

Received: 30 September 2022
Accepted: 29 November 2022
Published: 12 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Haptic interpretation is an important component of perception, providing real-time feedback on changes in the external environment, and has been widely used in practical applications, such as smart machines, wearable devices, and human–computer interaction [1,2]. The essence of haptic recognition is the recharacterization and feature extraction of pressure information to obtain multiple properties of the contact object, which is the basis for subsequent actions such as grasping, manipulating, and moving. The generic method is to consider the haptic information as a two-dimensional image. Therefore, visual data-based approaches are introduced for understanding haptic information. SIFT, SURF, chain code and other descriptors combined with clustering are applied for haptic recognition [3,4]. Pohtongkam et al. applied the BoW technique using SIFT for feature extraction and k-nearest neighbors (KNN) for evaluation to achieve object recognition by a tactile glove [5]. These local feature methods require manual feature design for specific task and are labor-intensive. Recent research has proved that convolutional neural networks (CNNs) are suitable for two-dimensional information and can automatically learn features and recognize objects [6,7]. Therefore, the haptic information can be ported to the CNN for automatic recognition. Gandarias et al. proposed classical transfer CNN models combined with transfer learning to identify large items [8]. Polic et al. designed a CNN encoder structure to reduce the dimensionality of the optical-based tactile sensor image output [9]. Cao et al. trained an end-to-end CNN for tactile recognition using residual orthogonal tiling and pyramid convolution ensemble [10].

Although the existing CNN methods can achieve better recognition performance than traditional artificial features, there are still the following challenges. Firstly, the size and pixels'

amount of haptic image depend on the density and area of the pressure sensor, and thus they are at least two orders of magnitude lower than visual RGB images [11,12]. Current mainstream CNNs are constructed for high-resolution visual images. The nonlinear fitting capability is enhanced by deepening the network [13]. However, the low-resolution haptic image limits the configured network depth, so the features extracted by shallow CNN are insufficient. Secondly, the inevitable nonideal effects of the sensing element itself reduce the accuracy of the original information mapping. Due to the flexible requirements and complex manufacturing process of the sensing elements, the pseudo-outputs caused by elastic coupling and restricted response range are inevitable, resulting in blurred pressure images [14]. In addition, the fineness of the haptic image relies on the sensitivity and response range of the sensor, which is much lower than that of the visual image. Different objects show similar shapes, so more distinguishable features are needed. The traditional shallow CNNs are not adequate in feature extraction capability and feature utilization efficiency.

To address the low-resolution and blur problem in haptic perception, a multiorder attentional spatial interactive convolutional neural network (MoAS-CNN) is proposed and a pressure information acquisition platform is built (flowchart of the overall framework is shown in Figure 1). The former improves the nonlinear fitting capability of the network by deepening the network and adding channel responses to enhance the representation of first-order high-level features; introducing spatial interactive second-order features to enhance the representation of edges and fusing multiorder features to enhance the efficiency of feature utilization. The latter validates the proposed method by constructing haptic letter shapes with complex edges.

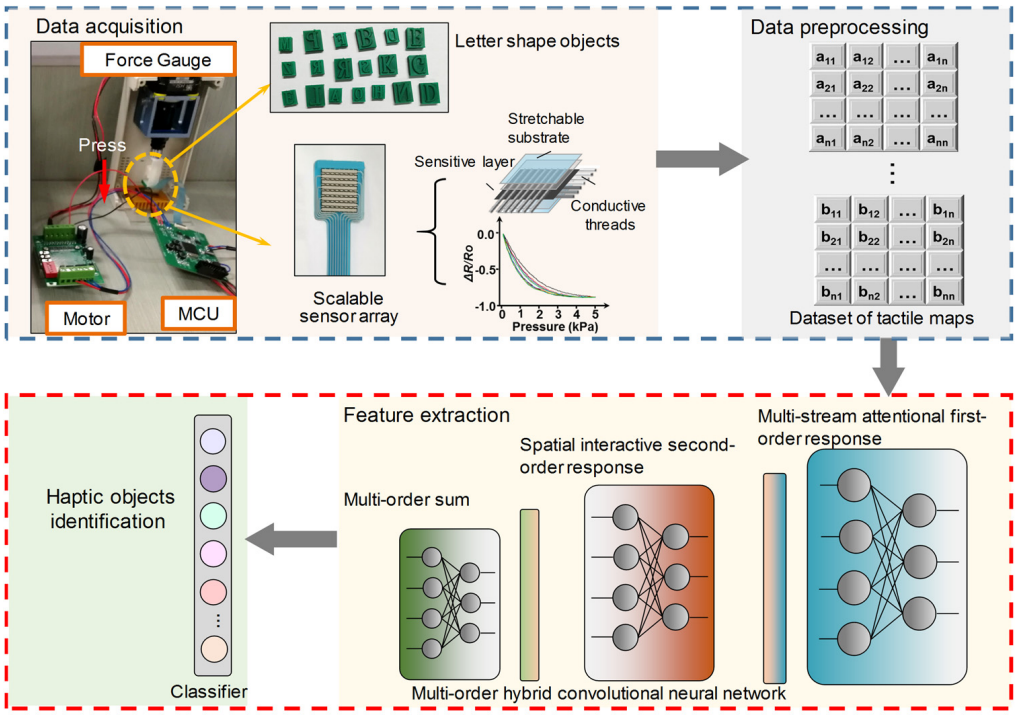

Figure 1. Flowchart of the overall framework.

The remainder of the paper is organized as follows. Section 2 discusses the nonideal characteristics about haptic images and the inadequacy of shallow CNN. In Section 3, the MoAS-CNN is proposed and each part is described. The Section 4 is dedicated to validate the proposed method using a three-scale sensor array and comparison experiments with other first-order and second-order CNNs. The discussion and conclusion are in Sections 5 and 6.

2. Problem of Insufficient Shallow CNN for Haptic Images

The original haptic information mapping of the sensing elements plays a crucial role in recognition. However, the following challenges remain. The density and area limitations as well as elastic coupling cannot be avoided due to the complex process and adhesion requirements of the sensor. In addition, the depth of field of an image depends on the sensitivity and response range of the sensing element. The resulting pixel and mapping quality of haptic images are both lower than those of visual images. The letter shape has complicated edges, which can sufficiently illustrate the issue. As shown in Figure 2a, the haptic images of 26 letter shapes are acquired by a 32 × 32 array sensor. It can be observed that the haptic images are low-resolution accompanied by blurring. Different categories show similar shapes, such as "V" and "Y," "O" and "D." The haptic image has only one channel and is grayscale, so the histogram is used as a quantitative measurement. As shown in Figure 2b, the grayscale distribution of different categories is in statistics, and the histograms of different letter shapes are easily confused. To further quantify the similarity, the Bhattacharyya coefficient is used as the proxy between each of the 26 letter shapes and is calculated as:

$$Cor(I_A, I_B) = \sum_{i=1}^{n} \sqrt{I_B(x,y) I_A(x,y)} \quad (1)$$

where $I\,(x, y)$ denotes the gray value of the haptic image at (x, y). A large positive number indicates a strong correlation between different pressure images and potential confusion. As shown in Figure 2c, different haptic shapes show strong positive interactions. Most of the correlation coefficients are in the range of 0.7–0.85, with certain categories reaching above 0.9, including K and X, G and Q, and O and U.

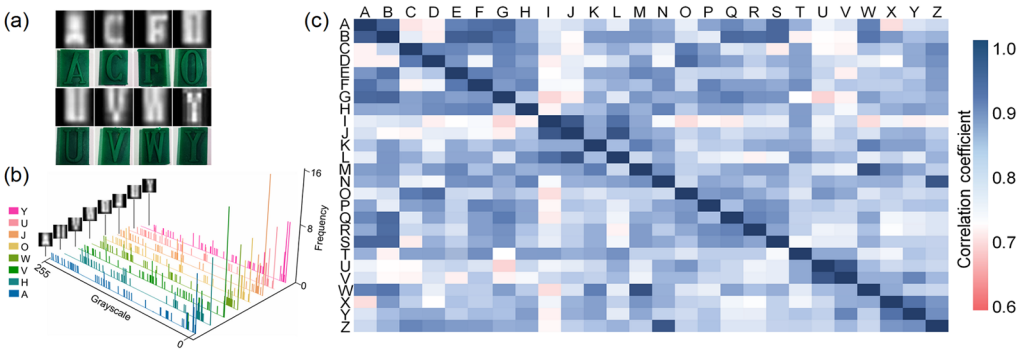

Figure 2. Nonideal effects of haptic images. (**a**)The letter-shape samples obtained by a 32 × 32 sensor array, (**b**) the grayscale distribution histograms of the samples, (**c**) the Bhattacharyya coefficient of the 26 letter shapes.

The current mainstream CNNs are designed for visual images that improve feature fitting by deepening the network. However, these nonideal effects of haptic images lead to a reduction in the differences between categories, so more distinguishable features are desired. Nevertheless, constrained by the sensor density and integrated area, the pixels of a haptic image are usually at 10^2 (RGB is above 10^4), making it impossible to improve nonlinear fitting by deepening the network, and only shallow networks can be configured [10,11]. Therefore, shallow CNNs with more powerful fitting ability need to be explored.

3. Haptic Recognition Method

3.1. MoAS-CNN Framework

Aiming to improve the feature description and utilization efficiency of the network, a shallow MoAS-CNN is constructed. Figure 3a illustrates the proposed structure, where spatial interactive-based second-order features enhance the nonlinear response of the network and a hybrid strategy of summing up the multiorder features makes the extracted

features fully utilized. It consists of three parts: first-order feature extractor, second-order feature generation, and multiorder feature hybridization.

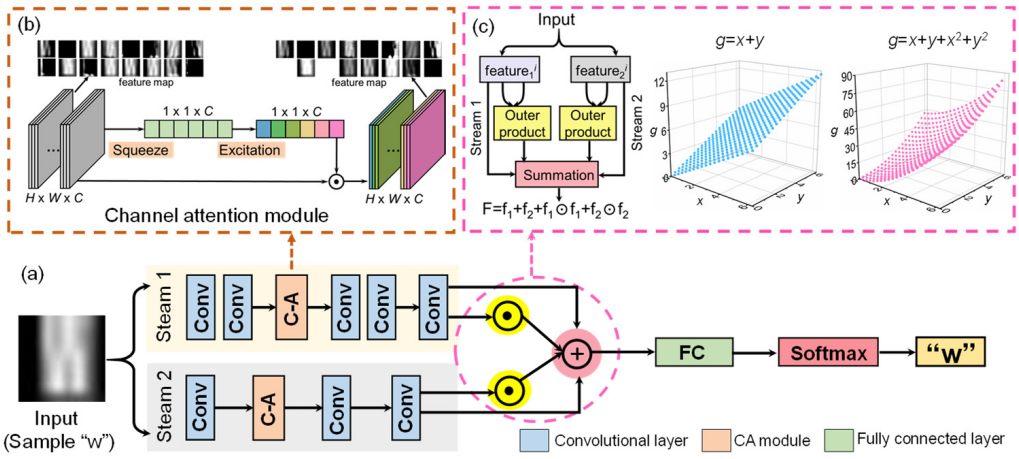

Figure 3. The MoAS-CNN framework. (**a**) The overall framework, (**b**) the squeeze–excitation channel attention module, (**c**) the spatial interactive feature generation and multiorder summation.

In particular, for low-resolution pressure inputs, the samples contain limited information and additional dimensionality reduction of the features is not expected. In our case, the pooling layers are removed and only the convolutional layers are retained to reduce the feature dimensionality loss [15,16]. An all-convolutional dual-stream neural network with the channel attention module inserted is constructed as a first-order feature extractor. The channel attention module based on squeeze and excitation operations is added for improving the first-order feature response. For second-order response, a cross-stream spatial interactive feature is generated to improve the feature nonlinear description. High-order features have been proved to be more sensitive to texture and edges [17,18]. Six $\{3 \times 3\}$ convolutional kernels (steam1) and three $\{5 \times 5\}$ convolutional kernels (steam2) are applied. Multiorder features are fused to enhance the utilization of different orders of features by summation. Features of different orders have different emphases, and complementary utilization promotes the overall nonlinearity of the network without wasting information [19]. This is beneficial for applications where the original information mapping is not sufficient or the network depth is limited.

3.2. All-Convolutional Neural Network-Based First-Order Feature Extractor with Channel Attentional Module Inserted

An asymmetric all-convolutional dual-stream CNN is configured as a feature extractor to automatically extract first-order features. CNN streams of different structures focus on respective priorities, which can fully mine image features. Mathematically, the feature maps at ith layer are calculated as:

$$f^i = \varphi(w^i \otimes f^{i-1} + b^i) \tag{2}$$

where φ denotes the activation function *ReLU*, and w and b represent the convolutional kernels and bias [20].

To highlight the emphasized parts of the features, the channel attention module including squeeze and excitation operations is inserted after the convolution layer [21]. As shown in Figure 3b, the feature maps are assigned scaling according to the channel importance to improve the feature representation. For a set of features $f \in \mathrm{R}^{h \times w \times c}$, the

squeeze operation is performed to obtain the global distribution Z{1 × 1 × *c*}, reflecting the features response over the channels. The specific mathematical expression is as follows:

$$\begin{aligned} Z &= \frac{GP(f^c)}{h \times w}, \\ GP(f^c) &= \sum_{i=1}^{h} \sum_{j=1}^{w} f^c(i,j) \end{aligned} \tag{3}$$

Here, the global average pooling is chosen to compress the feature into real numbers by spatial dimension. For excitation operation, Z is fed into two fully connected layers to further learn factor S:

$$S = \varphi(fc(Z)) \tag{4}$$

where φ denotes the activation function *ReLU* and *fc* represents the fully connected layer. These attention scalings are assigned to the initial feature map to obtain the rescaled feature map $\widetilde{f}^i$:

$$\widetilde{f}^i = S^i \times f^i \tag{5}$$

The attentional module is capable of suppressing the 2D features with lower response in the channel domain and instead increase the 2D features with higher response. After the squeeze and excitation, the feature maps are visualized in Figure 3b, and the "light and dark" changes of some feature maps can be clearly observed.

3.3. Spatial Interactive Second-Order Feature

To increase the nonlinear expression of the network, a cross-flow spatial interactive feature is proposed as a second-order response. In contrast to traditional second-order features captured by different channels at the same location, spatial interactive features in this work are generated by convolving single-channel feature maps at different streams with each other. Different stream branches focus on different extraction priorities, so the proposed method concerns more on the intrinsic relationship between the overall features in different streams. For tiny and low-resolution pressure inputs, further exploration of the interactions between different streams is necessary in the presence of network depth limitation.

The cross-flow spatial interactive feature generation is shown in Figure 4a. The extracted first-order features of stream1 and stream2 are represented as f^{stream1} and f^{stream2}. To make the interstream interaction more adequate, the original features of stream1 f^{stream1} are reconstructed without loss as $f^{\text{1-reconstruction}}$. Specifically, the f^{stream1} are split by interval sampling and stitched together, transforming the information on width and length to the channel dimension with no information loss (shown in Figure 5). In the case of this work, the dimension of the f^{stream1} {20 × 20 × 64} is adjusted to $f^{\text{1-reconstruction}}$ {5 × 5 × 1024}. The reconstructed feature of stream 1 $f^{\text{1-reconstruction}}$ and the features of stream2 f^{stream2} are subjected to an interactive operation, achieving a second-order feature.

The obtained second-order feature is performed in a sum pooling step to finally obtain the cross-flow spatial interactive feature $f^{\text{2-order}}$:

$$f^{\text{2nd-order}} = \sum_{i=1}^{nc} (f^{\text{1-reconstruction}} \otimes f^{\text{stream2}}) \tag{6}$$

To further improve the cross-flow spatial interactive feature representation, $f^{\text{2-order}}$:is normalized by element-wise signed square root followed by L_2 regularization as $f^{\text{2-order-norm}}$ [16]:

$$f^{\text{2nd-order-norm}} = \frac{sign(f^{\text{2-order}}) \times \sqrt{f^{\text{2-order}}}}{\left\| sign(f^{\text{2-order}}) \times \sqrt{f^{\text{2-order}}} \right\|_2} \tag{7}$$

where *sign* represent the symbolic function.

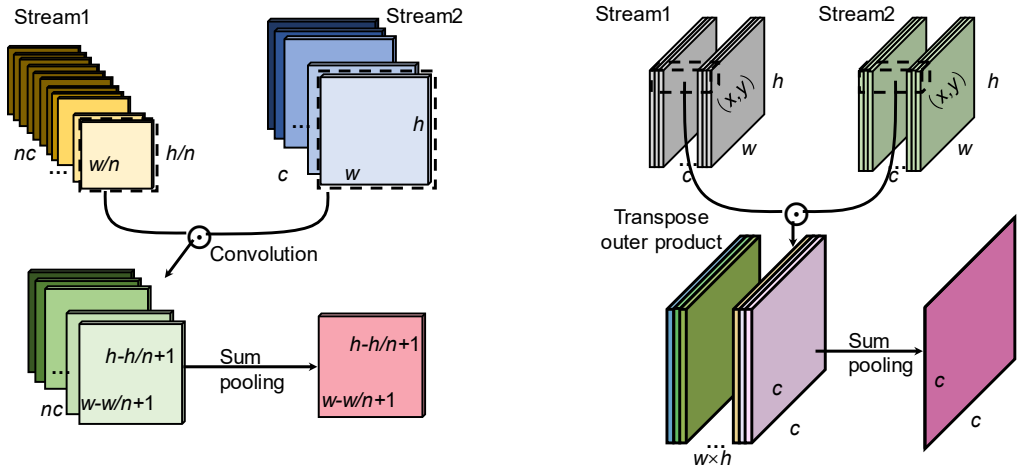

Figure 4. The generation of cross-flow spatial interactive feature. (**a**) The generation of proposed spatial interactive feature; (**b**) traditional second-order feature generation.

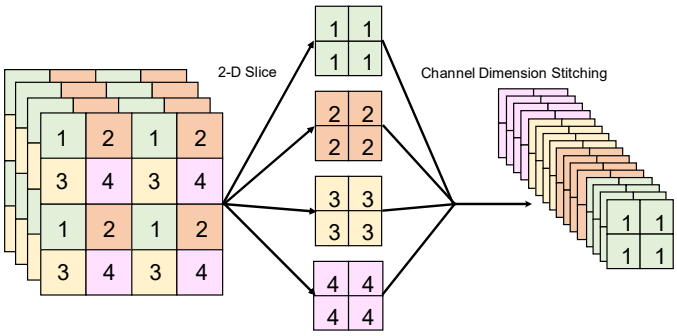

Figure 5. Schematic diagram of first-order feature reconstruction.

3.4. Multiorder Feature Hybrid

Different orders of features have different emphases and can be utilized in a more complementary way. Without adding additional parameters, the multiorder feature hybrid strategy is proposed to enhance the efficiency of feature utilization. The features of different orders are summed to obtain the fused features, as shown in:

$$y_{fusion} = f^{steam1} + f^{steam2} + g(f^{steam1} \odot f^{steam1}) + g(f^{steam2} \odot f^{steam2}) \quad (8)$$

where y_{fusion} stands for the fused feature and the $g(\bullet)$ is regularization (batch normalization here [22]). Intuitively, as shown in the upper right part of Figure 3c, only segmented linear functions are obtained when no second-order terms are added, which means that the nonlinearity appears only in a few 1D subspaces of the 2D plane R^2. However, if the second-order terms are added, nonlinearity exists $R^2 \doteq [0, \infty)^2$. The multiorder features allow the efficiency of feature utilization to be enhanced while keeping the network depth constant, facilitating the nonlinear fitting of the network.

4. Results

4.1. Experiment Setup

As shown in the upper part of Figure 1, a data-acquisition system containing three-scale pressure arrays, a force gauge, a motor driver, and a microcontroller is set up. To demonstrate the effectiveness of the proposed method, we choose letter shapes with complex contours as task targets. The stamp (0.8 cm × 1.0 cm, 1.0 cm × 1.5 cm) with raised letter shape is fixed on the force gauge to press the sensor array. The output matrix is normalized to the range 0–255 to form haptic grayscale images. The reliability of our

prototype allowed us to collect 500 samples for each letter-shape category using three scales of 16 × 16; 20 × 20; and 32 × 32 arrays, in a random pressure range of 0–5 kPa, angle and position. The data samples of each letter shape are shown in Figure 6. To avoid overfitting, data augmentation is applied to expand the sample quantity to 1500 per letter category. All the CNNs are constructed with MatConvNet framework of Matlab 2017 on an Intel Core i5-6500@3.2GHz CPU. The specific parameters of the network are set as shown in Table 1. The weights are initialized by the Xaiver initialization scheme and optimized by Adam algorithm with hyperparameters $\varepsilon = 10^{-8}$, 0.9, 0.999 [23,24]. The three-scale datasets are put into the MoAS-CNN with fivefold crossover to verify the performance of MoAS-CNN. The learning rate is 0.001, batch size 50, and training epoch 120.

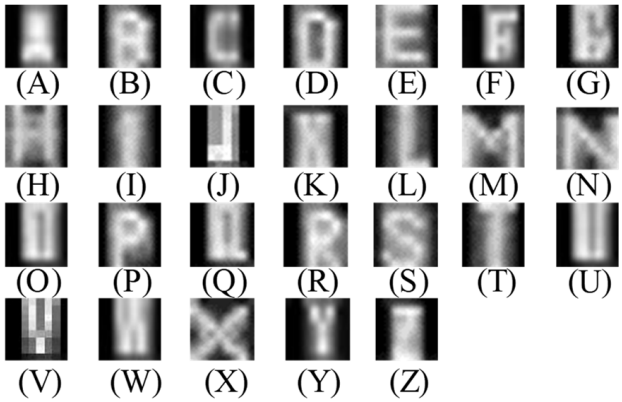

Figure 6. The dataset of 26 letter-shape samples (**A**–**Z**).

Table 1. Detailed parameters of CNNs.

Method	**MoAS-CNN**		**Bilinear CNN**		**Traditional CNN**
Layer	**Stream1**	**Stream2**	**Stream1**	**Stream2**	
Input	32 × 32 × 1		32 × 32 × 1		32 × 32 × 1
Conv1	3 × 3 × 8 3 × 3 × 16	5 × 5 × 16	3 × 3 × 8 3 × 3 × 16	5 × 5 × 16	3 × 3 × 8 3 × 3 × 16
CA module	Global Pooling FC: 1 × 1 × 16 FC: 1 × 1 × 16		Global Pooling FC: 1 × 1 × 16 FC: 1 × 1 × 16		/
Conv2	3 × 3 × 32 3 × 3 × 32	5 × 5 × 32	3 × 3 × 32 3 × 3 × 32	5 × 5 × 32	3 × 3 × 32 3 × 3 × 32
Conv3	3 × 3 × 64 3 × 3 × 64	5 × 5 × 64	3 × 3 × 64 3 × 3 × 64	5 × 5 × 64	5 × 5 × 64
second-order(output)	14 × 14 × 1		64 × 64 × 1		/
FC1	1 × 1 × 120		1 × 1 × 120		1 × 1 × 120
FC2	1 × 1 × 26		1 × 1 × 26		1 × 1 × 26

4.2. Performance of MoAS-CNN

The training and test process in each epoch are recorded. The training losses and test losses are shown in Figure 7a. It can be seen that the losses of all datasets decrease significantly and stabilize after 80 epochs. Accordingly, as shown in Figure 7b, the recognition accuracies achieved are 95.73% for 16 × 16, 98.37% for 20 × 20 and 98.65% for 32 × 32. This can be attributed to the fact that more information is captured through higher density and larger area, and thus the extracted features are more separable.

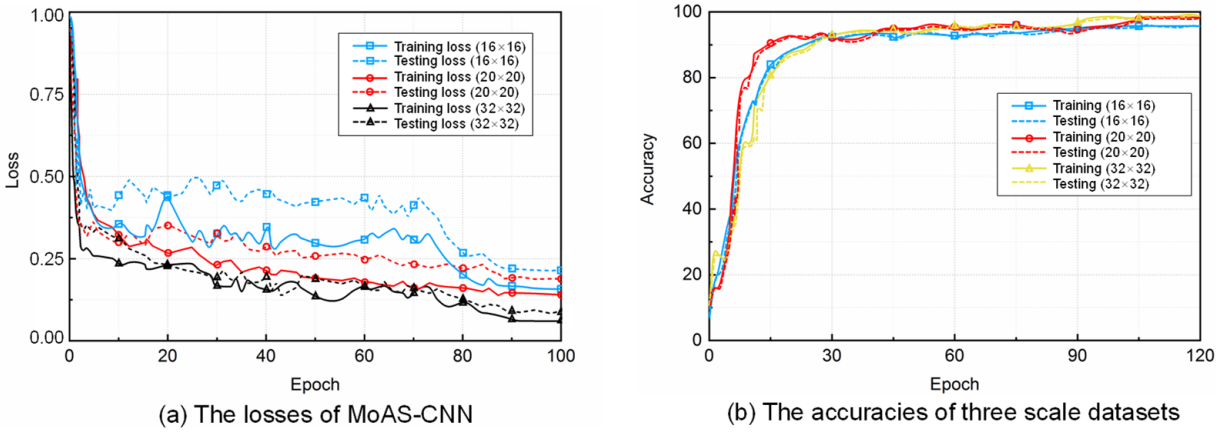

Figure 7. The performance of MoAS-CNN on three-scale datasets. (**a**)The losses in training and testing process, (**b**) the recognition accuracies in training and testing process.

We also record the gradients for all three datasets simultaneously, as shown in Figure 8a–c. The gradient values become larger in the backpropagation, and there is no gradient disappearance. Among them, the gradient value of Conv1 is the largest in the 32 × 32 dataset, making the loss converge quickly. There is no overfitting or underfitting, which indicates that our model and datasets are reasonable.

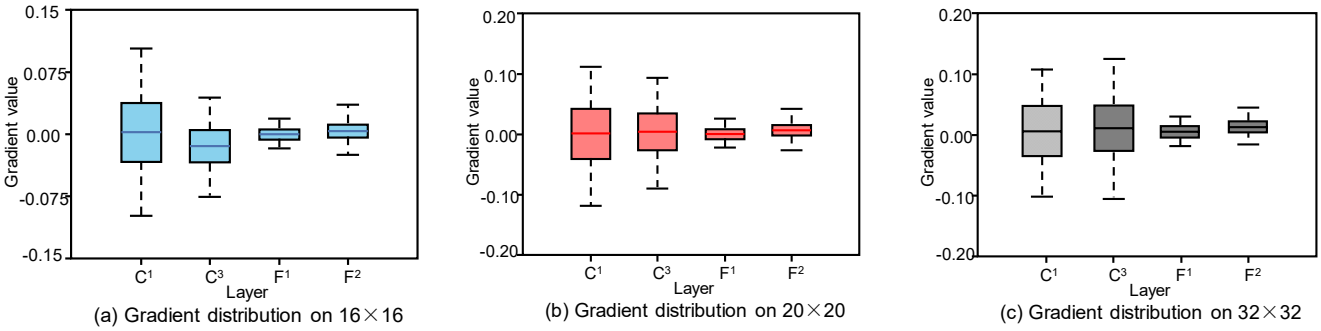

Figure 8. Gradient distribution on 16 × 16, 20 × 20, 32 × 32 datasets during training.

Figure 9 shows the test confusion matrix of the 32 × 32. Obviously, shapes with simple contours are more distinguishable, e.g., I. Conversely, complex contours are more prone to confusion, especially with similar categories, e.g., Q and G (mean Bhattacharyya correlation coefficient of 0.89).

4.3. Contribution of Spatial Interactive Second-Order Feature and Multiorder Hybrid

To further explore the contribution of the proposed spatial interactive second-order feature and hybrid order strategy, individual streams of MoAS-CNN as well as different structures are compared, and the results are shown in Figure 10. To validate the spatial interactive second-order features, we compared the recognition accuracies for f^{stream1} and $f^{\text{stream1}} + f^{\text{stream1}} \otimes f^{\text{stream1}}$ with f^{stream2} and $f^{\text{stream2}} + f^{\text{stream2}} \otimes f^{\text{stream2}}$, respectively. Due to the introduction of the spatial interactive features, the accuracy of stream1 increased by3.36%, 4.11%, and 1.96% on the three scales, and stream2 was 1.98%, 2.54%, and 1.75%. This fully demonstrates the effectiveness of the spatial interactive-based second-order feature.

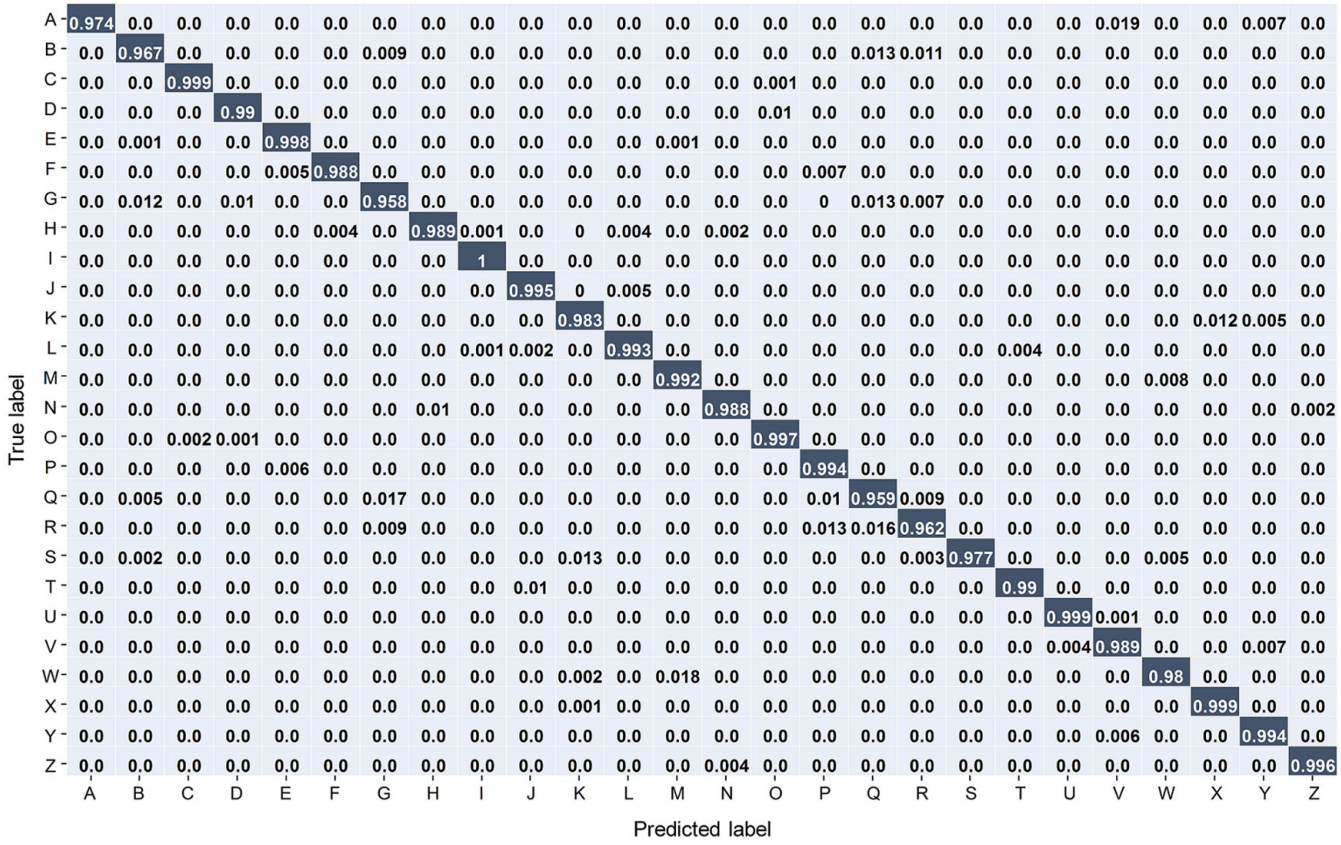

Figure 9. Confusion matrix of MoAS-CNN on 32 × 32 dataset.

To verify the effectiveness of the multiorder fusion features, the multistream first-order model $f^{\text{stream1}} + f^{\text{stream2}}$ and the multiorder hybrid $f^{\text{stream1}} + f^{\text{stream2}} + f^{\text{stream1}} \otimes f^{\text{stream2}}$ are compared. The results showed that the recognition accuracies were promoted by 2.26%, 2.54%, and 1.69%, respectively. Furthermore, it can be seen from the results of stream1 and stream2 that the accuracy is improved by replacing the large convolutional kernels with multiple layers of small convolutional kernels. This can be attributed to the increased depth of the network, where high-level features are more fully exploited.

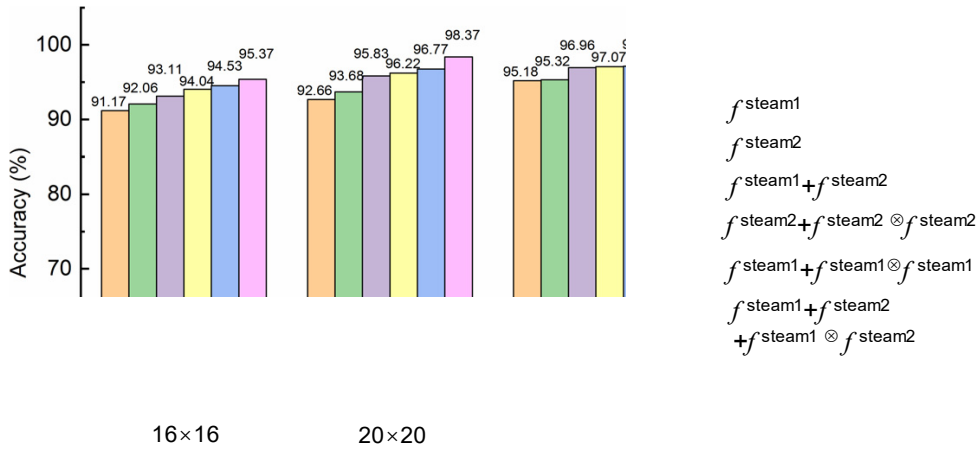

Figure 10. The contributions of cross-flow spatial interactive feature and multiorder hybrid strategies.

4.4. Performance Comparison

For a more comprehensive and fair comparison, a traditional first-order CNN, a bilinear based second-CNN, and a local feature SIFT method are chosen [16,25]. The bilinear CNN based on the traditional bilinear method with same parameters as the proposed method is constructed and the same for the first-order CNN (detailed in Table 1). The results are shown in Figure 11. It can be seen that the accuracies of MoAS-CNN are significantly higher than those of other methods on all sensor scales, especially for smaller-scale pressure arrays. This result illustrates the effectiveness of the proposed MoAS-CNN for low-resolution haptic image. For 16 × 16, the MoAS-CNN obtained 95.73%, while the traditional bilinear CNN was 92.46%, traditional CNN 89.79%, and local feature method 67.24%. Since the inputs of both 16 × 16 and 20 × 20 scales are too small to be applied in SIFT, which can only extract 3–5 key points, the recognition effect is very limited.

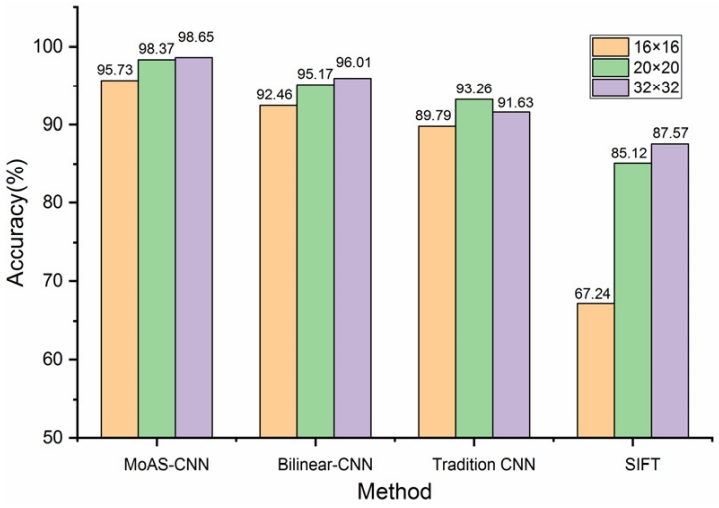

Figure 11. Performance comparison with other first-order, second-order and traditional local feature methods.

For the samples that the model failed to identify, the true labels and predicted probabilities are shown in Figure 12, marked in orange. These haptic images are collected in a random pressure range of 0.5–3 kPa, angle and position. The inference value of some letter categories with similar shapes are very close, indicating that the differences between features are not obvious. As a comparison, 10 volunteers (7 males and 3 females, age [20,30] are invited to identify the misclassified samples and the results are marked in green. It can be observed that some samples were not recognized by the neural network or humans, as shown in Figure 12a–f, and others were recognized by humans, but not by neural networks, as in Figure 12g–i. No volunteer can completely reclassify all failed samples. The misclassification can be mainly attributed to the following: firstly, the restricted sensor density leads to fine edges not being captured; secondly, the elastic coupling of the flexible sensors causes the unpressed pixels around the pressed pixels to be deformed, producing pseudo-outputs. Thirdly, features are not sufficiently mined through the MoAS-CNN and the features of similarly shape categories are not learned separately.

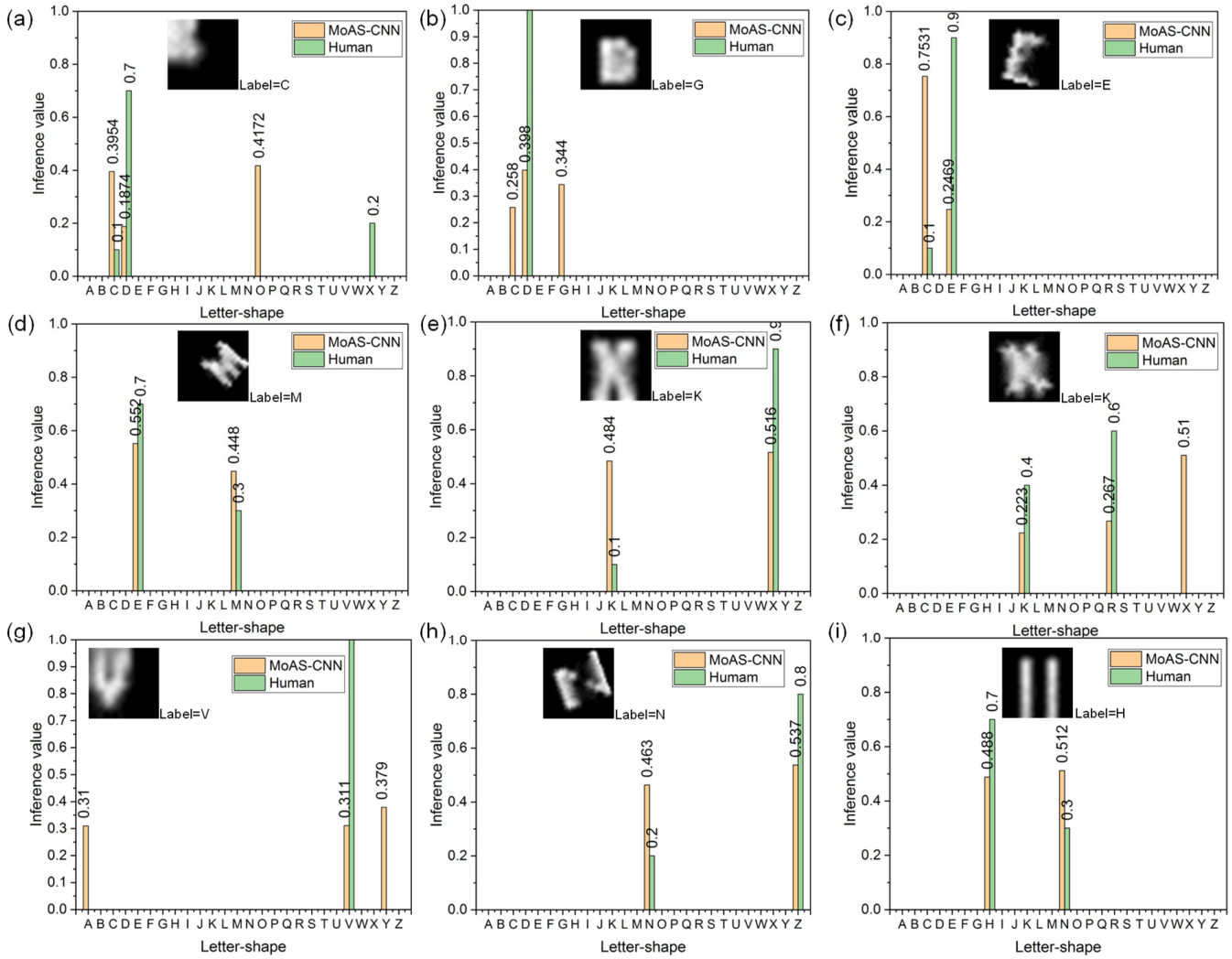

Figure 12. Comparison of recognition results between MoAS-CNN and human for some difficult samples. (**a**) Inference results of sample with label "C", (**b**) inference results of sample with label "G", (**c**) inference results of sample with label "E", (**d**) inference results of sample with label "M", (**e**) inference results of sample with label "K", (**f**) inference results of sample with label "K", (**g**) inference results of sample with label "V", (**h**) inference results of sample with label "N", (**i**) inference results of sample with label "H".

5. Discussion

Haptic technology provides real-time feedback on external force changes through flexible sensors mounted on or inside mechanical surfaces, providing an aid to scene understanding beyond vision. Therefore, recognition based on haptic information has become important for smart devices and has wide application prospects in human–computer interaction, intelligent machinery, and biomedicine. The intelligence level of human–computer interaction is improved by adding haptic perception. In addition, distribute pressure data feedback and analysis is important in many industrial sensing fields. Generally, the haptic information is converted into a grayscale image and then transferred to image-processing methods for subsequent recognition. However, compared with visual images, haptic images are still challenging because of their small dimensions, low resolution, and blurred edges. The experimental results show that MoAS-CNN can realize accurate haptic perception, and the highest accuracy of haptic letters with complex shapes was 98.65%. The cross-stream spatial interactive feature as a second-order response and multiorder feature fusion can significantly improve the extraction of haptic shapes by the network.

Although the proposed model has proven to be accurate and effective, it does have limitations. Through the analysis of the misclassified samples, some different classes with similar shapes cannot be clearly classified by the network. This may be attributed to two factors. Firstly, the feature extraction of the proposed method for haptic images is insufficient, and some information is lost in the extraction of high-level semantic information, especially for images with lower resolution, such as 16×16. Secondly, the characteristics of the haptic task itself, including the low resolution of the sensor array, the small number of pixels, and the blurred edge, lead to inaccurate mapping of the original shape, and the multiple meanings of individual images, resulting in lower-than-average recognition rates for some categories.

More importantly, smallness and low-resolution inputs are also present for practical vision tasks due to factors such as suboptimal raw data and environmental interference. Therefore, we tried to validate our method on another small general-purpose dataset CIFAR-10 [26], and the recognition accuracy was 98.81%. Compared with the current state-of-the-art results, it is lower than the largest-scale visual transformer methods [27,28] and essentially on par with the deeper CNN architecture [29]. This study is based on touch recognition based on a single image after touch completion. In the future, we plan to study recognition methods based on multiple dynamic touches and further improve the recognition performance of touch perception through reasonable optimization algorithms.

6. Conclusions

A framework called MoAS-CNN is proposed to address the challenge of low-resolution haptic recognition based on a pressure sensor array. The three contributions of our recognition model are to firstly apply a dual-stream CNN integrated with the channel attention module to automatically extract first-order features and increase the response and number of features. Secondly, a spatial interactive second-order feature is introduced to depict the second-order information to improve the feature extraction capability without additional feature downscaling. Thirdly, by exploring the complementarity of features of different orders, a multiorder hybrid strategy is developed to enhance the efficiency of feature extraction. To validate the model, a self-built acquisition platform based on a three-scale pressure array was built and haptic images of letter shapes (A–Z) with complex edges were collected. The results showed that 95.73%, 98.37%, and 98.65% accuracy was achieved at the scales of 16×16, 20×20, and 32×32, respectively. The accuracy of the proposed method has a significant advantage over traditional second- and first-order CNN-based methods, as well as manual feature methods. Furthermore, in addition to haptics, low-resolution inputs are common in practical applications because of the inherent limitations of various types of sensing elements and nonideal factors. Our approach provides a new general framework that can be easily extended to different systems.

Author Contributions: Conceptualization, K.W. and J.C. (Jie Chu); methodology, K.W., J.C. (Jie Chu) and J.C. (Jueping Cai); software, Y.C. and D.L.; validation, C.Z.; writing—original draft preparation, K.W. and J.C. (Jie Chu); writing—review and editing, J.C. (Jie Chu) and J.C. (Jueping Cai). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62274123), Natural Science Basic Research Plan in Shaanxi Province of China (2021ZDLGY02-01), and Wuhu-Xidian University Industry-University-Research Cooperation Special Fund (XWYCXY-012021003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this paper or are available from the corresponding authors on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Q.; Kroemer, O.; Su, Z.; Veiga, F.F.; Kaboli, M.; Ritter, J. A Review of Tactile Information: Perception and Action through Touch. *IEEE Trans. Robot.* **2020**, *36*, 1619–1634. [CrossRef]
2. Uddin, R.; Jamshaid, A.; Arfeen, A. Smart Design of Surgical Suture Attachment Force Measurement Setup Using Tactile Sensor. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 4001512. [CrossRef]
3. Luo, S.; Bimbo, J.; Dahiya, R.; Liu, H. Robotic Tactile Perception of Object Properties: A Review. *Mechatronics* **2017**, *48*, 54–67. [CrossRef]
4. Luo, S.; Mou, W.; Althoefer, K.; Liu, H. Novel Tactile-SIFT Descriptor for Object Shape Recognition. *IEEE Sens. J.* **2015**, *15*, 5001–5009. [CrossRef]
5. Pohtongkam, S.; Srinonchat., J. Object Recognition Using Glove Tactile Sensor. In Proceedings of the 2022 International Electrical Engineering Congress (iEECON), Khon Kaen, Thailand, 9–11 March 2022.
6. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
7. Li, P.; Wang, D.; Wang, L.; Lu, H. Deep visual tracking: Review and experimental comparison. *Pattern Recognit.* **2018**, *76*, 323–338. [CrossRef]
8. Gandarias, J.M.; Garcia-Cerezo, A.J.; Gomez-De-Gabriel, J.M. CNN-Based Methods for Object Recognition with High-Resolution Tactile Sensors. *IEEE Sens. J.* **2019**, *19*, 6872–6882. [CrossRef]
9. Polic, M.; Krajacic, I.; Lepora, N.; Orsag, M. Convolutional Autoencoder for feature extraction in tactile sensing. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3671–3678. [CrossRef]
10. Cao, L.; Sun, F.; Liu, X.; Huang, W.; Li, H. End-to-End ConvNet for Tactile Recognition Using Residual Orthogonal Tiling and Pyramid Convolution Ensemble. *Cognit. Comput.* **2018**, *10*, 718–736. [CrossRef]
11. Wang, S.; Xu, J.; Wang, W.; Wang, G.; Rastak, R.; Molina-Lopez, F.; Chung, J.; Niu, S.; Feig, V.R.; Lopez, J.; et al. Skin electronics from scalable fabrication of an intrinsically stretchable transistor array. *Nature* **2018**, *555*, 83–88. [CrossRef] [PubMed]
12. Song, L.; Zhu, H.; Zheng, Y.; Zhao, M.; Tee CA, T.; Fang, F. Bionic Compound Eye-Inspired High Spatial and Sensitive Tactile Sensor. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 7501708. [CrossRef]
13. Brahimi, S.; Aoun, N.B.; Amar, C.B. Improved Very Deep Recurrent Convolutional Neural Network for Object Recognition. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018.
14. Chen, M.; Kun, G.; Cheng, W.; Zhang, D.; Feng, W. Touchpoint-tailored ultra-sensitive piezoresistive pressure sensors with a broad dynamic response range and low detection limit. *Acs Appl. Mater. Inter.* **2019**, *11*, 2551–2558. [CrossRef] [PubMed]
15. Ruderman, A.; Rabinowitz, N.C.; Morcos, A.S.; Zoran, D. Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs. *arXiv* **2018**, arXiv:1804.044338.
16. Springenber, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806v3.
17. Lin, T.Y.; Roychowdhury, A.; Maji, S. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1309–1322. [CrossRef] [PubMed]
18. Carreira, J.; Caseiro, R.; Batista, J.; Sminchisescu, C. Semantic Segmentation with Second-Order Pooling. In Proceedings of the 12th European conference on Computer Vision(ECCV), Florence, Italy, 7–13 October 2012.
19. Akilan, T.; Wu, Q.; Safaei, A.; Jiang, W. A late fusion approach for harnessing multi-cnn model high-level features. In Proceedings of the 2017 IEEE International Conference on Systems, Man and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017.
20. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
22. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lile, France, 6–11 July 2015.
23. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.
24. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
25. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
26. The CIFAR-10 Dataset. Available online: https://www.cs.toronto.edu/~{}kriz/cifar.html (accessed on 11 November 2022).
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 2021 International Conference on Learning Representations (ICLP), Colombo, Sri Lanka, 20–27 September 2021.
28. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going deeper with Image Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
29. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 2019 International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.