information

MDPI

Article

Develop a Lightweight Convolutional Neural Network to Recognize Palms Using 3D Point Clouds

Yu-Ming Zhang [1], Chia-Yuan Cheng [1], Chih-Lung Lin [2,3,*], Chun-Chieh Lee [1] and Kuo-Chin Fan [1]

[1] Department of Computer Science and Information Engineering, National Central University, Taoyuan 320, Taiwan; k12501599501@gmail.com (Y.-M.Z.); ncuaipr35328@gmail.com (C.-Y.C.); jackcclee@cc.ncu.edu.tw (C.-C.L.); kcfan@csie.ncu.edu.tw (K.-C.F.)

[2] Department of Computer Science and Information Engineering, Hwa Hsia University of Technology, New Taipei 173, Taiwan

[3] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 100, Taiwan

* Correspondence: linclr@go.hwh.edu.tw

Abstract: Biometrics has become an important research issue in recent years, and the use of deep learning neural networks has made it possible to develop more reliable and efficient recognition systems. Palms have been identified as one of the most promising candidates among various biometrics due to their unique features and easy accessibility. However, traditional palm recognition methods involve 3D point clouds, which can be complex and difficult to work with. To mitigate this challenge, this paper proposes two methods which are Multi-View Projection (MVP) and Light Inverted Residual Block (LIRB).The MVP simulates different angles that observers use to observe palms in reality. It transforms 3D point clouds into multiple 2D images and effectively reduces the loss of mapping 3D data to 2D data. Therefore, the MVP can greatly reduce the complexity of the system. In experiments, MVP demonstrated remarkable performance on various famous models, such as VGG or MobileNetv2, with a particular improvement in the performance of smaller models. To further improve the performance of small models, this paper applies LIRB to build a lightweight 2D CNN called Tiny-MobileNet (TMBNet).The TMBNet has only a few convolutional layers but outperforms the 3D baselines PointNet and PointNet++ in FLOPs and accuracy. The experimental results show that the proposed method can effectively mitigate the challenges of recognizing palms through 3D point clouds of palms. The proposed method not only reduces the complexity of the system but also extends the use of lightweight CNN. These findings have significant implications for developing biometrics and could lead to improvements in various fields, such as access control and security control.

Keywords: palms recognition; multi-view projection; lightweight convolutional neural network

check for updates

Citation: Zhang, Y.-M.; Cheng, C.-Y.; Lin, C.-L.; Lee, C.-C.; Fan, K.-C. Develop a Lightweight Convolutional Neural Network to Recognize Palms Using 3D Point Clouds. *Information* **2023**, *14*, 381. https://doi.org/10.3390/info14070381

Academic Editors: Mohamed Hammad and Paweł Pławiak

Received: 26 April 2023
Revised: 24 June 2023
Accepted: 24 June 2023
Published: 3 July 2023

1. Introduction

With the rapid development of the information age, awareness of biometrics has become increasingly common in recent years. Many daily activities require the verification of personal identities, such as access control systems in sensitive places or epidemic control systems during the ravages of COVID-19. Therefore, it is necessary to build an ID recognition model based on a reliable token, such as the inherent component of human beings—the palm. Everyone has two palms which are unique components of the human body. The palms have rich features such as texture, finger length, and shape. Based on this simple intuition, the palm, represented in 3D point clouds, can best preserve biological characteristics and be used as a token for the ID recognition system. In addition to choosing a reliable token, a robust model is a cornerstone of the ID recognition system. Today, many cnn-based models, such as Vgg [1], ResNet [2], DenseNet [3], EfficientNet [4], and more, show extraordinary performance. While 3D point clouds are a better form of data, they also introduce additional complexity. For example, the 3D cloud point has complex properties such as disorder and lack of structure. In addition, the input of the 3D CNN is a sequence, and there are many possible combinations of the elements contained in the sequence, the sequence's length, and the sequence's order. It is far more complicated than a 2D image. To address this complexity, we propose the Multi-View Projection (MVP) method to project 3D palm data onto 2D images from several different views, just like humans observe their palms. Then, we propose Tiny-MobileNet(TMBNet), which combines advanced feature fusion and extraction methods. Finally, our experiments show a significant performance gap compared to the 3D CNN baselines, such as PointNet [5] and PointNet++ [6]. Overall, our proposed method TMBNet with MVP efficiently addresses the challenges of using 3D palms as a reliable token for an ID recognition system; It achieves better performance by projecting the 3D palms onto 2D images and combining advanced feature fusion and extraction methods than the classic 3D models.

2. Related Work

2.1. Overview for Palm Recognition

Previously, using Principal Component Analysis (PCA) to select critical features and then classify them by Support Vector Machine (SVM) was the common method; There are works of literature proposing the variants of PCA; For example, ref. [7] proposed the Gabor Wavelet with PCA to represent the 2D palms images; ref. [8] proposed the QPCA that is a multispectral version of PCA. Later, with the rapid development of deep learning, the literature [9] first used AlexNet to identify palms; some researchers focus on proposed new loss function [10,11], and they improve the performance of CNN at their time; there are some studies [12,13] presents the synthesized algorithm that combines palms data with other prior knowledge.

2.2. Overview of 3D Convolution Neural Networks

Today, some benchmark [14] for palm recognition is the form of point clouds, and in recent years, there has been much work to build a 3D CNN for 3D point clouds [5,15–17]. PointNet [5], and its derivative [6] are essential baselines in these 3D models. PointNet uses the symmetric function to solve the disorder caused by 3D point clouds and uses Multilayer Perceptrons (MLP) to extract high-level features; They propose a matrix network T-Net to attach at the beginning of the model for realignment features. When the input point clouds are aligned, sorted, and extracted, it goes through a Global Average Pooling (GAP) layer to get the final prediction. PointNet is a cornerstone for 3D point clouds, and after that, many studies have proposed novel methods based on it. Pointnet++ [6] has improved considerable performance through their designed local neighborhood sampling representation method and multi-level encoder-decoder combined network structure based on PointNet. Although these 3D CNNs have considerable performance, they are naturally more complex than 2D CNNs because of the negative properties of 3D point clouds, such as disorder and etc. In practice, the 3D CNNs hard to converge when training data is too few, so applying 3D data augmentation has become an often idea [18–20]. However, there is some trouble because these methods usually rely on point matching, which causes much computation.

2.3. Overview of 2D Convolution Neural Networks

The literature [21] uses AlexNet, VGG-16, GoogLeNet, and ResNet-50 to reach more impressive results than traditional methods in palm recognition tasks. In other words, they have been proven these classic 2D CNNs can achieve robust performance, such as VGG, ResNet, DenseNet, MobileNet, EffencienNet, and others. VGG [1] opened the era of widespread use of convolution layer with kernel size three by three. ResNet [2] proposes a skip connection to solve the problem, which is a nonlinear function to fit the identity function. DenseNet [3] chooses another way to achieve this purpose. They generate a few channels through a single convolution layer and then continue concatenating them to increase the channel gradually. They believe that they can pass features directly than skip connection. MobileNet [22] proposes the separable convolution to approximate convolution-3 × 3. They used it to build a lightweight CNN backbone with fewer parameters and low FLOPs than the other CNN backbone. MobileNetv2 [23] found that when the channel size is too small, adequate information will be lost because of dead cells due to ReLU [24]. To address it, they propose the Inverted Residual Block (IRB), which consists of a pointwise convolution (equal to convolution-1 × 1) and a separable convolution. The first pointwise convolution of IRB is designed to expand the channel for more redundancy to overcome the information lost. EfficienNet [4] built from NAS [25] technique proposes a comprehensive scale model from B0 to B7, no matter which scale is the leader at that time.

2.4. Projection Methods

As we just talked about, because of the harmful properties of 3D point clouds, there are studies proposing the projection method to project the 3D point clouds to the 2D data for reducing complexity. Some of them directly project the 3D point clouds into an image [15,26,27], and some methods convert it to Volume Pixel format [28]. The literature [26] has a conclusion that the collection of 2D images with different views can be highly informative for 3D shape recognition; the literature [27] hand over multiple groups of 2D images with different views to the learnable CNN to further strengthen the extracted features. Overall, in addition to directly processing the 3D point clouds as input of the model, it is also possible to project 3D point clouds to 2D format. However, dimensionality reduction will inevitably bring information loss. How to reduce the loss and maintain the richness of data is the main problem in this field.

3. Proposed Approaches

Preprocessing the 3D palm point clouds is difficult as they come with negative attributes such as disorder, scattered, and inconsistent data points. These attributes make it challenging to extract meaningful features for classification purposes. To address this issue, we propose a solution that involves projecting 3D plam to 2D images, simplifying the 3D point clouds by reducing their dimensionality. However, reducing the dimensionality of the 3D data may result in the loss of information that is essential for accurate classification. To address this issue, we propose a novel approach called Multi-View Projection (MVP). MVP aims to project the 3D palms into 2D images by imitating humans on how to view their palms. MVP enhances CNN's performance by generating robust augmented data from multiple views. We then propose a lightweight 2D CNN called TMBNet to reduce complexity further. TMBNet combines various advanced lightweight concepts based on MobileNetv2 [23], GhostNet [29], and Res2Net [30]. It has fewer layers and FLOPs than other models, making it more efficient and effective for processing large amounts of data. In summary, by employing MVP and TMBNet, our proposed method can achieve superior classification performance compared to existing 3D CNN.

3.1. Basic Projection (BP)

The Basic Projection method provides a simple and intuitive way to project 3D data onto a 2D plane for easy visualization and comparison during the experiment phase. This method involves a series of steps, starting with the min-max normalization of every point along the Z-axis to ensure that all values are within the same range. Next, the normalized Z-values are averaged to obtain an XY plane with a single representative value. This process is illustrated in Equation (1). Using the Basic Projection method, we can obtain a gray-scale image that comprehensively represents the 3D data. The BP method serves as a baseline in our research and allows us to compare the performance of our proposed Multi-View Projection (MVP) method with a straightforward and intuitive approach. In Figure 1, we provide two projected images by BP to show the effectiveness of this method. However, the BP method suffers from some drawbacks, such as the loss of valuable information and the inability to capture the object's depth. This motivates us to propose the MVP method to address these issues.

$$NZ_i = \frac{Z_i - Min(Z)}{Max(Z) - Min(Z)}, \; Z^c = \frac{\sum_{i=1}^{k} NZ_i^c}{k} \tag{1}$$

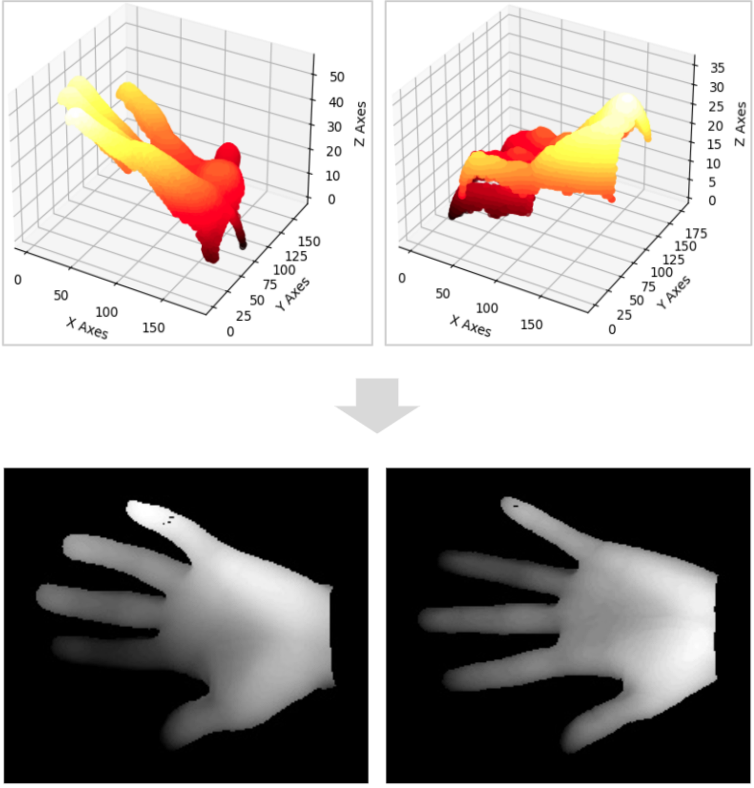

Figure 1. The above images are original 3D point clouds on PolyU-CHFD, and the bottom images are projected by Basic Projection (BP).

3.2. Multi-View Projection (MVP)

Multi-View Projection (MVP) is a powerful approach that simulates how humans view their palms from different angles to generate a wide range of images. The goal of MVP is to capture a comprehensive range of views and perspectives of the palm, which is difficult to achieve with other projection methods. By generating a large number of images from 3D point clouds of the palm, MVP enhances the robustness of 2D CNNs by introducing a greater degree of image diversity. To achieve this, the MVP process is broken down into three key steps: rotation, affine, and shear. Rotation involves rotating the palm around the y-axis at different angles, which simulates the human's natural viewing behavior. Affine transformation is applied to adjust the scale, orientation, and shape of the projected image to match the human palm's characteristics. Finally, shear transformation is used to correct the distortion of the image due to rotation and affine transformation. The MVP approach not only provides a more robust and accurate method for extracting various palm features but also offers a more intuitive and realistic method for simulating human vision. By incorporating multiple views and perspectives, MVP improves the classification performance, as demonstrated in our experiments. Figure 2 showcases three projected images by MVP, demonstrating the diversity and richness of the generated images.

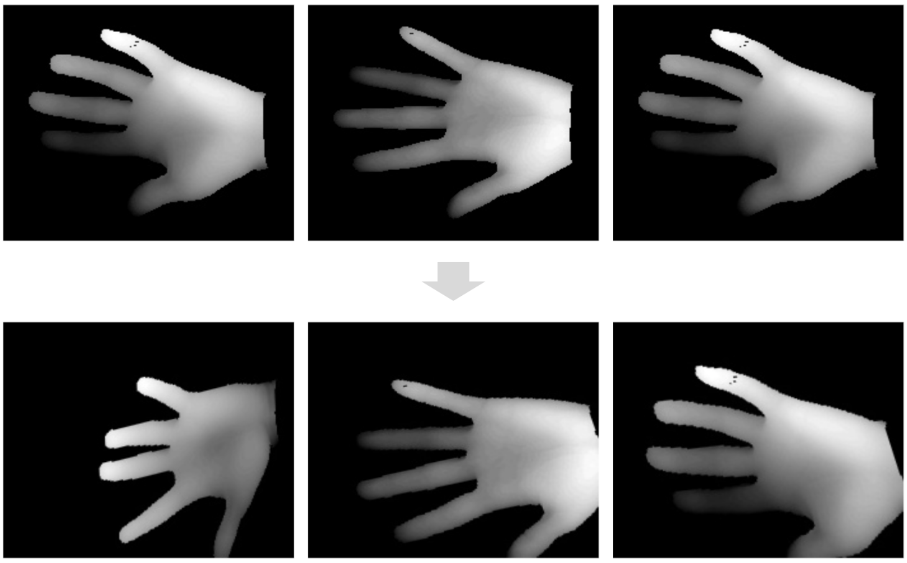

Figure 2. The above images are projected by Basic Projection (BP), and bottom images are projected by Multi-View Projection (MVP).

3.2.1. Rotation and Affine on XY Plane

To achieve a more efficient and structured approach for exploring the possible views of the palm, we utilize rotation and affine on the XY plane in our MVP process. The approach is aimed at converting the exhaustive view process into a spatial shift on the XY plane, thereby reducing the computational time and storage space required for the MVP process. To achieve this, we first set a range of plus or minus ten percent for affine and a range of plus or minus ten degrees for rotation. We then randomly apply affine and rotation to the image within the set range. This ensures that we can explore a comprehensive range of views while still keeping the process manageable and efficient. The use of affine in the MVP process is particularly effective in accounting for the varying distances between the palm and the camera in real-life scenarios. It allows for adjustments to be made to the size and orientation of the palm image, ensuring that the model can capture a wide range of perspectives and variations in the palm's appearance. Furthermore, the use of rotation in the MVP process enables us to capture different orientations of the palm, including rotations along the X, Y, and Z axes. This allows the model to capture variations in the palm's shape, texture, and features from different angles, providing a more comprehensive and robust dataset for training CNN. Overall, the use of rotation and affine on the XY plane is a critical component of our MVP approach, as it allows us to efficiently explore a comprehensive range of views while still capturing a wide range of variations in the palm's appearance.

3.2.2. Shear on X Axis and Y Axis

After applying spatial shift through random affine and rotation, the next step in the MVP process is shearing to mimic the angle of view of a human observer. Shearing involves distorting the 2D image along the X and Y axes to create the illusion of viewing the palm from a different angle. To shear the image along the X-axis, a random shear matrix $T_x(\theta)$ is generated as shown in Equation (2).

$$T_x(\theta) = \begin{bmatrix} 1 & tan(\theta) & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{2}$$

Once the matrix $T_x(\theta)$ is generated, it is multiplied with the transformed image to create a distorted image that simulates the effect of viewing the palm from a different angle. The degree of distortion along the X-axis is controlled by the θ parameter, which specifies the degree of shearing to be applied. Similarly, a shear matrix $T_y(\phi)$ is generated along the

Y axis as shown in Equation (3), and is multiplied with the transformed image by $T_x(\theta)$. The degree of distortion along the Y-axis is controlled by the ϕ parameter, which specifies the degree of shearing to be applied. The degree of distortion along both the X and Y axes is randomly selected within a set range. Through these steps, MVP can generate a diverse and comprehensive set of images that accurately capture the palm's many different angles and perspectives. By using shearing to mimic the angle of view of a human observer, MVP can produce images that are more realistic and useful for various applications.

$$T_y(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ tan(\phi) & 1 & 0 \end{bmatrix} \tag{3}$$

3.2.3. Primary Experiment for MVP

The experiment conducted to explore the capability of MVP involved training two classic 2D CNNs, VGG-16 and MobileNetv2, with both Basic Projection and MVP techniques. VGG-16 and MobileNetv2 represent the heavy but powerful type and the light but moderate type, respectively. The experiment results, as shown in Table 1, demonstrate that both VGG-16 and MobileNetv2 models achieve considerable gains by using MVP. Note that the accuracy indicates the correct rate of the prediction of the model in the PolyU-CHFD test set of 114 palm images compared with the ground truth. Interestingly, the gap between the two models based on Basic Projection is larger than between those based on MVP. This suggests that the effectiveness of MVP is remarkable and can significantly improve the performance of both heavy and light models. Furthermore, the experiment shows that with MVP, it is possible to build a lighter CNN than MobileNetv2 while still achieving similar performance. This indicates that MVP improves the performance of existing models and enables the development of more lightweight models that can save computation resources. Overall, the experiment provides strong evidence to support the effectiveness of MVP in enhancing the performance of 2D CNN CNNs and highlights its potential for enabling the development of more efficient and lightweight models in the future.

Table 1. The Comparison of Basic Projection and MVP on two classic 2D CNNs.

	Accuracy	**MFLOPs**
VGG-16 w/BP	74.85	15,670
VGG-16 w/MVP	98.82	15,670
MobileNetv2 w/BP	64.89	290
MobileNetv2 w/MVP	97.95	290

3.3. Tiny-MobileNet (TMBNet)

TMBNet is a newly proposed lightweight 2D CNN that is designed to be simpler and lighter in architecture, the overview as shown in Figure 3. It consists of a single layer of convolution-3 × 3 as the stem, five Light Inverted Residual Blocks (LIRB) to extract high-level features, and five pooling layers to downsample the image. Finally, the extracted features go through a fully connected layer to obtain the prediction. The LIRBs used in TMBNet are inspired by the inverted residual structure proposed in MobileNetv2. Each LIRB consists of a 1 × 1 convolution layer, a depthwise separable convolution layer, and another 1 × 1 convolution layer. Batch normalization and ReLU activation functions are also applied after each layer. The output of the first 1 × 1 convolution layer is channel-expanded before being fed into the depthwise separable convolution layer to enhance the network's representation ability. TMBNet's architecture is simpler and lighter compared to other 2D CNNs. The number of layers is reduced, and the size of each layer is optimized for efficient computation. The downsampling operation is also performed by pooling layers, which further reduces the computational cost. The smaller architecture of TMBNet results in a significant reduction in FLOPs (Floating Point Operations) compared to other models, which means it can achieve similar accuracy with much fewer computations. Table 2 shows the details of shape and channel. TMBNet's simplicity and efficiency make it an excellent candidate for applications with limited computational resources, such as edge devices and mobile devices. In the following sections, we will discuss the details of LIRB and analyze the FLOPs of TMBNet and other methods for objective comparison.

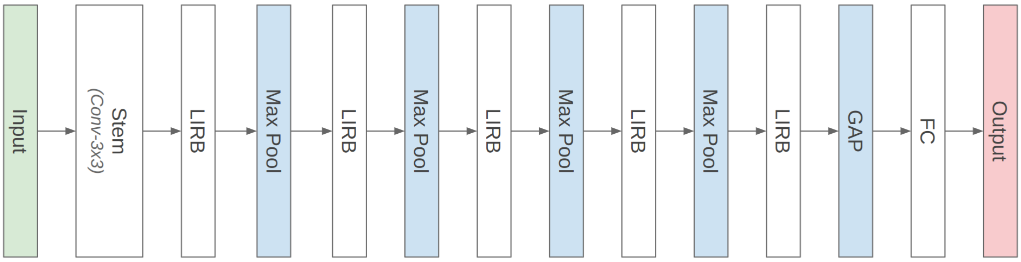

Figure 3. The architecture of Tiny-MobilNet (TMBNet). LIRB represents the proposed Light Inverted Residual Block, GAP means Global Average Pooling.

Table 2. The detail of TMBNet. Down Sampling indicates the downsampling rate of the input shape.

Input Shape	Layer	Output Channel	Down Sampling
224 × 224	Input	3	1
224 × 224	Conv-3 × 3	16	2
112 × 112	LIRB	32	1
112 × 112	Max Pool	32	2
64 × 64	LIRB	64	1
64 × 64	Max Pool	64	2
32 × 32	LIRB	128	1
32 × 32	Max Pool	128	2
16 × 16	LIRB	256	1
16 × 16	Max Pool	256	2
8 × 8	LIRB	512	1
8 × 8	Global Average Pool	512	global
1 × 1	FC	114	1
1 × 1	Output	114	1

3.3.1. Light Inverted Residual Block (LIRB)

The proposed Light Inverted Residual Block (LIRB) is a novel concept introduced in TMBNet. It is designed based on the Inverted Residual Block (IRB) proposed in MobileNetv2 and utilizes the Res2Net concept for channel expansion. LIRB is designed to be even more lightweight and faster than MobileNetv2 by stacking three layers of depthwise convolutions for channel expansion instead of using pointwise convolution. One unique feature of LIRB is the concept of Ghost mimics. This technique splits the required output channels into two halves. One half is generated by the Expansion Block and Separable Convolution, and the other half is generated by a simple linear transformation. These two halves are then concatenated together to form the final output channels. This technique helps to reduce computation while maintaining accuracy. Figure 4 illustrates the architecture of LIRB, which consists of a depthwise convolution layer, an Expansion Block, and a Separable Convolution layer. The Expansion Block expands the number of input channels, and the Separable Convolution layer performs a depthwise convolution followed by a pointwise convolution to produce the final output channels. The depthwise convolution layer is composed of three layers of depthwise convolutions for channel expansion, as mentioned earlier. To summarize, LIRB is a lightweight block that uses Res2Net for channel expansion and Ghost mimic for reducing computation while maintaining accuracy.

$$
\begin{aligned}
FLOPs_{lirb} &= 3 \times (H' \times W' \times 9 \times C) + (H' \times W' \times 9 \times 3C) \\
&+ (H' \times W' \times 3C \times 0.5C') + (H' \times W' \times 9 \times 0.5C') \\
&+ (H' \times W' \times C \times 0.5C') \\
FLOPs_{invrtrb} &= (H' \times W' \times C \times 2C) + (H' \times W' \times 9 \times 2C) \\
&+ (H' \times W' \times 2C \times C') \\
Let\ C =C',\ then \frac{FLOPs_{invrtrb}}{FLOPs_{lirb}} &= \frac{4C^2 + 18C}{1.5C^2 + 58.5C} \simeq 2.66667
\end{aligned} \tag{4}
$$

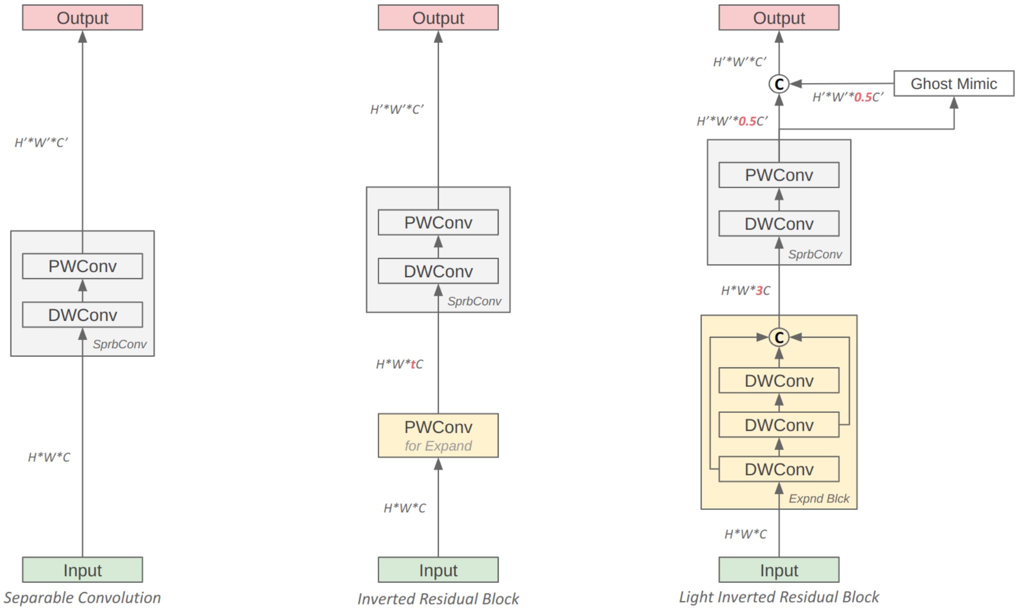

Figure 4. The architecture of Light Inverted Residual Block (LIRB). PWConv represents the pointwise convolution (or convolution-1 × 1), DWConv represents depthwise convolution, and Ghost Mimic means the cheap operation for linear mapping.

3.3.2. Analysis of FLOPs

Floating-point operations (FLOPs) is an important metric for measuring the computational complexity of a neural network. In this section, we will analyze the FLOPs of TMBNet. First, we analyzed the FLOPs of IRB the building blocks of MobileNetv2, and FLOPs of LIRB the building blocks of TMBNet. The results, shown in Equation (4), indicate that the FLOPs of LIRB are significantly lower than those of IRB, especially when the input and output channels are the same. This suggests that LIRB is more computationally efficient than IRB. Next, we compared the FLOPs of TMBNet with those of MobileNetv2 and VGG-16, as shown in Table 1. The results clearly demonstrate that TMBNet has significantly fewer FLOPs than both MobileNetv2 and VGG-16. For instance, when compared to MobileNetv2, TMBNet has 5.5 times fewer FLOPs. This means that TMBNet can perform the same amount of computation with much fewer operations, making it a much faster and more efficient model. In short, our FLOPs analysis shows that TMBNet is a highly efficient and lightweight model, with significantly fewer computational requirements than other popular models. This makes it an ideal choice for applications where computational resources are limited or speed is critical.

4. Experiments

4.1. PolyU-3D Contact-Free Hand Dataset (PolyU-CFHD)

The PolyU-3D Contact-Free Hand Dataset (PolyU-CFHD), proposed by the Hong Kong Polytechnic University [14], is a valuable resource for evaluating hand pose estimation algorithms. This dataset contains 570 hand data samples from 114 individuals, each with five palm images. Each hand data sample contains tens of thousands of points, with each point's X, Y, and Z coordinates representing its position in 3D space. As shown in Figure 5, the dataset's hand data samples have complex shapes and varied hand poses. To ensure a balanced distribution of data, we randomly selected 80% and 20% of the five images per single person for the training and testing sets, respectively. This resulted in a training set of 456 images and a testing set of 114 images. All accuracy measurements in this paper were computed using the test set of PolyU-CFHD. The visualization of 3D point clouds on PolyU-CFHD shown in Figure 5 highlights the complexity and diversity of the hand poses in this dataset.

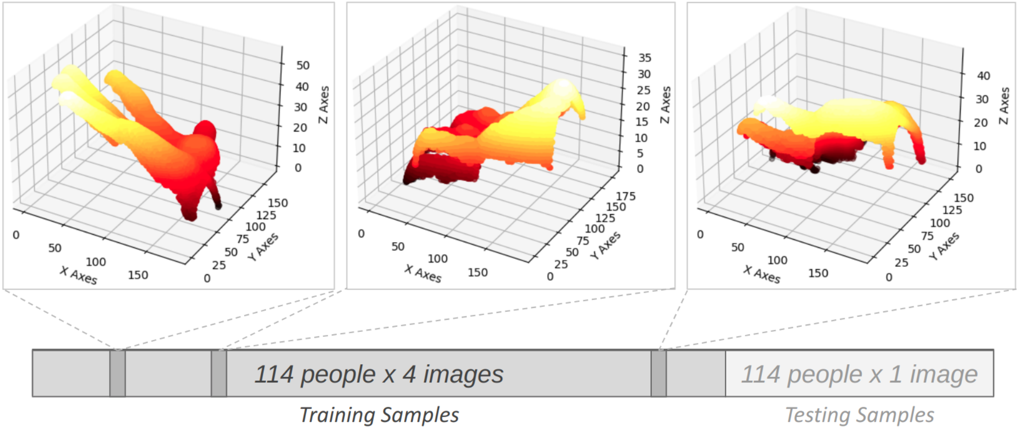

Figure 5. The original point clouds data in PolyU-CFHD. It contains 114 people, and each has five 3D point clouds palm data. We randomly split it into four for training and one for testing.

4.2. Classic 2D CNNs with MVP

In this section, we present the results of our experiments using MVP to improve the performance of several classic 2D CNN CNNs. To establish a baseline, we trained these CNNs using 2D images projected by Basic Projection. The purpose of this experiment was to compare the underlying capabilities of these CNNs with different computational complexities. As shown in the upper half of Table 3, we used heavy models like VGG-16 and ResNet-50 and lighter models like EfficienNet-B0 and MobileNetv2. The results indicate that the more complex and larger networks tend to perform better on PolyU-CFHD. For instance, VGG-16 and ResNet-50 achieved higher accuracy than EfficientNet-B0 and MobileNetv2, with accuracy gaps as high as 17% in some cases. To evaluate the effectiveness of MVP, we retrained these four CNNs using MVP. The lower half of Table 3 shows the performance improvement achieved by MVP. As we can see, the CNNs using MVP outperformed their counterparts without MVP by a significant margin, demonstrating that MVP is effective in reducing the feature loss caused by 3D point clouds to the 2D image. Note that accuracy refers to how correct the model's predictions are to the ground truth on the 114 test images. It also means that how powerful of model can predict the correct one from 114 possible individuals for the input of a palm data.

Table 3. The Comparison of PolyU-CFHD using BP and MVP, including four classical 2D CNN models and our TMBNet. The above part of the models are the results using BP, and the bottom part of the models are the results using MVP.

	Accuracy	**MFLOPs**
VGG-16	74.85	15,670
ResNet-50	72.61	3608
EfficientNet-B0	57.27	372
MobileNetv2	64.89	290
TMBNet	61.40	95
VGG-16 w/MVP	98.82	15,670
ResNet-50 w/MVP	98.53	3608
EfficientNet-B0 w/MVP	97.66	372
MobileNetv2 w/MVP	97.95	290
TMBNet w/MVP	95.61	95

4.3. Leave-One-Out Comparison of TMBNet

To further validate the robustness of 2D CNNs with the proposed MVP, we conducted ten independent experiments, each involving the random allocation of training and testing sets in a 4:1 ratio. In other words, out of the 114 individuals, only one image per person (out of a set of five) was retained as a testing sample, while the remaining four images were used for training purposes. Table 4 shows the results, where the mean accuracy represents the average accuracy across the ten independent experiments, the min accuracy represents the lowest accuracy achieved, and the max accuracy represents the highest accuracy achieved. We can observe that regardless of using VGG, ResNet, or our proposed TMBNet, both the min accuracy and max accuracy exhibit remarkable stability. This once again confirms the efficacy of our proposed MVP method in preserving the information lost during the transformation from 3D palm projects to 2D images, and in enhancing the performance of 2D CNNs as simpler classifiers.

Table 4. The leave-one-out comparison of classic 2D CNNs and TMBNet, all trained with the proposed MVP. The mean accuracy is the average result of ten runs.

	Mean Accuracy	**Min Accuracy**	**Max Accuracy**
VGG-16	98.78	97.60	99.33
ResNet-50	98.38	97.15	98.96
EfficientNet-B0	97.17	95.13	98.64
MobileNetv2	97.25	95.37	98.81
TMBNet	96.49	94.92	97.25

4.4. Comparison of TMBNet with 3D Baselines

We have already demonstrated the effectiveness of MVP in the 2D classic model experiments; now, we want to examine whether TMBNet w/MVP surpasses the 3D baseline. We use PointNet and PointNet++ as our baseline, the widely used 3D CNN models, and we compare their performance with our proposed TMBNet w/MVP. As shown in Table 5, we first find that PointNet++ beat PointNet by a minor margin; this phenomenon deserves our attention because the performance of PointNet++ surpassed PointNet on various large datasets by a significant margin in their studies. Considering the PolyU-CFHD is a small, single-class, palm-only dataset, it has more limitations of feature richness than other large datasets. So we believe that improving the performance cannot only rely on the feature extraction ability from models but should use multiple views to enhance the feature richness of single palm as MVP does. In short, instead of using a strong 3D extractor such as PointNet or PointNet++, it is better to use MVP to enhance the feature richness of a single palm and use the more simpler 2D CNNs. Then we can see that the proposed TMBNet w/MVP achieves 2.35% to 2.71% higher accuracy and 4.6 to 17.7× fewer FLOPs than baseline. This comparison shows the potential benefits of using 2D CNN CNNs with MVP to handle 3D point cloud data rather than relying on traditional 3D CNNs. Moreover, it demonstrates that the proposed MVP avoids the negative properties of the 3D palms in 2D image form and enables a straightforward model like TMBNet to achieve significant accuracy with tiny FLOPs. This result is particularly relevant for real-world applications where computational resources are limited and there is a need for efficient and effective models. Additionally, it opens up possibilities for using MVP in other types of 2D CNN CNNs for handling 3D palms data, potentially leading to more efficient and accurate models. Overall, the comparison between TMBNet w/MVP and PointNet confirms the advantages of our proposed method and highlights its potential for practical use in various applications.

Table 5. The comparison of PointNet and TMBNet with different input scales.

	Input Shape	Accuracy	MFLOPs
PointNet	-	92.90	440
PointNet++	-	93.26	1680
TMBNet w/MVP	224 × 224	95.61	95
TMBNet w/MVP	160 × 160	94.28	48
TMBNet w/MVP	120 × 120	93.47	26

4.5. Input Shape Reduction for More Smaller TMBNet

To explore the possibility of making TMBNet even more minor, we conducted experiments to reduce its input image size. This approach can sacrifice some accuracy while achieving even fewer FLOPs. The initial TMBNet model used a 224 × 224 input size. We reduced this input size to 160 × 160 and 120 × 120. Then training and testing the new models on the PolyU-CFHD dataset. As shown in Table 5, the accuracy of the 160 × 160 version decreased slightly to 94.28% compared to the initial model's accuracy of 95.61%, but it only required 48 MFLOPs. Furthermore, the 120 × 120 version achieved an accuracy of 93.47% with ultra-lightweight computation of 26 MFLOPs. Even though the accuracy of the 120 × 120 version is lower than the initial model, it still outperforms PointNet's accuracy of 92.90% with significantly fewer FLOPs. These results suggest that it is possible to reduce the input image size of TMBNet while still achieving high accuracy. Such a reduction can significantly reduce the computational complexity of the model, making it more suitable for resource-constrained environments. The experiments also demonstrate the effectiveness of our MVP method in reducing the feature loss caused by 3D point clouds, enabling a lightweight model like TMBNet to achieve high accuracy.

5. Conclusions

This paper has proposed a novel Multi-View Projection (MVP) method for enhancing the performance of 2D CNNs in human palms recognition tasks; MVP imitates human views on different angles for their palms, which enables the 2D CNNs to achieve significant accuracy with the weaker data type (2D images) than the more vital data type (3D point clouds). The experimental results have demonstrated the efficacy of MVP in various experiments on popular 2D CNNs, including VGG, ResNet, EfficientNet, and MobileNetv2, where the models get considerable improvement with MVP. Inspired by the success of MVP, we further proposed a more lightweight 2D CNN, Tiny-MobileNet (TMBNet), which performs impressive results on the human palms benchmark. TMBNet achieved a high accuracy of 95.61% with only 95 MFLOPs, surpassing the 3D baseline PointNet and PointNet++ with 2.35% to 2.71% accuracy margin while utilizing only a quarter of PointNet's computational complexity. Furthermore, we have explored the possibility of further reducing the input image size of TMBNet, where the ultra-lightweight version with an input size of 120 × 120 achieved an accuracy of 93.47% with only 6% of PointNet's FLOPs. Our research demonstrates the extremely lightweight TMBNet with MVP reaches high accuracy with much fewer computational resources than classic 3D methods in human palms recognition tasks, which makes the proposed methods suitable for resource-limited devices and real-time ID recognition systems. We believe our work can inspire further research in ID recognition based on human palms and promote the development of lightweight models for practical use.

Author Contributions: Conceptualization, Y.-M.Z. and C.-Y.C. and C.-L.L.; methodology, Y.-M.Z. and C.-Y.C.; software, Y.-M.Z.; validation, Y.-M.Z. and C.-Y.C.; formal analysis, Y.-M.Z.; investigation, Y.-M.Z.; resources, C.-L.L. and C.-C.L. and K.-C.F.; data collection, C.-Y.C.; writing—original draft preparation, Y.-M.Z. and C.-Y.C.; writing—review and editing, Y.-M.Z. and C.-L.L.; visualization, Y.-M.Z.; supervision, C.-L.L. and C.-C.L.; project administration, C.-L.L.; funding acquisition, K.-C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Science and Technology Council of Funder grant number (MOST 110-2221-E-008-057-MY3).

Data Availability Statement: The palm data set PolyU-CHFD used in this paper is provided by the Hong Kong Polytechnic University. This is the introduction of the data set and the download hyperlink: https://www4.comp.polyu.edu.hk/~csajaykr/Database/3Dhand/Hand3DPose.htm.

Conflicts of Interest: The authors declare no conflict of interest, and the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

TMBNet	Tiny-MobileNet
IRB	Inverted Residual Block
LIRB	Light Inverted Residual Block
MVP	Multi-View Projection
BP	Basic Projection
CNN	Convolutional Neural Network
PCA	Principal Component Analysis
SVM	Support Vector Machine
GAP	Global Average Pooling
FLOPs	floating-point operations
PolyU-CFHD	PolyU-3D Contact-Free Hand Dataset

References

1. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
4. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
5. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
6. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114.
7. Ekinci, M.; Aykut, M. Gabor-based kernel PCA for palmprint recognition. *Electron. Lett.* **2007**, *43*, 1077–1079. [CrossRef]
8. Xu, X.; Guo, Z. Multispectral palmprint recognition using quaternion principal component analysis. In Proceedings of the 2010 International Workshop on Emerging Techniques and Challenges for Hand-Based Biometrics, Istanbul, Turkey, 22 August 2010; IEEE: New York, NY, USA, 2010; pp. 1–5.
9. Dian, L.; Dongmei, S. Contactless palmprint recognition based on convolutional neural network. In Proceedings of the 2016 IEEE 13th International Conference on Signal Processing (ICSP), Chengdu, China, 6–10 November 2016; IEEE: New York, NY, USA, 2016; pp. 1363–1367.

10. Svoboda, J.; Masci, J.; Bronstein, M.M. Palmprint recognition via discriminative index learning. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: New York, NY, USA, 2016; pp. 4232–4237.
11. Zhong, D.; Zhu, J. Centralized large margin cosine loss for open-set deep palmprint recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1559–1568. [CrossRef]
12. Chen, W.; Yu, Z.; Wang, Z.; Anandkumar, A. Automated synthetic-to-real generalization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1746–1756.
13. Zhao, K.; Shen, L.; Zhang, Y.; Zhou, C.; Wang, T.; Zhang, R.; Ding, S.; Jia, W.; Shen, W. BézierPalm: A Free Lunch for Palmprint Recognition. In *Lecture Notes in Computer Science, Part XIII, Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022*; Springer: New York, NY, USA, 2022; pp. 19–36.
14. Kanhangad, V.; Kumar, A.; Zhang, D. A unified framework for contactless hand verification. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 1014–1027. [CrossRef]
15. Qi, C.R.; Su, H.; Nießner, M.; Dai, A.; Yan, M.; Guibas, L.J. Volumetric and multi-view cnns for object classification on 3d data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5648–5656.
16. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3D object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
17. Xu, C.; Wu, B.; Wang, Z.; Zhan, W.; Vajda, P.; Keutzer, K.; Tomizuka, M. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *Lecture Notes in Computer Science, Part XXVIII 16, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–19.
18. Chen, Y.; Hu, V.T.; Gavves, E.; Mensink, T.; Mettes, P.; Yang, P.; Snoek, C.G. Pointmixup: Augmentation for point clouds. In *Lecture Notes in Computer Science, Part III 16, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 330–345.
19. Kim, S.; Lee, S.; Hwang, D.; Lee, J.; Hwang, S.J.; Kim, H.J. Point cloud augmentation with weighted local transformations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 548–557.
20. Lee, D.; Lee, J.; Lee, J.; Lee, H.; Lee, M.; Woo, S.; Lee, S. Regularization strategy for point cloud via rigidly mixed sample. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15900–15909.
21. Fei, L.; Lu, G.; Jia, W.; Teng, S.; Zhang, D. Feature extraction methods for palmprint recognition: A survey and evaluation. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *49*, 346–363. [CrossRef]
22. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
24. Agarap, A.F. Deep learning using rectified linear units (relu). *arXiv* **2018**, arXiv:1803.08375.
25. Zoph, B.; Le, Q.V. Neural architecture search with reinforcement learning. *arXiv* **2016**, arXiv:1611.01578.
26. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953.
27. Li, L.; Zhu, S.; Fu, H.; Tan, P.; Tai, C.L. End-to-end learning local multi-view descriptors for 3d point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1919–1928.
28. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10529–10538.
29. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1580–1589.
30. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.