entropy

Article

Lightweight Deep Neural Network Embedded with Stochastic Variational Inference Loss Function for Fast Detection of Human Postures

Feng-Shuo Hsu [1,2], Zi-Jun Su [1,3], Yamin Kao [1], Sen-Wei Tsai [4], Ying-Chao Lin [5], Po-Hsun Tu [6], Cihun-Siyong Alex Gong [7] and Chien-Chang Chen [1,*]

[1] Bio-Microsystems Integration Laboratory, Department of Biomedical Sciences and Engineering, National Central University, Taoyuan 320317, Taiwan
[2] Department of Psychiatry, Taichung Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taichung 427213, Taiwan
[3] Department of Computer Science, College of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30093, Taiwan
[4] Department of Physical Medicine and Rehabilitation, Taichung Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taichung 427213, Taiwan
[5] Department of Neurological Institute, Taichung Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taichung 427213, Taiwan
[6] Department of Neurosurgery, Chang Gung Memorial Hospital, Linkou Branch, Taoyuan 33304, Taiwan
[7] Department of Electrical Engineering, College of Engineering, Chang Gung University, Taoyuan 33302, Taiwan
* Correspondence: gettgod@ncu.edu.tw

Citation: Hsu, F.-S.; Su, Z.-J.; Kao, Y.; Tsai, S.-W.; Lin, Y.-C.; Tu, P.-H.; Gong, C.-S.A.; Chen, C.-C. Lightweight Deep Neural Network Embedded with Stochastic Variational Inference Loss Function for Fast Detection of Human Postures. *Entropy* **2023**, *25*, 336. https://doi.org/10.3390/e25020336

Academic Editors: Nadia Kanwal and Mohammad Samar Ansari

Received: 17 January 2023
Revised: 7 February 2023
Accepted: 9 February 2023
Published: 11 February 2023

Abstract: Fusing object detection techniques and stochastic variational inference, we proposed a new scheme for lightweight neural network models, which could simultaneously reduce model sizes and raise the inference speed. This technique was then applied in fast human posture identification. The integer-arithmetic-only algorithm and the feature pyramid network were adopted to reduce the computational complexity in training and to capture features of small objects, respectively. Features of sequential human motion frames (i.e., the centroid coordinates of bounding boxes) were extracted by the self-attention mechanism. With the techniques of Bayesian neural network and stochastic variational inference, human postures could be promptly classified by fast resolving of the Gaussian mixture model for human posture classification. The model took instant centroid features as inputs and indicated possible human postures in the probabilistic maps. Our model had better overall performance than the baseline model ResNet in mean average precision (32.5 vs. 34.6), inference speed (27 vs. 48 milliseconds), and model size (46.2 vs. 227.8 MB). The model could also alert a suspected human falling event about 0.66 s in advance.

Keywords: bayesian neural networks; gaussian mixture model; human posture identification; integer-arithmetic-only; lightweight neural networks; self-attention; stochastic variational inference

1. Introduction

Timely detection of human fall events is vital in various care environments. Current technologies of fall detection include wearable devices [1,2], environmental sensing schemes [3], and vision-based methods [3–9]. The inconvenience of wearing sensors impedes relevant advances [4] and promotes the development of contactless smart sensors. Thus, vision-based methods have become mainstream. Skeleton-based [6,9] and image-based [3–5,7,8] posture detections are two primary strategies. Nevertheless, the high-cost apparatus for constructing human skeleton images hinders its development [3,4]. Image-based approaches employing deep neural networks with high structural complexities and computational costs require significant inference time and may jeopardize the detection performance of fall incidents.

Single-shot multi-box detector (SSD) is a one-stage approach for object predictions using multi-bounding boxes [10–12]. Compared to the YOLO series [13,14] or the two-stage method Faster-RCNN [15], SSD has a significantly higher speed of instant inspection. However, the conventional SSD employs VGG16 as its backbone [10], and the architecture still requires extensive model training and prolongs the inference time. Provided the heavy communication loads within the configuration of modern deep convolutional neural networks and unsatisfactory speed, lightweight neural networks emerged and have become the leading technique [16,17]. Among the lightweight neural networks, MobileNet adopts depth- and point-wise separable convolutional layers to decrease computational complexity and parameters [18–21]. MobileNetV2 adds layers of inverted residuals and linear bottlenecks, for which the feature extraction and data transmission occur in the high- and low-dimensional spaces, respectively [22,23]. With linear bottleneck layers as activations, MobileNetV2 further reduces parameters and computational costs. SqueezeNet [24] and deep compression [25] also open an avenue for model compression by effectively assigning 1×1 kernels (a point-wise-like layer structure) to replace the conventional ones. Integrating the architectures of depth-wise separable convolutions and ResNet bottleneck [23], the operation of channel shuffle proposed by ShuffleNet compresses the computational costs while preserving feature information [19]. ShuffleNet V2 [26] further verifies that the reduction in convolutional branches and the replacement of element-wise operations by layer concatenation could significantly raise the computational performance. Distinct from techniques of simplifying neural network architectures or optimizing computational procedures, the quantization training method of integer-arithmetic-only (IAO) algorithm [27,28] provides an alternative to reduce model sizes and accelerate inference by controlling weight bit-widths of convolutional kernels and activation functions. This algorithm also highlights that sophisticated deep convolutional neural networks are impractical baseline architectures because of overusing internal parameters and manually assigning channels in each layer. Under the demand of importing lightweight neural networks, utilizing quantization-aware training [27] in deep neural networks benefits inference performance.

With the combination of a lite SSD network, IAO algorithm, and lightweight neural networks, we established a new framework with much less model structural complexity and better inference capability. To compare the ability of different lightweight neural networks in terms of model size and inference speed, we employed them as backbones of the lite SSD network. In addition, to efficiently reinforce the capability of identifying objects with diverse scales, we embedded the feature pyramid network (FPN) [29] into the lite SSD network to extract local information. Moreover, the self-attention mechanism [30,31] and the residual blocks of ResNet [23] were adopted to process sequential bounding boxes (BBoxs) generated by the lite SSD network for parallelly extracting weighted centroid features of human postures at each time point. These feature maps then became inputs of the variational inference Gaussian mixture model [32] and backpropagation for further classification analysis. Therefore, based on the integrations of these techniques, this framework achieved a fast human posture classification with small model sizes and high inference speed. The key contributions of our study as listed as follows:

- Decreasing model sizes while increasing mean average precisions and inference speeds;
- Incorporating the self-attention mechanism for human posture prediction and data point clustering;
- Using a loss function constructed by Bayesian stochastic variational inference with the distributions rather than the coordinates of data points to reduce the computational complexity significantly and raise tolerance to outliers;
- Providing the probabilistic map to predict falling incidents in a timely manner;
- Validating that the types and observing directions of sensors for data acquisition would not affect the accuracy of the probabilistic map exhibition, i.e., highly compatible with various environments.

In the Materials and Methods section, we describe the datasets, data preprocessing operations, and the compositions of the neural network, including the backbone models and the baseline. We also explain how the proposed framework reduced computational complexity and detected small-size objects. A specific loss function is developed based on the theoretical foundation of Bayesian stochastic variational inference. By incorporating the self-attention mechanism and the backpropagation, the framework updates the loss function's statistical parameters according to the detected information. In Results and Discussion section, the performance comparison of the backbone models for fast object detection and prediction is exhibited. The model with the best performance is selected for our framework. The generalization capabilities of the loss function under different devices and environments are also demonstrated. Finally, we propose a probabilistic map for the prediction of human postures. In Conclusion section, we summarize the achievements of the proposed framework and the future recommendations.

2. Materials and Methods

Figure 1 illustrates the proposed framework and experimental procedures. For data preprocessing in part (a), we adopted the MS COCO dataset [33] and applied binary transform and data augmentation, including affine transformations, RGB correction, and intensity correction. For object detection in part (b), the input images were standardized before being sent into the lite SSD network with FPN and IAO using MobileNet, ShuffleNet, or SqueezeNet as the backbone. ResNet was also used as the baseline for comparison. For posture prediction in part (c), the locations and speeds of the extracted BBox centroids were the input features of the self-attention block. The feature vectors carrying centroids and clustering properties then delineated in the probabilistic map of human postures estimated by the Bayesian-based model.

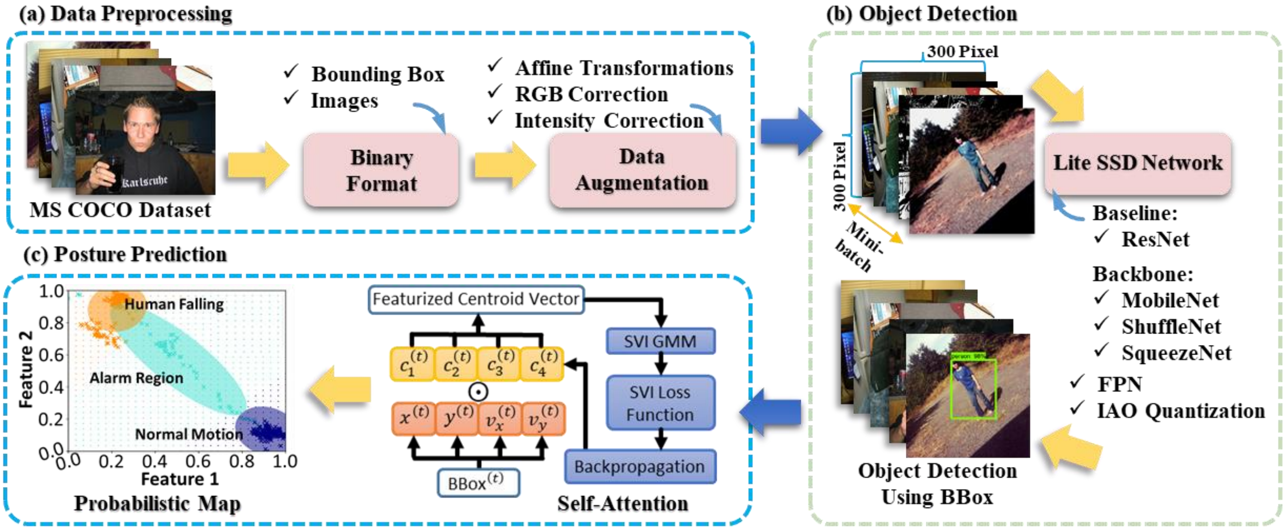

Figure 1. The proposed framework, including (**a**) data preprocessing using the MS COCO dataset for training and the ImageNet [34] for data augmentation, (**b**) establishment of a lite SSD network for fast human detection, and (**c**) the integration of statistic learning and self-attention mechanism for human posture prediction and clustering.

2.1. *Establishment of Binary Format and Data Augmentation*

The MS COCO, Pascal VOC2012 [35], and ImageNet [34] are open-access datasets extensively utilized for object detection. Among these, the MS COCO has the most detectable objects (i.e., BBox numbers) and the most balanced object sizes (equal portions of small, middle, and large objects). Hence, the MS COCO database is more in line with the daily environment and thus can achieve a better effect on deep learning model training on object recognition. We first transformed the labels and images into a binary data format to facilitate the reduction in loading time and the efficiency of parallel operations. Then we took the geometric (affine) transformations to avoid overfitting caused by the uneven distributions of image sizes and increase image content information. The affine transformations employed to the BBoxs and the images included rotations, flipping, random cutting, and deformations. We also randomly adjusted color brightness, saturation, hue, and contrast to simulate different real-world environments. To avoid overexposure or underexposure of the input images, we applied the mean values of RGB channels of the ImageNet dataset to whiten all input images.

2.2. Design of Lite SSD Network and Model Selection

As shown in the performance of lightweight neural networks presented by Lin et al. [29], the shallower the neural layers, the better detection of small objects, but the weaker information on locations. Conventional networks adopt featurized image pyramids that utilize multi-scale-fused features in training procedures to address this issue; however, it increases the inference time. Thus, we employed FPN to locate small objects in the shallow layers efficiently. Structurally, FPN extracts meaningful features with deep convolutional layers and then grasps better position information through up-sampling. It then fuses feature maps with the same size to preserve the original recognition scales. Meanwhile, to avoid the aliasing effect occurring in up-sampling procedures, we added a convolutional layer after the feature map fusing in the lite SSD network.

To pursue real-time object detection and implementation in lightweight neural networks, we considered the data-loading speed and storage size as vital factors for framework optimization. It is because the structural complexity of a neural network affects training efficiency and inference speed. In our experiments, we found that using FLOPs (floating-point operations per second) to evaluate the structural complexity might not reflect the actual inference speed. Reference [22] evidences the results. Thus, in addition to model size and inference speed, we adopted mean average precision (mAP) estimation [13] to evaluate the robustness of the backbones and the baseline. We also observed the performance of quantization-aware training of the IAO algorithm in this stage.

2.3. Theoretical Foundation of Stochastic Variational Inference Gaussian Mixture Model with Self-Attention Mechanism

As illustrated in part (c) of Figure 1, the self-attention mechanism extracts instant centroid locations (i.e., $x^{(t)}$ and $y^{(t)}$) and speeds (i.e., $v_x^{(t)}$ and $v_y^{(t)}$) at the time t from the detected BBox$^{(t)}$. The parameters $c_i^{(t)}$, $i = 1, 2, 3, 4$ are the corresponding weights to the instant centroid features. The symbol $\odot$ represents the operation of the Hadamard product between the vectors. The preprocessed sequential data generated vectors containing time and position information between BBox$^{(t)}$ and BBox$^{(t+20)}$, which became feature vectors for cluster analysis. The loss function derived from stochastic variational inference (SVI) and the backpropagation train the parameters used in the self-attention block and statistical distributions. Then, the results combined with the Gaussian mixture model (GMM) present the possible states of human motions in the probabilistic map.

The Bayesian neural network uses a set of variational distributions $q(\boldsymbol{z})$ to approximate the posterior distributions $p(\boldsymbol{z}|\boldsymbol{s})$. The logarithmic probability density function (PDF) of a sample $\ln p(\boldsymbol{s})$ can often be expressed as the linear combination of evidence lower bound (ELBO) and the Kullback–Leibler divergence (KLD) [36–38]:

$$\ln p(\boldsymbol{s}) = \int_{z} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{s},\boldsymbol{z})}{q(\boldsymbol{z})}\right) dz + \mathrm{KLD}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{s})), \tag{1}$$

where $\boldsymbol{s}$ and $\boldsymbol{z}$ represent sample data vectors and latent variable vectors, respectively. The goal of the variational inference is to achieve the maximization of ELBO and the minimization of KLD simultaneously through variational calculation under the condition that $\ln p(\boldsymbol{s})$ is a constant. Minimizing KLD means that the variational distribution should be similar to the posterior distribution, so that we only need to consider maximizing the ELBO under this constraint. The convenient way to find the extremum of the ELBO is to introduce the mean-field theory into the variational distribution [38]. However, this technique relies on taking all samples to update the variational distribution, computational complexity would arise. When the posterior distribution becomes more complicated, it also needs more iterations of variational distributions. All these operations lead to high computational costs and structural uncertainty under the circumstances of large datasets and complicated posterior forms [36,37].

We proposed a new technique based on the structure of stochastic variational inference to conquer these problems. It can resolve the issue of complicated posterior form and reduce computational costs through mini-batches. It also made the variational distribution approach the posterior distribution by maximizing the ELBO and adapting the ELBO into a tractable PDF:

$$\begin{aligned}\text{ELBO} &= \int_z q(z)\ln\left(\frac{p(s,z)}{q(z)}\right)dz \\ &= \int_z q(z)\ln p(s|z)dz + \int_z q(z)\ln p(z)dz - \int_z q(z)\ln q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\ln p(\boldsymbol{s}|\boldsymbol{z}) - \text{KLD}(q(\boldsymbol{z})||p(\boldsymbol{z})).\end{aligned} \tag{2}$$

The terms $p(\boldsymbol{s}|\boldsymbol{z})$ and $p(\boldsymbol{z})$ are variational likelihood and prior distribution, respectively. Notice that we used the discrete sampling form to replace the Monte Carlo integration in the first term of Equation (2). The parameter L is the sampling size. To integrate this result into the deep learning structure, we further modified the ELBO in Equation (2) as a loss function $\mathcal{L}(\boldsymbol{s},\boldsymbol{z})$ so that the backpropagation could sequentially update the parameters in the self-attention block and the GMM:

$$\mathcal{L}(\boldsymbol{s},\boldsymbol{z}) = -\left[\frac{1}{L}\sum_{l=1}^{L}\ln p(\boldsymbol{s}|\boldsymbol{z}) - \text{KLD}(q(\boldsymbol{z})||p(\boldsymbol{z}))\right]. \tag{3}$$

Equations (2) and (3) jointly show that ELBO is equivalent to the linear combination of variational likelihoods and the KLD is constructed by variational distributions and sample priors. Equation (3) implies that when the variational distribution $q(\boldsymbol{z})$ and prior distribution $p(\boldsymbol{z})$ gradually become similar during training, the KLD would also approach zero. Then, the logarithmic variational likelihood, the first term of Equation (3), would reach its maximum value due to obtaining the corresponding distributions of latent variables inputs $\boldsymbol{z}$. Since the KLD in Equation (3) is always positive, the logarithmic variational likelihood can be treated as the lower bound of the loss function. This equation is tractable and has a predictable lower bound. The belonging parameter distributions also can be updated in the training procedures. Thus, these elegant mathematical properties make it suitable to be a loss function. In other words, the loss function established from the ELBO in this study allows us to fuse the technique of backpropagation of deep neural networks with the statistical learning models for more complex analyses.

The prior and variational likelihood distributions were all Gaussian in the study. The relevant initial statistical parameters of the prior distribution $p(\boldsymbol{z})$ were the mean value $\mu_{prior,\ k} \sim Normal(0,1)$, the inverse covariance matrix $\Sigma^{-1}_{prior,\ k} \sim Wishart(3,\ \boldsymbol{I}_K/3)$, and the cluster weight $\alpha_{prior,k} \sim Dirchlet(2K,\ 2K)$. Then, the parameters of the variational likelihood distribution $p(\boldsymbol{s}|\boldsymbol{z})$ were $\mu_{var,\ k} \sim Normal(\boldsymbol{N}_1,\boldsymbol{N}_2)$, $\Sigma^{-1}_{var,\ k} \sim Wishart(\boldsymbol{W}_1,\ \boldsymbol{W}_2)$, and $\alpha_k \sim Dirchlet(\boldsymbol{D},\ \boldsymbol{D})$. The backpropagation updated the parameter vectors in these distributions, namely $\boldsymbol{N}_1$, $\boldsymbol{N}_2$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{D}$. The factor k was the index of cluster number K, and was assumed to be 2 or 3 in the SVI GMM training. Therefore, the variational likelihood has the form:

$$p(\boldsymbol{s}|\boldsymbol{z}) = GMM \sim \sum_{k=1}^{K=2,3}\alpha_k Normal\left(\boldsymbol{s}\middle|\mu_{var,\ k},\Sigma_{var,\ k}\right). \tag{4}$$

Please note that the proposed framework governed the training procedures and updated the statistical parameters of kth prior distribution $p(\boldsymbol{z})$ sequentially through mini-batches. The collected sequential data point distributions gradually fit the mean value $\mu_{prior,\ k}$, covariance matrix $\Sigma_{prior,\ k}$, and the cluster weight $\alpha_{prior,k}$ of the kth prior distribution $p(\boldsymbol{z})$ in the training procedures. Then, those optimized statistical parameters from prior distributions $p(\boldsymbol{z})$ consisted of and updated the parameters of the variational likelihood distribution $p(\boldsymbol{s}|\boldsymbol{z})$. Since this technique used only the distributions instead of the original position of data points, it reduced the computational complexity significantly and raised outlier tolerance. Not only could we provide the corresponding probabilistic map without losing the inference performance, but we could also estimate the posterior distributions $p(\boldsymbol{z}|\boldsymbol{s})$ by employing the outcome from Equation (4) directly:

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{\alpha_z Normal\left(\boldsymbol{s}\middle|\mu_{z,\ k}, \Sigma_{z,\ k}\right)}{\sum_{k=1}^{K=2,3} \alpha_k Normal\left(\boldsymbol{s}\middle|\mu_{var,\ k}, \Sigma_{var,\ k}\right)}. \tag{5}$$

the parameters α_z, $\mu_{z,k}$, and $\Sigma_{z,\ k}$ are the cluster weight, mean value, and covariance matrix of the data cluster constructed by the variational distribution $q(\boldsymbol{z})$, respectively.

3. Results and Discussions

3.1. Performance Comparison of the Object Detection Models

To fairly compare and inspect the capability of the proposed framework, we employed MoblieNet, ShuffleNet, and SqueezeNet as backbones of our lite SSD networks, in which FPN and IAO algorithms were incorporated to enhance small object detection, reduce model sizes, and raise the inference speed. We also adopted the ResNet as the baseline model for performance comparison. The backbone models were the main techniques employed for object detection, so their intrinsic performance indicated the general effectiveness. Table 1 summarizes the comparison results of these backbones incorporated with FPN and the IAO algorithm. The overall performance of mAP, inference speed, and model size reflected their potential of being the backbone model in the proposed framework.

Table 1. The performance comparison of backbone models.

Backbone	FPN	IAO	mAP	Inference Speed (mSec)	Model Size (MB)
ResNet			24.6	48	227.8
MobileNetV1			21.2	15	74.1
	✓ [a]		33.1	28	132.9
		✓	20.6	13	28.2
	✓	✓	32.5	27	46.2
MobileNetV2			23.0	17	174.9
	✓		35.6	29	320.1
		✓	22.7	12	94.9
	✓	✓	33.6	25	140.2
ShuffleNet V1 (Group = 4)			20.7	20	77.6 [b]
			20.7	16	77.6 [c]
		✓	20.3	– [e]	34.6 [d]
ShuffleNet V2			21.9	18	55.1
SqueezeNet			16.5	10	16.2

[a] The method was employed. [b] Using *For* loop iterations. [c] Using GPU parallel iterations. [d] Using TFlite framework. [e] Very poor inference speed.

ResNet, as the baseline, generated a fair mAP of 24.6 but a relatively slow speed of 48 mSec and a large model size of 227.8 MB. MobileNetV1 and MobileNetV2 had similar mAPs (32.5 and 33.6) and inference speeds (27 and 25 mSec), but MobileNetV1 had a much smaller model size than that of MobileNetV2 (46.2 vs. 140.2 MB). Because ShuffleNet divided different feature map channels into different groups and operated convolutions separately, general *For* loop iterations and single GPU parallel iterations were used to inspect its performance. Although *For* loop iterations and parallel convolution operations had faster inference speeds (20 and 16 mSec), their model sizes were not small enough. ShuffleNet did not effectively support TensorFlow Lite (TFlite) [26]; therefore, incorporating IAO using TFlite resulted in a poor inference speed. SqueezeNet had an extremely fast speed of 10 mSec and a tiny model size of 16.2 MB, but the worst mAP of 16.5. With the best overall performance, MobileNetV1 was selected as the backbone of the lite SSD network. The results listed in Table 1 also validate FPN's contribution to improving model accuracy and IAO's ability to accelerate inference speed.

3.2. Object Tracking and Human Posture Classification

There were 15 healthy subjects with a mean height of 158.6 ± 14.3 cm in our study. As shown in Figure 2, in-house-made 60 FPS (frame per second) videos were collected from each subject using a commercial webcam and a surveillance camera. We employed only low-resolution images in this study to achieve fast object detection. The two apparatuses were set at different heights to simulate different data acquisition environments with the camera at 3.1 m and the webcam at 1.6 m. The two data sources helped to validate whether the SVI GMM could map different data types to the same probabilistic map. Subjects were asked to rotate in place for 30 s to imitate the dizzy situation before falling onto the air mattress with consciousness. The protocol matched the requirement of [3].

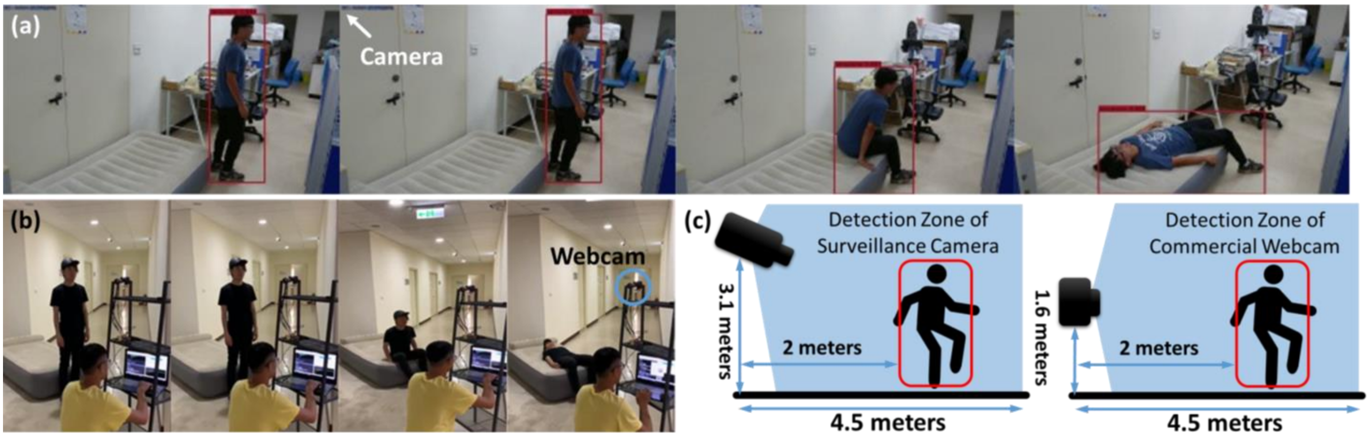

Figure 2. The demonstration of data acquisition from (**a**) a surveillance camera and (**b**) a commercial webcam. (**b**) The scene of inference performance testing, and thus there were no BBoxs on the subject. (**c**) The critical dimensions of the experimental environments using a surveillance camera and a commercial webcam.

To further explore the possible reduction in computational complexity of SVI GMM, the accuracies of classified results were analyzed by employing the diagonal and full covariances of the GMM. We initially assigned $K = 2$ in Equations (4) and (5) and used diagonal covariance to simplify the computational cost. Figure 3 demonstrates the corresponding variational likelihood map in (a) and the normalized feature map [39] in (b). The green and blue dot grids in Figure 3b represent the warning and normal regions, respectively. The cross markers represent the actual data points classified by the SVI GMM. Only two groups are delineated in Figure 3a,b since K equals 2; however, unclassified data points appear between those clusters. It implies that this dataset should have more than two groups [3,39]. Figure 4 shows the corresponding maps estimated by Equations (4) and (5) with $K = 3$ and diagonal covariance. The utilization of the oversimplified GMM covariance caused bizarre classified results. The group consisted of the unclassified data points, as those orange cross markers depicted in Figure 3b eventually dominated the classification. It also resulted in blurred group boundaries and reduced the maximum value in the variational likelihood map. In other words, utilizing the diagonal covariance of GMM with $K = 2$ or $K = 3$ would increase the uncertainty of data classification.

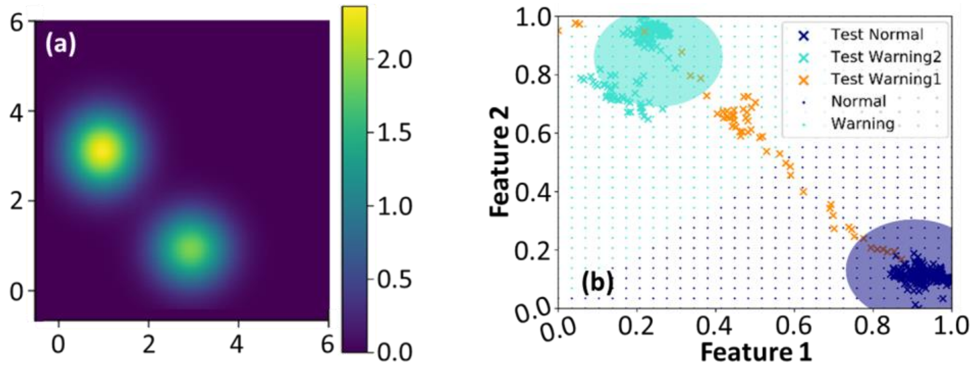

Figure 3. (**a**,**b**) The variational likelihood map and the normalized map in the feature space estimated by Equations (4) and (5) with $K = 2$ and diagonal covariance of GMM, respectively.

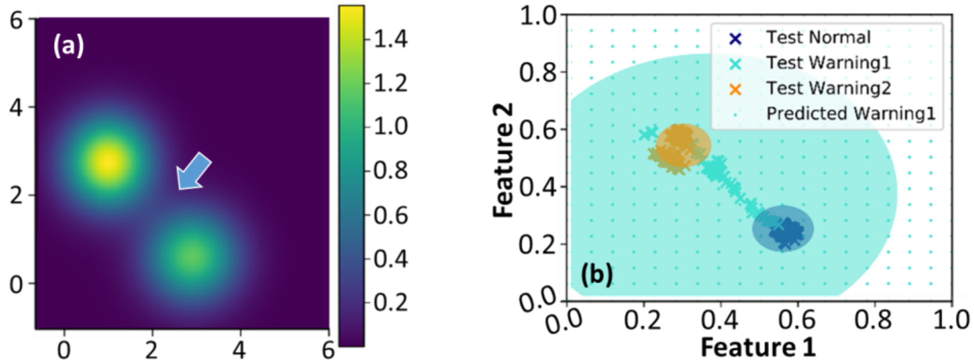

Figure 4. (**a**,**b**) The variational likelihood map and the normalized map in the feature space estimated by Equations (4) and (5) using $K = 3$ and diagonal covariance of GMM, respectively. The arrow depicted in (**a**) indicates that the boundaries of these two groups become blurred.

To reduce the classification uncertainty and reinforce the likelihood estimation, we eventually employed full covariance of the GMM and $K = 3$ in the proposed framework for human posture classification. Figure 5 illustrates the variational likelihood map and the corresponding normalized map in the feature space after the SVI GMM training. The intensity of the groups in the likelihood map became more concentrated. Predicted by the training datasets, the blue, green, and orange dot grids shown in Figure 5b indicate the regions of normal motion, transition warning, and falling, respectively. The cross markers represent the actual data points of normal motions, transition motions, and falling, respectively. Then the blue, green, and orange ellipse regions are the Eigen-matrices of the covariance of likelihoods corresponding to Figure 5a. These Eigen-matrices reflect the uncertainty of data variations and provide the visualization of the discriminant distributions. When a falling event occurs, the data points of posture features would sequentially distribute from the normal motion region through the transition region and then reach the falling regions. This procedure took about 0.66 s and underwent 40 extracted BBox centroid points. Table 2 lists the quantitative analysis of data point classification under the proposed framework. Table 3 lists the performance comparison between state-of-the-art techniques and the proposed framework. It should be emphasized that only the proposed framework inferred fast enough to generate alarm warnings before a human falling event happens.

Table 2. The quantitative results of the proposed framework.

True Positive	True Negative	False Position	False Negative
26,857	11,472	3015	1208

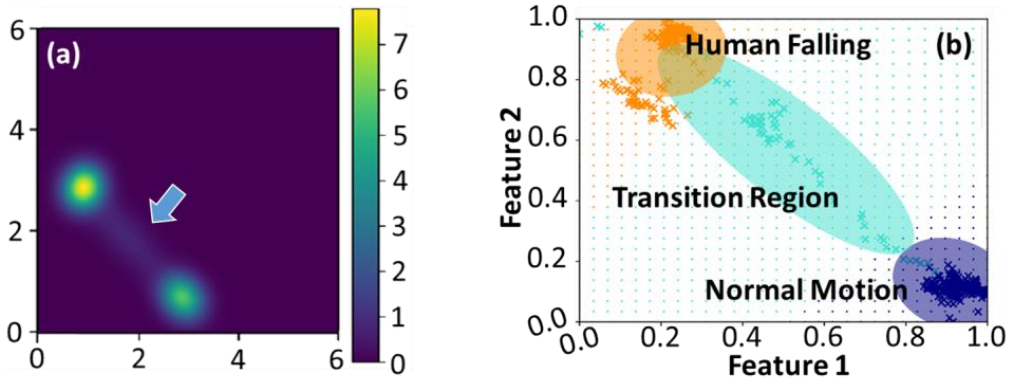

Figure 5. (**a**) The variational likelihood map estimated by Equation (4) and (**b**) a normalized map in the feature space. The colored regions in (**b**) exhibit the human posture classification. The arrow depicted in (**a**) implies there is a transition region between these two groups.

Table 3. The performance comparison between state-of-the-art techniques and the framework.

Source	Apparatus	Method	Accuracy	Alarm Timing [1]
Ref. [3]	Sensor Fusion	Machine Learning	0.90	+0.7 s
Ref. [4]	Vision-based Method	SpeedyAI, Inc.	0.89	+10 s
Ref. [6]	Vision-based Method	CNNs	0.98	– [2]
Ref. [7]	–	3D CNNs	0.99	–
This work	Vision-based Method	Lite SSD	0.90	−0.66 s

[1] + and −: The alarm will occur after and before the human falling events, respectively. [2] –: Not available.

Figure 6 exhibits the probabilistic maps established using the SVI GMM. This method mapped the BBox centroid points into three distinct predictive situations. Then, the SVI GMM endowed these points with their corresponding probability values. The centroid points were in the normal region of Figure 6a when the subject walked or stood normally. When the detected centroid points moved into the transition region (Warning1) of Figure 6b and migrated toward the falling region (Warning2) of Figure 6c, the system would generate alarm warnings immediately.

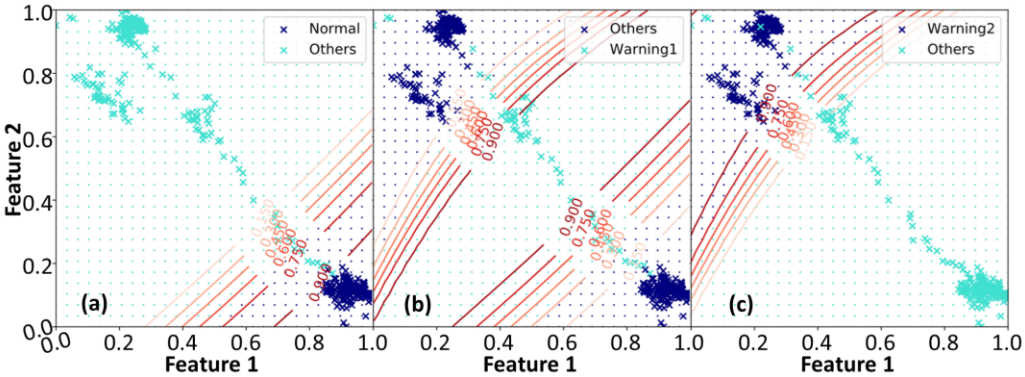

Figure 6. The demonstrations of the probabilistic maps constructed by the SVI GMM. (**a**) The distinct distributions of normal motion (the blue cross markers) and other situations (the green cross markers). (**b**) When the detected centroid points, as depicted by the green cross markers, were in the transition region, the proposed framework would generate the first alarm warnings. (**c**) When the detected centroid points, as depicted by the blue cross markers, were in the falling region, the proposed framework would generate second alarm warnings.

4. Conclusions

This article provides a new framework for lightweight deep neural network modeling, and it meets the demand for fast classifying of human posture images and subsequent warnings. This framework simultaneously achieves high mean average accuracy, high inference speed, and small model size of object detection tasks. The method uses a commercial webcam and a surveillance camera for data acquisition. It matches the requirement of contactless human posture detection. This method has a form of lite SSD network embedded with quantization-aware training and a self-attention mechanism, and thus it can reduce model sizes and raise the inference speed. The framework can fuse the information from images and corresponding sequential signals obtained from bounding boxes. The proposed method merges the techniques of statistical learning into deep learning. Hence, the trained parameters own their statistical meanings. The classified results of images and corresponding sequential signals could be mapped onto probabilistic maps directly. Therefore, this lightweight structure could quickly estimate the probability of human postures and generate alarms once the corresponding data points move into the warning regions. This method connects the loss function with the technique of stochastic variational inference. Thus, it endows the notions of probability to the classification inference. Since the framework has a superior achievement on inference speed and model size, it is a strong candidate for low-cost applications of edge computing and embedded systems. Furthermore, the framework can be the baseline for developing tiny machine learning (TinyML) techniques or other lite structural platforms. Therefore, we anticipate this framework can benefit the progress of contactless smart sensing and detection in biomedical AIoT developments.

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

Funding: This research was funded by the National Science and Technology Council of Taiwan, grant number 111-2221-E-008-087.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

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

References

1. Chander, H.; Burch, R.F.; Talegaonkar, P.; Saucier, D.; Luczak, T.; Ball, J.E.; Turner, A.; Kodithuwakku Arachchige, S.N.K.; Carroll, W.; Smith, B.K.; et al. Wearable Stretch Sensors for Human Movement Monitoring and Fall Detection in Ergonomics. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3554. [CrossRef] [PubMed]
2. Kerdjidj, O.; Ramzan, N.; Ghanem, K.; Amira, A.; Chouireb, F. Fall detection and human activity classification using wearable sensors and compressed sensing. *J. Ambient. Intell. Human Comput.* **2020**, *11*, 349–361. [CrossRef]
3. Hsu, F.-S.; Chang, T.-C.; Su, Z.-J.; Huang, S.-J.; Chen, C.-C. Smart Fall Detection Framework Using Hybridized Video and Ultrasonic Sensors. *Micromachines* **2021**, *12*, 508. [CrossRef] [PubMed]
4. Shu, F.; Shu, J. An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box. *Sci. Rep.* **2021**, *11*, 2471. [CrossRef] [PubMed]
5. Rastogi, S.; Singh, J. Human fall detection and activity monitoring: A comparative analysis of vision-based methods for classification and detection techniques. *Soft Comput.* **2022**, *26*, 3679–3701. [CrossRef]
6. Ding, W.; Hu, B.; Liu, H.; Wang, X.; Huang, X. Human posture recognition based on multiple features and rule learning. *Int. J. Mach. Learn. Cyber.* **2020**, *11*, 2529–2540. [CrossRef]
7. Alanazi, T.; Muhammad, G. Human Fall Detection Using 3D Multi-Stream Convolutional Neural Networks with Fusion. *Diagnostics* **2022**, *12*, 3060. [CrossRef]
8. Fei, K.; Wang, C.; Zhang, J.; Liu, Y.; Xie, X.; Tu, Z. Flow-pose Net: An effective two-stream network for fall detection. *Vis. Comput.* **2022**. [CrossRef]
9. Liu, J.; Wang, Y.; Liu, Y.; Xiang, S.; Pan, C. 3D PostureNet: A unified framework for skeleton-based posture recognition. *Pattern Recognit. Lett.* **2020**, *140*, 143–149. [CrossRef]

10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Lecture Notes in Computer Science*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37.
11. Araki, R.; Onishi, T.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. MT-DSSD: Deconvolutional Single Shot Detector Using Multi Task Learning for Object Detection, Segmentation, and Grasping Detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–30 June 2020.
12. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.-G.; Xue, X. DSOD: Learning Deeply Supervised Object Detectors from Scratch. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
16. Ansari, M.Y.; Yang, Y.; Balakrishnan, S.; Abinahed, J.; Al-Ansari, A.; Warfa, M.; Almokdad, O.; Barah, A.; Omer, A.; Singh, A.V.; et al. A lightweight neural network with multiscale feature enhancement for liver CT segmentation. *Sci. Rep.* **2022**, *12*, 14153. [CrossRef]
17. Li, W.; Liu, J.; Mei, H. Lightweight convolutional neural network for aircraft small target real-time detection in Airport videos in complex scenes. *Sci. Rep.* **2022**, *12*, 14474. [CrossRef]
18. 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.
19. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
20. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
21. Tan, M.; Le, Q.V. MixConv: Mixed Depthwise Convolutional Kernels. *arXiv* **2019**, arXiv:1907.09595.
22. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
23. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
24. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv* **2016**, arXiv:1602.07360.
25. Han, S.; Mao, H.; Dally, W.J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv* **2015**, arXiv:1510.00149.
26. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
27. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
28. Zhao, H.; Liu, D.; Li, H. Efficient Integer-Arithmetic-Only Convolutional Neural Networks. *arXiv* **2020**, arXiv:2006.11735.
29. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
30. Luong, M.-T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
32. Yao, L.; Ge, Z. Nonlinear Gaussian Mixture Regression for Multimode Quality Prediction With Partially Labeled Data. *IEEE Trans. Industr. Inform.* **2019**, *15*, 4044–4053. [CrossRef]
33. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755.
34. Jia, D.; Wei, D.; Richard, S.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
35. Everingham, M.; Eslami, S.M.A.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
36. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic Variational Inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
37. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.

38. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
39. Chen, C.-C.; Juan, H.-H.; Tsai, M.-Y.; Lu, H.H.-S. Unsupervised Learning and Pattern Recognition of Biological Data Structures with Density Functional Theory and Machine Learning. *Sci. Rep.* **2018**, *8*, 557. [CrossRef]

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.