electronics

Article

VHR-BirdPose: Vision Transformer-Based HRNet for Bird Pose Estimation with Attention Mechanism

Runang He [1], Xiaomin Wang [1], Huazhen Chen [2] and Chang Liu [3,*]

[1] School of Information Management, Beijing Information Science and Technology University, Beijing 100101, China; herunang@bistu.edu.cn (R.H.); wxm@bistu.edu.cn (X.W.)
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; huazhenchen@tju.edu.cn
[3] Institute of Applied Mathematics, Beijing Information Science and Technology University, Beijing 100101, China
* Correspondence: liu.chang.cn@ieee.org

Abstract: Pose estimation plays a crucial role in recognizing and analyzing the postures, actions, and movements of humans and animals using computer vision and machine learning techniques. However, bird pose estimation encounters specific challenges, including bird diversity, posture variation, and the fine granularity of posture. To overcome these challenges, we propose VHR-BirdPose, a method that combines Vision Transformer (ViT) and Deep High-Resolution Network (HRNet) with an attention mechanism. VHR-BirdPose effectively extracts features using Vision Transformer's self-attention mechanism, which captures global dependencies in the images and allows for better capturing of pose details and changes. The attention mechanism is employed to enhance the focus on bird keypoints, improving the accuracy of pose estimation. By combining HRNet with Vision Transformer, our model can extract multi-scale features while maintaining high-resolution details and incorporating richer semantic information through the attention mechanism. This integration of HRNet and Vision Transformer leverages the advantages of both models, resulting in accurate and robust bird pose estimation. We conducted extensive experiments on the Animal Kingdom dataset to evaluate the performance of VHR-BirdPose. The results demonstrate that our proposed method achieves state-of-the-art performance in bird pose estimation. VHR-BirdPose based on bird images is of great significance for the advancement of bird behaviors, ecological understanding, and the protection of bird populations.

Keywords: bird pose estimation; vision transformer; attention mechanism; Deep High-Resolution Network; Animal Kingdom dataset

check for updates

Citation: He, R.; Wang, X.; Chen, H.; Liu, C. VHR-BirdPose: Vision Transformer-Based HRNet for Bird Pose Estimation with Attention Mechanism. *Electronics* **2023**, *12*, 3643. https://doi.org/10.3390/electronics12173643

Academic Editor: Hamed Mozaffari

Received: 21 July 2023
Revised: 16 August 2023
Accepted: 24 August 2023
Published: 29 August 2023

1. Introduction

Pose estimation involves the automatic recognition and analysis of human or animal postures, actions, and movements using computer vision and machine learning techniques. In the context of 2D pose estimation [1], top-down methods approximate the object's position in the image, extract joint point position information, and generate a skeletal representation. The accurate positioning of the recognized object, the precise localization of the joint points, and the fidelity of the joint information directly impact the performance of pose estimation. Conversely, bottom-up methods directly detect all skeletal keypoints in the image, demanding precise joint localization and accurate skeleton information retrieval. Animal pose estimation holds significant importance and practical value in animal behavior research, animal conservation, agricultural production, animal health monitoring, and more [2–4].

Bird pose estimation holds immense significance and practical value in bird monitoring and protection [5]. In terms of bird behavior research, it enables researchers to gain deeper insights into bird behavior patterns, flight modes, and communication methods. Analyzing the gestures and movements of birds allows us to uncover various patterns related to their flight strategies, foraging behavior, courtship behavior, and breeding behavior. This knowledge contributes to a better understanding of the ecology, behavioral biology, and evolutionary biology of birds. In terms of bird protection, bird pose estimation plays a crucial role. By analyzing the postures and movements of birds, we can monitor their activity range, habitat utilization, and population status in real time. This information aids in the protection and management of bird habitats, as well as the assessment and monitoring of bird population numbers and distributions. Furthermore, it facilitates the implementation of timely protective measures to prevent and reduce threats and disturbances to birds. Bird pose estimation also facilitates the study of bird migration. By analyzing the postures and flight patterns of birds, we can unveil their migration routes, migration speeds, migration distances, and migration strategies. This information is valuable for understanding the migration ecology of birds, the adaptability of their migration behavior, and the impact of migration on bird populations. Moreover, it provides a scientific basis for the protection and management of birds.

Bird pose estimation is a crucial prerequisite for skeleton-based bird action recognition, behavior analysis, and object tracking [4]. However, it faces several challenges, including:

(1) Bird diversity and complexity: With various bird species having different body shapes, feather shapes, and flight modes, estimating bird poses is challenging [6]. The wide range of variations in bird postures, largely influenced by environmental factors, further complicates the task of developing accurate pose estimation algorithms for different bird species.

(2) Occlusion and posture change: Unlike human skeletons, bird skeletons are different and pose estimation often encounters issues such as changeable postures, non-rigid joint deformation, and partial occlusion of joints due to the uncontrollable photographing environment [5]. Overcoming these challenges requires solving the occlusion problem and establishing robust models for pose estimation [7].

(3) Data acquisition and annotation: Bird pose estimation requires a large number of annotated images for model training and evaluation, but collecting large-scale bird posture datasets is challenging [5]. The rapid movement, flying height, and concealment of birds make it difficult to collect data [8]. Additionally, accurate annotations of bird postures require professional knowledge and experience.

(4) Fine granularity of posture: Bird pose estimation demands precise positioning and tracking of multiple keypoints on the bird's body to obtain global posture information. However, bird posture has a complex structure and fine-grained changes [6], such as wing expansion, head rotation, and more. This necessitates pose estimation algorithms with high accuracy and robustness to capture subtle posture changes accurately.

To address the challenges and difficulties in bird pose estimation, we propose VHR-BirdPose, a bird pose estimation method that combines Vision Transformer (ViT) [9] and Deep High-Resolution Network (HRNet) [10] with an attention mechanism [11]. VHR-BirdPose effectively tackles the aforementioned issues and achieves outstanding performance in bird pose estimation. The key contributions of our method are as follows:

(1) Feature extraction using Vision Transformer: We introduce Vision Transformer as a feature extractor and utilize its self-attention mechanism to capture global dependencies in the images. This continuous global feature interaction allows the model to learn discriminative features from different regions, enabling better capturing of details and posture changes in bird images.

(2) Keypoint attention mechanism: We employ the attention mechanism to enhance the focus on bird keypoints, improving the accuracy of pose estimation. By dynamically adjusting the weights between image features and pose joints, the model can locate the most important regions in bird images and extract relevant features, leading to enhanced pose estimation accuracy.

(3) VHR-BirdPose model combining Vision Transformer and HRNet: Our proposed model integrates the advantages of HRNet, which extracts multi-scale features and

captures pose details through multiple parallel branches while preserving both high-resolution and low-resolution features. By combining HRNet with Vision Transformer, we leverage the attention mechanism of Vision Transformer to obtain richer semantic information while maintaining high-resolution features.

(4) Extensive experiments on the Animal Kingdom dataset [12]: We conduct thorough experiments to evaluate the performance of VHR-BirdPose. The results demonstrate that our model outperforms similar pose estimation methods, achieving state-of-the-art performance in the bird pose estimation task.

The code of VHR-BirdPose is available at https://github.com/LuoXishuang0712/VHR-BirdPose (accessed on 16 August 2023).

2. Related Work

There exists a wide range of deep learning methods for pose estimation, primarily utilizing Convolutional Neural Networks (CNNs) or Transformer-based approaches that incorporate attention mechanisms [11]. These methods often employ a common idea in pose estimation, which involves predicting joint positions using a deep learning network to generate heatmaps. Subsequently, computer vision techniques are utilized to convert these heatmaps into skeleton information, thus enabling pose estimation.

Since the emergence of the Transformer architecture, attention-based networks have been significantly improved and widely utilized in various computer vision tasks [13].

2.1. Pose Estimation Methods Based on HRNet

HRNet (High-Resolution Network) [10] is a successful deep learning method applied to human pose estimation. By employing a high-resolution multi-branch network structure and feature fusion strategy, pose estimation accuracy and robustness are significantly improved. When applied to animal pose estimation, the HRNet method draws inspiration from human pose estimation techniques and adjusts and optimizes them according to animal-specific characteristics.

The core concept of HRNet is to maintain high-resolution multi-scale features and fuse them using the multi-branch network structure, effectively leveraging multi-scale information for pose estimation [10,14]. HRNet consists of a base network and multiple branch networks. The base network extracts low-level features from images, whereas the branch networks extract high-level features at different scales. HRNet constructs a multi-level feature pyramid through several downsampling and upsampling operations, ensuring that each layer retains high-resolution features [15,16]. This approach incorporates both global and local features while preserving crucial details, thereby enhancing the accuracy of pose estimation.

Feature fusion plays a vital role in HRNet [14,16]. It involves merging features extracted from different branch networks. HRNet adopts a top-down feature fusion strategy, beginning with high-resolution features and gradually fusing them with low-resolution features layer by layer. This strategy preserves the high resolution of features and effectively utilizes multi-scale information to enhance pose estimation accuracy.

HRNet also utilizes a bottom-up pose estimation strategy, which starts with initial rough pose estimation results and gradually refines them through optimization and iteration. HRNet achieves pose estimation by employing both top-down pose estimation and bottom-up correlation embedding. In the top-down pose estimation stage, HRNet first detects human body instances in the image and then estimates the pose for each instance. In the bottom-up correlation embedding stage, HRNet further refines pose estimation by learning the correlation between different body keypoints, addressing challenges related to occlusion and overlapping in pose estimation.

The HRNet pose estimation method achieves accurate results by incorporating a high-resolution multi-branch network structure and effectively utilizing multi-scale features through feature fusion [16]. Animal pose estimation based on HRNet needs to consider the unique characteristics of animal poses. Challenges in animal pose estimation include extensive variations in animal poses, morphological differences among animals, occlusion, and distinct animal movement patterns. To mitigate these challenges, a combination of prior knowledge in animal pose estimation and specific data augmentation techniques is often employed to enhance the robustness and generalizability of pose estimation.

2.2. Applications of Vision Transformers in Pose Estimation

Vision Transformers have emerged as a great contributor to the Transformer structure from Natural Language Processing (NLP) to Computer Vision tasks [17]. Inspired by the Transformer's excellent performance in NLP, Dosovitskiy et al. [9] applied the classical Transformer structure to visual processing tasks by dividing images into grid-like patches, resulting in significant accuracy improvements in image classification. Since then, a series of Vision Transformer models have been developed for various visual processing tasks, including DeiT [18], which incorporates knowledge distillation to enhance training efficiency, and Swin Transformer [19], which introduces a multi-scale structure and sliding window attention mechanism. The remarkable performance of Vision Transformers in tasks like image recognition and semantic segmentation has inspired their further exploration and application in various specialized fields.

Yang et al. [20] introduced Vision Transformers to pose estimation, developing a pose recognition method that combines convolutional layers with Transformers. Li et al. [21] further modified the output of Vision Transformers for pose estimation by employing tokenization and generating heatmaps directly through fully connected layers. This approach unified the learning of key details and constraints. However, the application of Vision Transformers to bird pose estimation is still limited, and challenges persist in distinguishing the foreground from the background and the coarse granularity of joint recognition. To address these challenges, we propose a hybrid method that combines Transformers with HRNet, which effectively handles these issues.

2.3. Pose Estimation Based on Attention Mechanism

In pose recognition based on the attention mechanism, the network can pay more attention to the position and feature of keypoints by adaptively weighting the importance of keypoints, thus improving the accuracy and robustness of pose estimation. In traditional pose estimation methods, fixed weights or uniform distribution are usually used to model keypoints. However, in practical applications, the importance of different keypoints may be different, which means some keypoints contribute more to final posture output. Therefore, the attention-based method enables the network to automatically select and focus on the important keypoints by learning adaptive weights [22].

In the attention-based pose estimation methods, different attention mechanisms can be used to realize the weighting of keypoints. A common method is to use the self-attention mechanism, which assigns a weight to each keypoint by calculating the similarity between keypoints. In this way, the network can adaptively adjust attention according to the interaction between keypoints. Another method is to use the spatial attention mechanism, which assigns weights according to the positions of keypoints in image space. In this way, the network can pay more attention to important areas and keypoints in the image.

The pose estimation method based on the attention mechanism can also combine multi-scale information for pose estimation [23]. By introducing an attention mechanism at different scales, the network can pay attention to keypoints and features at different scales at the same time. This can help improve the multi-scale modeling ability of the network for posture and enhance the robustness and generalization ability of pose estimation [24].

The attention mechanism can be combined with the Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) to obtain more complex feature representation and time series modeling [13,25]. It can also be combined with the methods of Generative Adversarial Network (GAN) or Reinforcement Learning (RL) to achieve a more accurate and stable pose estimation.

3. Vision Transformer-Based HRNet with Attention Mechanism

3.1. Main Network Architecture

As illustrated in Figure 1, the VHR-BirdPose network comprises the attention branch and the HRNet branch. The attention branch is represented by the blue part of the network, whereas the yellow part represents the HRNet branch. These two branches operate in parallel. At the end of the model, the heatmap is generated through a feature fusion strategy. Following the first convolutional block, the attention branch and HRNet branch work independently, eventually merging their features before reaching the output layer. In practice, the merging strategy is implemented through direct addition.

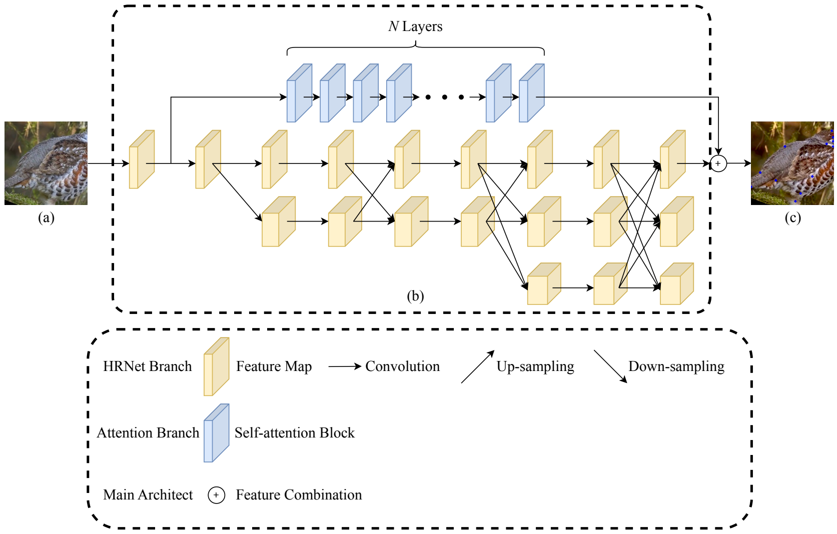


Figure 1. The main network structure of VHR-BirdPose. (**a**) Represents the input image; (**b**) represents the main network structures; (**c**) represents the output of the network covered on the original image. In part (**b**), the yellow-marked part represents the features extracted by the HRNet branch, and the blue part corresponds to the features generated by the self-attention blocks in the ViT branch. Following the initial convolutional block, the features are separately passed through the attention branch and the HRNet branch. Following N attention blocks, the features are merged and undergo a dimension reduction operation to ultimately produce the heatmap.

The Vision Transformer serves as the core feature extractor. Its self-attention mechanism learns the relationships among different regions of the image, enabling a better understanding of bird details and pose variations. By combining the Vision Transformer with HRNet, the attention mechanism of the Vision Transformer facilitates the extraction of richer semantic information while preserving high-resolution features.

Building upon the existing architecture, we have devised multiple variations of VHR-BirdPose in varying sizes, as outlined in Table 1. These variants have been tailored to suit diverse model requirements, and they differ in terms of attention branch depth, attention head count, and parameter quantity. The HRNet branch width remains consistent at 32 for all variants, making it suitable for a wide range of tasks.

Table 1. Illustration of VHR-BirdPose variants in different sizes.

Method	Layers	Self-Attention Heads	Width of HRNet Branch	#Param. [1]
VHR-BirdPose-S	6	12	32	106 M
VHR-BirdPose-B	12	12	32	181 M
VHR-BirdPose-L	12	16	32	181 M
VHR-BirdPose-XL	24	16	32	333 M

[1] The number of parameters.

3.2. Design of Attention Mechanism

In this approach, the utilization of a multi-head attention mechanism is employed to enhance the precision of pose estimation by emphasizing the keypoints of birds. By incorporating this attention mechanism, the model becomes capable of automatically learning the most crucial regions and features within bird images, thereby facilitating improved bird posture estimation.

The attention branch divides the image into smaller patches, each measuring 16 pixels in width and height. Position embeddings are incorporated at the initial Transformer attention block in the ViT architecture. Our method maintains the overall structure of ViT as the attention branch, employing N layers of ViT to direct attention towards the bird within the image. The image is split and restored by 2D convolution and 2D deconvolution after the input of $I^{att} \in \mathbb{R}^{B \times C_1 \times H \times W}$ and before the output of $O^{att} \in \mathbb{R}^{B \times C_1 \times H \times W}$. The split matrix $P \in \mathbb{R}^{B \times C_2 \times \frac{H}{d} \times \frac{W}{d}}, d^2 = 16, C_2 = C_1 \times d^2$ is a new input data which contains all image patches.

The attention branch consists of N consecutive multi-head attention blocks. At the beginning of the attention branch, there is a convolutional layer responsible for patch splitting, whereas a deconvolutional layer performs the inverse operation at the branch's end. Additionally, a 1 × 1 convolutional layer is applied to the feature maps generated by the attention branch, reducing their dimensionality in preparation for subsequent feature map merging operations. In our method, the number of layers and self-attention heads serve as hyperparameters, which are determined based on the model sizes. For instance, VHR-BirdPose-B includes 12 self-attention blocks and 12 self-attention heads. Refer to Table 1 for a comprehensive overview of the model's structure.

3.3. The HRNet Branch

HRNet is a pose estimation method based on Convolutional Neural Networks. Unlike traditional approaches that gradually decrease the resolution of the feature map, HRNet employs parallel branches to extract feature maps at different resolutions. It introduces a feature-sharing strategy, utilizing upsampling and downsampling operations to propagate information among feature maps of varying resolutions. This enables HRNet to preserve information from different image scales, maintain a global perception of large-scale structures, and enhance recognition accuracy for small-scale details. Notably, when dealing with intricate data such as limbs and faces, the feature sharing between maps of different scales ensures the preservation of multi-scale information. Compared to regular deep Convolutional Neural Network methods, HRNet effectively reduces parameters and network layers, improving training efficiency. In our method, we leverage HRNet as a parallel branch to construct the pose estimation network. The HRNet branch captures valuable information at different scales, leading to enhanced accuracy and robustness for the pose estimation task.

The HRNet branch in our approach follows the original HRNet-W32 network, maintaining the same feature map sizes. Specifically, the feature sizes of the three parallel branches are set to 64, 128, and 256, respectively. Similar to the ResNet architecture, the feature maps' size in the first stage undergoes a transformation to 64 after four residual blocks. Following that, a group of 3 × 3 convolutional layers further reduces the feature map size to 32.

The structure of the HRNet branch in our method adheres to the fundamental HRNet design principles. It leverages the HRNet branch's inherent capability to fuse information from different scales, utilizes pre-training weights to initialize parameters partially, and thereby enhances the training efficiency of our approach.

3.4. Training Process

Given that the HRNet branch in our approach can extract image features at various scales, the attention branch can learn to focus on critical regions in the spatial dimension. Furthermore, the animal pose estimation dataset is considerably smaller than its human counterpart, comprising images taken from different perspectives and under varying light conditions. Therefore, we employ scale-based data augmentation operations during the training phase to enhance our method's accuracy and robustness when dealing with images of varying scales and irregular resolutions.

Specifically, we randomly apply half-body clipping, image rotation, image clipping, and horizontal flipping. For half-body clipping, we select the upper and lower body based on the dataset type, and then randomly choose a half-body part with no fewer than two visible keypoints for clipping. If both parts have less than two keypoints, we do not perform the half-body clipping operation. Regarding image rotation, we determine a rotation factor F_R before augmentation and randomize the rotation angle R as follows:

$$R_{Rand} = F_R \times N(0,1), R = \begin{cases} 2F_R, R_{Rand} > 2F_R \\ R_{Rand}, -2F_R \leq R_{Rand} \leq 2F_R \\ -2F_R, R_{Rand} < -2F_R \end{cases} \tag{1}$$

where $N(0,1)$ represents a random value conforming to the standard normal distribution. In practice, the rotation angle is randomly chosen between $-2F_R$ and $2F_R$ for rotation. For the image clipping, a clipping factor F_S is set before augmentation, and the clipping ratio S is randomized as:

$$S_{Rand} = F_S \times N(0,1), S = \begin{cases} 1+F_S, S_{Rand} > 1+F_S \\ S_{Rand}, 1-F_S \leq S_{Rand} \leq 1+F_S \\ 1-F_S, S_{Rand} < 1-F_S \end{cases} \tag{2}$$

The clipping ratio for half-body clipping is randomly chosen between $1 - F_S$ and $1 + F_S$. As for horizontal image flipping, this operation randomly flips the image and keypoints' coordinates. However, instead of directly flipping the x and y coordinates, the coordinate flipping of keypoints involves exchanging the keypoints pairs that occupy the same vertical position but differ in horizontal position. The definition of keypoint pairs is determined by the dataset.

In the training process, the model is randomly initialized by default. Since our method retains the basic structure of HRNet and ViT, both branches can utilize the corresponding pre-training weights from ImageNet. Furthermore, in the experiments section, we compare the performance of our proposed VHR-BirdPose model when using the loaded pre-trained weights.

3.5. Loss Function

The loss function employed in our method is based on the Mean Square Error (MSE) between the heatmaps. To obtain the final loss value, we multiply the MSE value by the weights of the corresponding keypoints. The loss function is formulated as follows:

$$Loss_{MSE} = \sum_{i=1}^{n} \sqrt{\sum_{w}^{W_H} \sum_{h}^{H_H} (\bar{X}(w,h) - \hat{X}(w,h))^2 \times W_i} \tag{3}$$

where n is the number of keypoints, which is 23 in the bird pose estimation dataset used in this paper, W_H and H_H are the width and height of the heatmap, respectively, $\bar{X}$ is the predicted output heatmap, $\hat{X}$ is the ground truth heatmap, and W_i is the weight of the i-th keypoint.

4. Experiments

4.1. Dataset

We conducted experiments on the Animal Kingdom dataset [12] to train and evaluate our method. The Animal Kingdom dataset is a comprehensive dataset that provides annotations for various tasks, such as pose estimation, motion recognition, and video grounding. Specifically, for the pose estimation task, the dataset contains more than

33,000 annotated video frames, with 8524 frames belonging to the bird category. In our experiments, we used 6819 frames for training and 1705 frames for validation. Each annotated image in the dataset includes 23 pose joints, namely the middle head, left eye, right eye, front top of mouth, back left of mouth, back right of mouth, front bottom of mouth, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, middle back of torso, left hip, right hip, left knee, right knee, left ankle, right ankle, back top of tail, back mid of tail, and back end of tail.

4.2. Experimental Design

The original images in the dataset are RGB images with a resolution of 640 × 480 pixels. We use about 80% of images for training and 20% for testing. During the training process, the images are first clipped into square images with a width and height of 256 pixels, using the object as the center of the image. To further augment the dataset, random horizontal flips, rotations, scales, and half-body clippings are applied to the images. The image rotations are in the range of $[-30°, 30°]$ and the scales are in the range of $[0.75, 1.25]$.

During training, the initial learning rate is 0.001, and when the epoch reaches 170 and 200, the learning rate is multiplied by $\frac{1}{10}$. Adam optimizer is used for optimization and ReLU for activation. The default batch size of each GPU is 4. Only VHR-BirdPose-XL has a batch size of 2 for each GPU. All models are trained for 300 epochs. The channels of three stages in the HRNet branch are $[32, 64]$, $[32, 64, 128]$, and $[32, 64, 128, 256]$. The VHR-BirdPose model is implemented by PyTorch and the Vision Transformer branch is based on the Timm framework [26]. Training is based on a single V100 PCIE 32GB GPU platform.

PCK (Percentage of Correct Key points) [27] metrics are used to evaluate the accuracy of our method. The $PCK@\alpha$ of the k-th keypoint is:

$$PCK@\alpha(k) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1, |P_k^i - \hat{P}_k^i| \leq \alpha B_i \\ 0, |P_k^i - \hat{P}_k^i| > \alpha B_i \end{cases} \tag{4}$$

where N represents the batch size, P_k^i represents the ground truth of the k-th keypoint in the i-th image, $\hat{P}_k^i$ represents the prediction of the k-th keypoint in the i-th image, and B_i is the distance threshold of i-th image.

The distance threshold is used to determine whether the predicted keypoint falls within the allowable neighborhood of the ground truth. For $PCK@\alpha$, the distance threshold is formulated by:

$$B_i = Max(H_i, W_i) \tag{5}$$

where H_i and W_i represent the height and width of the heatmap of the i-th image, respectively.

Finally, the average PCK is used as the metrics for comparison between methods in the experiment, and the average $PCK@\alpha$ is formulated as:

$$PCK@\alpha_{Mean} = \frac{1}{N_k}\sum_{i=1}^{N_k} PCK@\alpha(i) \tag{6}$$

where N_k represents the number of keypoints in the skeleton. In practice, $\alpha = 0.05$ is taken.

4.3. Experimental Results

In this section, we compare the experimental results of VHR-BirdPose proposed in this paper with traditional pose estimation methods based on CNN structures, methods based on the attention mechanism, and methods that combine CNN with the attention mechanism. Additionally, we evaluate the performance of each comparison method based on the number of parameters and FLOPs (Floating Point Operations) to assess their performance with varying model sizes. To validate the effectiveness of VHR-BirdPose under pre-training, we also include the performance of VHR-BirdPose with pre-trained weights loaded. Let us briefly introduce the comparison methods:

(1) Cascade Pyramid Network (CPN) [28]: CPN addresses challenges like keypoint occlusion, complex background, and the inability to group the skeleton into a single object in multi-person pose estimation. It consists of a multi-layer stacked GlobalNet and a small-scale RefineNet. CPN adopts a top-down approach, where objects are first detected using bounding boxes, and then the keypoints within each bounding box are predicted.

(2) SimpleBaseline [29]: SimpleBaseline is a method designed for various computer vision tasks. It aims to provide a relatively simple baseline method for comparison in experiments. In this section, we compare the human posture estimation and tracking branches of SimpleBaseline.

(3) HRFormer [30]: HRFormer combines the high-resolution network HRNet with local self-attention. It introduces a feedforward neural network (FFN) to share information in unconnected image windows, improving the memory capacity and computational efficiency of the network. HRFormer is also a pose estimation method that combines CNN structures with the attention mechanism.

(4) ViTPose [31]: ViTPose utilizes ViT (Vision Transformer) as the backbone for pose estimation. It includes a variant called ViTPose+ [32], which is optimized for different types of pose estimation targets. ViTPose replaces the traditional pipeline used by CNN-based pose estimation methods and utilizes ViT with a decoder specifically designed for the pose estimation task. Therefore, ViTPose is a self-attention-based network.

Table 2 lists the experimental results between our proposed VHR-BirdPose and other methods on the Animal Kingdom dataset. The input image size for all methods is 256 × 256. Compared to the original HRNet-W32, VHR-BirdPose-B is improved by more than 0.3 percent under the $PCK@0.05$ evaluation metrics.

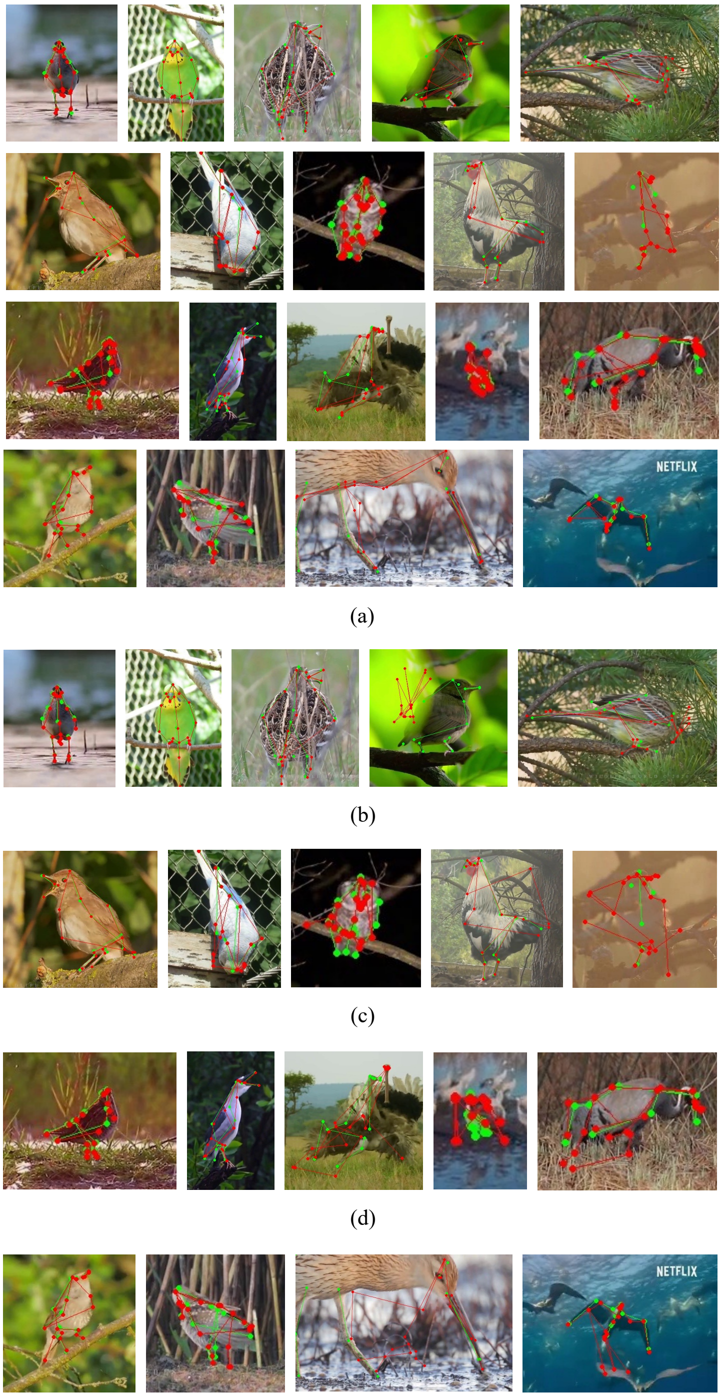
Table 2. Comparison of experimental results on Animal Kingdom dataset.

Method	Backbone	Pre-Train	#Param. [1]	GFLOPs	$PCK@0.05$
HRNet (2019) [14]	HRNet-W32	N	28M	9.49	89.01
CPN (2018, CVPR) [28]	ResNet101	N	27M	9.21	86.48
SimpleBaseline (2018) [29]	ResNet101	N	52M	16.51	88.22
HRFormer (2021) [30]	HRT-B	N	7M	33.44	89.93
ViTPose (2022) [31]	ViT-B	N	153M	125.04	83.13
VHR-BirdPose-S	VHR-BirdPose-S	N	106M	80.57	91.36
VHR-BirdPose-B	VHR-BirdPose-B	N	181M	134.50	89.31
VHR-BirdPose-L	VHR-BirdPose-L	N	181M	134.64	89.29
VHR-BirdPose-XL	VHR-BirdPose-XL	N	333M	242.64	87.68
VHR-BirdPose-S	VHR-BirdPose-S	Y	106M	80.57	90.33
VHR-BirdPose-B	VHR-BirdPose-B	Y	181M	134.50	86.87

[1] The number of parameters.

Our method performance is significantly better than other methods on the Animal Kingdom pose estimation problem. However, in our method, the small network performances are better than every bigger network and the accuracy decreases along with the increase in the model size. This may be caused by the limited dataset size that a large scale of parameters cannot fit data better. Furthermore, in the pre-trained VHR-BirdPose, all methods are worse than the identical method without pre-trained weight. This may be because the pose estimation task on birds is much more complex than on humans, so the pretained weight on the human pose estimation task cannot initialize our method, which for the bird pose estimation task worked well.

Figure 2 illustrates the visualized predicted keypoints and corresponding bird skeletons. These images are randomly sampled from the testing dataset. The predicted keypoints and skeletons are depicted in red, whereas the ground truth is represented by the green color. Each sub-figure indicates a specified method. It is evident that VHR-BirdPose effectively captures information across different scales. Moreover, when dealing with partially occluded objects in the image, VHR-BirdPose exhibits superior accuracy in predicting occluded keypoints, remaining unaffected by noise. These results exemplify the robustness of our method in complex scenarios.



(a)

(b)

(c)

(d)

(e)

Figure 2. Partial output of comparison methods on the test set. (**a**) Represents the output from VHR-BirdPose-B; (**b**) represents the output from HRNet-W32; (**c**) represents the output from CPN; (**d**) represents the output from ViTPose-B; (**e**) represents the output from HRFormer(HRT-B). Our method accurately detects keypoints that are occluded or hidden in complex backgrounds. Additionally, it preserves the background focus and detail, capturing features from ViT and HRNet, respectively.

4.4. Ablation Study

To assess the efficacy of various feature fusion methods between the attention branch and HRNet branch, we perform ablation studies on the Animal Kingdom dataset in this section. In particular, VHR-BirdPose utilizes direct addition to fuse the features from the two branches. Linear-VHR-BirdPose employs a linear layer to fuse the features generated by the two branches, whereas Cross-VHR-BirdPose utilizes a cross-attention block to fuse the features from both branches.

Table 3 presents the results of the ablation study conducted on the original VHR-BirdPose, Linear-VHR-BirdPose, and Cross-VHR-BirdPose models. To evaluate the impact of feature fusion strategies on models of varying sizes, each fusion method is applied to two models: VHR-BirdPose-S and VHR-BirdPose-B. The results indicate that directly adding the features generated by the two branches yields relatively improved performance in bird pose estimation with a lower parameter count. On the small model, the directly adding strategy obtains the best accuracy on the Animal Kingdom pose estimation problem. However, on the bigger model, the accuracy of the linear concat strategy is slightly higher than the directly adding strategy, which may be because the linear concat strategy can let two branches fit quicker than the directly adding strategy in the early stage, but tune slower in the last stage. Furthermore, during the training, we observed that both Linear-VHR-BirdPose and Cross-VHR-BirdPose exhibit smoother accuracy–epoch curves on the test dataset compared to the original VHR-BirdPose throughout the training process, suggesting that these two feature fusion approaches may offer greater stability.

Table 3. The results of the ablation studies.

Method	Feature Fusion Strategy	GFLOPs	#Param. [1]	*PCK*@0.05
VHR-BirdPose-S	Add	80.57	106M	91.36
Linear-VHR-BirdPose-S	Linear Concat	80.58	106M	89.94
Cross-VHR-BirdPose-S	Cross Attention	80.57	106M	88.21
VHR-BirdPose-B	Add	134.50	181M	89.31
Linear-VHR-BirdPose-B	Linear Concat	134.51	182M	89.65
Cross-VHR-BirdPose-B	Cross Attention	134.50	182M	85.13

[1] The number of parameters.

5. Discussion

We employed HRNet and ViT-based attention branches in parallel to extract image features, which are then fused using a feature fusion strategy. The two-branch design with attention mechanism effectively enhances pose estimation performance, with HRNet excelling in joint information extraction at different scales and ViT performing better in object separation and recognition. Feature fusion strategy significantly affects the method's ability and accuracy under the same network architecture, with various strategies, such as feature addition, linear layer, and cross-attention, available based on different network architectures. Attention mechanisms can be integrated into CNN in many ways, such as adding attention blocks following CNN layers and incorporating it into convolution operations. The method leverages the advantages of both HRNet and ViT, using network parallelism to reduce computational complexity. The design of network concatenation with residual connection effectively improves keypoint locating accuracy but requires more training data and time.

The attention blocks in our method learn image features based on parallel attention heads and position embeddings. In common sense, more attention heads may bring better performance, but in the comparison experiment results in Table 2, we found the performance of VHR-BirdPose-B and VHR-BirdPose-L with different attention heads are close. We considered that the greatest challenges from human pose estimation to bird pose estimation are the image quality and the relative size of the object in the images, which both can theoretically benefit from the additional attention heads. Due to the basically identical accuracy of these two methods, we guess that as we already reached the border of additional attention heads, more attention heads will not bring more benefits.Our method uses the position embedding design proposed in ViT by default, and inverting or transposing position embedding according to posture status may further improve pose estimation accuracy and efficiency but requires posture status estimation in advance.

Birds exhibit significant intra and inter-class variations, resulting in a diverse range of classes and features that pose challenges in establishing a universal standard. Despite achieving successful detection of the majority of bird species based on visual results, there remain specific objects that present significant challenges and cause confusion for the model. Recognizing bird species can provide valuable predefined features for refined posture recognition. In future work, we can enhance the balance of the training and testing data by incorporating class labels. Additionally, employing a multi-modal approach that combines text and image features can further augment the capabilities of our models. Leveraging the existing image features from the ViT branch, we can concatenate bird class names, bird descriptions within the images, and joint labels to generate text features using a text encoder like CLIP [33].

The Adam optimizer with a learning rate of 0.001 and reduced by $\frac{1}{10}$ when arriving at specific epochs is used in the training stage. The Adam optimizer is preferred for fast convergence due to the large number of parameters and small amount of images in the dataset. To address the problem of the Adam optimizer not reaching the global optimal solution at the end of training, we can use the SGD optimizer to continue learning with a small learning rate. Adjusting the learning rate in this way can effectively alleviate problems of unfitness and learning curve oscillation at the end of training. However, the learning rate at the end of training may be too small, leading to a local optimal solution. Exponential descent or logarithmic descent methods can be tried to address this issue. However, those methods still require fine-tuning of hyperparameters to avoid slow or fast descent, which remains a topic of concern.

6. Conclusions

Our novel pose estimation method, VHR-BirdPose, combines the strengths of Vision Transformer and HRNet models to effectively estimate bird pose. By incorporating the self-attention mechanism, our method accurately weights and attends to critical keypoints, enabling comprehensive modeling of the diversity and complexity of bird posture. This approach results in a significant improvement in the bird posture estimation task. Our method integrates an attention branch with the HRNet branch to extract high-resolution features and high-quality global features. Compared to traditional CNN-based methods, our proposed method excels in handling the diverse range of bird postures, leading to enhanced accuracy and robustness in pose estimation. During training, we employ various data augmentation techniques, suitable loss functions, and optimizers to improve model generalization and convergence speed. Leveraging the complementary advantages of Vision Transformer and HRNet, our proposed method achieves excellent performance on the Animal Kingdom dataset for bird pose estimation, surpassing HRNet with a $PCK@0.05$ score improvement of over 0.3 percentage points.

However, our method still faces several challenges. Annotating datasets and training models for the diverse and complex nature of bird postures is demanding. Additionally, occlusion, pose changes, and data imbalance can impact model performance. Future research should explore effective data annotation and training strategies to address these challenges.

The proposed VHR-BirdPose model showcases significant performance improvements by combining effective models and attention mechanisms. It provides a fresh and effective solution for bird pose estimation, with promising applications in bird behavior research, ecological monitoring, and other related fields. With ongoing research and further enhancements, we anticipate that the Vision Transformer-based HRNet method will play an increasingly influential role in pose estimation and lead to new breakthroughs in related research and applications.

Author Contributions: Methodology, H.C.; Software, R.H. and H.C.; Resources, X.W. and C.L.; Writing—original draft, R.H.; Writing—review & editing, X.W., H.C. and C.L.; Visualization, R.H.; Supervision, X.W. and C.L.; Project administration, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly funded by the National Natural Science Foundation of China (Grants No. 61931003, 62171044), the Natural Science Foundation of Beijing (Grants No. 4222004), QinXin Talents Cultivation Program (Beijing Information Science and Technology University) (Grants No. QXTCPB202105).

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from SUTD Computer Vision and Learning Group and are available https://github.com/sutdcv/Animal-Kingdom with the permission of SUTD Computer Vision and Learning Group.

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

References

1. Dang, Q.; Yin, J.; Wang, B.; Zheng, W. Deep learning based 2d human pose estimation: A survey. *Tsinghua Sci. Technol.* **2019**, *24*, 663–676. [CrossRef]
2. Perez, M.; Toler-Franklin, C. CNN-Based Action Recognition and Pose Estimation for Classifying Animal Behavior from Videos: A Survey. *arXiv* **2023**, arXiv:2301.06187.
3. Fang, C.; Zhang, T.; Zheng, H.; Huang, J.; Cuan, K. Pose estimation and behavior classification of broiler chickens based on deep neural networks. *Comput. Electron. Agric.* **2021**, *180*, 105863. [CrossRef]
4. Jiang, L.; Lee, C.; Teotia, D.; Ostadabbas, S. Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Comput. Vis. Image Underst.* **2022**, *222*, 103483. [CrossRef]
5. Badger, M.; Wang, Y.; Modh, A.; Perkes, A.; Kolotouros, N.; Pfrommer, B.G.; Schmidt, M.F.; Daniilidis, K. 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–17.
6. Liu, J.; Belhumeur, P.N. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2520–2527.
7. Yang, Q.; Shi, W.; Chen, J.; Tang, Y. Localization of hard joints in human pose estimation based on residual down-sampling and attention mechanism. *Vis. Comput.* **2021**, *38*, 2447–2459. [CrossRef]
8. Pereira, T.D.; Aldarondo, D.E.; Willmore, L.; Kislin, M.; Wang, S.S.H.; Murthy, M.; Shaevitz, J.W. Fast animal pose estimation using deep neural networks. *Nat. Methods* **2019**, *16*, 117–125. [CrossRef] [PubMed]
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
10. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
12. Ng, X.L.; Ong, K.E.; Zheng, Q.; Ni, Y.; Yeo, S.Y.; Liu, J. Animal kingdom: A large and diverse dataset for animal behavior understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19023–19034.
13. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *Acm Comput. Surv. (CSUR)* **2022**, *54*, 1–41. [CrossRef]
14. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2020; pp. 5386–5395.
15. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-hrnet: A lightweight high-resolution network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10440–10450.
16. Wu, H.; Liang, C.; Liu, M.; Wen, Z. Optimized HRNet for image semantic segmentation. *Expert Syst. Appl.* **2021**, *174*, 114532. [CrossRef]
17. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
18. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.

19. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
20. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11802–11812.
21. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11313–11322.
22. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
23. Li, J.; Su, W.; Wang, Z. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11354–11361.
24. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
25. Wang, X.; Tong, J.; Wang, R. Attention refined network for human pose estimation. *Neural Process. Lett.* **2021**, *53*, 2853–2872. [CrossRef]
26. Wightman, R. PyTorch Image Models. 2019. Available online: https://github.com/rwightman/pytorch-image-models (accessed on 19 April 2023).
27. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
28. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
29. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
30. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. Hrformer: High-resolution transformer for dense prediction. *arXiv* **2021**, arXiv:2110.09408.
31. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584.
32. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose+: Vision transformer foundation model for generic body pose estimation. *arXiv* **2022**, arXiv:2212.04246.
33. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.

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.