# Human Pose Estimation using SimCC and Swin Transformer*

Tongrui Li
*Web Technologies and Computer Simulation*
*Belarusian State University*
Minsk, Belarus
tongruili69@gmail.com

Sergey Ablameyko
*Web Technologies and Computer Simulation*
*Belarusian State University*
Minsk, Belarus
ablameyko@bsu.by

***Abstract*—2D Human Pose Estimation is an important task in computer vision. In recent years, methods using deep learning for human pose estimation have been proposed one after another and achieved good results. Among existing models, the built-in attention layer in Transformer enables the model to effectively capture long-range relationships and also reveal the dependencies on which predicted key points depend. SimCC formulates keypoint localization as a classification problem, dividing the horizontal and vertical axes into equal-width numbered bins, and discretizing continuous coordinates into integer bin labels. We propose a new model that combines the Swin Transformer training model to predict the bin where the key points are located, so as to achieve the purpose of predicting key points. This method can achieve better results than other models and can achieve sub-pixel positioning accuracy and low quantization error.**

***Index Terms*—Human Pose Estimation, Swin Transformer, SimCC**

## I. Introduction

Human pose estimation(HPE) is one of the key tasks in computer vision, which aims to identify different human instances in multimedia data and to locate a predefined set of human anatomy key points for each person. It has many important and promising applications, including behavioral action recognition, motion capture, human-computer interaction, and autonomous driving. Currently, 2D human pose estimation faces various challenges such as character entanglement, body size differences, and clothing. Based on this, a lot of work has been devoted to obtaining better feature representation and distinguishing the correct poses. However, these models generally suffer from high computational costs and limited generalization capabilities.

As the recognition effect of the DeepPose method proposed by Toshev et al. [1] is far better than that of traditional methods, many people began to shift the research on human pose estimation from traditional methods to deep learning methods. Current human pose estimation methods generally use deep convolutional neural networks to extract features to replace manual feature extraction.

According to the different representations of key points, detection methods can be divided into coordinate regression-based and heatmap-based detection methods. The method

This paper was funded by the China Scholarship Council.

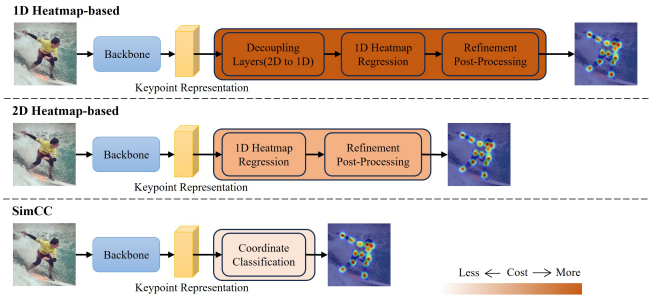


Fig. 1. Comparisons between the proposed SimCC and 2D/1D heatmapbased pipelines.

based on coordinate regression directly predicts the coordinates of key points of the human body in the image. However, the human pose estimation task is a highly nonlinear problem, so this type of method has obvious limitations and poor generalization ability. The method based on heat map detection aims to predict the approximate location of key points, using an improved deep learning network to generate accurate heat maps. The location of key points is represented by a two-dimensional Gaussian distribution centered on the key point location, which can be better represented The key point information of human body parts has better robustness. However, the prediction accuracy depends on the heat map resolution, the calculation amount is large, and the detection speed is slow.

In addition to the above two methods, the paper [2] proposes a new method for human pose estimation called SimCC. This method reconstructs the coordinate prediction problem into two classification tasks, targeting horizontal and vertical coordinates respectively. By evenly dividing each pixel into several intervals (larger than the original width and height), it achieves sub-pixel positioning accuracy and low quantization error. This approach eliminates the need for computationally expensive upsampling layers and additional post-processing, resulting in a simpler and more efficient HPE pipeline.

Recently, Transformer [3] and its various variants originating from natural language processing have become a new choice for various computer vision tasks. It has been widely used in target detection, semantic segmentation, video

understanding, and pose estimation compared with CNN.

Transformer has a larger receptive field, more flexible weight setting method and global modeling ability of features, and has the potential to provide higher quality feature input for downstream tasks. The traditional Transformer structure only generates output feature maps within a single scale, which cannot be directly used for human posture estimation tasks, and has high computational complexity and large memory consumption. And experiments have proven that using multi-head self-attention can improve performance [3].

In order to take advantage of Transformer's remote dependency capture capabilities and avoid excessive memory consumption, we choose Swin Transformer [4] as our backbone. Next, we fused Swin Transformer with SimCC and proposed a novel human pose estimation model. A convolutional layer, a fully connected layer and a , Gated Attention Unit(GAU) [17] are added between the backbone of Swin Transformer and SimCC, achieving better results compared with other Swin Transformer-based models.

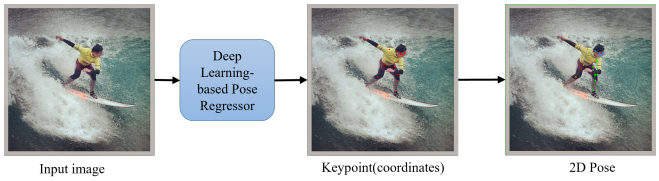


Fig. 2. Regression Methods.

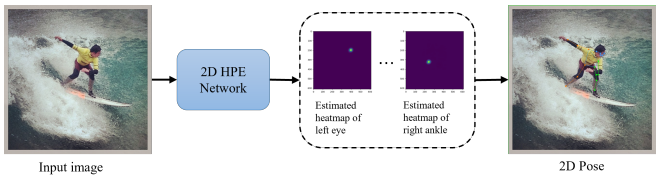


Fig. 3. Heatmap-based Methods.

## II. Related Work

### A. Human Pose Estimation

HPE methods based on deep learning have achieved many excellent performances. An efficient network structure not only has a small number of parameters and fast convergence speed, but also is easy to predict the location of key points. Therefore, many scholars have optimized and improved the network structure of deep convolutional neural networks applied to human posture estimation. Wei et al. [5] proposed a Convolutional Pose Machines (CPM) network, which uses a convolutional neural network to learn image texture information and spatial information. Prior to this, many scholars used convolutional neural networks to extract the texture of images. Information, using graphical models or other models to express the spatial relationship between various parts of the body, does not use both types of information at the same time. Wei et al. [5] use a convolutional neural network to learn these two features at the same time, making the learning effect better and having Helps end-to-end learning. After continuous refinement, a more accurate prediction value of the key point heat map will eventually be obtained. With the proposal of the residual network, Newell et al. [6] designed a stacked hourglass network. This network It is also a multi-stage structure, consisting of multiple stacked hourglass structures. Each hourglass structure contains the process from high resolution to low resolution and from low resolution to high resolution to estimate key points of human posture at different scales heat map information. On this basis, Yang et al. [7] added the pyramid residual module to enhance the robustness of the deep convolutional neural network to scale changes. In addition, Chu et al. [8] improved the residual unit to make branch filtering The device has a larger receptive field of view, and uses the improved residual unit structure to learn multi-scale features, further improving the accuracy of key point heat map prediction. Wang et al. [9] proposed a data enhancement method for learning random mixed images, which improves the robustness of key point detection in pose estimation under various damaged data (such as blur and pixelation).

### B. Scheme

2D HPE methods based on deep learning have achieved many excellent performances. Carreira et al. [10] proposed a general coordinate regression framework, using GoogleNet as the backbone network to jointly learn output features and input features, and model input features and output features at the same time. In order to make full use of the structural information inside the human body posture, Shuang et al. [11] proposed a structured perception regression method. This method uses re-parameterized bones instead of key points to express the human body posture. The bones have the intuition and stability of the human body. It can better express the human posture structure. The method proposed by Mao et al. [12] with the help of the attention mechanism in the converter can adaptively focus on the features most relevant to the target key points, which to a large extent solves the problems of previous regression-based methods feature misalignment problem and significantly improves performance. Lifshitz et al. [13] jointly generate the final pose estimation result through key point detectors and inference key point relationships. During key point detection, this method uses dilated convolution and deconvolution layers to improve the resolution of the feature map output by the model. This can effectively expand the convolution receptive field and improve the accuracy of key point heat map detection without increasing the number of model parameters.

### C. Transformer

Recently, Transformer and its variants have been used by researchers for human body pose estimation. For example, TransPose [14] uses the attention layer of Transformer to implicitly reveal the dependencies between key points as a model layer. Layer reasoning provides explanations for global spatial relationships. TokenPose [15] is inspired by the ViT

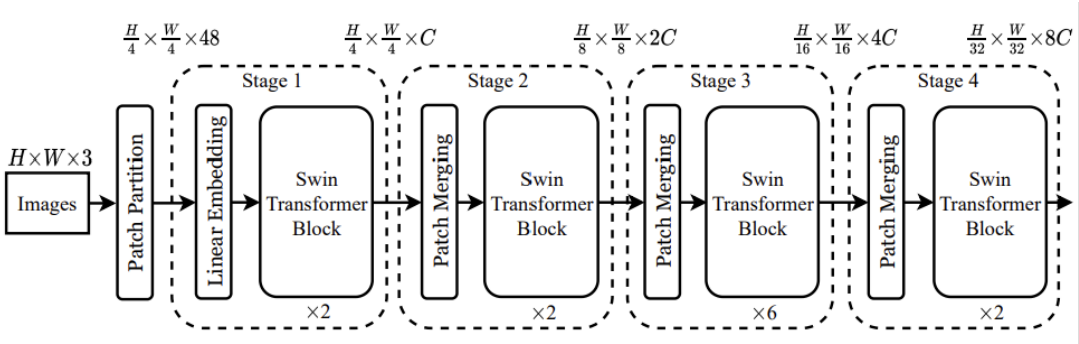


Fig. 4. The architecture of a Swin Transformer (Swin-T).

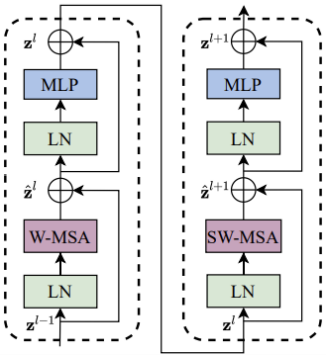


Fig. 5. Two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

(Vision Transformer) model, explicitly modeling key points into markers, and learning the constraint relationship between visual information and key points from the image. Both methods require a large number of Transformer encoders but do not consider low-resolution global semantic features. HRFormer (High Resolution Transformer) [16] uses multi-resolution architecture design and local window self-attention to achieve high-resolution feature representation, which has the characteristics of low memory and low computational cost. This method requires upsampling of low-resolution features, resulting in the loss of spatial semantic information.

The structure of Swin Transformer is similar to ResNet and consists of four stages. By limiting self-attention to non-overlapping local windows, Swin Transformer significantly reduces computational cost [4], making it suitable for downstream tasks. The shift window partitioning method is also applied to achieve information communication between those non-overlapping windows.

## III. Methodology

In this section, we elaborate on the entire model structure. The model mainly consists of two parts, one is the head composed of SimCC, and the other is the backbone composed of Swin Transformer. SimCC [2] provides a lightweight yet powerful baseline. On this basis, we adopt Gated Attention Unit (GAU) [17] to improve the feedforward network (FFN) of our Swin Transformer. Then we used pre-training and Adam optimization strategies to further improve model performance. The final model architecture is shown in Fig. 6.

### A. SimCC

SimCC treats the key point positioning task as two classification subtasks of the horizontal axis and the vertical axis, and represents the x and y coordinates of the 17 joint positions of human pose estimation as two independent one-dimensional vectors. Divides the horizontal and vertical axes into equal-width numbered bins, and discretizes continuous coordinates into integer bin labels. The model is then trained to predict the bins where the key points are located. The SimCC structure is very simple, using only a 1 × 1 convolutional layer to convert the features extracted from the backbone into vectorized keypoint representations, and using two fully connected layers respectively to perform classification. Through a large number of bins, the quantization error can be reduced to the sub-pixel level, thereby achieving sub-pixel positioning accuracy.

### B. Module

*a) Pretraining:* Pretraining the backbone using a heatmap-based approach can improve model accuracy, so our model uses the published pretrained weights of the original Swin Transformer model pretrained on ImageNet22K.

*b) Optimization Strategy:* We choose Adam Optimizer as the optimizer. Adam Optimizer has the advantages that the size of parameter update does not change with the scaling of the gradient size, the boundary of the step size when updating parameters is limited by the setting of the step size of the hyper parameter, and does not require a fixed objective function.

*c) Self-attention module:* We adopt a variant of transformer, Gated Attention Unit (GAU) [17], which has faster speed, lower memory cost and better performance than ordinary transformer. Specifically, GAU uses Gated Linear Unit (GLU)to improve the feed forward network (FFN) in the transformer layer. The attention mechanism form:

$$\begin{aligned} U &= \phi_u\left(XW_u\right) \\ V &= \phi_v\left(XW_v\right) \\ O &= (U \odot AV)W_o \end{aligned} \tag{1}$$

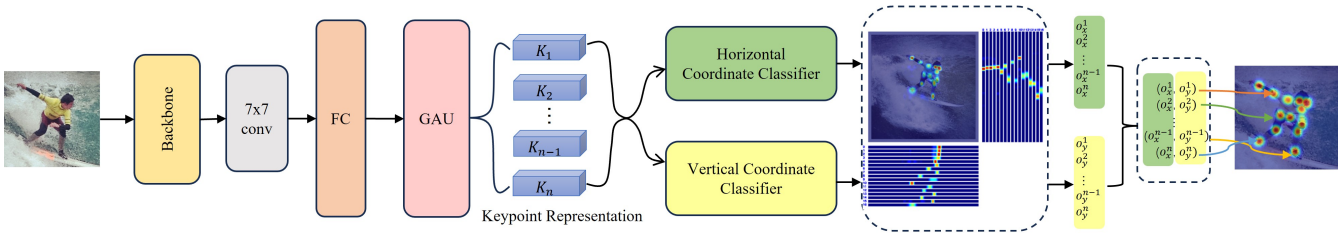


Fig. 6. SimCC pipeline.

where $\odot$ is the pairwise multiplication (Hadamard product) and $\phi$ is the activation function. In this work we implement the self-attention as follows:

$$A = \frac{1}{n} relu^2\left(\frac{Q(X)K(Z)^T}{\sqrt{s}}\right), Z = \phi_z(XW_z) \tag{2}$$

where s = 96, Q and K are simple linear transformations, and $relu^2$ $(\cdot)$ is ReLU then squared.

*d) Backbone:* The Swin Transformer network extracts the internal structure information of the image block through the internal Block. The self-attention in the Block obtains the hyper spectral image by calculating the score of the feature map matrix information representing one band and the feature map matrix information representing other bands the relationship between each band. The Swin Transformer network model contains 4 stages. Each stage consists of a Patch merging and several Swin Transformer Blocks. Each stage will reduce the resolution of the input feature map and expand the receptive field layer by layer like CNN. Among them, Patch merging The module performs down sampling before the start of Stage to reduce the image resolution.

Swin Transformer Block consists of multi-layer perceptron (MLP), layer normalization (layerNorm), window multi-head self-attention layer (W-MSA) and sliding window multi-head self-attention layer(SW-MSA), shown in Fig. 5. MLP consists of input layer, hidden layer and output layer, used for tensor reshaping. LayerNorm is used to normalize the data, that is, calculate the mean and variance on each sample. W-MSA is used for tensor reshaping calculate attention under one window. In order to better interact with other windows, SW-MSA is introduced in Swin Transformer. Both use global context information to encode each band to capture the interaction between each band of the hyper spectral image relationship. The number of Blocks contained in each layer of Swin Transformer is an integer multiple of 2, one layer is provided to W-MSA, and one layer is provided to SW-MSA. The increase in the number and value of Blocks in Swin Transformer will improve the classification accuracy to a certain extent. However, considering the model size and computational complexity as well as the experimental hardware, the number of Blocks used in each layer of the Swin Transformer network structure in this article are 2, 2, 6, and 2 respectively.

## IV. Experiments

### A. Dataset

he COCO dataset [18] contains more than 200,000 images and 250,000 human body instances with 17 key points. The model was trained only on the COCO train2017 data set without additional training data, and was tested on the val2017 data set and test2017 data set. These three sub-datasets contain 57,000, 150,000 and 5,000 images respectively.

### B. Evaluation metric

Object Keypoint Similarity (OKS) is to calculate the similarity between the predicted human body key points and the real human body key points. Its calculation equation is as shown in:

$$AP = \frac{\sum_p \delta(OKS > s)}{\sum_p 1} \tag{3}$$

Average precision (AP) is an indicator that measures the accuracy of key points, and is calculated as shown in (4). mAP (MeanAveragePrecision) is to calculate the mean value of AP of all key points.

$$OKS = \frac{\sum_i exp(-\frac{d_i^2}{2s^2k_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{4}$$

where $d_i^2$ is the Euclidean distance between the $i$-th predicted keypoint coordinate and the corresponding groundtruth, $v_i$ is the visibility flag of the keypoint, s is the object scale, and $k_i$ is a keypoint-specific constant.

### C. Settings

The experiment was completed on the Google colab platform. The PyTorch 2.0.1 deep learning framework was built in the Ubuntu 20.04 system. The Python language version was 3.10.12 and the GPU was NVIDIA A100-SXM4-40GB. We set the batch size to 32, the number of epochs to 200, the sliding window size to 7×7, and the initial learning rate experiment to 5e-5.

### D. Results

We compare our results with other human pose estimation models, including Residual Steps Network (RSN), Swin-T, and Residual Network (ResNet), and show the results in Table 1. As shown in the table, when the input image size

TABLE I
COMPARISON ON THE COCO VALIDATION SET*

| Method | Backbone | Scheme | Input Size | AP | $AP^{50}$ | $AP^{75}$ | AR | $AR^{50}$ |
|---|---|---|---|---|---|---|---|---|
| ResNet 50 | ResNet 50 | Heatmap | 256x192 | 0.715 | 0.897 | 0.791 | 0.771 | 0.935 |
| ResNet 50 | ResNet 50 | Heatmap | 384x288 | 0.724 | 0.899 | 0.794 | 0.777 | 0.936 |
| RSN-18 | RSN-18 | Heatmap | 256x192 | 0.704 | 0.887 | 0.781 | 0.773 | 0.927 |
| RSN-50 | RSN-50 | Heatmap | 256x192 | 0.724 | 0.894 | 0.799 | 0.790 | 0.935 |
| ResNet 50 | ResNet 50 | SimCC | 256x192 | 0.721 | 0.897 | 0.798 | 0.781 | 0.937 |
| ResNet 50 | ResNet 50 | SimCC | 384x288 | 0.735 | 0.899 | 0.800 | 0.790 | 0.939 |
| Swin-T | Swin-T | Heatmap | 256x192 | 0.724 | 0.901 | 0.806 | 0.782 | 0.940 |
| **Ours** | Swin-T | SimCC | 256x192 | 0.733 | 0.908 | 0.807 | 0.790 | 0.942 |

*Results on COCO val2017 with detector having human AP of 56.4 on COCO val2017 dataset.

is consistent, our model achieves better results in both AP and AR. Compared with the original Swin-T model, AP has increased by 0.009 and AR has increased by 0.008.

## V. CONCLUSION

In this paper, we propose a new method for 2D human pose estimation that integrates Swin transformer and SimCC. This model takes advantage of SimCC's advantage over heatmap-based representation in terms of model performance, combined with Gated Attention Unit, and is modified from the original Swin transformer model. Experimental results show that this model is better than the original swin transformer model. Current lightweight work on human pose estimation models can significantly reduce the model's inference cost and increase the inference speed. The next step will be to study the application of knowledge distillation in human posture estimation to strike a balance between computational cost and high performance to adapt to the requirements of limited computing resources.

## ACKNOWLEDGMENT

The support provided by China Scholarship Council (CSC) during 1 y study of Tongrui Li to Belarusian State University is acknowledged.

## REFERENCES

[1] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1653-1660, doi: 10.1109/CVPR.2014.214.

[2] Y. Li et al., "SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation," in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 89–106.

[3] A. Vaswani et al., "Attention is all you need," arXiv e-prints, p. arXiv:1706.03762, Jun. 2017, doi: https://doi.org/10.48550/arXiv.1706.03762.

[4] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[5] S. -E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional Pose Machines," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4724-4732, doi: 10.1109/CVPR.2016.511.

[6] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," arXiv e-prints, p. arXiv:1603.06937, Mar. 2016, doi: https://doi.org/10.48550/arXiv.1603.06937.

[7] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," arXiv e-prints, p. arXiv:1708.01101, Aug. 2017, doi: https://doi.org/10.48550/arXiv.1708.01101.

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille and X. Wang, "Multi-context Attention for Human Pose Estimation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5669-5678, doi: 10.1109/CVPR.2017.601.

[9] J. Wang, S. Jin, W. Liu, W. Liu, C. Qian and P. Luo, "When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 11850-11859, doi: 10.1109/CVPR46437.2021.01168.

[10] J. Carreira, P. Agrawal, K. Fragkiadaki and J. Malik, "Human Pose Estimation with Iterative Error Feedback," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4733-4742, doi: 10.1109/CVPR.2016.512.

[11] X. Sun, J. Shang, S. Liang and Y. Wei, "Compositional Human Pose Regression," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2621-2630, doi: 10.1109/ICCV.2017.284.

[12] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "TFPose: Direct human pose estimation with transformers," arXiv e-prints, p. arXiv:2103.15320, Mar. 2021, doi: https://doi.org/10.48550/arXiv.2103.15320.

[13] I. Lifshitz, E. Fetaya, and S. Ullman, "Human Pose Estimation Using Deep Consensus Voting," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 246–260.

[14] S. Yang, Z. Quan, M. Nie and W. Yang, "TransPose: Keypoint Localization via Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 11782-11792, doi: 10.1109/ICCV48922.2021.01159.

[15] Y. Li et al., "TokenPose: Learning Keypoint Tokens for Human Pose Estimation," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 11293-11302, doi: 10.1109/ICCV48922.2021.01112.

[16] Y. Yuan et al., "HRFormer: High-resolution transformer for dense prediction," arXiv e-prints, p. arXiv:2110.09408, Oct. 2021, doi: https://doi.org/10.48550/arXiv.2110.09408.doi:10.48550/arXiv.2110.09408.

[17] L. Xue, X. Li, and N. L. Zhang, "Not All Attention Is Needed: Gated Attention Network for Sequence Data," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 6550–6557, Apr. 2020, doi: https://doi.org/10.1609/aaai.v34i04.6129.

[18] T. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755.