Honorable professors, good morning. I am Mingchen Wang, a student of Master of Science in Intelligent Technology enrolled in September 2023. My supervisor is Prof. Liu Xin. The topic of today's presentation is Research on the structure of offline handwritten signature verification models based on transformer, which is a cooperative project with Zhuhai Kingsware Information Technology Co. Ltd. during the internship in 2023, as a way to carry out the subsequent research. The next section is divided into five parts.

First of all, offline handwritten signature is a specific application in biometric technology. Biometric technology is a technology for personal identification or verification based on individual physiological features such as fingerprints, face, iris, etc. or behavioral features such as voiceprints and handwriting. The technology is widely used in enterprise security authentication, financial transactions, access control systems and other security fields. For example, the unlocking of cell phones and bank card transfers in daily life all involve this kind of authentication technology. This research topic is mainly based on the handwritten signature that contains the behavioral characteristics of personal handwriting in order to carry out research. Before understanding the handwritten signature, it is necessary to understand the difference between identification or verification. Biometric technology is mainly used for identification and verification, identification is the user to provide personal physiological or behavioral characteristics to the system, the system will be based on these characteristics to determine which user is registered in the system. That is, to solve the problem of “who are you”. Verification is based on the identification of the need for users to declare their identity, the system will be based on these characteristics to determine whether the user is true, that is, to solve the problem of “you are not xxx”. Therefore, handwritten signature verification is to judge whether the information provided is true or not by the handwriting of the individual writing a handwritten signature. Inside the handwritten signature verification task, it will be divided into offline and online according to the signature device. Online means that the user writes his/her personal signature using digital devices, such as digital plates used at bank reception desks, tablets and electronic pens in daily life learning. Offline refers to users writing only through pen and paper. The former in the collection of data need to strictly require the model and batch of digitizing equipment, and the user's pen posture, the angle of contact with the writing pad, the pressure has a great deal to do with different personal writing habits and equipment will lead to the collection of handwritten signature samples need to spend extra time for data pre-processing. The latter process of collecting data is only related to the user's pen and is not affected by other influences, and the process of holding a pen and writing personal signatures on paper is non-invasive, i.e., there is no problem of inconsistent and intermittent writing handwriting due to the network or device in the middle of the process. Therefore scholars are more likely to study offline handwritten signature verification. In the previous offline handwritten signature verification models or algorithms, most of the scholars take the traditional machine learning approach in order to accomplish the task. However, the traditional machine learning approach has a drawback that it relies heavily on manually designed features, i.e., the process of how images are converted into feature vectors. This process requires a lot of experience in image processing to be able to extract better quality handwritten signature verification. And the offline handwritten signature verification task is essentially a multi-task. It consists of two sub-tasks: Writer-Dependent, defined as WD, and Writer-Independent, defined as WI. In the WD task, it is necessary to compare the provided reference signature, defined as reference signature, with the provided signature, defined as Query signature, to determine whether the provided signature is genuine or fake. signature is genuine or fake. On the other hand, in the WI task, there is no need to provide the reference signature, and only the provided query signature is used to identify whether the signature is a forgery or not. Therefore, the traditional machine learning approach based on the above offline handwritten signatures would be costly in terms of manually designing features and multi-tasking.

The second part will be carried out next, describing the research objectives and the significance of the research. Based on the above research background, the study of offline handwritten signature verification model can learn to some extent how deep learning can innovate and improve with this direction. The Offline Signature Verification Transformer, abbreviated OSVTF, shown in the upper right corner of the page is the Transformer-based model structure proposed in this paper. In terms of research, it is divided into three stages: Stage 1, reproduces the relatively novel deep learning offline handwritten signature verification models in the current academic world, analyzes and summarizes the advantages and disadvantages of them, so as to summarize a set of design and improvement schemes applicable to OSVTF; Stage 2, according to the above reproduced conclusions, so as to propose a Transformer-based offline handwritten verification model, which can combine with convolutional neural network localization, and is able to combine with the convolutional neural network localization. model is capable of combining the OSV model structure with the local feature learning capability of convolutional neural network and the global feature learning capability of Transformer; Stage 3, for the multi-subtasks of WI and WD, the existing traditional machine learning methods are improved so as to be able to identify whether the handwritten signatures are forgeries or not in a better quality and efficient way. In summary, there are four main implications of studying this topic: first, proposing a higher accuracy OSV model can, to a certain extent, make the advantages of deep learning methods clearer, so as to better carry out the research of subsequent updated models or algorithms. Second, adopting deep learning methods can save most of the costs of manually designing features, and can have more costs to research high-quality and efficient model structures or algorithms. Third, less inference cost, when the number of model parameters is not much different, can be GPU hardware computing, thus saving more resources out of parallel computing other tasks, in order to achieve more efficient completion of various tasks. Fourth, the image feature research based on handwritten signature images, to a certain extent, can affect the OCR related research work, for the OCR field to bring a certain degree of methodological innovation.

The third part will be a timeline related literature review for offline handwritten signatures and deep learning aspects of Transformer. Firstly offline handwritten signature verification has been researched in the 1980s, a time period when feature-based template matching methods were used. As early as 1989 Plamondon et al. proposed a geometric feature based method for handwritten signature verification. In their paper they investigated verification methods based on geometric features and Euclidean distances. These early methods focused on basic features such as the overall shape of the signature, mural thickness and slant.In 1991 Kashi et al. proposed a shape-matching based verification method for handwritten signatures in their paper. The method focuses on signature matching by comparing the geometric features of the signatures. Immediately following in the early 1990s to the beginning of the 21st century were structural features and texture analysis methods. in 1997 Justino et al. introduced Hidden Markov Models for signature verification in their paper, which utilized the statistical and structural features of signatures for matching. In the following year, Sabourin et al. proposed a signature verification method based on wavelet transform, which improves the verification system's ability to perceive the signature details by decomposing the frequency features of the signature image.In the 21st century until the beginning of 2010, scholars are adopting the statistical model and machine methods for offline handwritten signature verification research. Among them, Coetzer, Vielhauer, Ferrer, etc. are all based on Hidden Markov, Support Vector Machines and other statistical machine learning models for offline handwritten signature verification, in which their methods invariably use a variety of data preprocessing methods to improve the verification accuracy, such as Discrete Radon Transform, Discrete Wavelet Transform and feature fusion. Until the introduction of deep learning methods in 2012, Hafemann et al. for the first time introduced CNN to offline signature verification tasks, which can automatically learn high-level features of signatures and no longer rely on manual feature extraction. As a result, scholars proposed a contrast loss-based twin network, which optimizes the loss function to enable the network to more accurately distinguish between genuine and forged signatures.In 2020, Soleimani et al. evaluated a variety of deep learning models, including CNNs, Siamese Networks, and GANs, and also proposed a model that combines a GAN and a twin network. for generating forged signature samples to augment the training dataset. In the latest phase i.e. 2020 to the future, there is a gradual move towards self-supervised learning with multimodal fusion. In 2021 Kim et al. proposed a self-supervised learning based approach using SimCLR model for pre-training to reduce the dependence on large amount of labeled data.In 2022 Siddiqi et al. even proposed a multimodal signature verification approach combining offline and online signature data to improve the robustness and adaptability of the model. On the other hand deep learning is also developing in parallel, Alex et al. in 2012 proposed AlexNet convolutional neural network in ImageNet competition, its high quality performance makes the academic community set off a convolutional neural network deep learning frenzy, more and more deep convolutional neural networks are proposed, such as He et al. in 2016, the residual structure CNN, which is capable of retaining a certain preserve some of the features of the image to a certain extent. Not only the image field, but also the NLP field is developing rapidly.In 2017, Vaswani et al. proposed the Transformer model for machine translation, which is an Encoder-Decoder structure based on the mechanism of multi-head attention. The structure does not take any convolution, pooling operations and simply goes through various linear layers and linear computations in order to achieve the machine learning translation task. Compared with the convolutional operation of CNN, its multi-attention mechanism can better focus on the global characteristics of feature vectors, which can achieve high quality accuracy and robustness in the translation task with different lengths and phrases. Thus, in the field of CV, Dosovitskiy et al. proposed Vision Transformer, or ViT for short, in 2021. The principle of ViT is to patch and flatten the image, treat each patch as a single character vector of word vectors, and the whole image as word vectors to enter the Transformer Encoder. The whole image is used as a word vector to enter the Transformer Encoder, and then an MLP is subsequently added to accomplish the image classification task of ImageNet. The experiments of ViT have proved that the model with better focus on global contextual features can achieve better results in the case of training convergence, while the CNN can achieve better results in the case of training for a certain period of time. However, the Transformer-based ViT model is not as fast as CNN in the inference operation time, the subsequent development of deep learning with the hardware iteration upgrade Transformer's inefficient inference cost is also ignored, and more Transformer-based visual deep learning models have been put forward one after another, such as DETR and Deformable for target detection. DETR, the former is similar to ViT, but DETR adopts a complete Transformer Encoder-Decoder structure, which performs Transformer inference after the image has been extracted by CNN to extract the multi-channel feature maps, and finally goes through the MLP to reason about the relative positional coordinates of the objects in the image and their categories. Deformable DETR adds multi-scale features on the basis of DETR, and adopts FPN-Style to input all the feature maps of different sizes obtained from multiple convolutional blocks in the inference process of CNN as backbone into the Transformer Encoder. But getting multiple sizes of feature maps will lead to an increase in the computational cost inside the Transformer Encoder, so the authors proposed the variable attention mechanism in the paper. On the basis of the multi-attention mechanism, K-nearest neighbor sampling is taken to optimize the attention computation of the whole feature vector in the multi-attention computation of Transformer Encoder, and the attention computation is carried out for some of the points of special attention, and in this way the attention computation is carried out for some of the more concerned features. The performance of DETR and other non-end-to-end target detection models is achieved by utilizing shorter time in the thesis experiments. Therefore Deformable DETR's multi-scale end-to-end target detection structure will be one of the important references.

The fourth section will introduce the basic framework of the OSVTF model in the research topic and detail the individual components and dataset sources. The first is the overall architecture of OSVTF, and the training inference process is shown here. Firstly, a pair of handwritten signature images are input, and a pair of flat tokens and features are obtained after backbone and encoder with shared weights. Where the pair of tokens will be subjected to FC loss computation and the features will be rearranged and enter the convolution module. After the convolution module, two branches of inference will be carried out: one of them will carry out global average pooling to get the convolutional flat features for FC loss computation; the other branch will directly enter into the decoder computation to get the flat decode features, which will be used for FC loss computation. That is, the whole model architecture during the training process will collect flat encode, convolutional, and decode features for FC loss computation.The backbone part of OSVTF is the backbone network part of CNN, which uses the convolutional part for extracting the multichannel feature maps, and here we take the ResNet-50 as an example.After the backbone inference, there is then the Encoder, which is the backbone of TransOSV model architecture. In the TransOSV model structure, the Encoder architecture of ViT is adopted directly, and the multichannel feature maps are patched and then flattened to obtain 2D feature vectors with the same shape as the word vectors, which are used for the subsequent computation of multi-head attention. Scaled Dot Product Attention is taken in Multihead Attention to compute the attention weights of individual heads. A random parameter with the same shape as the word vector is generated in the ViT Encoder and accumulated in the feature vector for attention weight calculation. The purpose of this step is to be able a learnable parameter to speed up model convergence. In subsequent experiments multiple sizes of feature maps of backbone are stitched together to perform multi-scale multi-head attention computation. In the decoder, TransOSV's Contrast based Part Decoder is directly followed, and the cross attention decoder is different from the traditional Transformer decoder, which adds a new learnable parameter on top of the input features in the same way as ViT. This decoder adds a new learnable parameter to the input features as the ViT, in order to pay more attention to the contrast between reference and query.The decoder first undergoes a multi-attention process, which assigns attention weights to the input feature vectors, in order to strengthen the obvious features, and weaken the weaker ones. Subsequently, a cross-attention computation is performed based on the above added science family parameters, as a way to cross-compare the features of the REFERENCE and QUERY to get the final DECODER features. At the loss function for model training, two loss functions are taken: sparsity loss and focal contrast loss. sparsity loss is proposed only for the linear mapping in the decoder.

Sparsity loss is proposed only for the linear mapping in the decoder, which can accelerate the convergence speed of the decoder. The focal contrast loss is a training loss function for one-handed handwritten signature features based on the contrastive loss for evaluating the difference between two objects and the improved double marginal loss in CaP. Where D() is denoted to calculate the distance between the samples, here the Euclidean distance is taken. alpha 1 and alpha2 denote the two margin values, overline K and overline V denote the scaling factors, and the above are the hyperparameters of the training process. The datasets are BHSig-B, BHSig-H and CEDAR, which are publicly available datasets for offline handwritten signature verification, where B and H of BHSig denote the writing languages Bengali and Hindi, and the BHSig series of datasets are the datasets of handwritten signatures of 100 Bengali and 160 Hindi users released by the Indian Institute of Technology, Guwahati, India. Both datasets each user provided 24 authentic signatures and 30 forged signatures.CEDAR dataset is developed and published by Center of Excellence for Document Analysis and Recognition. The CEDAR dataset was developed and published by the Center for Excellence for Document Analysis and Recognition. Each user signed 24 authentic signatures and 24 forged signatures.

Finally, the fifth part will show the existing and expected results, as well as some problems encountered. The first part is the reproduction of some OSV models, trying to reproduce the TransOSV based on Transformer, due to the difference of training equipment, the original author took 8 NVIDIA 3090 graphics cards to train in the case of batch size 64, the reproduction process is to take 1 4070S for reproduction and training, a certain number of epoch training. The performance difference between BHSig-B and BHSig-H datasets is around 20% to 30%. The SGD optimizer with the same learning rate and decay rate as provided in the original article is adopted, and adjusting the optimizer scheme should be able to narrow the gap around 10%. Subsequently, we will focus on optimizing the model architecture for the multi-scale and encoder parts. Secondly, this research topic is originated from the related cooperation project between the company where the internship is conducted and the undergraduate college, which first submitted the application for the project of R&D project for Guangdong enterprise science and technology specialists in late 2021, and then subsequently joined the project team and worked on the Chinese handwritten signature data collection and model deployment in 2022. After the success of the subsequent application in 2023, it is estimated that the cumulative annual new output value will be about 3 million yuan, and will have a broader economic prospect by combining with the company's other AI products to form an automated audit solution based on business scenarios. Finally, in mid-March 2023, we submitted ICNCC-2023 and included the related article in ACM: Offline Signature Verification Using a 2D Attention Encoder-Decoder Network. there are about three issues in the research process: the first one is for image multiscale. The first issue is the image multi-scale problem. For some handwritten signature images with width and height less than 224 pixels in ImageNet image classification task, whether the multi-scale method similar to Deformable DETR can get more effective model quality improvement, and the current ViT-based encoder must specify the image input with uniform size, so multi-scale is the key to optimize the model structure. image features is the main direction for subsequent model structure optimization. The second issue is that the encoder adopts flatten features for multi-attention computation, whether we can find a solution to compensate for the loss of some features in the decoding process after the image features are flattened. The third issue is the training environment, it is difficult to reach the deep learning model training environment with a 16G NVIDIA 4070S graphics card alone, and we need to rent a server for model training to get better model performance. Finally, for the expected results, we hope to get a better model performance than TransOSV after adopting the CNN+Transformer approach, and have a more stable generalization, which is expected to work on Chinese handwritten signatures. For the optimized model architecture, in the feature extraction part, we hope to give some ideas to the image research related to the OCR field, so as to broaden the idea of designing the OCR model architecture. At the same time, I also hope to realize the deployment of OSVTF in the cooperation project, and combine it with other AI products to realize the automatic audit solution corresponding to business scenarios in a more stable way.

appreciate all the professors for listening.