1st Place Solution to the 1st SkatingVerse Challenge

Tao Sun*, Yuanzi Fu*, Kaicheng Yang*, Jian Wu*, Ziyong Feng
DeepGlint
{taosun,yuanzifu,kaichengyang,jianwu,ziyongfeng}@deepglint.com

***Abstract*— This paper presents the winning solution for the 1st SkatingVerse Challenge. We propose a method that involves several steps. To begin, we leverage the DINO framework to extract the Region of Interest (ROI) and perform precise cropping of the raw video footage. Subsequently, we employ three distinct models, namely Unmasked Teacher, UniformerV2, and InfoGCN, to capture different aspects of the data. By ensembling the prediction results based on logits, our solution attains an impressive leaderboard score of 95.73%.**

I. INTRODUCTION

The 1st SkatingVerse Challenge, held as part of the SkatingVerse Workshop, is affiliated with the prestigious 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG). In contrast to previous tasks that may lack practical applicability, such as fine-grained action segmentation and assessment, this challenge focuses on the construction of a comprehensive dataset comprising 1,687 continuous videos from figure skating competitions. The primary objective is to foster the development of algorithms capable of accurately analyzing each action depicted in these videos. The challenge dataset consists of 19,993 video clips for training and 8,586 video clips for testing. It encompasses a wide range of 28 distinct categories of figure skating actions, and Figure 1 illustrates six representative examples of these actions.

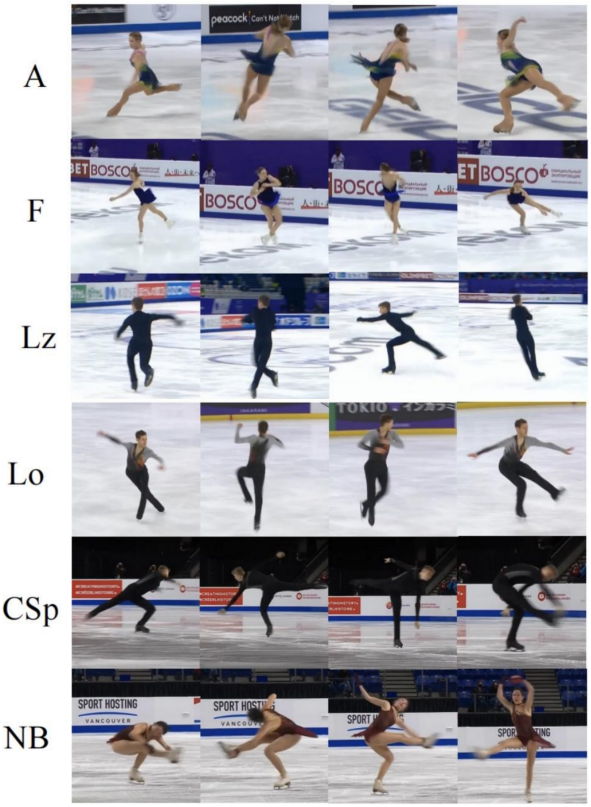

Fig. 1: Examples of figure skating actions: A (Axel); Lo (Loop); F (Flip); CSp (Camel Spin); Lz (Lutz); NB (No Basic)†.

II. METHOD

A. Dataset Pre-processing

In order to enhance the model's attention toward figure skating actions, we performed human body detection on the original video frames. To accomplish this, we utilized FFmpeg to extract video frames and employed the DINO framework [11] for extracting bounding boxes corresponding to human detections in each frame. These individual bounding box results from all frames within a video were then consolidated to generate the ultimate detection box for that specific video. Finally, using FFmpeg, we crop the raw video clips based on the combined bounding box information obtained from the human detection process.

B. Model Structure

The architecture of our proposed method is illustrated in Figure 2. To begin, we employ the DINO model [11] to extract precise human detection bounding boxes, facilitating the generation of cropped frames. Subsequently, we conduct fine-tuning on two widely-used general video pre-training models, namely Unmasked Teacher [5] and UniformerV2 [4]. Furthermore, we leverage ViTPose [10] to extract human skeleton sequences, which are then utilized to train the InfoGCN model [1] for accurate action predictions.

1) Unmasked Teacher: Unmasked Teacher(UMT) [5] is a two-stage training-efficient pretraining framework that enhances temporal-sensitive video foundation models by incorporating the advantages of previous approaches. In Stage 1, the UMT exclusively utilizes video data for masked video modeling, resulting in a model that excels at video-only tasks. In Stage 2, the UMT leverages public vision-language data for multi-modality learning, enabling the model to handle complex video-language tasks such as video-text retrieval and video question answering.

In this paper, we initially fine-tune the pre-trained UMT-

*These authors contributed equally in this work.
†Examples are from the challenge website:https://skatingverse.github.io/

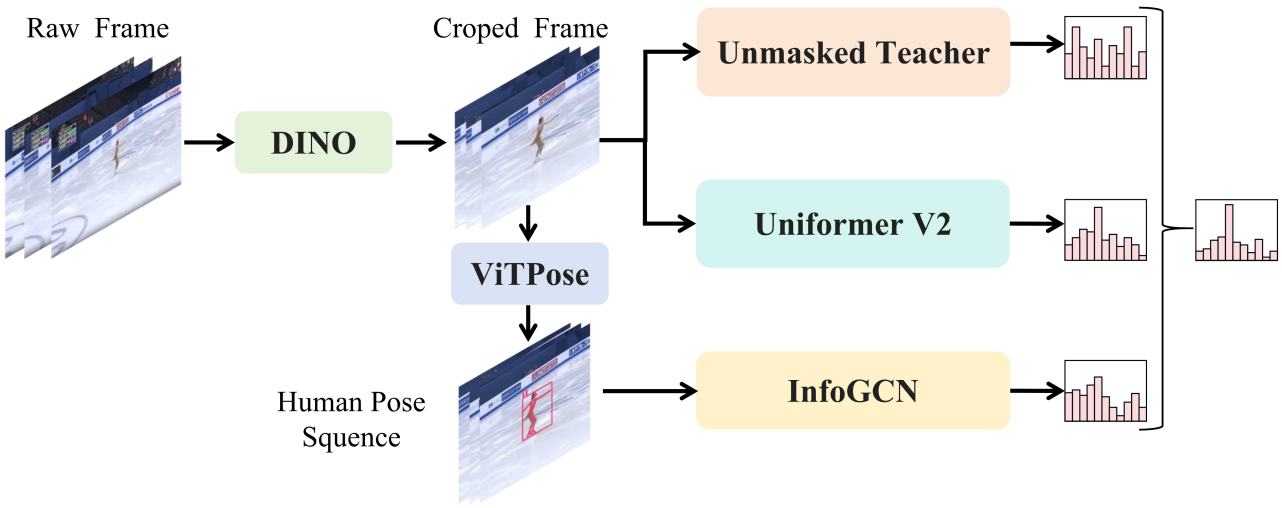

Fig. 2: The overall architecture of our solution.

L16 model (which is pre-trained and fine-tuned on Kinetices710 with 8×224^2 input images) for 50 epochs using 16×224^2 input images. Subsequently, based on the obtained fine-tuned model weights, we perform weight interpolation and further fine-tune the model for 10 additional epochs. This fine-tuning is conducted separately using both 16×448^2 images and 32×224^2 images. To obtain the final prediction result of UMT, we aggregate the predicted probabilities from the three models and apply the softmax function.

2) UniformerV2: UniformerV2 [4] is a versatile approach for building a robust collection of video networks. It combines the image-pretrained Vision Transformers (ViTs) with efficient video designs from UniFormer [6]. The key innovation in UniformerV2 lies in the inclusion of novel local and global relation aggregators. These aggregators seamlessly integrate the strengths of both ViTs and UniFormer, achieving a desirable balance between accuracy and computational efficiency.

In this work, we conduct fine-tuning on the UniFormerV2-L14 model(we use the CLIP [8] model weight pre-trained on LAION400M [9]) for 30 epochs, utilizing 32×224^2 input images.

3) ViTPose & InfoGCN: To obtain additional skating-related information from human skeleton sequences, we first use ViTPose [10] to extract the human skeleton sequence for each cropped frame. ViTPose utilizes plain and non-hierarchical vision transformers as backbones to extract features for individual person instances. It also incorporates a lightweight decoder for efficient pose estimation. ViTPose offers great flexibility in terms of attention type, input resolution, pre-training, fine-tuning strategies, and the ability to handle multiple pose tasks.

Following the extraction of the human skeleton sequence for each cropped frame using ViTPose, we obtain a sequence of length 320 for each video. To accurately predict the categories of figure skating actions, we train the InfoGCN [1] model from scratch over the course of 300 epochs. This training process enables the model to effectively analyze and classify the various figure skating actions depicted in the videos.

III. Experiments

A. Implementation Details

All experiments in this study are conducted using the PyTorch 2.0 framework, and the training process is performed on a powerful setup consisting of 8 NVIDIA A100 (80G) GPUs. The input image size is set to 224 × 224 pixels. To optimize the training process, we utilize the AdamW optimizer [7]. The specific hyperparameters employed for fine-tuning UMT, UniformerV2, and InfoGCN are provided in Table I, offering detailed insights into the configuration choices.

To expedite the training process, we incorporate FlashAttention2.0 [2], which leads to a notable 2 to 3 times acceleration in training speed. Additionally, we employ gradient accumulation, further enhancing the efficiency of the training process. During model inference, we adopt Test-Time Augmentation (TTA), which involves conducting 5 temporal and 3 spatial samplings, resulting in a total of 15 predictions. These predictions are subsequently fused together to generate the final prediction result, ensuring robustness and accuracy in the model's output.

B. Evaluation Metric

In this challenge, the performance of the model is assessed based on the mean accuracy of the categories, denoted as *Mean*. The number of correctly predicted samples, referred to as M, represents the instances where the category with the highest confidence from the model aligns with the actual category. The total number of samples is denoted as N. For each category with N_i samples, if the number of correctly predicted samples is M_i, the mean accuracy *Mean* is calculated using the following formula:

$$Mean = \frac{1}{l} \sum_{i=1}^{l} \frac{M_i}{N_i} \tag{1}$$

TABLE I: Detail hyperparameters used in training UMT, UniformerV2, and InfoGCN.

(a) Unmasked Teacher-L16 with 16×224^2 input images.

Hyperparameter	Value
Adam β_1	0.9
Adam β_2	0.98
Adam ε	10^{-6}
Weight decay	0.05
Drop path	0.2
Layer decay	0.85
Batch size	128
Learning rate	2e-4
Warmup epochs	5
Training epochs	50
Gradient accumulation	2
GPU	8×A100(80G)

(b) UniformerV2-L14 with 32×224^2 input images.

Hyperparameter	Value
Adam β_1	0.9
Adam β_2	0.999
Adam ε	10^{-8}
Weight decay	0.05
Drop path	0.2
Batch size	16
Learning rate	2e-5
Warmup epochs	5
Training epochs	30
GPU	4×A100(80G)

(c) InfoGCN with input human skeleton sequences.

Hyperparameter	Value
Adam β_1	0.9
Adam β_2	0.999
Adam ε	10^{-8}
Weight decay	0.0005
Batch size	128
Learning rate	1e-3
Warmup epochs	5
Training epochs	300
GPU	1×A100(80G)

C. Experiment Result

Tab. II showcases the experimental results on the leaderboard, highlighting the performance of three distinct models. Notably, both Unmasked Teacher (UMT) and UniformerV2 demonstrate a considerable improvement in performance compared to InfoGCN. This notable enhancement can be attributed to the fact that UMT and UniformerV2 are video foundation models, benefiting from pre-training on extensive video datasets.

TABLE II: Summary of experimental results on the leaderboard.

Method	Online Score
Unmasked Teacher [5]	94.55
UniformerV2 [4]	95.02
InfoGCN	92.03

D. Model Ensemble

To further improve the leaderboard score, we adopt two simple ensemble strategies. The first approach involves voting among all the model predictions, with UniformerV2's prediction being selected in case of a tie. The second approach involves weighted aggregation of the predictions, where the weights for UMT, UniformerV2, and InfoGCN are determined using the softmax function with weights [94.5, 95.0, 92.0], respectively. The ensemble results are shown in Tab. III

TABLE III: Summary of ensemble results on the leaderboard.

Method	Online Score
Model Voter	95.67
Model Weighted Summation	95.73

REFERENCES

[1] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20186–20196, 2022.

[2] T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023.

[3] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[4] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer, 2022.

[5] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023.

[6] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022.

[7] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[9] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[10] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

[11] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.