Fact sheet: CVPR 2021 TinyAction Challenge

I. Team details

- Team leader name: ALONG
- Username on Codalab: LoveLetter
- Team members : LiuCen, Yunbo Peng, Yue Lin
- Team affiliation: Netease Games AI Lab
- Team leader phone number: +86-18698676651
- Team leader email: liucen05@163.com

II. Contribution details

A. Title of the contribution

We use the two-stage method to get the results, which are super-resolution module and video classification module.

B. Introduction and Motivation

Our approach consists of two main stages: Video super resolution and action classification. For video super-resolution steps, we use the BasicVSR [1]. It presents an informationrefill mechanism and a coupled propagation scheme to facilitate information aggregation and leads to state-of-the-art performance. The second stage, action classification network, we use SlowFast [2] and TANet [3] to train several models, we believe that ensemble the strongest models can get more better results. Please see Fig 1 for our workflow.

C. Detailed method description

Our team analyzes the dataset and finds it contains arbitrary sized low-resolution videos which ranged from 10x10 pixels to 128x128 pixels. We first scale the video with a height of less than 96 to a size of 64, and then use the video super-resolution model mentioned above to expand it to a size of 128, and for other videos to scale it to the size of 128. We think that high performance can be obtained by a ensemble of various SOTA models, so we use the above datasets to train SlowFast [2] and TANet [3] and fuse the prediction scoces of all networks to get the final results. Another important thing is that the pretrained model on Kinects-400 can impove the training results. We also use multigrid [4] to train our network. The training stage is conducted on a 8 x 2080Ti GPUs with 12GB GPU memory footage for each GPU.

TABLE I
Results obtained by the proposed approach.

Method	F1-Score
Bilinear Interpolation + SlowFast(r101)	0.392355
BasicVSR + SlowFast(r101)	0.424861
BasicVSR + TANet(r50)	0.411005
Ensemble	0.442072

D. Challenge results and final remarks

Table I shows our experimental results.

III. Additional method details

Please reply if your challenge entry considered (or not) the following strategies and provide a brief explanation.

- Did you use any kind of depth information (directly, such as RGBD data, or indirectly such as 3D pose estimation trained on RGBD data), either if during training or testing stage? ( ) Yes, (X) No
- Did you use pre-trained models? (X) Yes, ( ) No
  Yes, BasicVSR models pretrained on REDS4 and slowfast models pretrained on Kinects400
- Did you use external data? ( ) Yes, (X) No
- Did you use other regularization strategies/terms? ( ) Yes, (X) No
- Did you use handcrafted features? ( ) Yes, (X) No
- Did you use any face / hand / body detection, alignment or segmentation strategy? (X) Yes, ( ) No
- Did you use any fusion strategy of modalities? ( ) Yes, (X) No
- Did you use ensemble models? (X) Yes, ( ) No
  Yes.
- Did you use any spatio-temporal feature extraction strategy? (X) Yes, ( ) No
  Yes, slowfast and TANet.
- Did you use any bias mitigation technique (e.g. rebalancing training data)?

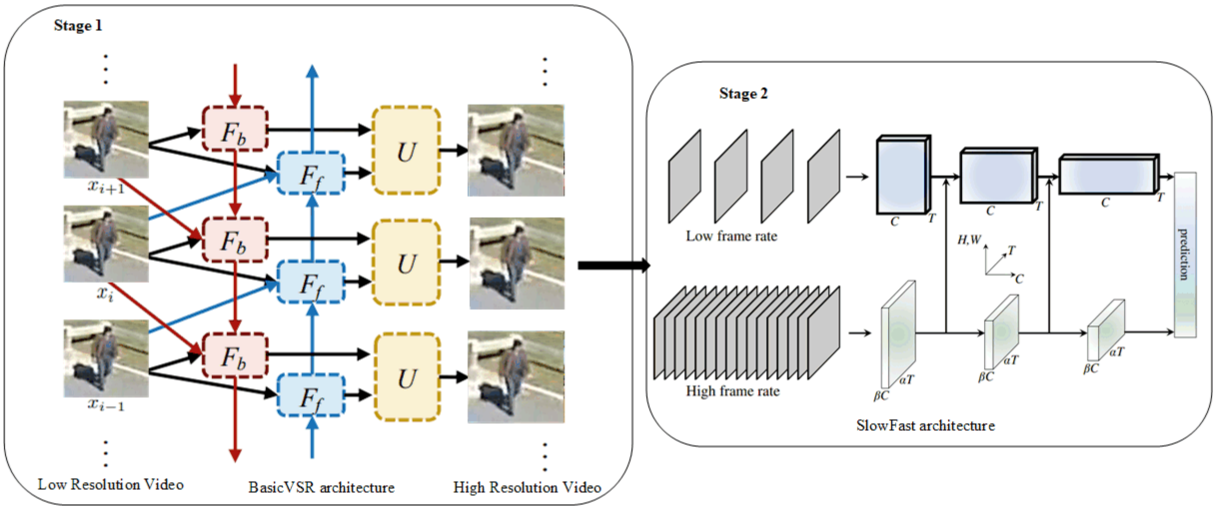

Fig. 1. Workflow of our method

( ) Yes, (X) No

IV. Code repository

Still finishing, will come soon.

References

[1] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2021.

[2] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2019, pp. 6201–6210.

[3] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," arXiv preprint arXiv:2005.06803, 2020.

[4] C.-Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krähenbühl, "A Multigrid Method for Efficiently Training Video Models," in CVPR, 2020.