# Video Scene Parsing with Transformer Based Vision Encoder

Peng Liu Quanlong Zheng Qiuju Dai Tong Yang
OPPO Research
{alren.liu, zhengquanlong, daiqiuju, yangtong}@oppo.com

## Abstract

*In this paper, we use a simple image segmentation method to solve the video scene parsing task and get a good result. Unlike many CNN-based methods, we use a transformer based encoder to extract features. The extracted features are first fed to a multi-scale fusion module to obtain fused features, which are then fed to a segmentation head to obtain the final segmentation result. Combining with a multi-scale testing scheme, our method achieves the fourth place on the VSPW 2021 Challenge with a mIoU score of 55.62% on testing set.*

## 1. Introduction

Scene parsing is a very important problem in the field of computer vision. Although there are some image-based segmentation datasets, video scene parsing dataset is still lacking. VSPW [8] is the first large-scale video scene parsing dataset with high resolution and frame rate. It also contains 125 categories which cover most scenes.

Video scene parsing a multi-class video semantic segmentation task. Most segmentation network architectures are designed based on CNN encoder with a semantic head(Current state-of-the-art semantic method on Cityscapes [3] is HRNet [9] with OCRhead [11]). In recent years, with the emergence of transformer, many transformer backbones for visual tasks, such as Swin-Transformer [7], CSWin-Transformer [5] and BEiT [1], show excellent performance on vision task. In our method, we select BEiT as the backbone for feature extraction.

## 2. Semantic Segmentation with Transformer based vision encoder

### 2.1. Basic Network Architecture

As shown in Figure 1, the proposed network consists of a BEiT encoder, feature fusion module and a segmentation head. According to Unified Perceptual Parsing Network (UPerNet)[10], we set a Pyramid Pooling Module (PPM) [12] appended on the last layer of the BEiT network. Then 4 different-level features are fed into a top-down branch in the Feature Pyramid Network (FPN) [6]. At the end of the network, a simple segmentation head is used to generate segmentation score maps.

### 2.2. BEiT

BEiT stands for Bidirectional Encoder representation from Image Transformers. It applies the ideas of BERT [4] into the vision field. Experimental results on image classification and semantic segmentation show that the pre-trained BEiT models outperform previous pre-training methods with a large margin. Hence, we use BEiT-large as our backbone in our experiment.

### 2.3. Pyramid Pooling Module (PPM)

Contextual information plays a very important role on segmentation task. By enhancing the context information, it can achieve better segmentation results at different scales. As shown in Figure 2, PPM is a relatively good way to make full use of global information

### 2.4. feature pyramid networks(FPN)

For the low-level feature, semantic information is relatively small, but the target location is accurate; As for the high-level feature, semantic information is richer, but the target location is relatively rough. In addition, although some algorithms use multi-scale feature fusion, they generally use the fused features to make predictions.

The difference in FPN [6] simultaneously uses the high-resolution of low-level features and the high-semantic information of high-level features, and achieves the effect of prediction by fusing the features of these different layers. And the prediction is performed separately on each fused feature layer, which is different from the conventional feature fusion method.

### 2.5. Loss Function

Our loss function contains two terms: a feature loss $L_{feat}$ by comparing the images in feature space and a segmentation loss $L_{seg}$ for final result. We apply lovasz loss [2] for the segmentation loss and Cross-entropy loss for the

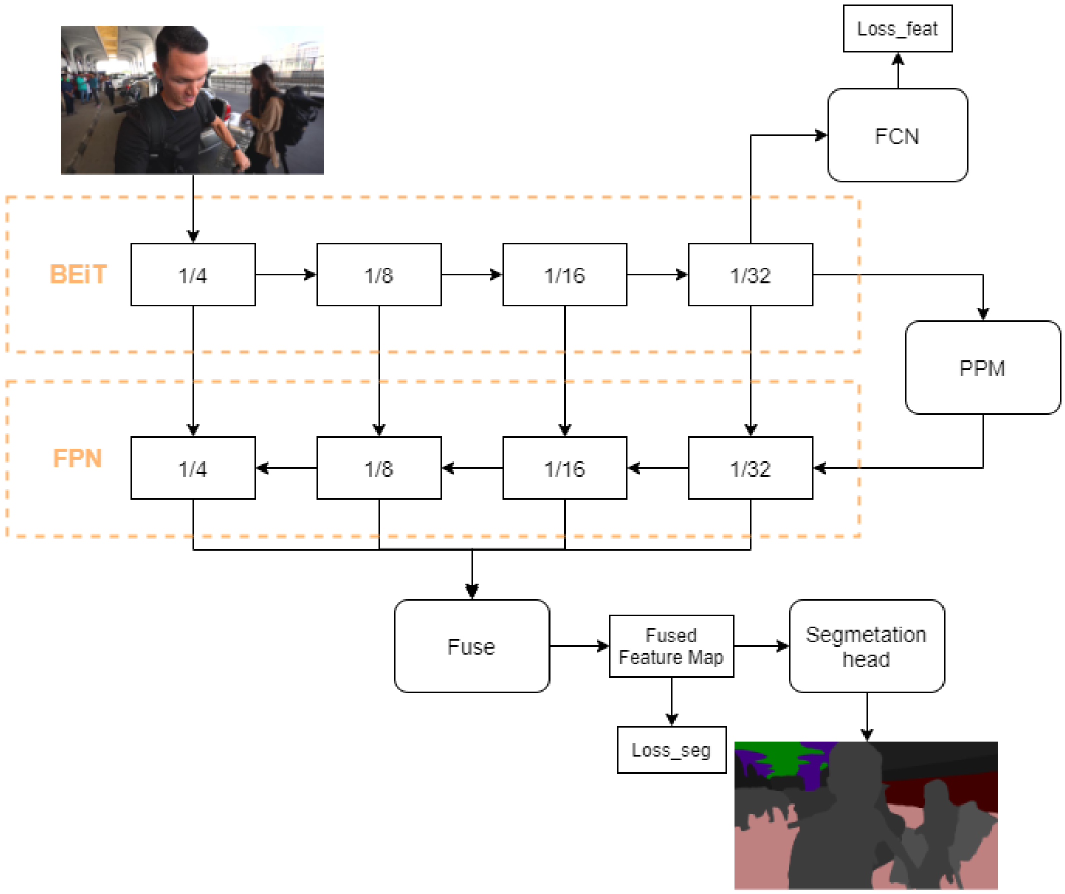

Figure 1. Basic Network Architecture.

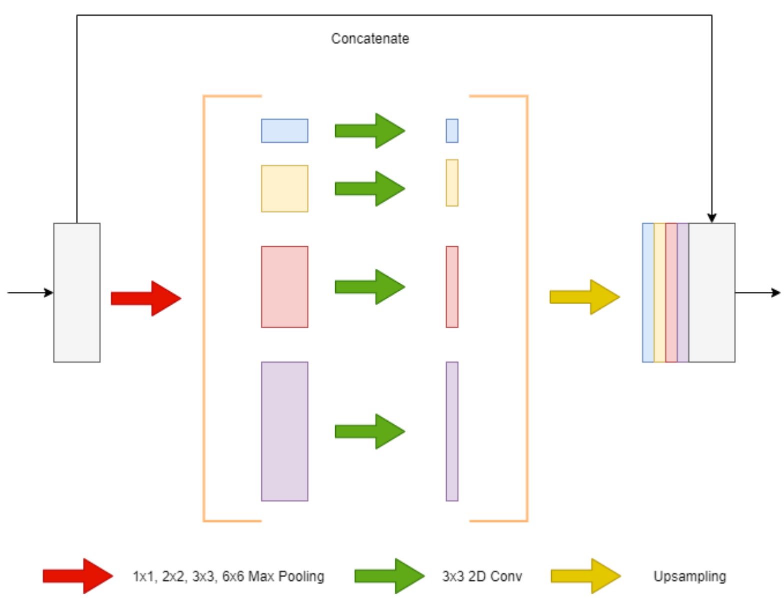

Figure 2. Pyramid Pooling Module (PPM)

feature loss. The weight of $L_{seg}$ is set to 1, and the weight of $L_{feat}$ is 0.4.

# 3. Experiments

## 3.1. Training Details

For VSPW 2021 Challenge, we use BEiT-large as our backbone and get the ImageNet-22k pre-trained model from [1]. The images are augmented by random cropping ($480 \times 480$), random scaling in the range of $[0.8, 2]$, and random horizontal flipping.

For optimization, we use AdamW with $\beta_1 = 0.9$ , $\beta_2 = 0.99$ and weight decay=0.05. The initial learning rate is set to 2e-5. The poly learning rate policy with the power of 1.0 is used for dropping the learning rate. We implement our models with Pytorch framework and train them for 160K itertions with batch size of 24 on 8 Tesla V100 GPUs and syncBN. It costs 0.91s using one single Tesla V100 GPU for processing per image with 480×853 pixels.

## 3.2. Testing Details

We used flip and multi-scale testing with scales=[0.75, 1.0, 1.5, 1.75]. The multi-scale and flip testing get a +0.80 mIoU improvement comparing with single-scale testing.

# References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[8] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4133–4143, 2021.

[9] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[10] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[11] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[12] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.