A Stronger Multi-task Foundation Model For Intelligent Transportation

Yantian Wang*
Wuhan University
wytcsuch@whu.edu.cn

Defang Zhao*
Tongji University
zhaodefang18@gmail.com

Abstract

Different from the previous single-task "training from scratch" scheme, the unified foundation model promotes knowledge communications by optimizing multiple tasks simultaneously. Due to its remarkable feature representation and robust generalization ability, the unified large models have been applied in many visual scenes. In this paper, we propose a stronger foundation model and exploit it in the intelligent transportation industry for classification, segmentation, and detection tasks. Specifically, we apply Swin Transformer which is friendly for dense recognition tasks as our backbone and adopts Mask2Former and DINO as segmentation and detection heads respectively. For classification, we construct a Cross-Level Feature Fuse Module (CLFFM) to utilize multi-scale tokens. In addition, a two-stage optimization method based on an uncertainty weighting strategy is proposed, encouraging the model to learn a general and robust feature representation in early training. Furthermore, a universal data augmentation pipeline called DAPMT is designed, aiming to prevent the model from overfitting. Our method finally won third place in the CVPR 2023 1st Foundation model challenge.

1. Introduction

The past few years have witnessed the prosperity of deep learning, many milestone works have been proposed and made remarkable achievements in many vision fields, *e.g.*, image classification, object detection, and image segmentation. However, most of these works are based on Single-Task Learning which heavily depends on the distribution of training data, thus may cause a poor generalization ability.

Recently, with the development of computer processing capabilities and the enrichment of application scenarios, the AllinOne model which aims to solve multiple tasks in a unified architecture has become an urgent demand. In natural language processing (NLP), large-scale language models (LLMs) have entered the mainstream, which can handle a variety of natural language processing tasks including summarizing, translating, recognizing, predicting, generating text and other forms of content, based on large transformer models pretrained from the massive corpus.

Inspired by the significant success of multi-task large models in NLP, great progress has also been made in computer vision. UFO [1] proposes a novel training and deploying paradigm, which trains a supernet and extracts a dedicated sub-network for each specific downstream task by trim, thus leading to a big convenience for flexible deployment. Based on UFO, Open-TransMind [2] proposes a multi-task foundation model, which aims to use one unified model to handle the classification, segmentation, and detection tasks simultaneously in the transportation scene and has won a superior performance, compared with learning each task individually. However, it still needs to improve in some aspects, such as small object segmentation and detection, the weak effect caused by few shot data in some scenes, and unsatisfying generalization.

In this work, we construct a stronger foundation model based on Open-TransMind. First, considering the importance of multi-scale information in dense prediction tasks, we employ a hierarchical Swin Transformer [3] instead of ViT [4]. Secondly, we specially design the structure for classification and segmentation. For classification, we construct a cross-level feature fusion module (CLFFM) to enhance the detail representation. For segmentation, we adopt Mask2Former [5], which can use multi-scale high resolution features and is convenient to solve all the segmentation tasks in a unified manner. Thirdly, to prompt more robust training, a two-stage optimization method is proposed which can automatically weight each task based on homoscedastic task uncertainty. Moreover, we present a universal data augmentation pipeline named DAPMT (Data Augmentation Pipeline for Multi-Task) which includes three different augmentation pools: color, space, and noise, with the purpose of strengthening the generalization and robustness in various scenes, especially a few shot scenes. The experiments show that our model achieves excellent performance with scores of 70.45%, 95.60%, and 95.45% in segmentation, classification and detection re-

*Co-first authors with equal contributions.

spectively, which outperforms Open-TransMind and proves the effectiveness.

2. Methodology

In this section, we introduce the overall framework of our foundation model. As illustrated in Figure. 1, It consists of a shared backbone for extracting multi-scale features and three prediction heads responsible for each task respectively.

2.1. Backbone

ViT has achieved comparable or superior performance over CNNs in many image classification tasks, however, its columnar structure is not suitable for dense task prediction, such as object detection and semantic segmentation. Thus, in this work, we adopt Swin Transformer as our backbone. It is a hierarchical structure that is flexible to model at various scales and has linear computational complexity, benefitting from its shifted windowing scheme.

2.2. Detection Module

Currently, more and more transformer-based object detection methods have been proposed, among which DINO [6] has achieved excellent performance in both accelerating training convergence and detection results, due to its improvements of comparative denoising training, mixed query selection and look forward twice. Thus we select DINO as our detection head.

2.3. Segmentation Module

Open-TransMind uses a progressive upsampling head (PUPHead) that still considers semantic segmentation as a per-pixel classification task. However, several works [5] have proven that the mask-level classification-based approach can be easily expanded to any segmentation tasks (panoptic, instance, or semantic segmentation), without changing structure or loss, which is consistent with our goal of constructing a unified multi-task large model. Thus, we adopt mask2fromer, a universal image segmentation architecture that contains a pixel decoder and a Transformer decoder with masked attention, in our model. It also proposes an efficient strategy to exploit high-resolution features for small object segmentation, which fits well with our multi-scale backbone.

2.4. Classification Module

Open-TransMind adopts a linear-projection layer and a fully connected layer as feature decoders for classification tasks, which are unable to utilize multi-scale features. Thus, we design a new structure called Cross-Level Feature Fusion Module (CLFFM) to replace the original linear projection layer.

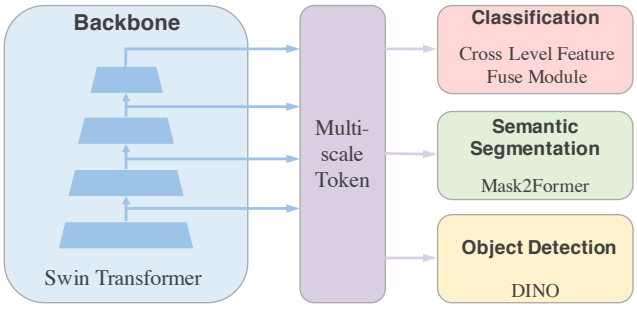

Figure 1. The overall framework of our method.

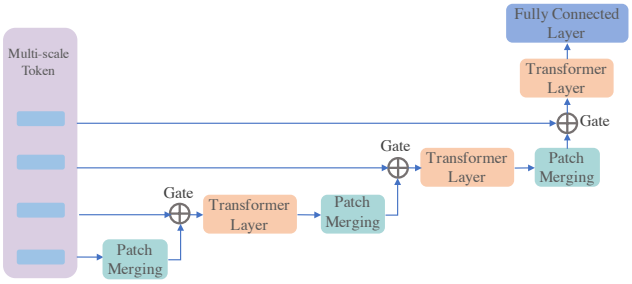

Figure 2. The structure of the Cross-Level Feature Fusion Module.

As illustrated in Figure. 2, we adopt a learnable gate to automatically aggregate features from adjacent layers and then a multi-head self-attention module (MHSA) is used to mine detailed information from the fused token. While effectively utilizing multi-level features, the proposed structure also strengthens the isolation between classification and other branches, which is more conducive to classification learning.

2.5. Two-Stage Optimization Objectives

In general, the performance of multi-task models strongly depends on the relative weights of different tasks, so it is of great importance to weigh multiple loss functions adaptively. To this end, Lukas *et al.* [7] proposed a method that can automatically adjust the loss weight of each task by considering the homoscedastic uncertainty. The calculation formula is as follows:

$$L_T = \sum_{\tau \in T} \frac{1}{2\sigma_\tau^2} + log(1 + \sigma_\tau^2) \quad (1)$$

where $\tau \in T$ represents individual tasks, and L_τ denotes the loss of each task. σ is the homoscedasticity that denotes the uncertainty of each task. Larger homoscedasticity indicates that the task is more difficult to optimize.

In this work, considering that the optimization objective of the multi-task large model focuses on different aspects in

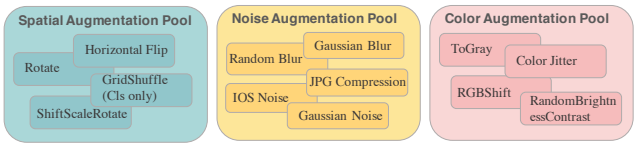

Figure 3. Illustration of three data augmentation pools.

early and middle-later training periods, we propose a two-stage optimization method, which can be explained in the formula (2). In the early period, influenced by data quality and optimization difficulty, there usually shows an obvious loss imbalance across multiple tasks, which may lead to biased learning. In this condition, the dynamic loss weighting principle illustrated above can reduce the weight of the task that is hard to optimize, promoting the model to learn a general and robust representation from easier tasks. However, starting from the mid-term, the model should pay more attention to difficult tasks, in which case dynamic loss weighting should be turned off. It should be noted that we generally define the first 1/5 of the total training iters or epochs as the early period.

$$L_T = \begin{cases} \sum_{\tau \in T} \frac{1}{2\sigma_\tau^2} + log(1+\sigma_\tau^2), \ iter \in early\ period \\ \sum_{\tau \in T} L_\tau, \ iter \in middle - later\ period \end{cases} \tag{2}$$

2.6. Data Augmentation Pipeline for Multi-Task

Furthermore, we propose a universal data augmentation pipeline for multi-task training(DAPMT) which can be easily transferred to other unified model training. Specifically, we split the commonly used data augmentation methods into three pools, called spatial augmentation pool, noise augmentation pool, and color augmentation pool, as shown in Figure 3. The spatial augmentation pool includes rotation, scale-shift rotation, horizontal flip, and grid shuffle (for classification tasks only). The noise augmentation pool includes Random Blur, Gaussian Blur, JPG Compression, IOS Noise, Gaussian Noise, and more. The color augmentation pool includes RGB shift, Random brightness and contrast adjustments, Color jitter, and more. During training, the pipeline randomly selects one augmentation method from each pool and then composes them into a new augmentation operation, which effectively extends the distribution of the training data.

Notably, for different application scenarios, we can flexibly adjust each data augmentation pool. Considering the traffic scene in this work, we add mosaic and random weather (rain, fog) augmentation methods to the spatial and color augmentation pools respectively, aiming to improve the performance in segmenting and detecting small objects as well as adaptability under different weather conditions.

3. Experiments

3.1. Settings

Datasets. Same with Open-TransMind, we use Stanford Car, TT100K, and BDD100K datasets for image classification, object detection, and semantic segmentation tasks respectively. The Standford Car dataset consists of 196 car classes with 16,185 images, of which 8,144 images are used for training and left 8,041 images for testing. Tsinghua-Tencent 100K (TT100K) is a large benchmark for traffic-sign detection and classification, which provides 100,000 images containing 30,000 traffic-sign instances. We use detection-annotated samples, which consist of 6,107 train images and 3,067 test images. BDD100K is a large-scale diverse driving video dataset with annotations for ten tasks. In this work, we use samples for the semantic segmentation task, containing 19 classes, including 7,000 images for training and 1,000 for testing.

Evaluation Metrics. We employ Acc, mIoU, and mAP as metrics to evaluate classification, segmentation and detection, respectively. And we calculate the average of these three metrics as a global metric to measure the overall effectiveness of the foundation model.

Experiment Settings. We initialize the backbone with the parameters pretrained on imagenet-22k. We train our model for 80 epochs on eight A100 GPUs with batch sizes of 8, 64, and 8 for segmentation, classification, and detection tasks respectively. In the first 20 epochs, we turn on the dynamic loss weighting. For the last 10 epochs, we stop the mosaic augmentation in the segmentation and detection tasks for a better fit on the distribution of the training data. Besides, we use the AdamW optimizer with an initial learning rate of 0.0001 and a weight decay of 0.0001. For the learning rate, we employ a linear warmup for the first 200 iterations and then gradually decrease the learning rate using the CosineAnnealingLR strategy. Additionally, we set the input size of detection to 1184 for better performance in detecting small objects.

3.2. Experiments Results

To validate the effectiveness of our improvements, we conduct multiple comparison experiments, and the results are shown in Table 1. The first two rows of the Table list the scores of the baseline with different ViT scales.

We first replace the backbone with Swin Transformer Large and win scores of 60.94%, 94.71%, and 84.10% in three tasks, which completely exceed the baseline with ViT-Base, especially in segmentation and detection tasks, as shown in the third row. Based on the third row, we further adopt mask2former and CLFFM, instead of the original PupHead structure and linear-projection layer, which makes great progress of 5.92% and 0.49% in segmentation and classification tasks respectively, as illustrated in

Table 1. Comparison of our method with Open-TransMind on three tasks. Ours-Large indicates the model with a higher input resolution. Ours-Large* means we adopt the two-stage optimization method during training.

Method	Backbone	Seg Head	Cls Head	Det Head	Seg(mIoU%)	Cls(acc%)	Det(mAP%)	Avg
Open-TransMind	ViT-Base	PupHead	Linear-Projection	DINO	55.13	91.64	76.90	74.56
	ViT-Huge	PupHead	Linear-Projection	DINO	64.80	95.96	83.24	81.33
Ours	Swin-Large	PupHead	Linear-Projection	DINO	60.94	94.71	84.10	79.92
	Swin-Large	Mask2Former	CLFFM	DINO	66.86	95.20	84.66	82.24
Ours-Large	Swin-Large	Mask2Former	CLFFM	DINO	69.52	95.72	94.83	86.69
Ours-Large*	Swin-Large	Mask2Former	CLFFM	DINO	70.45	95.60	95.45	87.17

the fourth row. The experiments above reveal that both the stronger backbone and individual task head are crucial for the unified large model. Then, to improve the detection performance of small targets, we scale up the input resolution in the detection task from 640 to 1184 and employ a mosaic augmentation method. From the results listed in the second-to-last row of the table, we can observe a significant gain of 10.17% in mAP. Notably, the mIoU of the segmentation task also increased by 2.66%, indicating that the strategies used in the detection task are also beneficial for segmentation, which obviously exhibits the advantages of the knowledge reference mechanism between different tasks in a unified model. Besides, we further apply the mosaic strategy in segmentation and the two-stage optimization method is also adopted, which is present in the last row of the table. It can be seen that by aggregating all these improvements our model acquire the best performance of 70.45%, 95.60%, and 95.45% in segmentation, classification, and detection, respectively, which even surpasses the ViT-Huge based baseline and win significant advantages. The extensive experimental results demonstrate the effectiveness and superiority of our foundation model.

4. Conclusion

In this paper, a stronger unified foundation model is proposed to improve the performance of multiple visual tasks in the intelligent transportation industry. We apply Swin Transformer as our hierarchical feature extractors and introduce Mask2Former and DINO in segmentation and detection tasks respectively. For the classification task, We design a new structure called CLFFM, which introduces a learnable Gate and MHSA modules to fuse and mine detailed information from multi-scale features. In addition, we adopt a two-stage optimization method that encourages robust learning at the beginning of training by balancing multi-task loss dynamically and then focusing on difficult tasks in the mid-term. Moreover, we introduce DAPMT, a universal pipeline with abundant data augmentation combinations, which can be easily adjusted and injected into other multi-task training.

A series of experiments show that our model substantially outperforms the baselines. And we also hope that our work can serve as a new starting point, attracting more interest in universal model improvements.

References

[1] Teng Xi, Yifan Sun, Deli Yu, Bi Li, Nan Peng, Gang Zhang, Xinyu Zhang, Zhigang Wang, Jinwen Chen, Jian Wang, Haocheng Feng Lufei Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Ufo: Unified feature optimization. In *ECCV*, 2022. 1

[2] Yifeng Shi, Feng Lv, Xinliang Wang, Chunlong Xia, Shaojie Li, Shujie Yang, Teng Xi, and Gang Zhang. Open-transmind: A new baseline and benchmark for 1st foundation model challenge of intelligent transportation. In *CVPR*, pages 6327–6334, 2023. 1

[3] 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. In *ICCV*, pages 10012–10022, 2021. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 1, 2

[6] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[7] Lukas Liebel and Marco Korner. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1907.03892*, 2018. 2