OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track: 3D-Pose-Estimation Track
- Team name: USTC-IAT-United
- Team leader name: Jun Yu
- Team leader address, phone number, and email: Department of Automation, University of Science and Technology of China, Hefei, Anhui Province, China; +86-13856070316, harryjun@ustc.edu.cn
- Rest of the team members: Keda Lu, Mohan Jing, Yaohui Zhang
- Team website URL: `https://auto.ustc.edu.cn/2021/0510/c25977a484905/page.htm`
- Affiliation: University of Science and Technology of China
- User names on the OOD-CV Codalab competitions: USTC-IAT-United

- Link to the codes of the solution(s): Please see our attachment.

2 Contribution details

Method	acc@pi/6(val)	acc@pi/18(val)	mederr(val)	acc@pi/6(test)
Resnet50(baseline)	50.29	19.74	29.36	44.51
Resnet50-pose(Ours)	51.59	26.75	27.65	45.21
Resnet50-pose+wea(Ours)	51.84	26.83	27.17	44.91
Resnet50-pose+jitter(Ours)	52.06	27.21	27.42	**45.25**

Table 1: Comparison between our proposed method and baseline method.

Method	Resnet50-pose	Resnet50-pose+wea	Resnet50-pose+jitter
OOD(Mean)	45.21	44.91	45.25
IID	70.22	69.81	69.64
shape	55.11	55.24	54.18
pose	18.36	16.62	16.13
context	48.32	48.63	49.02
texture	54.57	55.40	57.15
occlusion	35.76	33.89	35.49
weather	59.25	59.64	59.44

Table 2: The evaluation indicators of our method in the final phase.

- Title of the contribution: Resnet50-pose: A visual model suitable for pose estimation tasks

- General method description: The improvement of the pose estimation task for this competition is quite difficult, and we refer to the baseline to convert it into a classification task, i.e., to classify the 3 angle values.

  We improve it from the model and data perspectives respectively. We try many different backbones and find that there is no enhancement

or even a decreasing trend on the metrics, then we modify the model based on resnet50 and proposed the resnet50-pose structure to better adapt to the pose estimation task. Specifically, after using resnet50 to extract features from the input images, we design 3 heads based on multi-layer perceptron (MLP) to perform classification tasks on the 3 target values, sharing the input feature maps but not the weights of the heads. In addition, we introduce several data augmentation methods and evaluate them, and finally confirm that the color jitter and the "weather" disturbance in the corruption can effectively improve the robustness of the model. As shown in Table 1 and Table 2, this is the result of our scheme for the **1st** in this track.

- References:

1. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

2. Pavlakos, Georgios, et al. "Coarse-to-fine volumetric prediction for single-image 3D human pose."Â Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

3. Martinez, Julieta, et al. "A simple yet effective baseline for 3d human pose estimation."Â Proceedings of the IEEE international conference on computer vision. 2017.

4. Habibie, Ikhsanul, et al. "In the wild human pose estimation using explicit 2d features and intermediate 3d representations."Â Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

5. Li, Shichao, et al. "Cascaded deep monocular 3d human pose estimation with evolutionary training data."Â Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

6. Xu L, Ouyang W, Bennamoun M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4310-4319.

- Representative image / diagram of the method(s): Figure 1 shows our

method.

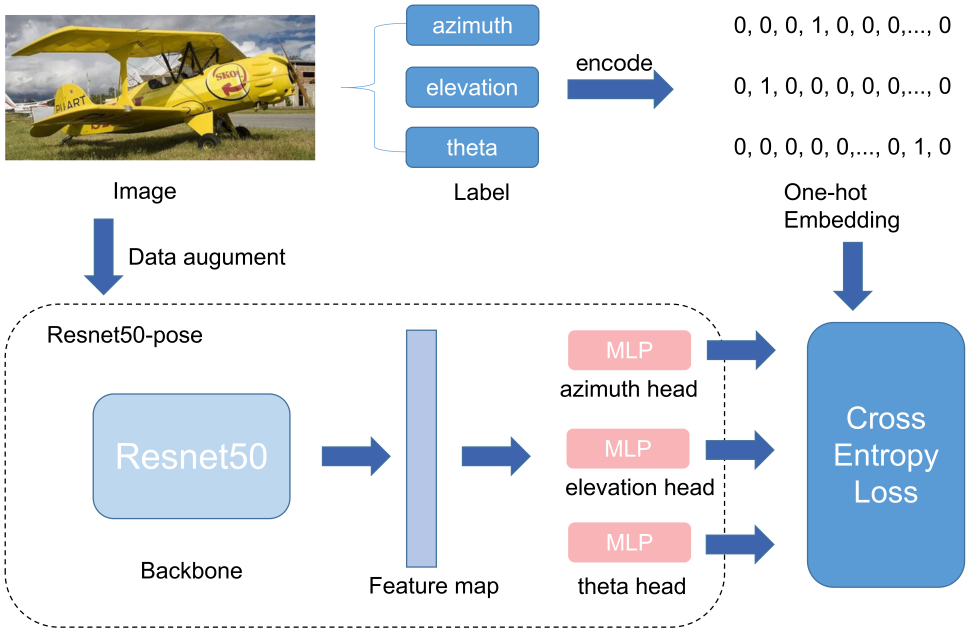

Figure 1: The overall framework diagram of our proposed approach.

3 Global Method Description

- Total method complexity: With 1 GeForce RTX 3090 graphics cards, it takes a total of 3 hours to train 100 epochs and only 10 minutes to infer. The time required can be further reduced by using devices with stronger arithmetic power.

- Model Parameters: ResNet50 param count: 23M

- Run Time: With 1 GeForce RTX 3090 graphics cards, it takes about 2 minutes to train one epoch and at least 20 epochs to train. Inference

takes only about 1 minutes.

- Which pre-trained or external methods / models have been used: Pre-trained ResNet50 using ImageNet-1K dataset.

- Training description: For model selection, we used ResNet50 as the backbone.

  For training, we choose SGD optimizer and use MultiStepLR to adjust learning rate for training.

- Testing description: For testing, We infer the final test set directly using the best results from the developmental phase.

- Quantitative and qualitative advantages of the proposed solution: We use ResNet50-pose to classify the 3 target parameters, and the model balances training speed and accuracy.

- Results of the comparison to other approaches (if any): None

- Novelty of the solution and if it has been previously published: A model structure purposely proposed for the pose estimation task.

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any).: In this method, we do not use the strategy of model fusion.

- What was the benefit over the single method?: None

- What were the baseline and the fused methods?: None

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements): This method is implemented in python. 1 GeForce P 40 graphics cards which are used for parallel training and testing. Each graphics card occupies approximately 24 GB of video memory for training and 24 GB for testing.

- Human effort required for implementation, training and validation?: We need to perform deep exploratory data analysis at the beginning of the project implementation, but our approach does not require Human effort for training and validation, and the approach can be deployed end-to-end.

- Training/testing time? Runtime at test per image: In the case of 1 GeForce RTX 3090 graphics card training, it should last for up to 3 hours for 100 epochs. In the case of 1 GeForce RTX 3090 graphics card testing, it takes about 30 minutes to infer, the inferring speed is about 10 img/s, and the time required for each image test is 0.2 s.

- Comment the efficiency of the proposed solution(s)?: Our proposed method uses only ResNet networks, which can be trained with few parameters and are fast.

6 Other details

- General comments and impressions of the OOD-CV challenge.:

  In addition to the effective solutions mentioned above, we also tried more solutions in this competition, such as replacing the backbone with more than 10 different structures and adjusting the number of classified bins, but the result seems to be not very satisfactory. In addition, we also tried to use MCTformer to simulate the generation of "Occlusion" data by obtaining foreground images from Imagenet-1K, but the training results were not satisfactory either.

At the pose estimation track, I think the task is quite difficult. The number of participants is low and therefore the track is not well studied. We believe that more and more people will pay attention to this track.

- Other comments: For OOD data, competition tracks such as action recognition, image segmentation, etc. can also be set up to attract more researchers to participate. We hope that the efforts of the organizers will contribute to the further development of this field.

  We guarantee that our experimental results are fully reproducible under the pre-trained model using only training data and Imagnet-1k. If the organizers encounter any problems during the reproduction process, please feel free to contact us.