

# OOD-CV Challenge Report

October 12, 2022

## 1 Team details

- Challenge track: Classification 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, Hao Chang, Mohan Jing, Xiaohua Qi, Liwen Zhang, Zhihong Wei, Ye Yu, Fang Gao
- 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

Table 1: The effect of different backbones(development phase).

Backbone	Img size	IID Top-1	OOD Top-1(mean)
Resnet-50	224	86.84%	81.72%
Efficientnet-B2	224	86.40%	79.38%
Efficientnet-B3	300	87.21%	81.46%
ConvNeXt-B	224	88.53%	85.76%
ConvNeXt-L	224	88.91%	86.43%
DeiT-L	384	89.87%	88.36%
VOLO-D5	512	90.02%	88.63%

- Link to the codes of the solution(s): <https://github.com/wujiekd/ECCV2022-OOD-CV-Challenge-Classification-Track-USTC-IAT-United>

## 2 Contribution details

- Title of the contribution : Look beyond the nature of the data: Data-centric approach to solving OOD problems
- General method description: We will describe our approach in 3 stages. The evaluation results in the first two stages are the results of the tests on the development phase, and in the third stage we perform further optimization for the test set in the final phase.

### 2.1 Stage 1

#### 2.1.1 Selecting model

First, the organizers restrict the use of pre-trained models trained only with ImageNet-1K, so we exclude a part of pre-trained models based on large-scale datasets such as ImageNet-22K, such as Swin Transformer. In fact, we try the performance of the Swin Transformer-L model and could already achieve 89.58% without any trick in the development

Table 2: The effect of different augmentation method(development phase).

Augmentation	IID Top-1	OOD Top-1(mean)
RandAugment	+0.14%	-0.19%
Augmix	-0.23%	-0.28%
Random Erasing	-0.18%	+0.34%
Mixup	-0.13%	+0.51%
Cutmix	-0.11%	+1.18%
Cutmix+Random Erasing	-0.16%	<b>+1.38%</b>

phase. Therefore limiting the use of pre-trained models with larger datasets to do transfer, OOD presents us with more challenges. We choose different training strategies for the two types of models, using an initial learning rate of 1e-2 for the CNN series and 3e-4 for the Transformer series. We use SGD with momentum optimizer. We use Mutilstep to adjust the learning rate, and sets [40,80], and trains 120 epochs. We try and evaluate a series of CNN or Transformer based models, as shown in the Table 1. It can be concluded that ConvNeXt-L, DeiT-L and VOLO-D5 perform significantly better than other models.

### 2.1.2 Data augmentation

We try various automatic data augmentation strategies, as well as some general data augmentation methods, as shown in Table 2. It can be seen that Cutmix+Random Erasing can significantly improve the OOD score, and we decided to use this augmentation combination. In addition, we also try to simulate the test images in OOD scenes with corruption, we divide corruption into four groups, namely weather, digital, noise and blur. Adding Gaussian noise, which is additive noise, and some blur operations hardly improve the generalizability of OOD in realistic scenes. As shown in Table 3, where the combination of Weather + Digital works better. In the testing stage, we evaluate the effect of different Test Time Augmentation(TTA) such as TenCrop and FiveCrop on the effect after using the above identified augmentation and corruption, as shown in Table 4.

Table 3: The effect of different corruption method(development phase).

corruption	IID Top-1	OOD Top-1(mean)
Weather	+0.12%	+0.05%
Weather and Digital	+0.35%	<b>+0.13%</b>
Weather, Digital and Noise	+0.26%	+0.12%
Weather, Digital, Noise and Blur	+0.21%	+0.10%

Table 4: The effect of different TTA(development phase).

Backbone	TTA	IID Top-1	OOD Top-1(mean)
DeiT-L	None	90.02%	90.42%
DeiT-L	FiveCrop	90.33%	<b>92.01%</b>
DeiT-L	TenCrop	90.12%	91.96%
ConvNeXt-L	None	90.25%	88.46%
ConvNeXt-L	FiveCrop	89.59%	<b>91.31%</b>
VOLO-D5	None	90.89%	90.51%
VOLO-D5	FiveCrop	90.08%	<b>91.65%</b>

### 2.1.3 Adding modules

As can be seen, the IID is close to the limiting threshold of 91.1, and we try to use Exponential Moving Average (EMA) to mitigate the overfitting, and we add EMA on all three models. as shown in Table 5, the Transformer family of models brings very weak improvement, while the CNN family of models brings a significant improvement.

There are many challenges between the train and test sets of OOD classification dataset, such as unseen distribution and domain shift. Thus we can solve the task which does not have suitable training data to ensure generalization by exploring sample relationships. Among recent data scarcity learning methods, sample relationships have been inten-

Table 5: The effect of EMA or BF(development phase).

Backbone	OOD Top-1(EMA)	OOD Top-1(BF)
DeiT-L	+0.04%	+0.22%
ConvNeXt-L	+0.49%	+0.18%
VOLO-D5	+0.02%	+0.28%

Table 6: The effect of Model Ensemble(development phase).

Backbone	Weight	IID Top-1	OOD Top-1(Mean)
DeiT-L(BF)	0.4	90.33%	92.01%
ConvNeXt-L(EMA)	0.4	91.09%	91.79%
VOLO-D5	0.2	90.08%	91.65%
Ensemble		91.01%	<b>92.98%</b>

Table 7: The effect of Model Ensemble(development phase).

Backbone	shape	pose	texture	context	weather	occlusion
Ensemble	86.74%	93.02%	96.46%	91.35%	96.58%	93.71%

sively explored using an explicit scheme from either regularization or knowledge transfer. Specifically, a simple yet very effective way is to directly generate new data samples from existing training data, such as mixup, cutmix, copy-paste, crossgrad. Another approach is not to explore sample relationships from the input but to enable the neural network itself to explore sample relationships, such as BatchFormer, which explores sample relationships from a batch perspective. Therefore, we use BatchFormer(BF) to help explore the association between the samples and improve the robustness of the model to identify OOD data. BatchFormer is a model suite that easily loads the overall architecture of the model on which we add BatchFormer to all three models. As shown in Table 5, the scores of each model showed some improvement.

#### 2.1.4 Model Ensemble

Since EMA brings a relatively weak boost to Transformer, we use it only in ConvNeXt-L. We weighted the logits results of the best three model outputs to obtain the best scores for the first stage, as shown in Table 6. As shown in Table 7, we present the results of the best ensemble model in each OOD metric evaluation.

## 2.2 Stage 2

### 2.2.1 Post-processing

We perform an exploratory analysis of the confusion matrix obtained from the fused logits of the individual image outputs by image category calculation. We find several obvious problems that Chair is easily misclassified as Sofa or Dining table, therefore, we can post-process these two categories from the fused logits. The specific approach can be seen in Algorithm 1, where we take Sofa and Chair as examples and correct the labels according to fused logits.

---

**Algorithm 1** Post-processing

---

```
Get all samples with predicted label Sofa as  $D_{Sofa}$ 
for  $D_i$  in  $D_{Sofa}$  do
   $L \leftarrow D_i$  output on the 3 models with fused logits
   $\alpha \leftarrow$  a parameter  $> 1$ 
  if  $L_{Chair} \times \alpha \geq L_{Sofa}$  then
    label = Chair
  end if
end for
```

---

### 2.2.2 Detection

As shown in Table 7, we analyze the best results from the previous stage and find that among the 6 categories of data, the scores for shape and context are lower compared to the other 4 categories, so we focus on these 2 categories. For the shape type, we find that sofa and chair have very similar shape and texture, but sofa have the distinctive feature that they have more than two positions, so we use the detected bounding boxes to help us classify them. Without using any additional dataset, we train a Cascade-RCNN based detection model using only the OOD training set and label information. We correct the predictions of the model based on the aspect ratio of the detection frames.

As shown in Table 8, under the conditions of using post-processing as well as detection for assistance, we achieve the best result **93.64%** in the development phase and the 1st in the Codalab list.

Table 8: The best result in the development phase.

Method	IID Top-1	OOD Top-1(Mean)
Ensemble	91.01%	92.98%
Ensemble+Post-processing+Detection	90.95%	<b>93.64%</b>

Table 9: The effect of Style transfer(development phase).

Backbone	IID Top-1	OOD Top-1(Mean)	Context Top-1
ConvNeXt-L(EMA)	91.09%	91.79%	90.65%
ConvNeXt-L(EMA)+Style transfer	90.74%	<b>92.25%</b>	93.29%

### 2.2.3 Style transfer

As we stated earlier, the context still has a large room for improvement. In the context category of misclassified images, mainly by some crashes in the ocean are misclassified as Boat. obviously, our model mainly focuses on the ocean rather than Airplane itself. As shown in Figure 1, we select some shark images from Imagenet-1K as the ocean background, and then obtain the Airplane in the ocean image by color space conversion of the Airplane and Shark images by traditional machine learning methods, which was significantly improved in the test of ConvNeXt-L, as shown in Table 9. Unfortunately, the method is not used in the final solution because no data of this type exist in the final phase.



Figure 1: Styer transfer of traditional machine learning method.

Table 10: The effect of Model Ensemble(final phase).

Method	IID Top-1	OOD Top-1(Mean)
DeiT-L(BF)+ConvNeXt-L(EMA)+VOLO-D5	91.04%	83.23%
1st round of Pseudo-labeling(Replace ConvNeXt-L)	91.04%	84.12%
1st round of Pseudo-labeling(Replace DeiT-L)	90.86%	84.50%
2st round of Pseudo-labeling(Replace VOLO-D5)	91.04%	<b>85.61%</b>

Table 11: The final results on each OOD index(final phase).

shape	pose	texture	context	weather	occlusion
85.21% (1)	90.32% (1)	68.93%(13)	89.52% (1)	89.20% (1)	97.75% (2)

## 2.3 Stage 3

### 2.3.1 Iterative Pseudo-labeling

As shown in Table 10, we first obtain the optimal results by processing using the best solution of the development phase. Then we perform iterative Pseudo-labeling training. We output the prediction confidence of each image for the above best results, and images with confidence  $> 0.5$  are selected and add to the training set to retrain DeiT and ConvNeXt; then these two models replace the original model to output the new prediction confidence, and images with confidence  $> 0.8$  are selected and add to the training set to retrain VOLO. Finally, the VOLO is replaced by the original one and then ensemble to output the optimal result.

### 2.3.2 Customized post-processing

In the final phase, where the style transfer in the Stage 2 is not valid due to changes in the distribution of the test set, we perform post-processing on the more confused categories in pursuit of higher scores. Our final result ranks 2nd in Codalab, and the final average OOD score is **86.82%**. The specific indicators are shown in Table 11.



- Description of the particularities of the solutions deployed for each of the tracks : This project describes the solution for the classification track only; our team’s solutions for the detection track and the pose estimation track will be presented in two other reports. It is worth noting that we used part of the solution for the detection track to effectively make a significant improvement in the classification track.
- References:
  1. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C] International Conference on Machine Learning. PMLR, 2021: 10347-10357.
  2. Yuan L, Hou Q, Jiang Z, et al. Volo: Vision outlooker for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
  3. Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.
  4. Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[J]. arXiv preprint arXiv:1811.12231, 2018.
  5. Li Y, Yu Q, Tan M, et al. Shape-texture debiased neural network training[J]. arXiv preprint arXiv:2010.05981, 2020.
  6. Hou Z, Yu B, Tao D. BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7256-7266.
  7. Tang Z, Gao Y, Zhu Y, et al. Selfnorm and crossnorm for out-of-distribution robustness[J]. 2020.
  8. Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations[J]. arXiv preprint arXiv:1903.12261, 2019.
  9. Zhao B, Yu S, Ma W, et al. OOD-CV: A Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural

Images[C] Proceedings of the European Conference on Computer Vision (ECCV), 2022.

- Representative image / diagram of the method(s): As shown in Figure 2, this is the overall framework diagram of our approach.

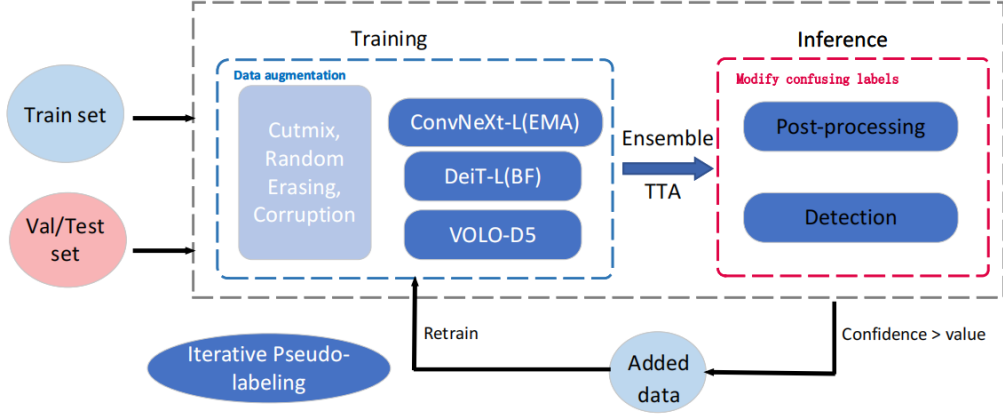


Figure 2: The overall framework diagram of our proposed approach.

### 3 Global Method Description

[\* Indicates method used in competition test results.]

- Total method complexity: The project requires the training of three classification models as well as a detection model, where the classification model requires two iterations and the total complexity should be determined by twice the VOLO of the classification model with the largest parameters.
- Model Parameters: ConvNeXt-L param count: 196M, DeiT-L param count: 310M, VOLO-D5 param count: 294M.

- Run Time: In the case of 4 A30, the training takes about 100 hours and the inference takes only half an hour. The training time can be reduced to less than 48 hours when resources allow.
- Which pre-trained or external methods / models have been used: Only the pre-trained model in the ImageNet-1K dataset was used for the experiments.
- Training description : We first train the ConvNeXt-L, DeiT-L, and VOLO-D5 models, then output the fused and post-processed results as Pseudo-labeling (confidence  $> 0.5$ ), then retrain the ConvNeXt-L and DeiT-L models, replace the original Convnext and Deit models for fusion and post-processing, output the Pseudo-labeling once more (confidence  $> 0.8$  ), and finally retrain the VOLO-D5 model, and finally fuse and post-process it once more.
- Testing description: We infer the 3 models obtained by Pseudo-labeling training and just perform post-processing.
- Quantitative and qualitative advantages of the proposed solution : The effect of our approach has been quantified and analyzed in detail in Chapter 2.
- Results of the comparison to other approaches (if any) : The effect of our approach has been quantified and analyzed in detail in Chapter 2.
- Novelty of the solution and if it has been previously published: First we improve the effect from the model. We use models based on different architectures of CNN or Transformer for fusion, which ensures that the model has both global and local inductive bias, which can greatly improve the robustness of the model.  
Secondly, we use BatchFormer to help explore the association between samples and improve the robustness of the model to recognize OOD

data. Exploring invariant features between images belonging to the same category also helps in robust representation learning.

Finally, our innovations focus on deeper mining of image data, leading to three targeted approaches: using detection to aid classification tasks, style migration based on traditional machine learning methods, and post-processing based on obfuscated category data.

## 4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any).: The fusion method we chose is the fusion of the output layers, where the logits layers of the three models are weighted and fused.
- What was the benefit over the single method? : The model structures we choose are based on CNN or Transformer, respectively. The information of these two types of structures for images is not exactly intersecting, for example, CNN focuses more on local information, while Transformer focuses more on global information, so the fusion can bring a qualitative improvement.
- What were the baseline and the fused methods? : The baseline is a single CNN model, ConvNeXt-L, and the fusion is performed by weighting ConvNeXt-L, DeiT-L, and VOLO-D5 in the ratio of 0.35, 0.35, and 0.3.

## 5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) : Project language: Python language  
Implementation details: four Nvidia A30s with 24G of video memory per gpu. CPU memory is 64G. Convnext is trained in parallel with two cards, and Deit and VOLO are trained in parallel with four cards.

- 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 : Training time: In the case of 4 A30, it takes up to 100 hours of training, if there are more devices, the fastest training can be completed in 48 hours. Test time: In the case of 4 A30, it takes only 30 min to infer the final stage of the dataset, with an inference speed of about 10 imgs/s and a time of 0.1s per image tested.
- Comment the efficiency of the proposed solution(s)? : We believe that the solution is still very effective. First, we integrate the most effective CNN and Transformer family of representative models from different architectures, and achieve excellent results with only two Pseudo-labeling iterations of the model without applying additional datasets.

## 6 Other details

- General comments and impressions of the OOD-CV challenge. : First of all, we find the OOD-CV challenge very interesting and valuable in solving the current interference with tasks such as recognition and detection in real-life scenarios, and the organizers are very nice and prompt in responding to any questions we had.
- Other comments: I hope the organizers will carefully review the code and submissions to determine the winner. Our solution was 93.68 in the development phase (without the use of style transfer and Pseudo-labeling), which was 1st in the development phase, and in the final phase, we rank 2nd. Therefore we question the solution of the winning team in the final phase. At the same time, we are willing to explore the

nature of OOD data to solve this problem, and we hope our solution will be published in IJCV, and we look forward to the pronouncement of the organizer and the chairman. 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.