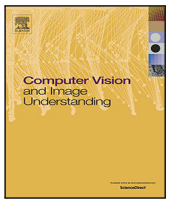
Computer Vision and Image Understanding 271 (2026) 104885

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



MOSAIC: Maximizing out-of-distribution sensitivity via aligned image classification

Hussain Ahmad Madni [a,*,1], Rao Muhammad Umer [c,1], Carsten Marr [d,e,f,g,h,2], Gian Luca Foresti [a,b,2]

[a] Department of Mathematics, Computer Science and Physics (DMIF), University of Udine, Via delle scienze, 206, 33100, Udine, Italy
[b] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT), Consiglio Nazionale delle Ricerche (CNR), Corso Duca degli Abruzzi, 24, 10129, Torino, Italy
[c] Institute of AI for Health, Helmholtz Munich, Germany
[d] Computational Health Center & Helmholtz AI, Helmholtz Munich, Neuherberg, Germany
[e] Department of Medicine III, LMU University Hospital, LMU Medizin, LMU Munich, Germany
[f] Department of Physics, Ludwig-Maximilians-Universität München, Munich, Germany
[g] German Cancer Consortium (DKTK), partner site Munich, Germany
[h] Munich Center for Machine Learning (MCML), Munich, Germany

ARTICLE INFO

Communicated by Nicu Sebe

Keywords:
Classification
Data augmentation
Data-driven
Domain generalization
Out-of-distribution
Test time learning
Transformer

ABSTRACT

Out-of-Distribution (OOD) classification is a domain generalization task in computer vision. Deep learning models are typically developed and tested under the implicit assumption that training and test data are drawn independently and identically distributed (IID) from the same distribution. Overlooking OOD images can lead to poor performance under unseen or adverse viewing conditions, which are common in real-world scenarios. In this work, the proposed solution can be described as a data-driven approach to solve the OOD classification task in computer vision. The proposed approach consists of three stages, a training stage for exploiting labeled source data with different data augmentation strategies using powerful pretrained vision transformer models, an intermediate stage for weighted model ensemble and post-processing strategies, and finally an inference stage for exploiting unlabeled target data by using test-time learning. The proposed data-driven approach enhances the OOD generalization ability of deep models that withstand shifts in nuisances such as shape, pose, context, texture, occlusion, and weather in OOD or rare scenarios. Extensive data-augmentation strategies are used to improve the OOD generalization of deep models across various nuisances. The effectiveness of the proposed approach is evaluated using two standard computer vision benchmarks: ROBIN and a test set provided by the OOD-CV Challenge 2023. The experimental results show that the proposed approach demonstrates a performance improvement of 2.73% the ROBIN test set and achieves accuracy of 94.04% for the Challenge test set in terms of OOD robustness evaluation with classification accuracy. Furthermore, the proposed solution has secured a position within the top three OOD-based rankings on the OOD-CV Challenge Image Classification Leaderboard, 2023.

1. Introduction

In the last decade, deep learning methods have emerged due to their impact and significant improvements in various fields. These methods have been used for different tasks, such as detection, segmentation, and classification. Usually, deep learning methods require a large amount of data, assuming that it is Independent and Identically Distributed (IID) taken from the same distribution (Yang et al., 2024; Kuan and Mueller, 2022). However, in real-world environments, data is commonly non-Independent and Identically Distributed (non-IID) obtained from different distributions. Thus, deep learning models are far from ideal performance in vision tasks when compared to human-level performance. The limited performance of deep models is due to adverse and unusual environmental conditions, referred to as nuisances, applied to training data. However, these realistic conditions cannot fool the human observer. For example, describing an object with a rare and unusual context, pose, weather conditions, shape, and texture is a challenging task for deep learning methods. In such a scenario of out-of-distribution (OOD) data, the robustness and generalization of deep

* Correspondence to: Department of Mathematics, Computer Science and Physics (DMIF), University of Udine, Italy.
E-mail address: hussain.madni@uniud.it (H.A. Madni).
[1] These authors contributed equally to this work.
[2] These authors jointly supervised this work.

https://doi.org/10.1016/j.cviu.2026.104885
Received 23 August 2024; Received in revised form 18 June 2026; Accepted 22 July 2026
Available online 27 July 2026

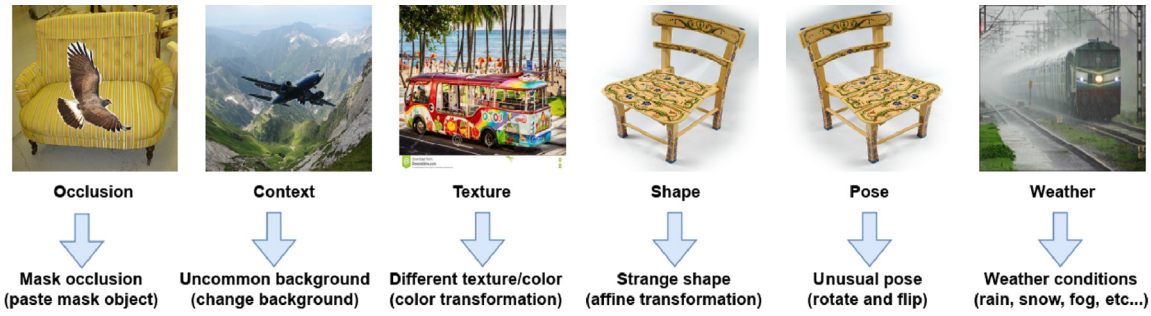


Fig. 1. Data augmentation with different nuisances in the training dataset.

learning models are a challenging classification task that has not been fully solved.

Many existing methods have been introduced that transform training data into real-world scenarios to improve the generalization and robustness of deep models. Such methods can be divided into three categories: (1) Some existing methods based on benchmarks (Hendrycks et al., 2021b; Hendrycks and Dietterich, 2019; Zhang et al., 2024) and algorithms (Ye et al., 2022; Osman and Shehata, 2022; Yan et al., 2024) train deep models on one dataset and test on a different dataset to evaluate the robustness of the trained model. (2) Some approaches generate artificial perturbations known as nuisances, such as partial occlusion (A. Wang et al., 2020), synthetic noise (Hendrycks and Dietterich, 2019), and weather (Michaelis et al., 2019) as a measure of robustness. (3) Some existing approaches record objects for the acquisition of nuisance annotations using synthetic datasets (Kortylewski et al., 2018) or laboratory datasets (Borji et al., 2016). However, such controlled object recordings are feasible for a small number of objects and do not guarantee the transformation of the output to real-world applications.

It is important to train a baseline model using OOD data with various nuisances for robustness and generalization. For this, many data augmentation strategies have been introduced to cope with different domain shifts and distributions. However, unlike traditional data augmentation methods, a mask-level copy-paste data augmentation approach is used in which the foreground or background of an actual image is replaced to generate more data, known as occlusion and context augmentation, respectively. If the background of an image is replaced, the background is known as a task-unrelated object, while the foreground is called task-related object, and vice versa. More specifically, MCTformer (Xu et al., 2022) is used as a weekly supervised semantic segmentation to extract foreground objects from images in the ROBIN dataset (Zhao et al., 2022) and ImageNet-1K dataset (Russakovsky et al., 2015). Then, augmentation of these foreground objects is performed using color affine transformation against texture, pose, and shape domain shifts. Moreover, extracted task-related foreground objects are pasted onto the task-unrelated objects against context domain shift, and task-unrelated foreground objects are pasted onto task-related objects in terms of occlusion domain shift as shown in Fig. 1. Finally, taking advantage of test-time domain adaptation from learning of noisy labels as introduced in SFOD (Li et al., 2021) and SSNLL (Cubuk et al., 2018), the performance of the model is improved. Thus, after pre-training of the model, pseudo-labels (i.e., noisy labels) produced during interference are used on target data for noisy-label learning step.

The proposed approach is evaluated with convolutional networks and non-convolutional transformer models on OOD benchmarks, ROBIN (Zhao et al., 2022) dataset, OOD Track-1 and Track-2 test sets (Classification Track-1 and Track-2 leaderboard)[3] from OOD-CV Challenge 2023 with respect to each nuisance. An incremental approach is used including strong augmentation, pseudo-labeling, model ensemble, model fusion, and test-time learning to train and evaluate a strong generalized model over multiple nuisances. The authors also participated as a team named 'raoumer' in the OOD-CV Challenge 2023 for classification Track-1[4] and Track-2[5] organized by ICCV 2023. Track-1 is based on the ImageNet-1K classification, while Track-2 is based on the ImageNet-22K Self-supervised pretrained classification. Quantitative results and comparisons with other methods have been presented for the Track-1 and Track-2 test sets as given in Tables 4 and 5, respectively. This work has secured a position within the top five rankings for Track-1 and top three for Track-2 based on OOD performance.

This paper presents a novel data-driven approach for OOD image classification that integrates strong data augmentation, model ensemble, post-processing, and test-time learning into a unified three-stage framework. Unlike existing methods that address OOD robustness through a single component or synthetic perturbations, the proposed approach explicitly models individual nuisance factors such as shape, pose, context, texture, occlusion, and weather using mask-level copy–paste augmentation. Furthermore, the method leverages unlabeled target-domain data via ensemble strategies and pseudo-label-based test-time learning, leading to improved robustness under diverse real-world distribution shifts. Extensive evaluations on the ROBIN benchmark and the OOD-CV Challenge 2023 datasets demonstrate the effectiveness and competitiveness of the proposed approach.

Our main contributions are given as follows:

- A data-driven approach is proposed to solve the out-of-distribution classification problem in computer vision.
- The proposed approach solves the out-of-distribution problem by generalizing the labeled source-domain data via strong data augmentation against different nuisances to train a well-generalized model.
- The pre-trained model is extended to adapt to the unlabeled target-domain data via model ensemble, post-processing strategies, and test-time learning.
- The proposed approach is evaluated on out-of-distribution benchmarks, ROBIN dataset (see Table 1) and the OOD Track-1 and Track-2 test sets provided by the OOD-CV Challenge 2023, as given in Table 4 and 5. The experimental results show the superior efficiency of the proposed approach over existing methods.

2. Related works

2.1. Synthetic data for model robustness

Many existing methods (Michaelis et al., 2019; Kurakin et al., 2016; Hendrycks and Dietterich, 2019) have used synthetic data for the

[3] https://www.ood-cv.org/challenge.html

[4] https://tinyurl.com/8kpmasc5

[5] https://tinyurl.com/2s3jkne2

robust analysis of deep learning models. ImageNet-C (Hendrycks and Dietterich, 2019) uses synthetic noise to perturb ImageNet (Deng et al., 2009) test data with motion blur, JPEG compression, and Gaussian noise. Similarly, image perturbation with the addition of synthetic noise to the Pascal-VOC (Everingham et al., 2015) and COCO (Lin et al., 2014) test sets has been used in Michaelis et al. (2019) for the object detection task. In addition to perturbations caused by image processing pipelines, another study (Geirhos et al., 2018) examines the influence of shape and texture biases on deep learning models employing images with artificially replaced textures. To improve robustness and counteract such texture changes and synthetic image noises, linear combinations of original images with strongly augmented images or style transfer (Gatys et al., 2016) as augmentation (Geirhos et al., 2018) have shown a significant influence. However, real-world and 3D nuisances, such as novel poses or shapes of objects, cannot be adopted using synthetic image perturbations. Furthermore, strong augmentation (Hendrycks et al., 2019b) and style transfer (Gatys et al., 2016) are not enough to help with pose and shape variations. Additionally, such benchmarks are constraint-centric and work for a single task. For example, COCO-C (Michaelis et al., 2019)is used to only evaluate the object detection task, and ImageNet-C (Hendrycks and Dietterich, 2019) is used to evaluate only the classification robustness. Another method, DomainBed (Gulrajani and Lopez-Paz, 2020), has been introduced for the generalization of the distribution domain in the classification task. However, the effectiveness and robustness of the proposed method is evaluated in various nuisances such as pose, shape, texture, weather conditions, occlusion, and shape.

2.2. Robustness on real-world data

Many recent methods (Recht et al., 2019; Zhang et al., 2024, 2022) have relied on the collection of images or images from the real world to evaluate deep learning models. A new ImageNet-V2 (Recht et al., 2019) test data set downloaded from Flickr has been created for ImageNet (Deng et al., 2009), which degrades the model performance due to the distribution shift in real images influencing model learning. Using an adversarial filtration approach that excludes all images correctly classified by a predetermined ResNet-50 model (He et al., 2016), ImageNet-A (Hendrycks et al., 2021b) curated a novel test set and demonstrates that these adversarially filtered images can be transferred to other architectures, resulting in a significant performance decrease. ImageNet-A (Hendrycks et al., 2021b) is unable to avoid the nuisance element, while it importantly evaluates the robustness of real-world images. Recently, four out-of-distribution benchmark test datasets have been collected in ImageNet-R (Hendrycks et al., 2021a) with distribution shifts in camera parameters, texture, blur, and geo-location. This method demonstrates that not a single approach can enhance the efficiency of the model in all nuisances. Another method (Tang et al., 2022) explains model learning from unbalanced data with invariant features. In this paper, a robust benchmark is proposed that complements existing datasets by separating and evaluating individual out-of-distribution nuisance factors associated with semantic aspects of an image such as context object, object texture, weather conditions, and shape that enable the proposed method to analyze robustness in different vision tasks because of abundant annotation of data.

2.3. Architectural changes and augmentation to improve robustness

Different approaches, such as Mohseni et al. (2021), have been proposed to reduce the gap between the performance of the model in the real world scenario and given datasets. Such robustness-improving approaches are based on either architectural amendments or data augmentation. For example, adversarial training using image mixtures (Erichson et al., 2022; Hendrycks et al., 2019b; Yun et al., 2019), augmentation in feature space (Hendrycks et al., 2021a), perturbation at training time (Wong et al., 2020), image stylizations (Geirhos et al., 2018), and strong data augmentation (H. Wang et al., 2021; Cubuk et al., 1805) are possible approaches in data augmentation for robustness evaluation. These data augmentation methods have shown effectiveness in model performance with synthetic perturbed images (Hendrycks et al., 2019b; Geirhos et al., 2018). Architectural changes to the model have been implemented to incorporate additional inductive biases for robustness evaluation. In Xie et al. (2019), denoise is performed to improve the adversarial robustness with feature representation. Some nuisances, such as occlusion, can be handled with analysis-by-synthesis methods (Kortylewski et al., 2021; A. Wang et al., 2021) using top-down feedback and a generative object model (Xiao et al., 2020). Moreover, emerging transformer architectures for vision tasks (Shao et al., 2021; Liu et al., 2021; Dosovitskiy et al., 2020) have shown amazing robustness compared to CNNs (Mahmood et al., 2021; Bhojanapalli et al., 2021). The robustness of a model has also been improved by object-centric representations (Wen et al., 2022; Locatello et al., 2020), such as self-supervised representations of out-of-distribution examples (Cui et al., 2022; Zhu et al., 2021; Hendrycks et al., 2019a; Zhao and Wen, 2020).

In summary, existing approaches for improving robustness primarily focus on either data augmentation or architectural modifications, with each addressing only a subset of OOD challenges. While these methods have demonstrated improvements under specific distribution shifts, their effectiveness often varies across different nuisance factors and real-world scenarios. This highlights the need for a more comprehensive strategy that can jointly address multiple sources of distribution shift. Motivated by this observation, we propose a data-driven approach that integrates nuisance-aware augmentation, ensemble learning, post-processing, and test-time adaptation to enhance OOD robustness across diverse conditions.

3. Methodology

The complete setup of the proposed approach with the training and inference stages is shown in Fig. 2. All the steps taken in the proposed approach during the end-to-end training and testing of the model have been described as follows:

3.1. Data augmentation

As shown in Stage 1 of Fig. 2, data augmentation is performed to improve the performance of the classifier in the OOD scenario. Different augmentation methods, including occlusion, context, texture, shape, pose, and weather are used, as shown in Fig. 1. Among all augmentation methods, only texture, shape, pose, and weather augmentation are performed during model training. However, occlusion and context augmentation are performed in the pre-processing stage. For context and occlusion, MCTformer (Xu et al., 2022) is used to generate class-specific attention maps, which are used to segment objects from the source images as shown in Fig. 3. As shown in the upper and lower rows of Fig. 3, two random images are taken from the dataset, and extract objects using class-specific attention maps and generated masks for semantic segmentation. These segmented objects are copied from the source images and pasted on the target images using masks generated by the attention maps to generate augmented images. Thus, context and occlusion are performed with mask-level copy-paste data augmentation. ImageNet (Russakovsky et al., 2015) is used for the context and occlusion augmentation using MCTformer (Xu et al., 2022), where ImageNet-1k images are used as background in context augmentation and as foreground objects in occlusion augmentation. Thus, in context augmentation, each background image is task-unrelated to the training data, and in occlusion augmentation, the foreground objects are task-unrelated to the training data. Moreover, color and random affine transformations are used to perform the augmentation against pose, shape, and texture of foreground objects. Furthermore, other augmentation strategies are stacked such as weather shift, CutMix (Yun

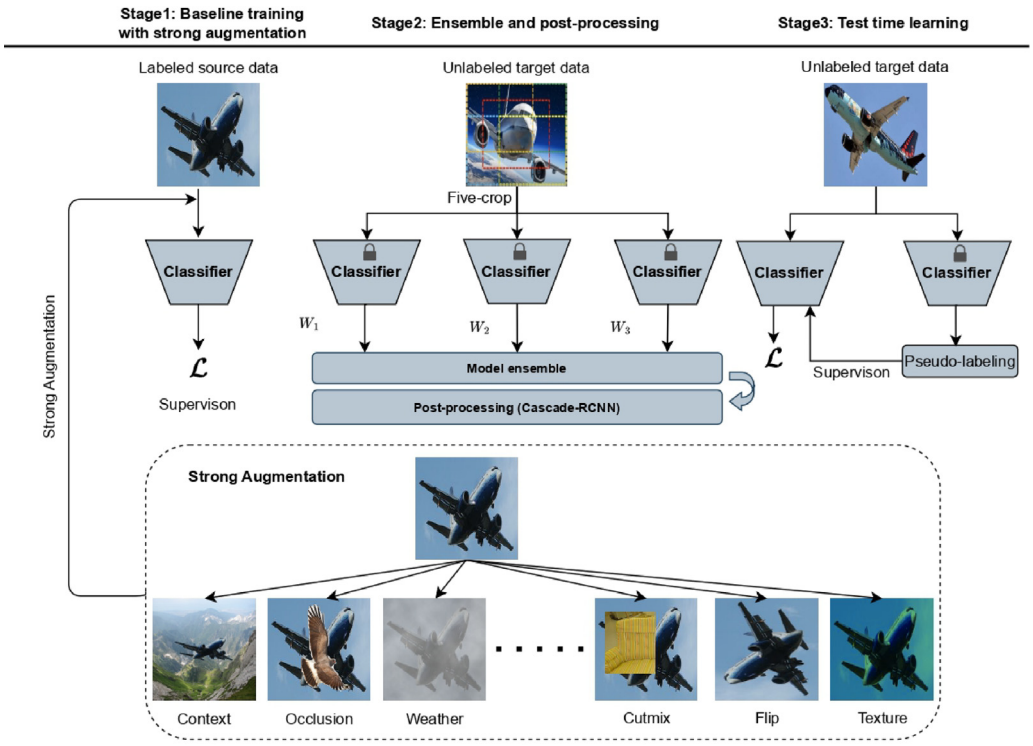


Fig. 2. An overview of the proposed data-driven approach, which consists of three stages: (1) a strong baseline model training with strong augmentation on labeled source data, (2) ensemble, fusion, and post-processing on unlabeled target data, (3) test-time learning with pseudo-labeling on unlabeled target data.

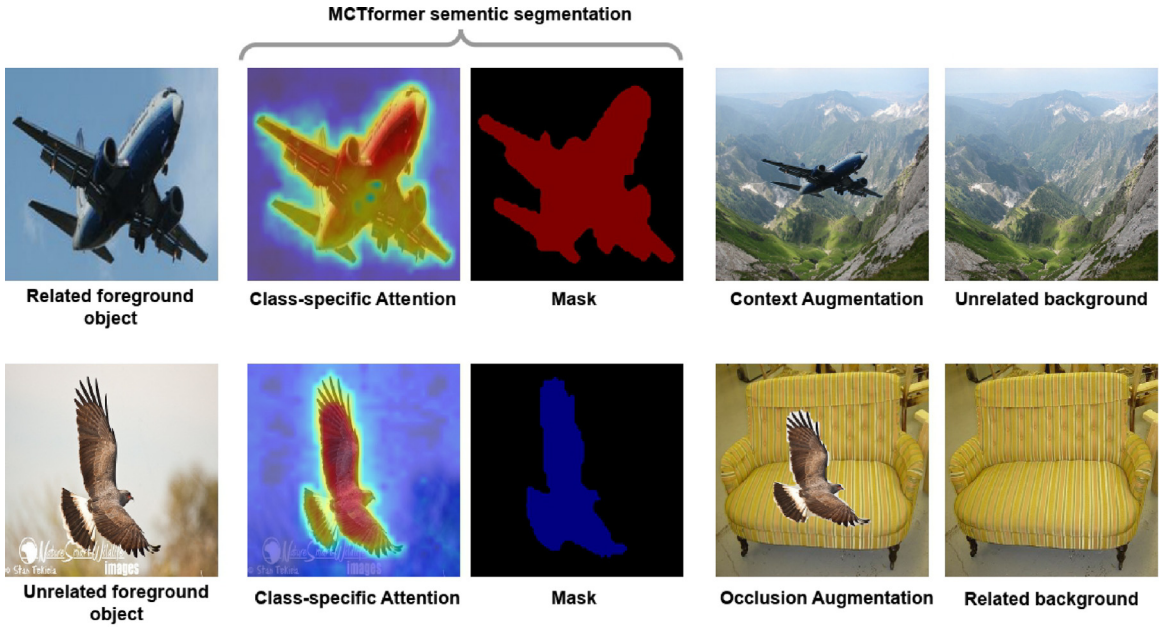


Fig. 3. Context and occlusion augmentation. Objects from the images are extracted using class-specific attention maps and masks produced for the semantic segmentation as introduced in Xu et al. (2022). Then, using these masks, objects are copied from one image and pasted on the other image to perform occlusion and context augmentation.

et al., 2019), and AutoAug (Cubuk et al., 1805) to produce a strong generalized pre-trained model. Thus, source domain data is generalized and adaptive to the target domain by a strong data augmentation, as shown in Fig. 4.

3.2. Ensembles strategies

As shown in Stage 2 of Fig. 2, model-ensemble strategy is applied, where the final logits of the models are weighted and aggregated to

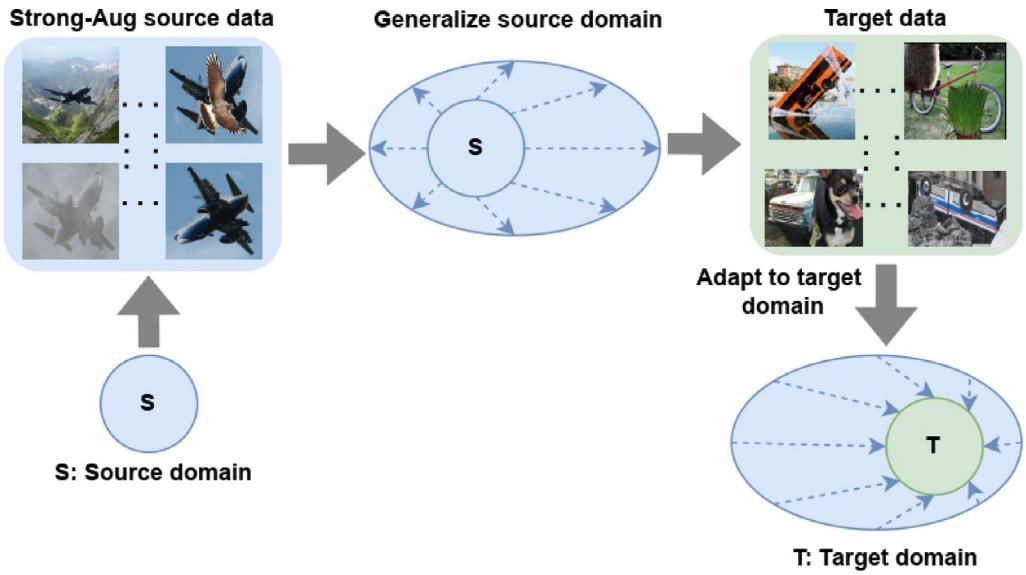


Fig. 4. Transformation of source domain data by a strong augmentation to be adaptive to target domain data through generalization.

improve the model performance. In this process, a weight as a fraction of 1 is assigned to each model based on its individual performance, and then all models are aggregated by adding their logits that are multiplied with their individual and assigned weights. In addition, Test Time Augmentation (TTA) such as FiveCrop for the individual model prediction is used before performing model ensemble. Exponential Moving Average (EMA) is also used only for the convolutional family of models to make them robust to data shift during training, while omitting the non-convolutional (i.e., transformers) family of models as these models already have a strong inductive bias for robustness.

The ensemble weights are determined empirically based on the validation performance of individual models on the ROBIN benchmark. For each test image, logits from all selected models are aggregated using weighted averaging, and the final prediction is obtained from the combined score vector. Test-Time Augmentation (FiveCrop) is applied prior to ensembling to improve prediction stability under distribution shifts.

3.3. Post-processing

After the model ensemble step, during post-processing, as shown in the bottom of Stage 2 in Fig. 2, model scores are analyzed in terms of OOD accuracy against nuisances in the test data. For each nuisance, the categories causing lower scores are filtered. In this case, two categories, chair and sofa, are the cause of the lower performance of the model. Thus, for such categories, post-processing is utilized as used in Guo et al. (2023) to further improve the model performance, where the labels of the filtered categories are corrected with their confidence prediction score as they are mostly misclassified to each other. Moreover, predictions of the model are corrected against given nuisances such as shape, pose, and context, on the basis of the aspect ratio of the detection frames by the cascade RCNN-based (Cai and Vasconcelos, 2018) detection model.

The post-processing rules are derived from an analysis of nuisance-specific confusion matrices on the validation set. Categories exhibiting frequent mutual misclassification, particularly chair and sofa, are corrected using confidence-based label refinement. In addition, object aspect-ratio information obtained from the Cascade R-CNN detector is used as an auxiliary cue to adjust predictions in cases affected by shape, pose, or context shifts.

3.4. Test time learning

As shown in Stage 3 of Fig. 2, noisy label learning (i.e., pseudo-labeling) is performed on unlabeled test data. In this step, the test images with confidence score > 0.5 are selected and added to the training set to retrain the models with higher performance in the previous training. More specifically, the classifier predicts labels (i.e., pseudo labels) for the unlabeled test dataset. These predicted pseudo labels, selected based on a confidence threshold, are then used as part of the training dataset to enhance the classifier's performance.

During test-time learning, pseudo-labels are generated for unlabeled target-domain samples and filtered using a confidence threshold of 0.5. The selected samples are incorporated into the training set, and the classifier is further fine-tuned for a small number of adaptation steps. This procedure enables the model to better align with the target-domain distribution while avoiding excessive influence from unreliable pseudo-labels.

4. Experiments

4.1. Datasets

Description of the training, validation, and test data is given as follows.

4.1.1. Train and validation data

The training dataset consists of a fixed set of 10 object categories, including aeroplan, bus, car, train, boat, bicycle, motorplane, chair, dining table, and sofa, from the PASCAL VOC 2012 (Everingham et al., 2010) and ImageNet (Russakovsky et al., 2015) datasets, provided by OOD-CV challenge 2023. The model's performance is evaluated on the OOD benchmark ROBIN (Zhao et al., 2022) validation set during the development phase of the challenge (i.e., for both Track-1 and Track-2).

4.1.2. Test data

The test data consist of six nuisances (i.e., shape, pose, context, texture, occlusion, and weather), where each test sample is subjected to an OOD shift in a specific nuisance. The performance of the model is evaluated on test sets during the development and final testing phases in both tracks (i.e., Track-1 and Track-2) of the OOD-CV Challenge 2023. The proposed approach is also evaluated using ROBIN (Zhao et al., 2022) test set to compare with the existing OOD based methods.

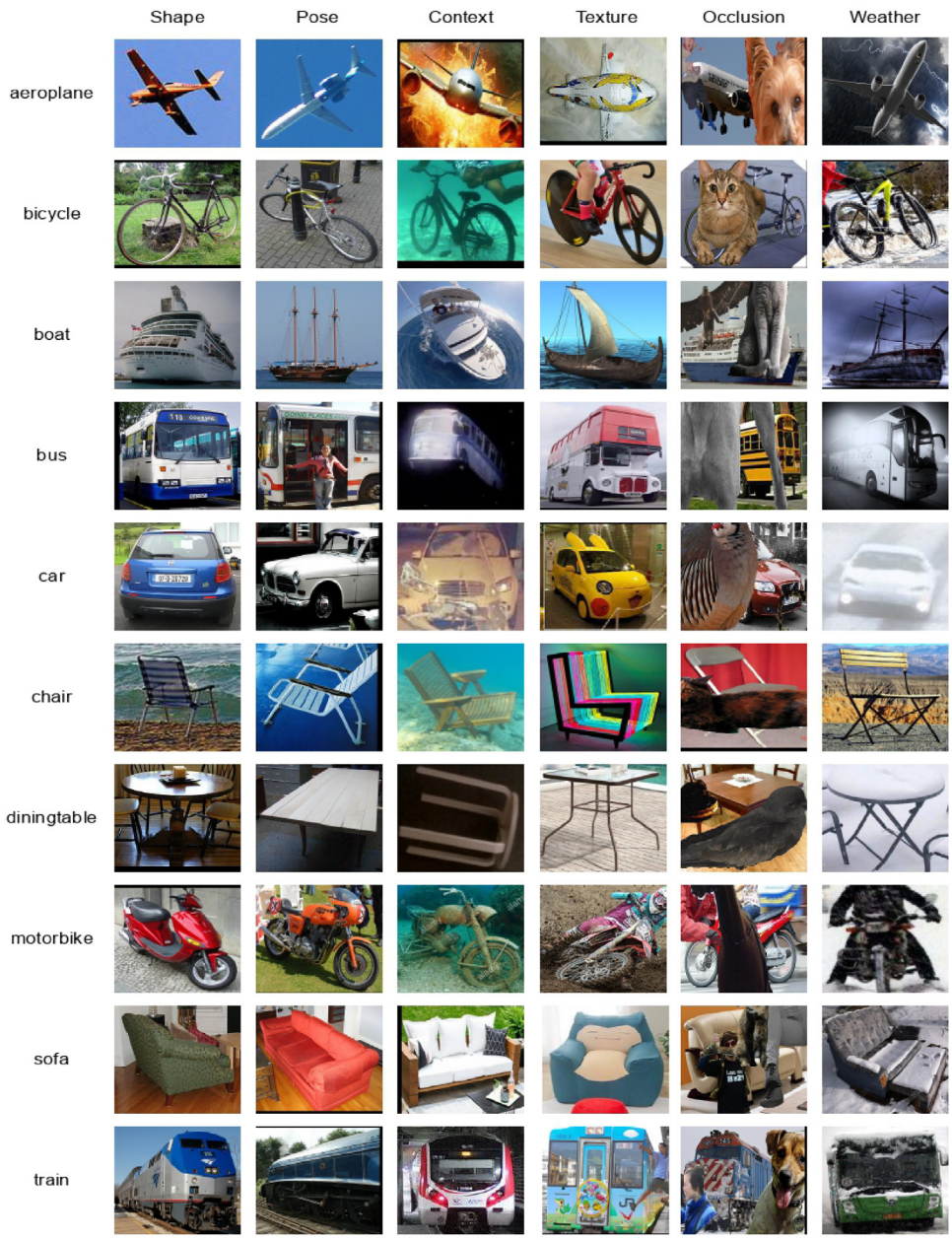


Fig. 5. Some examples from all categories ROBIN (Zhao et al., 2022) dataset, where each image of a category is visualized with a nuisance.

The test set contains 10 categories affected by different nuisances. Thus, some examples from all categories, where each image in the test set reflects a nuisance as visualized in Fig. 5.

4.2. Training description

In the experiments, transformer models are selected because of their global attention mechanism making them flexible, scalable, and generalized across diverse data and various vision tasks. Moreover, non-convolutional models have recently emerged and showed superior performance as compared to convolution-based models. However, convolution-based models are also used to validate and compare the performance of the proposed method. In all the experiments, pre-trained models trained with ImageNet-1k (Russakovsky et al., 2015) dataset are used in all the experiments for the ImageNet-1k classification track (i.e., Track-1) and with ImageNet-22k (Russakovsky et al., 2015) dataset for self-supervised pre-train classification track (i.e., Track-2). Five models are selected for fine-tuning on the given training data, two from CNN-based architecture (i.e., ResNet50 (He et al., 2016), ConvNeXt-L (Liu et al., 2022b)), and three from non-convolutional transformers-based architectures (i.e., DeiT-L (Touvron et al., 2021), VOLO-D5 (Yuan et al., 2022), and Swinv2-B (Liu et al., 2022a)). Different training strategies are chosen for these two types of models using an initial learning rate of $1e^{-2}$, for CNN-based models, and $3e^{-4}$ for transformer models. In addition, cross-entropy loss with label smoothing is used as the training loss function and a stochastic gradient descent (SGD) optimizer with momentum.

All models are trained for 300 epochs using a batch size of 128. We use the SGD optimizer with momentum 0.9, weight decay 2×10^{-5}, and an initial learning rate of 0.05. The learning rate is scheduled using a cosine decay scheduler with 3 warm-up epochs and a minimum learning rate of 1×10^{-6}. Label smoothing with a factor of 0.1 is applied during training. Augmentation methods, including random resized cropping, horizontal and vertical flipping, color jittering, and CutMix augmentation, are implemented. All experiments are conducted using PyTorch on an NVIDIA GeForce RTX 3090 GPU with 24 GB of

Table 1
Stage 1 (baseline model training) results (Top1 accuracy) on the OOD benchmark ROBIN (Zhao et al., 2022) dataset with convolutional and non-convolutional transformer models evaluated for both IID and OOD performance.

Methods	IID Top-1	OOD Top-1						
		Shape	Pose	Context	Texture	Occlusion	Weather	Avg.
ResNet50 (He et al., 2016)	76.88	72.17	75.72	75.84	87.97	86.57	85.38	80.61
ConvNeXt-L (Liu et al., 2022b)	81.68	77.88	80.56	85.54	89.56	89.82	90.50	88.19
DeiT-L (Touvron et al., 2021)	**82.79**	80.98	84.64	88.34	93.78	**95.89**	95.34	89.83
VOLO-D5 (Yuan et al., 2022)	81.20	**81.86**	**86.32**	**91.00**	**93.98**	92.78	**95.99**	**90.32**
Swinv2-B (Liu et al., 2022a)	81.55	78.70	83.81	87.12	93.45	91.68	94.58	88.22

Table 2
The effect of model ensemble, test time learning, and post-processing on the performance of the model generalization evaluated on the ROBIN benchmark test set in terms of IID and OOD average accuracy. The pre-trained models are only trained on ImageNet-1k.

Method	Weight	IID Top-1	OOD Top-1 (Avg.)
DeiT-L (Touvron et al., 2021)	0.4	**82.79** %	89.83%
VOLO-D5 (Yuan et al., 2022)	0.4	81.20%	**90.32** %
Swinv2-B (Liu et al., 2022a)	0.2	81.55%	88.22%
Ensemble		82.50%	**92.78** %
DeiT-L (test time learning)		**82.83** %	**91.21** %
Swinv2-B (test time learning)		81.62%	88.36%
DeiT-L (Ensemble + Post-processing)		**82.91** %	**93.05** %

Table 3
The effect of model ensemble and post-processing on the model performance evaluated on OOD-CV Challenge 2023 test set in terms of IID and OOD average accuracy when the pre-trained models are trained on ImageNet-1k, ImageNet-21k and ImageNet-22k.

Method	Weight	IID Top-1	OOD Top-1 (Avg.)
ConvNeXt-L (22k)	0.1	81.16%	89.83%
DeiT-L (21k)	0.4	83.55%	92.66%
VOLO-D5 (1K)	0.2	81.20%	90.32%
Swinv2-L (22k) (Liu et al., 2022a)	0.3	82.48%	90.72%
Ensemble		83.51%	93.87%
Ensemble + Post-processing		**83.74** %	**94.04** %

memory. To improve memory efficiency, gradient accumulation with 8 accumulation steps is employed during training.

4.3. Quantitative results

First, the proposed approach is evaluated for IID Top-1 accuracy and OOD Top-1 accuracy with respect to all the nuisances given in the ROBIN test dataset, as given in Table 1. The accuracy of each convolutional and non-convolutional model is measured for IID and OOD against each nuisance and the average accuracy of all nuisances in the last column. From these initial results, Top three (i.e., transformers) models are selected with higher performance for inclusion in the weighted model ensemble strategy to improve the OOD accuracy. Each model is assigned a weight value in between 0 and 1 to make a total 1 according to its performance score. Finally, these weighted model logits are ensemble for the final output. The ensemble strategy does not improve the IID accuracy but significantly improves the OOD average accuracy from 90.32% to 92.78% as given in Table 2.

Furthermore, the test-time learning method is used to improve the generalization ability of the model with pseudo-labels. In test-time learning, the model is retained according to a threshold value of its confidence score higher than 0.5. As shown in Table 2 (i.e., lower half), test-time learning does not contribute significantly to the performance improvement in the case of IID as it only improves the performance of DeiT-L from 82.79% to 82.83% and that of Swinv2-B from 81.55% to 81.62%. However, it significantly improves the performance of the model in the OOD scenario, improving the accuracy of DeiT-L from 89.83% to 91.21% and that of Swinv2-B slightly from 88.22% to 88.36%. As OOD is more important for the model generalization, the contribution of test-time learning with the DeiT-L model is of great importance.

In the last step, the proposed approach is evaluated on the ROBIN test set using post-processing with ensemble and fusion strategies. The accuracy of the DeiT-L model is measured using these strategies as given in the bottom (i.e., last) row of Table 2, and the performance increase is observed from 82.50% to 82.91% and from 92.78% to 93.05% for IID and OOD respectively.

The proposed approach is also evaluated on the test data provided by the OOD-CV Challenge 2023. Convolutional ConvNeXt-L(22k) and non-convolutional transformer models, including DeiT-L(21k), VOLO-D5(1k), and Swinv2-B(22k) are also used in the experiments. ConvNeXt-L(22k) and Swinv2-B(22k) are pre-trained on ImageNet-22k while DeiT-L(21k) and VOLO-D5(1k) are pre-trained on ImageNet-21k and ImageNet-1k respectively. In such a scenario, the weights are assigned to these 4 models for the ensemble strategy. The average performance of each model is given in Table 3. The IID Top-1 accuracy and the average OOD Top-1 accuracy are measured for each model. Ensemble and post-processing strategies are applied to improve the performance of the proposed approach. The results show the increase in both IID and OOD Top-1 accuracy when applying ensemble and post-processing in the bottom of Table 3.

4.3.1. ImageNet-1k classification track (challenge track-1)

The transformer models pre-trained on ImageNet-1k are used, and fine tuning is performed for the ImageNet-1k classification track of the challenge. The Top-1 accuracy of the IID and each nuisance is computed to compare the results of the proposed approach with other methods, as shown in Table 4 taken from the leaderboard-ranking of the challenge.

Table 4
Final results (Top1 accuracy) calculated on the OOD Phase-2 test set (Classification Track-1 of Challenge leaderboard) with each nuisances.

Team Name	IID Top-1	OOD Top-1						
		Shape	Pose	Context	Texture	Occlusion	Weather	Avg.
USTC-IAT-United	0.93	0.87	0.90	0.90	0.77	0.98	0.90	$0.89_{(1)}$
Zhongqiang.Zhang	0.97	0.85	0.89	0.91	0.78	0.97	0.88	$0.88_{(2)}$
_Shaun	0.97	0.84	0.88	0.85	0.85	0.96	0.86	$0.87_{(3)}$
aiisfun	0.96	0.86	0.84	0.87	0.85	0.95	0.86	$0.87_{(3)}$
Yuqi_Li	0.95	0.86	0.89	0.90	0.73	0.98	0.87	$0.87_{(3)}$
jfdklsajflasdjfa	0.96	0.85	0.89	0.89	0.73	0.96	0.87	$0.86_{(4)}$
king	0.92	0.86	0.88	0.88	0.72	0.98	0.87	$0.86_{(4)}$
Dark_House	0.93	0.85	0.94	0.88	0.67	0.98	0.86	$0.86_{(4)}$
akun	0.97	0.84	0.78	0.85	0.87	0.93	0.86	$0.85_{(5)}$
LeeeeL	0.96	0.83	0.85	0.82	0.85	0.92	0.82	$0.85_{(5)}$
raoumer	**0.92**	**0.82**	**0.87**	**0.86**	**0.73**	**0.95**	**0.84**	$\mathbf{0.85_{(5)}}$

Table 5
Final results (Top1 accuracy) on the OOD Phase-2 test set (Classification Track-2 of Challenge Self-supervised pretrain leaderboard) with each of the nuisances during the test phase of the Challenge.

Team Name	IID Top-1	OOD Top-1						
		Shape	Pose	Context	Texture	Occlusion	Weather	Avg.
ZK_	0.94	0.90	0.93	0.93	0.86	0.98	0.90	$0.92_{(1)}$
_Shaun	0.96	0.87	0.92	0.90	0.87	0.96	0.89	$0.90_{(2)}$
Yuqi_Li	0.95	0.85	0.89	0.89	0.75	0.97	0.87	$0.87_{(3)}$
raoumer	**0.94**	**0.85**	**0.90**	**0.89**	**0.75**	**0.97**	**0.85**	$\mathbf{0.87_{(3)}}$
aiisfun	0.96	0.83	0.80	0.85	0.89	0.90	0.84	0.85
LeeeeL	0.95	0.80	0.85	0.81	0.83	0.95	0.80	0.84
Marry	0.90	0.82	0.89	0.85	0.62	0.97	0.85	0.83
Bee_666	0.91	0.79	0.84	0.81	0.81	0.92	0.81	0.83
king	0.89	0.76	0.86	0.82	0.60	0.96	0.80	0.80
jackeyguo	0.92	0.79	0.82	0.78	0.69	0.90	0.76	0.79

The last column of the table shows the average Top-1 accuracy of all OOD nuisances. The proposed method is in the top 5 rankings for OOD average accuracy in Track-1 with the team name as 'raoumer'. The results given in Table 4 have been sorted by the OOD average accuracy in descending order.

4.3.2. ImageNet-22k self-supervised pretrain classification track (challenge track-2)

During the self-supervised pre-train classification track of the challenge, the models pre-trained on the ImageNet-22K (Russakovsky et al., 2015) dataset are fine-tuned. The proposed approach is evaluated on the test set provided by the challenge and compare the Top-1 accuracy of the IID and each nuisance with the average accuracy of the OOD, as shown in Table 5.

4.3.3. Comparison with existing methods

To evaluate the proposed approach and compare with the existing similar OOD classification methods, further experiments are performed using ROBIN (Zhao et al., 2022) OOD benchmark dataset, having different distribution and domain shifts. Some examples of the test set with different nuisances are shown in Fig. 5. The comparison results of the proposed approach with existing methods are summarized in Table 6. The results witnessed the leading performance of the proposed method in IID and all nuisances except texture. The reason behind the lower accuracy for texture is due to complex visual patterns of foreground objects that undergo significant alterations. Thus, normal patterns extracted as features from training images cannot represent such complex patterns. However, this problem can be solved by combining the proposed approach with an alternate augmentation method. Moreover, OOD top-1 accuracy is higher than other baselines throughout the nuisances and in the average.

4.4. Discussion and analysis

4.4.1. Ablation study

The proposed method is an incremental approach involving pre-processing and post-processing steps. For the selection of suitable backbone, different experiments are performed with top convolutional and non-convolutional models that have been used in computer vision. Two different benchmark datasets are used to evaluate the best selected model, as given in Table 1, Table 2, and Table 3. In Table 1, DeiT-L (Touvron et al., 2021) performs well in case of IID, whereas VOLO-D5 (Yuan et al., 2022) performs better in OOD scenario. However, while using ensemble and post-processing, the overall performance of DeiT-L (Touvron et al., 2021) is better in both IID and OOD scenarios, as given in Table 2. Therefore, DeiT-L (Touvron et al., 2021) is selected as the backbone that performs better compared to other convolutional networks and transformers, and evaluate the proposed approach on other dataset provided by OOD-CV Challenge 2023, as summarized in Table 3. The better performance of the proposed approach is observed in Table 3;, however, the competition results are also presented from the OOD-CV Challenge leaderboard sorted by OOD performance in Tables 4 and 5, where the proposed method is in the top 5 and 3 ranks, respectively.

The superior performance of non-convolutional transformers (i.e., DeiT-L in this case) over convolutional models is due to their architecture. Transformers process image patches, known as tokens, to extract features. Their robustness comes from the self-attention mechanism, which effectively aggregates image information. Unlike convolutional architectures, transformers employ a layered, non-convolutional structure that examines the average distance among learned weights. The comparison of attention distances and their variability across layers reveals a near-uniform distribution throughout the network's depth. This unique characteristic improves the ability of transformers to capture contextual relationships for image interpretation, setting them apart from traditional convolution-based architectures.

4.4.2. Hardware and complexity

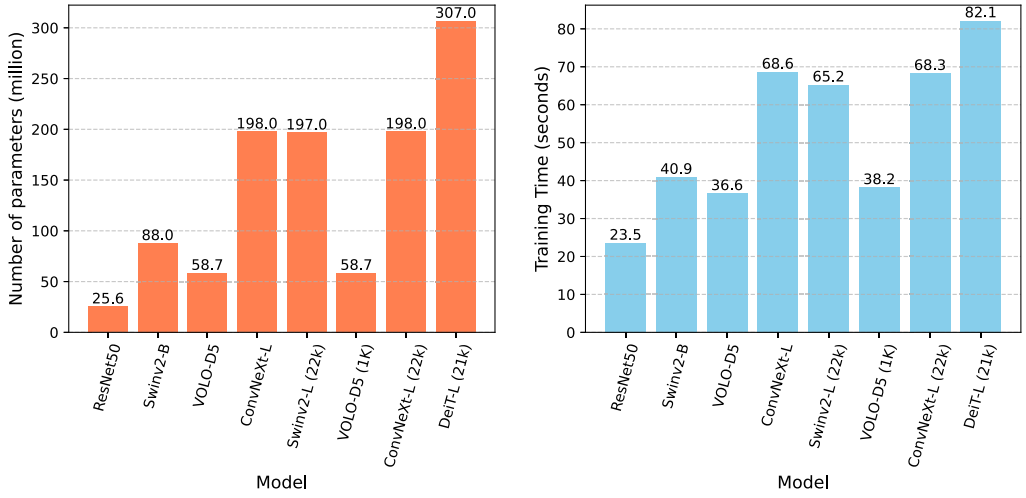
The proposed approach is applied using the PyTorch library for all experiments on NVIDIA GeForce RTX 3090 GPU and i7-8700 CPU installed in windows-11 operating system, where GPU has built-in 24 GB memory and CPU has 50 GB memory.

The performance of each model is evaluated based on the time taken during training. For this, an approximate number of parameters is presented for each model and compare both convolutional neural networks

Table 6
Final results (Top1 accuracy) on the OOD benchmark ROBIN (Zhao et al., 2022) dataset comparing the proposed approach with existing methods under both IID and OOD settings.

Methods	Eval. Protocol	IID Top-1	OOD Top-1						
			Shape	Pose	Context	Texture	Occlusion	Weather	Avg.
FACT (Xu et al., 2023)	SI	0.93	0.83	0.88	0.82	0.91	0.67	0.85	0.83
CutMix (Yun et al., 2019)	SI	0.90	0.82	0.88	0.80	0.89	0.61	0.84	0.81
MixUp (Zhang et al., 2018)	SI	0.90	0.82	0.88	0.79	0.89	0.58	0.82	0.80
ERM (Vapnik, 2013)	SI	0.91	0.83	0.89	0.78	0.91	0.67	0.82	0.82
RSC (Huang et al., 2020)	SI	0.90	0.83	0.88	0.82	0.91	0.69	0.85	0.83
JiGen (Carlucci et al., 2019)	SI	0.90	0.79	0.83	0.84	0.89	0.62	0.81	0.80
MASF (Dou et al., 2019)	SI	0.85	0.79	0.87	0.85	0.87	0.59	0.80	0.80
Epi-FCR (Li et al., 2019)	SI	0.87	0.80	0.86	0.86	0.79	0.63	0.82	0.79
MMD-AAE (H. Li et al., 2018)	SI	0.89	0.81	0.88	0.82	0.80	0.67	0.80	0.80
CIDDC (Y. Li et al., 2018)	SI	0.86	0.83	0.84	0.86	0.85	0.64	0.77	0.80
StableNet (Zhang et al., 2021)	SI	0.88	0.82	0.88	0.86	0.85	0.65	0.80	0.81
MSAM (Li et al., 2023)	SI	0.91	0.80	0.86	0.85	0.86	0.68	0.81	0.81
MiRe (Chen et al., 2022)	SI	0.87	0.83	0.82	0.87	0.85	0.61	0.80	0.80
EISNet (S. Wang et al., 2020)	SI	0.89	0.78	0.87	0.83	0.86	0.65	0.78	0.80
CFFDL (Li et al., 2024)	SI	0.89	0.80	0.87	0.84	0.81	0.68	0.79	0.80
MMLD (Matsuura and Harada, 2020)	SI	0.88	0.83	0.88	0.87	0.87	0.64	0.82	0.82
MOSAIC (Stage-1)	SI	0.93	0.83	**0.90**	**0.88**	0.74	**0.85**	**0.85**	**0.84**
MOSAIC (Stage-1 + Stage-2)	SI-E	0.93	**0.84**	**0.90**	**0.88**	0.74	**0.90**	**0.85**	**0.85**
MOSAIC (Full)	TTA	**0.94**	**0.85**	**0.90**	**0.89**	0.75	**0.97**	**0.85**	**0.87**

SI: standard inference without test-time data; **SI-E**: standard inference with ensemble; **TTA**: test-time adaptation using unlabeled test data.



(a) Number of parameters in millions for each model used in the experiments.

(b) Average time taken by each model during one training epoch.

Fig. 6. Comparison of convolutional neural network and non-convolutional transformer model based on number of parameters and training time. (a) Approximate number of parameters (in million) for each model trained ImageNet-1K and Imagenet-22K as given on x-axis. (b) Training time (in seconds) of each model during one training epoch, where models are pretrained on ImageNet-1K and ImageNet-22k as shown on x-axis.

and non-convolutional transformer models, as shown in Fig. 6-(a). Moreover, the experiments are performed using the ROBIN dataset to measure the training time during one training epoch for each model, as shown in Fig. 6-(b).

Table 7 summarizes the computational complexity of the proposed system by reporting the number of parameters, single-forward FLOPs, and per-image inference latency (batch size = 1) for each backbone classifier and for their ensemble. The ensemble configuration includes all four backbone models operating simultaneously, with ensemble weights affecting only score aggregation. Cascade R-CNN is reported separately as an optional post-processing module applied after classification since its FLOPs depend on image resolution and proposal count, measured latency is reported instead. Test-time learning (TTL) introduces no additional parameters and performs a small number of adaptation steps offline, and therefore does not affect standard inference-time latency. This analysis highlights the trade-off between computational cost and robustness gains offered by the full system. However, our primary objective is to maximize robustness under challenging distribution shifts, rather than to optimize for minimal system complexity or inference latency. Accordingly, the proposed system is not intended to be directly comparable to single-model or efficiency-focused OOD methods.

4.4.3. Qualitative analysis and benchmark scope

While Class Activation Maps (CAMs) and feature heatmaps can provide intuitive qualitative insight into model behavior, their interpretation in OOD settings is highly dependent on the underlying architecture. In particular, attention maps produced by transformer-based models are fundamentally different from CAMs derived from convolutional networks, making direct visual comparison across heterogeneous methods unreliable. To nevertheless provide qualitative insight into the proposed approach, Fig. 7 presents attention visualizations of the top-performing DeiT-L model under different nuisance conditions on the ROBIN test set. The visualizations demonstrate that, despite

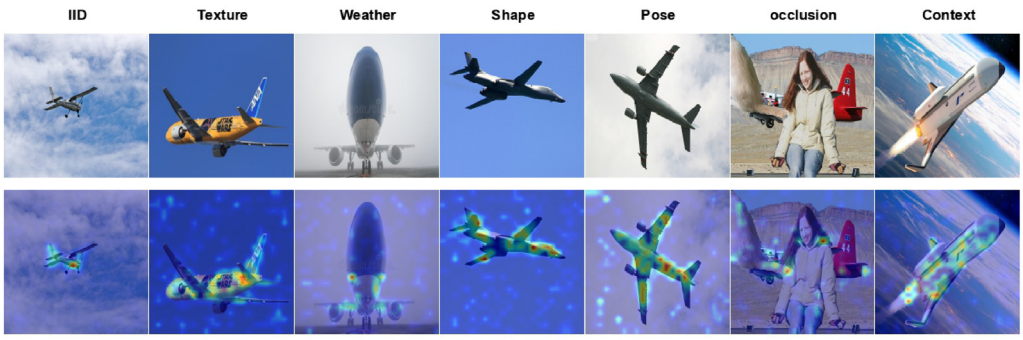


Fig. 7. Attention visualizations of the top-performing DeiT-L model under each nuisance condition are shown. The top row presents input images from the ROBIN test set, while the bottom row shows the corresponding attention maps, illustrating the model's focus on semantically relevant regions under OOD shifts.

Table 7
A summary of parameter count, single-forward FLOPs, and per-image latency (batch size = 1) for individual backbone classifiers and their ensemble (all four models used simultaneously; ensemble weights affect score aggregation only). Cascade R-CNN is applied as a post-processing module after classification and is reported separately; its FLOPs depend on image resolution and proposal count, and therefore measured latency is reported. Test-time learning (TTL) introduces no additional parameters and performs N = 4 adaptation steps (12×forward FLOPs) offline, without affecting standard inference latency.

Inference	Model	Parameters (M) ↓	FLOPs (G) ↓	Latency (ms) ↓
Baseline	ConvNeXt-L	197.77	34.40	10.73
	DeiT-L	306.87	59.70	12.69
	VOLO-D5	58.74	70.29	26.65
	Swinv2-L	196.68	35.10	26.95
Ensemble	Ensemble	994.28	199.49	77.02
Post-processing	Cascade RCNN	41.76	Resolution-based	27.19
TTL	–	–	–	Offline

severe distribution shifts, the model consistently attends to semantically meaningful object regions rather than spurious background cues. The method emphasizes that these attention maps are not intended for method-level comparison with the baselines in Table 6, as most of those approaches are CNN-based or rely on standard augmentation strategies and do not employ attention mechanisms. Therefore, the primary evaluation of robustness in this work remains quantitative and nuisance-specific, as supported by controlled benchmarks and extensive experimental results.

While detection-based correction can further improve robustness in some cases, it also introduces practical constraints. The detection-based post-processing relies on the availability of reliable object detections and assumes a degree of alignment between the detector's label space and the classification categories. Consequently, its effectiveness may degrade in challenging OOD scenarios where object detection becomes unreliable. For this reason, Cascade R-CNN is treated as an optional and replaceable heuristic rather than a core component of the proposed approach. The primary robustness gains of the proposed method stem from the training-stage nuisance-aware learning, and the overall framework does not depend on any specific detector architecture; the post-processing step can be omitted or substituted without affecting the applicability of the method.

5. Conclusion

A data-driven approach has been introduced to solve an out-of-distribution classification problem in the computer vision domain. The proposed approach is divided into three stages. In the training stage, we exploit labeled source data with different data augmentation by using pre-trained vision transformer models. In the intermediate stage, we exploit the unlabeled target data using model ensemble strategies and post-processing techniques. In the final stage, test-time learning is performed during inference. The proposed approach is evaluated on standard benchmark computer vision datasets and achieve better results in terms of OOD Top-1 accuracy. The proposed approach deepens our understanding of the robustness of the model under different domain shifts in the real-world scenario.

CRediT authorship contribution statement

Hussain Ahmad Madni: Writing – original draft, Visualization, Methodology. **Rao Muhammad Umer:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Carsten Marr:** Writing – review & editing, Supervision, Investigation. **Gian Luca Foresti:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper was partially supported by the FVG Project "Supporting the diagnosis of rare diseases (MR) through artificial intelligence" (2023-27) (Project A with CUP: F53C22001770002 and Project B with CUP F53C22001780002). This paper was partially supported by University of Udine project on "Piano Stretegico Dipartimentale on Artificial Intelligence" (PSD-AI) (2022-25) project at the University of Udine. C. Marr received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreements 866411, 101113551, and 101213822) and support from the Hightech Agenda Bayern.

Data availability

Data will be made available on request.

References

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A., 2021. Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10231–10241.

Borji, A., Izadi, S., Itti, L., 2016. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2221–2230.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T., 2019. Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2229–2238.

Chen, C., Tang, L., Liu, F., Zhao, G., Huang, Y., Yu, Y., 2022. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. Adv. Neural Inf. Process. Syst. 35, 33302–33315.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 1805. Autoaugment: Learning augmentation policies from data. arxiv 2018. arXiv preprint arXiv:1805.09501.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2018. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501.

Cui, Q., Zhao, B., Chen, Z.-M., Zhao, B., Song, R., Zhou, B., Liang, J., Yoshie, O., 2022. Discriminability-transferability trade-off: An information-theoretic perspective. In: European Conference on Computer Vision. Springer, pp. 20–37.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B., 2019. Domain generalization via model-agnostic learning of semantic features. Adv. Neural Inf. Process. Syst. 32.

Erichson, N.B., Lim, S.H., Xu, W., Utrera, F., Cao, Z., Mahoney, M.W., 2022. NoisyMix: Boosting model robustness to common corruptions. arXiv preprint arXiv:2202.01263.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. 111, 98–136.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88, 303–338.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.

Gulrajani, I., Lopez-Paz, D., 2020. In search of lost domain generalization. arXiv preprint arXiv:2007.01434.

Guo, Y., Shi, X., Chen, W., Yang, S., Xie, D., Pu, S., Zhuang, Y., 2023. 1St place solution for ECCV 2022 OOD-CV challenge image classification track. arXiv preprint arXiv:2301.04795.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al., 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349.

Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.

Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019a. Using self-supervised learning can improve model robustness and uncertainty. Adv. Neural Inf. Process. Syst. 32.

Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B., 2019b. Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D., 2021b. Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271.

Huang, Z., Wang, H., Xing, E.P., Huang, D., 2020. Self-challenging improves cross-domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, pp. 124–140.

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T., 2018. Empirically analyzing the effect of dataset biases on deep face recognition systems. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2093–2102.

Kortylewski, A., Liu, Q., Wang, A., Sun, Y., Yuille, A., 2021. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. Int. J. Comput. Vis. 129, 736–760.

Kuan, J., Mueller, J., 2022. Back to the basics: Revisiting out-of-distribution detection baselines. In: ICML.

Kurakin, A., Goodfellow, I., Bengio, S., 2016. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., Zhuang, Y., 2021. A free lunch for unsupervised domain adaptive object detection without source data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (10), pp. 8474–8481.

Li, J., Li, Y., Wang, H., Liu, C., Tan, J., 2023. Exploring explicitly disentangled features for domain generalization. IEEE Trans. Circuits Syst. Video Technol. 33 (11), 6360–6373.

Li, H., Pan, S.J., Wang, S., Kot, A.C., 2018. Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5400–5409.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D., 2018. Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 624–639.

Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.-Z., Hospedales, T.M., 2019. Episodic training for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1446–1455.

Li, S., Zhao, Q., Sun, B., Wang, X., Zou, Y., 2024. Domain generalization via causal fine-grained feature decomposition and learning. Comput. Electr. Eng. 119, 109548.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022a. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T., 2020. Object-centric learning with slot attention. Adv. Neural Inf. Process. Syst. 33, 11525–11538.

Mahmood, K., Mahmood, R., Van Dijk, M., 2021. On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7838–7847.

Matsuura, T., Harada, T., 2020. Domain generalization using a mixture of multiple latent domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (07), pp. 11749–11756.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W., 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484.

Mohseni, S., Wang, H., Yu, Z., Xiao, C., Wang, Z., Yadawa, J., 2021. Practical machine learning safety: A survey and primer. 4, arXiv preprint arXiv:2106.04823.

Osman, I.I., Shehata, M.S., 2022. Few-shot learning network for out-of-distribution image classification. IEEE Trans. Artif. Intell. 4 (6), 1579–1591.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V., 2019. Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning. PMLR, pp. 5389–5400.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Shao, J., Wen, X., Zhao, B., Xue, X., 2021. Temporal context aggregation for video retrieval with contrastive learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3268–3278.

Tang, K., Tao, M., Qi, J., Liu, Z., Zhang, H., 2022. Invariant feature learning for generalized long-tailed classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022. Springer Nature Switzerland, Cham, pp. 709–726.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Wang, A., Kortylewski, A., Yuille, A., 2021. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. arXiv preprint arXiv:2101.12378.

Wang, A., Sun, Y., Kortylewski, A., Yuille, A.L., 2020. Robust object detection under occlusion with context-aware compositionalnets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12645–12654.

Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., Wang, Z., 2021. Augmax: Adversarial composition of random augmentations for robust training. Adv. Neural Inf. Process. Syst. 34, 237–250.

Wang, S., Yu, L., Li, C., Fu, C.-W., Heng, P.-A., 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In: European Conference on Computer Vision. Springer, pp. 159–176.

Wen, X., Zhao, B., Zheng, A., Zhang, X., Qi, X., 2022. Self-supervised visual representation learning with semantic grouping. Adv. Neural Inf. Process. Syst. 35, 16423–16438.

Wong, E., Rice, L., Kolter, J.Z., 2020. Fast is better than free: Revisiting adversarial training. arXiv preprint arXiv:2001.03994.

Xiao, M., Kortylewski, A., Wu, R., Qiao, S., Shen, W., Yuille, A., 2020. Tdmpnet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, pp. 447–463.

Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K., 2019. Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 501–509.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D., 2022. Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4310–4319.

Xu, Q., Zhang, R., Fan, Z., Wang, Y., Wu, Y.-Y., Zhang, Y., 2023. Fourier-based augmentation with applications to domain generalization. Pattern Recognit. 139, 109474.

Yan, Z., Ruan, D., Wu, Y., Huang, J., Chai, Z., Han, X., Cui, S., Li, G., 2024. Contrastive open-set active learning based sample selection for image classification. IEEE Trans. Image Process.

Yang, J., Zhou, K., Li, Y., Liu, Z., 2024. Generalized out-of-distribution detection: A survey. Int. J. Comput. Vis. 132 (12), 5635–5662.

Ye, N., Li, K., Bai, H., Yu, R., Hong, L., Zhou, F., Li, Z., Zhu, J., 2022. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7947–7958.

Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S., 2022. Volo: Vision outlooker for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk management. In: 6th Int. Conf. Learning Representations. ICLR, pp. 1–13.

Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z., 2021. Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5372–5382.

Zhang, D., Fu, Y., Zheng, Z., 2022. UAST: Uncertainty-aware siamese tracking. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 162, PMLR, pp. 26161–26175.

Zhang, D., Xiao, X., Zheng, Z., Jiang, Y., Yang, Y., 2024. Probabilistic assignment with decoupled IoU prediction for visual tracking. IEEE Trans. Circuits Syst. Video Technol.

Zhao, B., Wen, X., 2020. Distilling visual priors from self-supervised learning. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, pp. 422–429.

Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., Kortylewski, A., 2022. OOD-CV: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In: European Conference on Computer Vision. pp. 163–180.

Zhu, R., Zhao, B., Liu, J., Sun, Z., Chen, C.W., 2021. Improving contrastive learning by visualizing feature transformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10306–10315.