Multi-teacher Invariance Distillation for Domain-Generalized Action Recognition

Jongmin Shin[1], Abhishek Maiti[2], Yuliang Zou[3], and Jinwoo Choi[1(✉)]

[1] Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea
jinwoochoi@khu.ac.kr
[2] IIIT Delhi, New Delhi 110020, India
[3] Virginia Tech, Blacksburg, VA 24061, USA

Abstract. In this work, we tackle the problem of domain-generalized action recognition, i.e. we train a model on a source domain and then test the model on other unseen target domains with different data distributions. Generalizing across different domains often requires distinct representational invariances and variances, which makes domain generalization even more challenging. However, existing methods overlook the nuanced requirements of representational invariances/variances across different domains. To this end, we propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR), a method to learn multiple representational invariances/variances tailored to the unique characteristics of diverse domains. MIDAR comprises two key learning stages. First, we learn multiple teacher models to specialize in distinct representational invariances/variances. Then, we distill the knowledge of teachers to a student model through the adaptive reweighting (ARW) layer, which determines the ratio of supervision from different teachers. We validate the proposed method on public benchmarks. The proposed method shows favorable performance compared to the existing methods across multiple domains on public benchmarks.

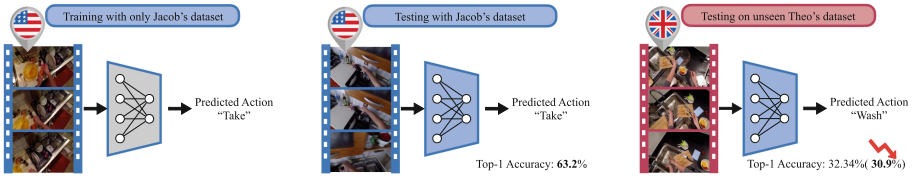
Keywords: Action Recognition · Domain Generalization · Knowledge Distillation · Self-Supervised Learning · Invariance

1 Introduction

The rapid progress in action recognition [6,15,16,26,41,52] has significantly improved the ability of video models to understand human actions in videos. Despite the great progress, most action recognition models often suffer from performance degradation on the test datasets with different distributions from the training dataset [9–11,33]. This performance drop is evident in domain generalization [47], highlighting the vulnerability of action recognition models to distribution shifts. As shown in Fig. 1 (a) and (b), training and testing in the same dataset, e.g. Jacob's kitchen dataset, allows the model to correctly recognize the action '*Take*'. However, as depicted in Fig. 1 (c), testing the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78110-0_8.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15329, pp. 116–132, 2025.
https://doi.org/10.1007/978-3-031-78110-0_8



(a) Model training (b) In-domain testing (c) Out-of-domain testing

Fig. 1. (Single-source) domain generalization. (a) We train a video recognition model on a source domain (e.g., Jacob's Kitchen); (b) When we test the model on the same data distribution, the model performs reasonably well; (c) However, when the model is evaluated on an unseen target domain (e.g., Theo's Kitchen), the performance drops significantly due to domain shifts.

model on a dataset with a distribution shift from the training data, e.g. trained on Jacob's kitchen dataset and test on Theo's kitchen dataset, significantly degrades the model performance from 63.2% to 30.9%. The model fails to recognize the action '*Take*' and misclassifies it as '*Wash*'. A desired model would not suffer from this performance drop across domains.

We hypothesize that we can enhance the generalization performance of a model by learning multiple representational invariances/variances. We empirically find that the beneficial invariances/variances depend on the source and target distributions. In Table 1, we show domain generalization performance of a few models: i) a baseline TSM [26] that does not explicitly learn any representational invariances, ii) a color-invariant TSM, iii) a temporal-invariant TSM, iv) an color&temporal-invariant TSM, all evaluated on the EPIC-KITCHENS dataset [12]. Please refer to Sect. 3.1 for detailed information on model training procedures. We find that the color and temporal-invariant model outperforms the baseline, whereas the color&temporal-invariant model underperforms the baseline. The results indicate that the effectiveness of specific invariances/variances depends on the source and target distributions. We could expect improved generalization performance if we can appropriately learn to incorporate multiple invariances/variances.

Table 1. Baseline Domain Generalized Action Recognition Performance. We show the domain generalization accuracy of models with distinct representational invariances and a model naively learned multiple invariances. We use the TSM model with a ResNet-50 backbone.

Method	Average Accuracy
Baseline Model	37.07 ± 3.39
Color Invariant Model	37.83 ± 3.65
Temporal Invariant Model	38.36 ± 2.73
Color&Temporal Invariant Model	35.34 ± 5.38

Multi-source domain generalization [2,24,25,40] might be a solution to learn multiple invariances. However, it is impractical for video action recognition as collecting and labeling multiple video action recognition datasets is labor-intensive and costly. Single-source domain generalization methods [5,8,42,43,47,53] could learn representational invariances in image recognition. However, we empirically find these methods struggle with the temporal dimension critical for video data. RADA [47] learns invariances by adversarial perturbations. They perturb

the data distribution of the source domain to cover the unseen target domain. However, they do not learn diverse invariances e.g. temporal and order variance/invariance, which may be beneficial in some domains. A naive approach for learning multiple invariances could be training a model with multiple tasks, each responsible for a specific type of invariance. However, we empirically find this approach results in inferior performance even compared to the baseline without any invariances in Table 1. Models with only a single type of invariance, e.g. color invariance, show improved domain generalization performance (38.36% *vs.* 37.07%). However, a model naively trained with both color and temporal invariance learning heads underperforms compared to the baseline (35.34% *vs.* 37.07%). The observations underscore the importance of a nuanced approach to learning multiple representational invariances and variances to achieve robust performance across diverse domain generalization scenarios.

In this work, we introduce Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR) to address the challenge of learning multiple invariance/variance. Our approach involves two stages. In the first stage, we train multiple teacher models, each specializing in a different representational invariance or variance. In the next stage, we distill the knowledge from multiple teachers into a student model. MIDAR adaptively reweighs the supervision from multiple teachers, allowing the student model to learn distinct representational invariance/variance. We validate the effectiveness of the proposed method on public benchmarks. MIDAR shows favorable performance compared to the existing methods.

To summarize, we make the following contributions.

- We introduce MIDAR, a new training method addressing the challenge of learning multiple representational invariances/variances for domain-generalized action recognition.
- We introduce the Adaptive Reweighting layer to adjust the contribution of multiple teachers, allowing the student model to leverage the diverse representational invariance/variance of each teacher.
- We conduct extensive experiments on the Epic-Kitchens benchmark to validate MIDAR. Our findings indicate that MIDAR's approach to learning diverse representational invariance/variance outperforms current SOTA methods like RADA, which rely on adversarial perturbation.

2 Related Work

2.1 Video Action Recognition

2D CNNs [26,41,52], 3D CNNs [6,15,38], and two-stream CNNs [16,34] are popular techniques to recognize human actions from videos. More recently, Transformer-based methods have shown great performance [3,4,14,18,31,44,46]. Despite the great recent advances in action recognition, we find that state-of-the-art action recognition methods still suffer from cross-domain generalization: a model trained on one domain shows poor performance on other domains with different data distributions. In this work, we tackle the domain-generalized action recognition task to address the challenge.

2.2 Domain Generalization

Recently, domain generalization has drawn significant attention from the community since training and test data usually have different distributions in practice. Broadly, there are two principal categories of approaches in the domain generalization literature: i) feature-based domain generalization, and ii) data-based domain generalization. Feature-based domain generalization methods [2,5,24,25,42] aim to learn domain-invariant representations to enhance the generalization performance of models. On the other hand, data-based domain generalization methods [21,39,40] augments training data to generate adversarial samples and synthetic data with different styles and scenes that bridge the gap between source and target domains. These works have shown great progress in domain-generalized image recognition. However, domain generalization for video recognition is still under-explored. To the best of our knowledge, there is only one work on domain-generalized action recognition: Robust Adversarial Domain Augmentation (RADA) [47]. RADA learns domain invariant video representation by training on the perturbed data and adversarial examples. Our work is on domain-generalized action recognition as well. In contrast to RADA, the proposed method learns the nuanced requirements of the representational invariances across different domains, thus offering a novel approach to this challenging problem.

2.3 Knowledge Distillation

Knowledge distillation is a popular technique to transfer knowledge from one model to another model. We can categorize knowledge distillation into three groups: i) response-based, ii) intermediate, iii) relation-based, and iv) multi-teacher knowledge distillation. Response-based knowledge distillation methods [19,51,54] encourage the student model to mimic the output of the teacher. In intermediate knowledge distillation [1,29], the student model aims to learn the same feature representation as the feature representation of the teacher. In Relation-based knowledge distillation [30], the student model mimics the relative distance and angle between data points in the feature space of the teacher model. In Multi-teacher knowledge distillation [27,48,49], a student model learns from the combined knowledge of multiple teacher models, leveraging diverse representations. In this work, we leverage knowledge distillation techniques to address the challenge of domain generalization. We learn a student model using multiple teachers, each of which specializes in distinct representational invariances. We dynamically adjust the contribution of different teachers by learning an adaptive re-weighting layer.

3 Method

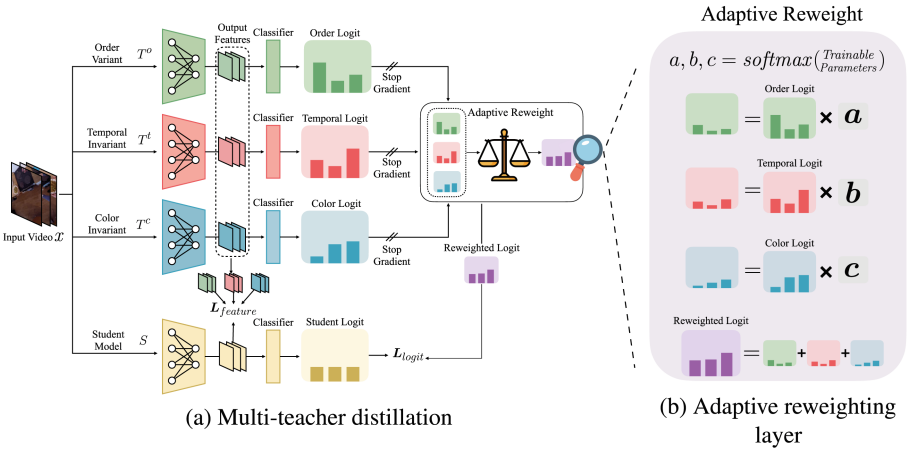
We propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR). As shown in Fig. 2, we employ multiple teachers each with expertise in distinct representational invariance/variance. Our objective is to distill a broad spectrum of invariances/variances, including order variance, temporal invariance, and color invariance, into a student model. This knowledge distillation process encompasses both the feature representations and the output logits of these teacher models. We

propose an adaptive reweighting layer to dynamically adjust the contribution from each teacher based on the data. In the following subsections, we provide detailed descriptions of each component of MIDAR. We describe the training process of teacher models in Sect. 3.1. Then we illustrate the proposed multi-teacher distillation framework in Sect. 3.2. Finally, we describe the proposed adaptive reweighting method in Sect. 3.3.

3.1 Training Teacher Models

Color Invariant Teacher. Color invariance is desirable in many action recognition scenarios. For example, a model should be able to correctly recognize the 'playing tennis' action regardless of whether the tennis court is green grass or brown mud. To learn color invariance, we employ color jittering augmentation during the color invariant teacher training process. Given an input video, we randomly jitter the brightness, contrast, saturation, and hue of each frame. Following prior works [17,32,55], we employ a temporally coherent color jitter augmentation, i.e. we use the same color jittering across all the frames within an input video.

We employ supervised contrastive learning (SCL) [23] for color invariant teacher training. We empirically find that SCL is beneficial for color invariance learning, compared to using the cross-entropy loss. In the SCL framework, we define any pair of videos from the same action class as a positive pair, regardless of color augmentation. We define any pairs from different action classes as negative pairs. We define the SCL loss for learning color invariance as follows:



(a) Multi-teacher distillation

(b) Adaptive reweighting layer

Fig. 2. Overview. (a) We use a multi-teacher distillation framework to distill multiple representational (in)variances into a student model. Both features and the logits are distilled from each teacher to the student model. (b) For logit distillation, we propose an adaptive reweighting layer to adjust the impacts of each teacher. Specifically, we assign one learnable parameter for each teacher so that the distillation strength of each teacher is dynamically adjusted during training.

$$L_{\text{SCL}} = \sum_{i \in B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \tag{1}$$

where B denotes the set of all input data within a minibatch while each i-th instance is an anchor. $A(i)$ denotes the set of all input data within a mini-batch except the i-th instance, i.e. $A(i) \equiv B \backslash \{i\}$. The set of all positive pairs $P(i)$ contains samples with identical class labels to i-th instance, including color-augmented samples. In (1), we scale the similarity between the anchor embedding $\mathbf{z}_x$ and any positive sample embedding $\mathbf{z}_p$ by the temperature hyperparameter τ.

In SCL, a model learns to align positive pairs, each consisting of different augmentations. As a result, a teacher model trained with SCL has specialized expertise, i.e. color invariance in our case, that could be generalizable across different domains [37].

Temporal Invariant Teacher. Unlike image data, video data has an additional temporal dimension. The same human action might have different speeds, durations, or temporal patterns across different domains. Consequently, to robustly recognize human actions in various domains, we desire a temporal invariant model [13,17,35,55]. To learn temporal invariant representations to a teacher model, we employ three temporal augmentations [55] that have shown significant performance improvement: T-Half, T-Drop, and T-Reverse. For example, let us assume we have 4 frames with indices $[1, 2, 3, 4]$. Then the T-Half augmentation repeats the first or the second half of the video only: e.g. $[1, 2, 3, 4] \rightarrow [1, 2, 1, 2]$ or $[1, 2, 3, 4] \rightarrow [3, 4, 3, 4]$. The T-Half augmentation encourages the model to be robust to the partial temporal occlusion. The T-Drop augmentation drops random frames in the video, substituting them with the previous frame: e.g. $[1, 2, 3, 4] \rightarrow [2, 2, 4, 4]$. The T-Drop augmentation encourages the model to be invariant to the speed of the action. The T-Reverse augmentation inverts the order of the video frames, e.g. $[1, 2, 3, 4] \rightarrow [4, 3, 2, 1]$. Following the prior work [55], we randomly select one temporal augmentation for augmenting each video. We employ (1), supervised contrastive learning with these temporal augmentations [55]. We empirically find that using the SCL loss is beneficial for temporal invariance learning compared to using the standard supervised training with the cross-entropy loss for the prediction.

Order Variant Teacher. To distinguish fine-grained actions with subtle differences, e.g. opening and closing a door, a model needs to be sensitive to the temporal order of events. To encourage a model to be sensitive to the order of temporal events, we employ a self-supervised task: video clip order prediction [45]. In this task, we shuffle the clips sampled from an input video. Then we input the shuffled clips into a model. The model should predict the correct chronological order of the clips. Predicting the correct temporal sequence of video clips encourages the model to specialize in order *variance*. Through this task, a model better understands temporal relationships and dependencies between different temporal segments of the actions. Order variance is beneficial for domain generalized action recognition since the order of the action often does not change across people or locations. Furthermore, order variance is desirable as learning order variant representation is learning action representations robust to scene distribution shift across domains [11].

To learn order variance, we define the order variance (OV) loss as follows:

$$L_{\text{OV}} = -\sum_{i=1}^{C!} y_i \log f_o(\psi). \tag{2}$$

Here, the model takes the concatenated input $\psi = (\phi_1, ..., \phi_C)$, where each ϕ_i is a feature vector of i-th clip in the input video. The model predicts a probability distribution across $C!$ possible temporal orders of the input clips, where C denotes the number of clips in an input video. y_i is the i-th element of $\mathbf{y}$, while $\mathbf{y}$ is the ground truth one-hot vector of length $C!$ with the correct clip order of the input video.

3.2 Distilling Invariances from Multiple Teachers

In Table 1, we observe that the naive learning of multiple representational invariances degrades the domain generalized action recognition performance (i.e. 35.34 ± 5.38 vs. 37.07 ± 3.39). To address this challenge, we propose Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR). MIDAR has a multi-teacher knowledge distillation architecture [20,22,27] comprising teacher models with expertise in order variance (Ω_O), temporal invariance (Ω_T), and color invariance (Ω_C). As depicted in Fig. 2 (a), each teacher contributes distinct expertise to the learning process.

Both the student model, π, and the teacher models, Ω_O, Ω_T, and Ω_C, take the same input RGB video, $I \in \mathbb{R}^{M \times H \times W \times C}$, where each video contains M frames with the height of H pixels, width of W pixels, and C channels. We employ the feature-space distillation loss, L_{feature}. We compute L_{feature} as the mean squared error between the feature vectors of the student and teacher models as follows:

$$L_{\text{feature}} = \sum_{t \in \{O,T,C\}} \left(\frac{1}{n} \sum_{i=1}^{n} (\Omega_t(i) - \pi(i))^2 \right). \tag{3}$$

Here, t is the teacher model index and $\Omega(i)$ and $\pi(i)$ denote the i-th feature of the teacher and student model, respectively. By using the L_{feature}, we effectively guide the student model to mimic the expertise of the teachers.

Moreover, we employ the Kullback-Leibler (KL) Divergence loss, L_{KL}, in MIDAR for the output-space distillation as follows:

$$L_{\text{KL}} = \sum_{k=1}^{K} P_{\Omega}(k) \log \left(\frac{P_{\Omega}(k)}{Q_{\pi}(k)} \right). \tag{4}$$

Here P_{Ω} denotes the output probability of the adaptive reweighting (ARW) layer and k is the action category index for K action categories. Q_{π} is the output probability of the linear action classifier for the student model. The output-space distillation encourages the student model to mimic the prediction of the teacher model.

3.3 Learning to Re-weight Multiple Teachers

We introduce an adaptive reweighting (ARW) layer in MIDAR to reflect the nuanced influence of multiple teacher models for learning a student model. The ARW layer takes the softmax probability of each teacher Ω_O, Ω_T, and Ω_T and outputs the single softmax probability vector P_Ω. We define adaptive reweighting operation as follows:

$$P_\Omega = \sum_{t \in \{O,T,C\}} \frac{\exp(\alpha_t)}{\sum_{i \in O,T,C} \exp(\alpha_i)} f_t(\Omega_t). \tag{5}$$

Here, f_t is a linear action classifier for the teacher t. α_i's are learnable parameters for the adaptive reweighting. We set the same number of parameters the same as the number of teachers. By (5), we get the final reweighted probability, P_Ω, that aggregates the nuanced contributions of all the teacher models, as illustrated in Fig. 2 (b).

During training, the parameters α_i's are continuously updated, leading to dynamic adjustments of the contribution of each teacher: more effective teachers get higher weights and less effective ones get lower weights. The learnable parameter α_i dynamically adjusts the student model's focus on multiple invariances and variances. The balance is crucial for enhancing the student model performance, as it allows for a more nuanced understanding that could be beneficial in multiple domains. The proposed ARW layer enables the student model to effectively extract the diverse representational invariances/variances of the teacher models.

We define the total loss function of MIDAR as follows:

$$L = L_{\text{CE}} + L_{\text{feature}} + L_{\text{KL}}. \tag{6}$$

The total loss function consists of three components. First, L_{CE} is the standard cross-entropy loss to learn action categories. L_{feature} aligns feature representations of the student model with the feature representations of the teacher models. L_{KL} guides the student model to mimic the adaptively re-weighted predictions of the teacher models.

4 Experimental Results

4.1 Experimental Setup

Dataset. To evaluate the effectiveness of MIDAR, we use the EPIC-KITCHENS-55 dataset [12]. EPIC-KITCHENS-55 is a large-scale egocentric action recognition dataset consisting of multiple domains. We use the subset for evaluating domain generalization methods, following the experimental protocol in a prior work [28]. The subset comprises three domains, D1, D2, and D3, which results in six domain generalization settings: D1→D2, D1→D3, D2→D1, D2→D3, D3→D1, and D3→D2. The subset consists of 8 action classes across all the domains: put, take, open, close, wash, cut, mix, and pour. Each domain has different actors and kitchen environments but the same action categories. The subset consists of $10,094$ videos in total.

Evaluation Metric. We evaluate the effectiveness of Multi-teacher Invariance Distillation for domain-generalized Action Recognition (MIDAR) by adopting the standard evaluation protocols across benchmarks [28]. For the Epic-Kitchens benchmark, in our protocol [28], we select the model that demonstrates the highest in-domain performance and evaluate the cross-domain performance of the model. We measure the model's performance using the averaged Top-1 accuracy and the standard deviation across six different cross-domain generalization settings.

Implementation Details. Here, we provide details of our training setup and implementation. For additional information, please refer to the supplementary materials. *Base setting.* We employ Temporal Shift Module (TSM) [26] with a ResNet-50 backbone as the base model, unless we specify another model. From each video, we sample 8 frames to construct an input clip. The initial learning rate is 0.0075. We train models for 150 epochs.

Teacher Model Training. For color and temporal invariant teacher models, we build the models upon the SimSiam [7] architecture. We use the supervised contrastive loss Eq. (1) as a loss function to train the color-invariant and temporal-invariant teacher models, with the temperature τ set to 0.3. For the order variant teacher, we implement the video clip order prediction (VCOP) [45] pre-text task, processing 3 clips of 8 frames each, with an inter-clip interval of 8 frames. We train the model for 800 epochs. We attach a linear classifier on top of the backbone. Then we train the model end-to-end just like other teacher models.

Student Model Training. When training the student model, we freeze the weights of all the teacher models. The learning rate is set to 0.005. For the adaptive reweighting layer, each trainable parameter α_t is initially set to an equal value of 1. This initialization strategy ensures that, before updating the trainable parameters, each value post-softmax normalization approximates 0.3333, thereby providing a fair starting point.

Please see the supplementary materials for details on the model training and inference.

Baseline. To establish a baseline for domain generalization performance, we train a TSM with a ResNet-50 backbone on one domain of the benchmark dataset. Subsequently, we evaluate the trained model on another domain of the dataset. We repeat the same process for all six settings in the EPIC-KITCHENS dataset. During training, we do not apply any learning technique that encourages domain-invariant representations. To establish a baseline for domain generalization performance, we train a TSM with a ResNet-50 backbone on one

Table 2. Individual Invariant/Variant Model Performance. We show the domain generalization performance of individual invariant/variant models. Every model is equipped with the TSM with a ResNet-50 backbone.

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
Color Invariant Model	37.83 ± 3.65
Temporal Invariant Model	**38.36 ± 2.73**
Order Variant Model	37.38 ± 3.54

domain of the EPIC-KITCHENS dataset [28]. Subsequently, we evaluate the trained model on another domain of the EPIC-KITCHENS dataset. We perform the same process for all six settings in the EPIC-KITCHENS dataset. During training, we do not apply any learning technique that encourages domain-invariant representation learning.

4.2 Individual Invariance/Variance Model Performance

We first study the effectiveness of each model with distinct representational invariances/variances by comparing the domain generalization performance of each model with the baseline performance. As shown in Table 2 each representational invariant/variant model outperforms the baseline. The temporal invariant model shows the most improvement of 1.29 points and the color invariant model shows an improvement of 0.76 points compared to the baseline. The order variant model achieves a marginal improvement of 0.31 points compared to the baseline. These results indicate that incorporating individual representational invariances/variances could improve the domain generalization performance, but the improvement is not very significant. As shown in Table 1, naively learning multiple invariances, e.g. learning the temporal and the color invariant representations simultaneously, results in inferior performance compared to the baseline without learning any invariances: 35.34% *vs.* 37.07%. Therefore, we need a nuanced approach, such as MIDAR, to learn multiple representational invariances/variances to achieve superior domain generalization performance.

Table 3. Effect of Distillation in Domain Generalization. We compare the performance of the logit-space distillation, the feature-space distillation, and both the logit-space and faeture-space distillation. We employ the temporal invariant model as a teacher in this experiment.

Method	Logit	Feature	Top-1 Accuracy
Baseline	–	–	37.07 ± 3.39
Temporal Invariant Teacher	–	–	38.36 ± 2.73
Student	✓		38.78 ± 2.85
		✓	35.28 ± 4.13
	✓	✓	$\mathbf{38.93 \pm 3.61}$

Table 4. Effect of Multi-Teacher Distillation. We compare the domain generalization performance of students learned from different combinations of teachers. Properly using all three teachers shows the best domain generalization performance.

Method	Color	Temporal	Order	Top-1 Accuracy
Baseline	–	–	–	37.07 ± 3.39
Single Teacher Distillation		✓		38.93 ± 3.61
Two-Teacher Distillation	✓	✓		38.64 ± 2.48
	✓		✓	37.20 ± 5.30
		✓	✓	40.03 ± 3.29
Three-Teacher Distillation	✓	✓	✓	$\mathbf{41.12 \pm 2.61}$

Table 5. Ablation study. We conduct experiments with different distillation methods to validate the effect of each distillation strategy, logit, feature, and multi-teacher distillation.

(a) How to aggregate multiple teacher outputs?

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
Correct Teacher	38.55 ± 1.74
Most Confident Teacher	38.99 ± 3.76
Lowest Cross-Entropy Teacher	38.89 ± 2.81
Average of Teachers	38.70 ± 3.74
Adaptive Reweighting (Ours)	41.12 ± 2.61

(b) Which loss to distill features?

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
CORAL Loss [36]	40.12 ± 3.47
Huber Loss	39.02 ± 3.87
MSE Loss	41.12 ± 2.61

(c) Multi-teacher distillation method

Method	Top-1 Accuracy
Baseline	37.07 ± 3.39
KD [19]	38.45 ± 3.86
FiTNet [1]	37.38 ± 4.34
Average [48]	38.68 ± 3.70
Ours	41.12 ± 2.61

4.3 Distillation for Learning Multiple Invariances/variances

Is Distillation Beneficial in Domain Generalization? We empirically find that distillation is beneficial in domain generalization. In Table 3, compared to the temporal invariant teacher model, a student model learned by the logit-space distillation shows an improved performance of 38.78%. A student model learned by the feature-space distillation shows inferior performance compared to the teacher. However, a student model learned by both logit and feature-space distillation shows the best performance of 38.93%. The results demonstrate that distillation is beneficial in domain generalization even when we have a single teacher only. In the remaining experiments, we distill both features and logits.

Is Multi-Teacher Distillation Beneficial for Learning Multiple Invariances/variances? We investigate the effect of multiple teachers in Table 4. We can observe a trend: as the number of teachers increases, the student model demonstrates improved performance. Specifically, the student model, which learns from both the temporal-invariant and order-variant teachers, achieves an accuracy of 40.03%, surpassing the single teacher distillation with an accuracy of 38.93%. Furthermore, when the student model learns from the knowledge distillation of the color invariant, the temporal invariant, and the order variant teachers simultaneously, the student model achieves the best accuracy of 41.12%. The results underscore the significance of leveraging multiple teachers to enrich the knowledge of the student model and subsequently enhance the domain generalization.

4.4 Ablation Study

We conduct ablation experiments to explore the various design choices of the multi-teacher distillation strategy to improve the domain generalization performance. Here, we conduct all experiments with multi-teacher distillation that encompasses all invariant and variant teacher models.

How to Aggregate the Output of Multiple Teachers? Since we have multiple teachers, how to aggregate the output of multiple teachers is an important design choice. In Table 5 (a), we compare five logit-space distillation methods. i) Correct Teacher: we select the correctly predicted teachers for the distillation. We average the prediction vectors if multiple teachers agree, and we discard the sample if all predictions are incorrect. ii) Most Confident Teacher: we select the teacher with the highest softmax probability among all the teachers for the distillation. iii) Lowest Cross-Entropy Teacher: we choose the prediction from the teacher with the minimum cross-entropy loss for the distillation. iv) Average of Teachers: we average the predictions of all the teachers for the distillation. v) Adaptive reweighting (ours): we dynamically adjust the contribution of each teacher by Eq. 5. The results demonstrate that the adaptive reweighting outperforms the other compared methods, achieving 41.12% which is 2.13 points higher than the second-best method, Most Confident Teacher. The results suggest that our adaptive reweighting is more effective in leveraging multiple teachers to improve domain generalization.

Which Loss for Feature-Space Distillation? Here, we compare three loss functions for the feature-space distillation in MIDAR. i) The CORAL (CORelation ALignment) loss [36]: we align the second-order statistics of feature distributions by minimizing the difference in their covariance matrices. The CORAL loss is typically applied for domain adaptation. We employ the CORAL loss to align student features with the teacher features to tackle the problem of domain generalization. ii) Huber loss is a hybrid loss function that is a combination of both Mean Squared Error (MSE) and Mean Absolute Error (MAE). Huber loss aims to mitigate the influence of outliers during distillation. Also, huber loss offers a balance between sensitivity to data variance and robustness to outliers. iii) Mean Squared Error (MSE) loss: a loss function that minimizes the average of the squares of the errors. As shown in Table 5 (b), employing CORAL loss outperforms Huber loss with a margin of 1.1 points (40.12% vs. 39.02%). However, using MSE loss outperforms CORAL loss with a margin of 1.0 points (41.12% vs. 40.12%). The results suggest that the MSE loss is more effective for feature-space distillation.

Comparing Multi-teacher Distillation Methods. We conduct a comparative analysis of MIDAR against established distillation techniques. We replace the proposed multi-teacher distillation method with the following methods and compare the domain generalization performance. i) KD [19], which distills the average predictions from multiple teachers, ii) FitNet [1], which distills their average features, and iii) Average [48], which distills both averaged features and predictions. Table 5(c) shows that MIDAR achieves the best accuracy of 41.12%. Compared to MIDAR, FitNet shows a 3.74-point drop (41.12% *vs.* 37.38%) and KD shows a 2.67-point drop (41.12% *vs.* 38.45%). Average shows a 2.44-point drop (41.12% *vs.* 38.68%). The results showcase the effectiveness of our multi-teacher distillation approach in enhancing domain generalization.

4.5 Comparison with Existing Domain Generalization Methods

We compare the domain generalization performance of MIDAR with existing single-source domain generalization methods in Table 6. Please see the supplementary materials for details of the results. We compare MIDAR with four image-based methods extended to the video domain. i) Mixup [50]: we blend each video with a randomly chosen video in the batch and set the mixup ratio as 0.2. ii) Mixstyle [53]: we integrate a Mixstyle module into the ResNet backbone of TSM. Mixstyle mixes the statistics, i.e. mean and standard deviation, of feature maps from different instances during the training process. Mixstyle incorporates a new style in the feature space and encourages the model to learn domain generalizable features. iii) JiGen [5] recognizes action and simultaneously solves jigsaw puzzles to understand spatial correlations. Solving jigsaw puzzles acts as a regularization for the classification task. The shared feature embedding between the classification and the jigsaw puzzle tasks allows the model to generalize across domains. iv) EISNet [42] enhances generalization performance by multi-task learning from both extrinsic and intrinsic supervisions. EISNet employs momentum metric learning for domain-invariant yet class-specific features and solves jigsaw puzzles. For JiGen and EISNet baselines, we apply consistent augmentation across all frames in a video clip to maintain temporal coherence. Additionally, we compare MIDAR with RADA [47][1], the state-of-the-art video-based domain generalization method.

Table 6. Comparison with the state of the arts on EPIC-Kitchens. We compare the domain generalization performance (top-1 accuracy) of our model with several image-based methods (Mixup [50], Mistyle [53], JigSen [5], EISNet [42]) and a recently proposed video-based method (RADA [47]).

Method	Backbone	Average Accuracy
Baseline	TSM	37.07 ± 3.39
Mixup [50]	TSM	37.54 ± 4.69
Mixtyle [53]	TSM	36.88 ± 5.18
JiGen [5]	TSM	38.59 ± 6.14
EISNet [42]	TSM	37.52 ± 1.31
RADA [47]	APN [47]	40.52 ± 3.23
Ours	TSM	**41.12 ± 2.61**

All the compared methods employ the TSM [26] as a backbone except RADA. RADA [47] is equipped with the Adversarial Pyramid Network (APN) backbone. We select the learning rate with the highest performance for each method. We use the learning rates of 0.01, 0.001, 0.005, and 0.0075 for Mixup, Mixstyle, JiGen, and EISNet respectively. For RADA we use the learning rate of 0.001. As shown in

Table 7. Effect of using different backbones: ResNet-50 *vs.* ResNet-101 on the EPIC-Kitchens dataset. We compare the domain generalization performance (top-1 accuracy) of our model with the video-based method(RADA [47]).

Method	Model	Backbone	Average Accuracy
RADA [47]	APN [47]	ResNet-50	40.52 ± 3.23
Ours	TSM	ResNet-50	**41.12 ± 2.61**
RADA [47]	APN [47]	ResNet-101	43.08 ± 4.27
Ours	TSM	ResNet-101	**43.54 ± 5.59**

[1] The TSM backbone employed by MIDAR has 4.8 million *fewer* parameters than the APN used by RADA.

Table 6, JiGen outperforms other image-based methods with an average accuracy of 38.59% exceeding Mixstyle by 1.71 points, Mixup by 1.05 points and EISNet by 1.07 points.

However, JiGen shows inferior performance compared to the video-based method RADA by 1.93 points. MIDAR surpasses RADA by 0.60 points and Jigen by 2.53 points, resulting in the best accuracy as well as relatively lower standard deviation. The results indicate that MIDAR shows favorable performance across various domain shifts and demonstrates the effectiveness in various domain generalization scenarios.

For a fair comparison, we evaluate MIDAR and RADA using the same ResNet-50 and ResNet-101 backbones in Table 7. As shown in Table 7, with a stronger ResNet-101 backbone, MIDAR shows an improvement of 0.46 points over RADA with ResNet-101 backbone on the Epic-Kitchens dataset. The favorable performance of MIDAR on the different backbones underscores its effectiveness.

5 Conclusions

In this paper, we tackle the problem of domain-generalized action recognition, which is a challenging, yet relatively under-explored problem. We study a wide spectrum of representational invariance/variance learning which is often beneficial in the context of domain-generalized action recognition. We empirically find that naively learning multiple invariances leads to even inferior domain generalization performance compared to the baseline without learning any representational invariances. To tackle this challenge, we introduce MIDAR, an innovative multi-teacher distillation approach that learns nuanced influence from multiple teachers with distinct representational invariances/variances. We propose an adaptive re-weighting layer to learn such nuanced influence from multiple teachers as well as to incorporate both feature-space and output-space distillation. The empirical results on the challenging EPIC-Kitchens dataset with a moderate size demonstrate that MIDAR generalizes across different domains compared to the existing domain generalization methods. Our future work is overcoming this limitation. We plan to improve MIDAR's adaptability to various data scales. Moreover, we plan to apply MIDAR to Transformer architectures and tailor MIDAR to leverage the representational invariance and variance of Transformers.

Acknowledgement. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) under grant RS-2024-00353131, RS-2021-II212068 (Artificial Intelligence Innovation Hub), and RS-2022-00155911 (Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)). Additionally, it was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. 2022R1F1A1070997).

References

1. Adriana, R., Nicolas, B., Ebrahimi, K.S., Antoine, C., Carlo, G., Yoshua, B.: Fitnets: hints for thin deep nets. In: ICLR (2015)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: ICCV (2021)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)
5. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
8. Cheng, S., Gokhale, T., Yang, Y.: Adversarial bayesian augmentation for single-source domain generalization. In: ICCV (2023)
9. Choi, J., Huang, J.B., Sharma, G.: Self-supervised cross-video temporal learning for unsupervised video domain adaptation. In: ICPR (2022)
10. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: WACV (2020)
11. Choi, J., Sharma, G., Schulter, S., Huang, J.-B.: Shuffle and attend: video domain adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 678–695. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_40
12. Damen, D., et al.: Scaling egocentric vision: the epic-kitchens dataset. In: ECCV (2018)
13. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: TCLR: temporal contrastive learning for video representation. CVIU **219**, 103406 (2022)
14. Fan, H., et al.: Multiscale vision transformers. In: ICCV (2021)
15. Feichtenhofer, C.: X3d: expanding architectures for efficient video recognition. In: CVPR (2020)
16. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
17. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: CVPR (2021)
18. Herzig, R., et al.: Object-region video transformers. In: CVPR (2022)
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
20. Hu, C., et al.: Teacher-student architecture for knowledge distillation: a survey. arXiv preprint arXiv:2308.04268 (2023)
21. Jackson, P.T., Abarghouei, A.A., Bonner, S., Breckon, T.P., Obara, B.: Style augmentation: data augmentation via style randomization. In: CVPR Workshop (2019)
22. Kaplun, G., Malach, E., Nakkiran, P., Shalev-Shwartz, S.: Knowledge distillation: Bad models can be good role models. In: NeurIPS (2022)
23. Khosla, P., et al.: Supervised contrastive learning. In: NeurIPS (2020)
24. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
25. Li, Y., et al.: Deep domain generalization via conditional invariant adversarial networks. In: ECCV (2018)
26. Lin, J., Gan, C., Han, S.: Tsm: temporal shift module for efficient video understanding. In: ICCV (2019)

27. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing **415**, 106–113 (2020)
28. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (2020)
29. Nie, X., Li, Y., Luo, L., Zhang, N., Feng, J.: Dynamic kernel distillation for efficient pose estimation in videos. In: ICCV (2019)
30. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)
31. Patrick, M., et al.: Keeping your eye on the ball: trajectory attention in video transformers. In: NeurIPS (2021)
32. Qian, R., et al.: Spatiotemporal contrastive video representation learning. In: CVPR (2021)
33. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: temporal contrastive video domain adaptation with background mixing. In: NeurIPS (2021)
34. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS (2014)
35. Singh, A., et al.: Semi-supervised action recognition with temporal contrastive learning. In: CVPR (2021)
36. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation. In: Domain Adaptation in Computer Vision Applications, pp. 153–171 (2017)
37. Tong, Y., et al.: Quantitatively measuring and contrastively exploring heterogeneity for domain generalization. In: Proceedings of SIGKDD (2023)
38. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
39. Volpi, R., Murino, V.: Addressing model vulnerability to distributional shifts over image transformation sets. In: ICCV (2019)
40. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NeurIPS (2018)
41. Wang, L., et al.: Temporal segment networks for action recognition in videos. TPAMI **41**(11), 2740–2755 (2018)
42. Wang, S., Yu, L., Li, C., Fu, C.-W., Heng, P.-A.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 159–176. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_10
43. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV (2021)
44. Wu, C.Y., et al.: Memvit: memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR (2022)
45. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
46. Yan, S., et al.: Multiview transformers for video recognition. In: CVPR (2022)
47. Yao, Z., Wang, Y., Wang, J., Philip, S.Y., Long, M.: Videodg: generalizing temporal relations in videos to novel domains. TPAMI **44**(11), 7989–8004 (2021)
48. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: Proceedings of SIGKDD (2017)
49. Zhang, H., Chen, D., Wang, C.: Confidence-aware multi-teacher knowledge distillation. In: ICASSP (2022)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
51. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR (2022)
52. Zhou, B., Andonian, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (2018)

53. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)
54. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. In: NeurIPS (2018)
55. Zou, Y., Choi, J., Wang, Q., Huang, J.B.: Learning representational invariances for data-efficient action recognition. CVIU **227**, 103597 (2023)