Improving Knowledge Distillation Under Unknown Covariate Shift Through Confidence-Guided Data Augmentation

Niclas Popp[1,2], Kevin Alexander Laube[1], Matthias Hein [⋆2], and Lukas Schott[⋆⋆3]

[1] Bosch Center for Artificial Intelligence
{niclas.popp, kevinalexander.laube}@de.bosch.com
[2] Tübingen AI Center, University of Tübingen
matthias.hein@uni-tuebingen.de
[3] Aleph Alpha
lukas.schott@aleph-alpha.com

Abstract. Large foundation models trained on extensive datasets demonstrate strong zero-shot capabilities in various domains. Knowledge distillation has become an established tool for transferring knowledge from foundation models to small student networks when data and model size are constrained. However, the efficacy of distillation is often hampered by limited training data coverage. This can result in a covariate shift between training and test data, which in turn can lead the student to exploit spurious features or even shortcut learning. We address this problem by introducing a novel diffusion-based data augmentation strategy that generates images by maximizing the disagreement between the teacher and the student, effectively creating challenging samples that the student struggles with, thus mitigating the problem of covariate shift. Experiments demonstrate that, compared to state-of-the-art diffusion-based data augmentation baselines, our approach is best or second-best in sample mean accuracy and improves the worst group and mean group accuracy on CelebA-HQ, SpuCo Birds and BAR as well as the spurious score on Spurious ImageNet under covariate shift.

Keywords: Knowledge Distillation · Synthetic Data Augmentation · Diffusion Models · Image Classification

1 Introduction

Foundation models have demonstrated strong zero-shot capabilities and distributional robustness across a wide range of data domains [21, 47]. However, training these models demands massive datasets and computational resources, and their deployment in resource-constrained environments is often infeasible due to their scale. Achieving similar generalization and robustness with substantially

⋆ Joint senior author.
⋆⋆ Joint senior author, work done while at Bosch Center for Artificial Intelligence.

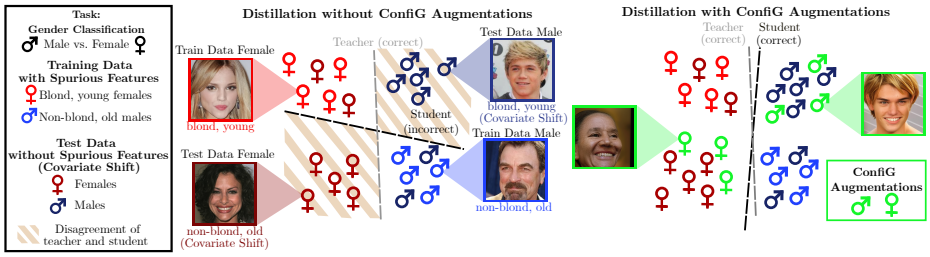

Fig. 1: Knowledge Distillation under Covariate Shift with ConfiG Data Augmentations. Our goal is to maximize student performance even on groups that are fully absent from the training data. **Left:** Without having seen similar samples, the student is incapable of correctly classifying the unseen groups in the test set. **Right:** Leveraging the discrepancy between a robust teacher and the biased student, ConfiG generates and adds challenging samples to the training set. Further training improves the student's performance on previously missing groups by removing shortcuts.

smaller models and less training data remains an open research area. Knowledge distillation (KD) [19] offers a way to transfer knowledge from a pre-trained teacher model to a smaller student model. Compared to supervised training with only the hard ground-truth labels, distillation has been shown to improve both performance and generalization of the student model [16, 38, 44]. The reason why knowledge distillation is believed to improve generalization is that the output of the teacher contains more information than hard labels (referred to as "dark knowledge"). However, the availability of task-specific training data is often a crucial limitation on the knowledge that can be transferred to the student. Especially in the limited-data regime, the training data might not sufficiently cover the data distribution that the student will face during deployment.

In this work, we investigate the challenge of covariate shift between the training and test data distributions in knowledge distillation. Covariate shift refers to a situation where the distribution of the input features (covariates) changes between the training and test datasets while the conditional distribution of the output given the input remains the same. In particular, covariate shift between training and test data can lead to "spurious features" or even severe "shortcut learning" [12] which enables correct decision making on the training but not on the test data. An example for this setting is illustrated in Fig. 1. The task is gender classification on a training data subset of CelebA-HQ, which only contains young and blond females as well as non-blond, old males. At test time, however, the student also encounters non-blond females or blond males. Without training data from these groups, the student fails to learn the correct decision boundary but instead relies on spurious features such as hair color or age. Assuming the pre-trained teacher disentangles spuriously correlated covariates and features that are predictive for the target classes, our goal is to distill a student that is similarly capable.

Several recent works improve model performance in the limited-data regime through diffusion-based data augmentations [17, 20, 54, 59, 60, 68, 72]. These approaches can be interpreted as a form of implicit distillation, where a generative

model is used to provide additional knowledge in the form of augmented training samples. However, by only looking at existing training samples or the student model, the augmented samples cannot effectively target the student's reliance on spurious features. To this end, we introduce ConfiG, a Confidence-Guided diffusion-based data augmentation method that leverages the disagreement between teacher and student to identify and mitigate student biases without requiring explicit knowledge of what these biases are. As illustrated in Fig. 1, ConfiG generates synthetic data augmentations in the regions of disagreement (dashed), such as women with brown hair or blond, young men. Further training on the real training images together with the augmented samples then leads the student to better approximate the decision function of the teacher, reducing the reliance on spurious features. Our contributions are summarized as follows:

1. We demonstrate the problem that, while knowledge distillation from a robust teacher improves overall student performance, under unknown covariate shift the student still performs poorly on test samples from unseen groups.
2. We introduce ConfiG, a data augmentation framework for knowledge distillation that guides a diffusion model to generate samples that target covariate shift through maximizing teacher and student disagreement.
3. We provide empirical and theoretical motivation for ConfiG and demonstrate its superior performance compared to prior diffusion-based augmentations under unknown covariate shift in knowledge distillation on CelebA-HQ [23,34], SpuCo Birds [22], BAR [41], and a subset of ImageNet [42,50].

2 Related Work

Knowledge distillation (KD) is a widely used training framework [14,36] to transfer knowledge from one model (teacher) to another (student). In contrast to empirical risk minimization, in pure response-based KD the student is trained to match the predictive distribution of the teacher [36]. If the teacher is sufficiently accurate, doing so provably improves the generalization of the student over training with hard labels [38,45,52,57]. While there exist a lot of KD variants, simply matching the responses of student and teacher [19] has been shown to be highly effective for image classification [6,16]. We adopt response-based knowledge distillation purely based on the soft labels of the teacher and investigate the influence of the training data.

Synthetic data augmentation is a rising trend to mitigate data scarcity and bias. Unlike standard augmentations like CutMix [65] or MixUp [67], generative models like GANs [31, 37, 69, 71] or diffusion models can perform semantic augmentation, e.g. changing the background or attributes like hair color [4,5,8,13, 26,32,35,39,58,63,64,70]. This can be viewed as indirect knowledge distillation, where the generated data represents the implicit knowledge of the generative teacher. In our work, we combine this implicit knowledge distillation together with explicit knowledge distillation by guiding the augmentation process based on the confidences of teacher and student. The assumption of also having an explicit teacher for the classification task is usually not stronger in practice, as models like CLIP [21,47] are trained on similar datasets as diffusion models.

Group Robustness without Annotations Common methods for achieving group robustness rely on labels for majority and minority groups in either the training [1, 25, 28, 51] or validation data [29, 33]. Recent approaches improve group robustness without requiring explicit group labels [24, 41, 46, 55, 66], but require all relevant groups to be represented in the dataset. This assumption is still suboptimal in practice as relevant groups may be *unknowingly absent*. Our proposed method is designed to improve robustness even for groups that are not represented at all.

3 Methods

We introduce ConfiG, a Confidence-Guided data augmentation method for knowledge distillation. We first review the concept of empirical distilled risk minimization for knowledge distillation and subsequently describe the methodology of ConfiG based on guided diffusion.

3.1 Background: Empirical Risk and Empirical Distilled Risk

We consider classification with training data $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ sampled from a distribution $\mathcal{P}_{\text{train}}$ and test data $\mathcal{D}_{\text{test}} \sim \mathcal{P}_{\text{test}}$. Both share a common discrete label space $[L]$. A student model $\mathbf{f}$ and a teacher model $\mathbf{t}$ produce predictions in form of probability distributions over the label space. For a given input $\mathbf{x}$ with ground-truth label y, we denote the predicted probability assigned to the correct class y by the student and teacher as $\mathbf{f}(\mathbf{x})_y$ and $\mathbf{t}(\mathbf{x})_y$ respectively. We refer to these values as the *confidences*. The true test risk is defined as

$$\begin{aligned}\mathcal{R}_{\mathcal{P}_{\text{test}}}(\mathbf{f}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_{\text{test}}}[\ell(\mathbf{f}(\mathbf{x}), y)] = \mathbb{E}_{x\sim\mathcal{P}_{\text{test},\mathbf{x}}}[\mathbb{E}_{y|\mathbf{x}}[\ell(\mathbf{f}(\mathbf{x}), y)]] \\ &= \mathbb{E}_{\mathcal{P}_{\text{test},\mathbf{x}}}[\mathbf{p}^*(\mathbf{x})^\top \mathbf{l}(\mathbf{f}(\mathbf{x}))]\end{aligned} \tag{1}$$

where $\mathbf{p}^*(\mathbf{x}) = [\mathbb{P}(y|\mathbf{x})]_{y\in[L]}$ is the data-generating conditional distribution over the labels, and $\mathbf{l}(\mathbf{f}(\mathbf{x})) = [\ell(\mathbf{f}(\mathbf{x}), y)]_{y\in[L]}$ denotes the vector of losses for each possible label. Similarly, the true train risk is

$$\mathcal{R}_{\text{train}}(\mathbf{f}) = \mathbb{E}_{(x,y)\sim\mathcal{P}_{\text{train}}}[\ell(\mathbf{f}(\mathbf{x}), y)] = \mathbb{E}_{\mathcal{P}_{\text{train},\mathbf{x}}}[\mathbf{p}^*(\mathbf{x})^\top \mathbf{l}(\mathbf{f}(\mathbf{x}))]. \tag{2}$$

In practice, the true data distribution is unknown, and models are instead trained by minimizing the empirical risk on the training data. In standard empirical risk minimization (**ERM**) the student model $\mathbf{f}$ is trained to minimize the empirical approximation of $\mathcal{R}_{\text{train}}$ using the labels of the train samples. The empirical risk on $\mathcal{D}_{\text{train}}$ is defined as

$$R_{\text{train}}(\mathbf{f}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{f}(\mathbf{x}_i), y_i) = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_{y_i}^T \log(\mathbf{f}(\mathbf{x}_i)), \tag{3}$$

where we choose $\ell(\cdot,\cdot)$ to be the cross-entropy loss. In knowledge distillation, the student is trained by empirical distilled risk minimization (**EDRM**) which

optimizes the cross-entropy loss between the teacher and student predictive probabilities:

$$R^D_{\text{train}}(\mathbf{f}) = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{t}(\mathbf{x}_i)^\top \log(\mathbf{f}(\mathbf{x}_i)). \tag{4}$$

Here, $\mathbf{t}(\mathbf{x}_i)$ refers to the vector with the predicted probability for each class by the teacher at $\mathbf{x}_i$. The student is trained to match these soft targets provided by the teacher instead of the hard labels. The generalization error of the student under this training scheme is defined as

$$\Delta = \mathcal{R}_{\text{test}}(\mathbf{f}) - R^D_{\text{train}}(\mathbf{f}). \tag{5}$$

Using a teacher which approximates $\mathbf{p}^*$ closely has been shown both theoretically and empirically to improve the generalization error if $\mathcal{P}_{\text{train}} = \mathcal{P}_{\text{test}}$ [38, 44, 56, 57]. This corresponds to the setting *without* covariate shift where the input distribution for training and test data is the same. In the case of covariate shift, the input marginal distributions $\mathcal{P}_{\text{train},\mathbf{x}}$ and $\mathcal{P}_{\text{test},\mathbf{x}}$ are different and only the data-generating conditional distribution $\mathbf{p}^*$ is the same at train and test time.

3.2 Confidence-Guided Data Augmentations using Diffusion Models

For knowledge distillation under covariate shift, we consider the case where $\mathcal{P}_{\text{train}} \neq \mathcal{P}_{\text{test}}$. While the error between the distilled risk and the true train risk $\mathcal{R}_{\text{train}} - R^D_{\text{train}}$ has been studied in prior works [38, 57], in the case of covariate shift this does not correspond to the true generalization error. Therefore, we first decompose Δ into two terms

$$\Delta = \Omega + \Psi \tag{6}$$

where $\Psi = \mathcal{R}_{\text{train}}(\mathbf{f}) - R^D_{\text{train}}(\mathbf{f})$ and $\Omega = \mathcal{R}_{\text{test}}(\mathbf{f}) - \mathcal{R}_{\text{train}}(\mathbf{f})$. While Ψ depends on how well the teacher approximates $\mathbf{p}^*$ on $\mathcal{P}_{\text{train}}$ [15, 38], Ω depends on the distribution of the training and test data. The goal of our data augmentation scheme is to reduce $|\Omega|$.

Motivation: Teacher-Student Disagreement under Covariate Shift We assume access to a teacher that approximates $\mathbf{p}^*(\mathbf{x})$ well on both $\mathcal{P}_{\text{train}}$ and $\mathcal{P}_{\text{test}}$. In Fig. 2, we distill a student on training data where certain

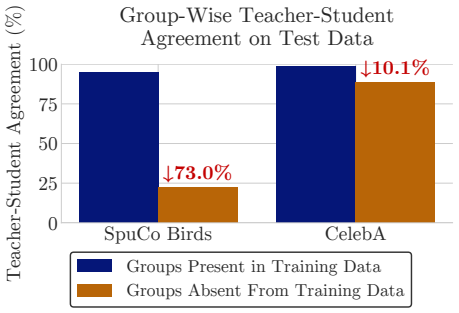

Fig. 2: For a student distilled only on the biased real training data (see Sec. 4), the agreement between teacher and student is lower on groups which are not present in the training data.

Table 1: The vast majority of test images where student and teacher disagree belong to groups not encountered in training. The students are the same as in Fig. 2

Group-Wise Disagreements on Test Data			
Dataset	Total Disagreements	On Groups in Trainset	On Groups Not in Trainset
CelebA	2351	95 (**4%**)	2256 (**96%**)
SpuCo	828	49 (**6%**)	779 (**94%**)

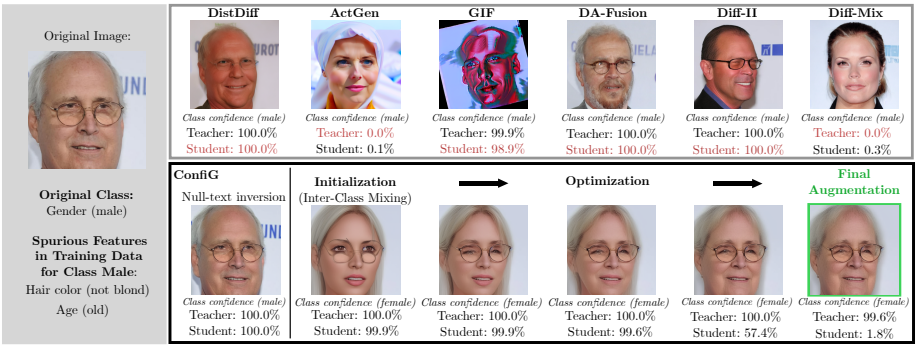

Fig. 3: Illustration of ConfiG and Other Diffusion-Based Augmentation Methods: Left: the task is gender prediction (male/female) where all men in the training data are old and not blond and all women in the training data are young and blond. **Right:** Existing augmentations yield high confidences for "male" or "female" for both the teacher and student. In contrast, ConfiG maximizes the difference between teacher and student and thus generates a challenging example for the student - in this case an old and not blond female face which has attributes that the student picked up as spurious features from the training data.

groups from the test data are fully absent and analyze the agreement of teacher and student on test data. For instance, in CelebA-HQ, where the task is gender classification (female/male), our training data only includes female samples that are young and blond. The test data, however, also includes females who are older or non-blond. We observe that the agreement between teacher and student is high for groups that are present in both the training and test data but lower for groups absent from the training data (Fig. 2). This indicates that the performance of the student is negatively influenced by covariate shift and does not accurately recover the teacher's decision boundary in these groups (see Fig. 1). Tab. 1 shows that the majority of samples where teacher and student disagree on the test set belongs to groups which are absent from the training data. Given a sample from the test set where teacher and student disagree, there is high empirical likelihood that this sample belongs to a group which is not present in the training data.

Latent Optimization: ConfiG aims to generate augmented images that reduce reliance on spurious features or dataset-specific biases without requiring explicit knowledge of what these features are. Therefore, we make use of the observation that teacher and student predictions tend to diverge on samples from groups that are absent from the training data. We use a diffusion model and real training images to generate augmented samples on which the student has low confidence in the target class while the teacher's confidence remains high. Maximizing the teacher's confidence ensures that the causal, class-relevant features are preserved and the image label remains correct. Simultaneously minimizing the student's confidence yields images that include non-causal (potentially spurious) features which lead the student into making incorrect decisions. For this purpose, we use the DiG-IN guided diffusion framework [2, 3] based on Stable Diffusion [49]. To simplify the notation below, we define $\mathcal{M}_{\epsilon_\theta}$ as the application of a noise prediction

model ϵ_θ, iteratively mapping a latent variable $\mathbf{z}_t$ to its final representation $\mathbf{z}_T$. For an image $\mathbf{x}$ with corresponding original label y, we first encode the image $\mathbf{z}_T = \mathcal{E}(\mathbf{x})$ and apply null-text inversion [40] to obtain a latent code $\mathbf{z}_0$ and null-text $\emptyset$ suitable for reconstructing $\mathbf{z}_T \approx \hat{\mathbf{z}}_T = \mathcal{M}_{\epsilon_\theta}(\mathbf{z}_0, \emptyset, C)$ where C is a conditioning signal from a CLIP text encoding. We further need to decode the latent code back to images $\mathbf{x} \approx \hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_T)$, as input for the teacher $\mathbf{t}$ and student $\mathbf{f}$. We then minimize the loss:

$$\min_{\mathbf{z}_0, \emptyset, C} \mathbf{f}(\mathcal{D}\big(\mathcal{M}_{\epsilon_\theta}(\mathbf{z}_0, \emptyset, C)\big))_y^\gamma + \Big(1 - \mathbf{t}\big(\mathcal{D}(\mathcal{M}_{\epsilon_\theta}(\mathbf{z}_0, \emptyset, C))\big)_y\Big)^\gamma \tag{7}$$

As Problem (7) is differentiable with respect to $\mathbf{z}_0$, the null-text $\emptyset$ and the conditioning C, we can solve it using a stochastic optimizer such as Adam [27]. In practice, we observe that the student often is highly confident on the initial image with small gradients [2]. We alleviate this issue by adding the nonlinear transformation $(\cdot)^\gamma$ to improve optimization speed. Empirically, we find $\gamma = 2$ to perform well. In addition, we follow DiG-IN and use a foreground-aware distance regularization. An example of the optimization process is visualized in Fig. 3. The closest baseline is ActGen [20], which can generate augmentations that minimize the cross-entropy loss of the auxiliary student model given a real image and a target class. However, it lacks explicit class guidance apart from the text prompt and the real image. As we show in Fig. 3, ActGen may thus modify the causal attribute (i.e., gender) rather than only the spurious features.

Warm-starting the Optimization through Inter-Class Mixing: DiffMix [60] has demonstrated that inter-class mixing of training samples greatly benefits the effectiveness of diffusion-based data augmentations. For ConfiG, we combine inter-class mixing with latent optimization. For every augmentation we sample a random target class, which is constrained to be *different* from the original class. We warm-start the optimization process by first performing null-text inversion using the original image label, followed by a version of prompt-to-prompt editing [18] modified for DiG-IN using the target class label. The latent of the resulting image is taken as the initialization for the optimization problem of Equation (7) which is optimized with respect to the target class. In Fig. 4 we visualize three example augmentations after only inter-class mixing and of ConfiG (with additional subsequent latent optimization).

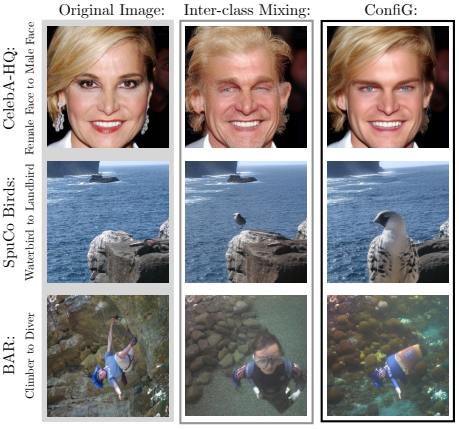

Fig. 4: Images from CelebA-HQ, SpuCo Birds, and BAR. Example augmentations show inter-class mixing alone versus the full ConfiG pipeline, which combines this mixing with latent optimization.

Auxiliary student To generate images with ConfiG, we require a student net-

work to optimize Equation (7). We fine-tune an auxiliary student on only the real training data, where it picks up the spurious features and biases as demonstrated in Fig. 2. It is then used to generate the synthetic augmentations, which replace real samples with probability $\alpha = 0.5$ in the training of more robust students.
Theoretical Motivation We provide a theoretical investigation of the effect of using ConfiG augmentations in App. A.

4 Experiments

In our experiments, we empirically demonstrate the effectiveness of ConfiG to improve knowledge distillation under covariate shift. We first outline the experimental setup (Sec. 4.1) before discussing the results on datasets with group shift (Sec. 4.2), Spurious ImageNet (Sec. 4.3) and the ablations (Sec. 4.4).

4.1 Experimental setup

Datasets. Our experiments are based on four datasets for image classification. We use CelebA-HQ [23, 34], SpuCo Birds [22] and BAR [41] to investigate group shift as an example for covariate shift. We subsample CelebA-HQ and SpuCo Birds to obtain a training dataset with 250 samples per class, *fully* discarding certain groups from the training data. For *CelebA-HQ* we only retain training images of young, blond female celebrities as well as old, non-blond male celebrities for training. The descriptive feature is the gender, other features are spuriously correlated only in the training split. The test set spans all combinations of age (old/young), hair color (blond/non-blond) and gender (female/male) which leads to eight test groups. *SpuCo Birds* includes waterbirds or landbirds over water or land backgrounds, with bird type as the descriptive feature. For SpuCo Birds, we only retain training images of waterbirds with water background and landbirds with land background such that the background becomes a binary spurious feature. The test split also contains waterbirds with land background and vice versa. The *Biased Activity Recognition* dataset (BAR) contains images of six sport activities (climbing, diving, fishing, pole vaulting, motorsports racing, throwing) where the training images are biased towards distinct places per class. We sample 100 images per class for training. The test images feature the same sport activities but were taken at different places which have no overlap with the places for the training set. Our fourth dataset is a subset of *ImageNet* [50]. We subsample a challenging training dataset with 100 training images for each of the 100 classes contained in Spurious ImageNet [42]. We denote this setting by *ImageNet-100*. Further details on training and test data selection can be found in App. B.
Diffusion-Based Data Augmentation. The diffusion-based augmentations are combined with the real training sets. Following DA-Fusion, Diff-Mix and Diff-II, we replace a natural image by a synthetic augmentation with probability $\alpha = 0.5$. The hyperparameters of ConfiG can be found in App. B. For ConfiG we generate one synthetic augmentation per image. To ensure a fair comparison with the baselines, we adjust the number of synthetic augmentations generated

by each method such that the computational cost of data generation, measured in FLOPs, is approximately matched. More details about the computational cost of ConfiG and the multipliers used for all baselines and datasets can be found in App. C. Experiments with more ConfiG augmentations per image can be found in Sec. 4.4 and App. F.

Baselines. We compare ConfiG to six state-of-the-art diffusion-based data augmentation methods: GIF [68], DistDiff [72], DA-Fusion [54], Diff-Mix [60], Diff-II [59] and ActGen [20]. All methods generate augmentations by applying a diffusion model to augment real images. To ensure a fair comparison, we follow the baselines [54, 68, 72] and use Stable Diffusion 1.4 [49] as implicit teacher for all methods. In App. E we demonstrate that ConfiG also works with modern single-step diffusion models [48] for faster generation.

ActGen and ConfiG use a student model for generating data augmentations. For this purpose, we use the setup introduced in Sec. 3.2 and first distill an auxiliary student (which can pick up spurious features due to the limited coverage of the training data) and then use this student to generate augmentations with the goal of getting close to the teacher model even outside the domain covered by the training data. In this way, the obtained student is more robust against picking up spurious features as shown by a significantly improved worst group accuracy. Further details on the baselines are listed in App. N.

Models and Training Hyperparameters. For our main experiments we use a ViT-T [9] student pre-trained on ImageNet-21k using the AugReg recipe [53]. Ablations with different student and teacher models are in App. G. We use a CLIP [21, 47] ViT-L/14 pre-trained on DataComp-XL [11] as teacher. For CelebA-HQ and BAR we directly use classnames to obtain zero-shot predictions from the teacher. For SpuCo Birds we first use fine-grained bird categories and then aggregate into waterbirds and landbirds to obtain robust predictions (see App. B for details). The standard, non-synthetic augmentations used for training on CelebA-HQ, SpuCo Birds and BAR are random resized crops by default. Ablations with stronger, non-synthetic augmentations can be found in Sec. 4.4. On ImageNet-100, we follow the setup of VanillaKD [16] and use a BEiTv2-B [43] teacher trained on the full ImageNet dataset together with the same "A1" augmentations [61] for our experiments. A detailed overview of all hyperparameters for training and standard augmentations can be found in App. B.

Evaluation Metrics. We consider three evaluation metrics for SpuCo Birds and CelebA-HQ. Given a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ composed of K disjoint groups $\{\mathcal{D}_{\text{test}}^1, ..., \mathcal{D}_{\text{test}}^K\}$, the sample mean accuracy (**SMA**) is the average accuracy independent of groups:

$$\mathbf{SMA}(f) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_{\text{test}}} \mathbb{1}(\text{argmax } \mathbf{f}(\mathbf{x}_j) = y_j) \tag{8}$$

while the group mean accuracy (**GMA**) computes the average accuracy when groups are weighted equally:

$$\mathbf{GMA}(f) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{|\mathcal{D}^i_{\text{test}}|}\sum_{(\mathbf{x}_j,y_j)\in\mathcal{D}^i_{\text{test}}}\mathbb{1}(\operatorname{argmax}\mathbf{f}(\mathbf{x}_j)=y_j) \tag{9}$$

The worst group accuracy (**WGA**) is the minimum accuracy on a group:

$$\mathbf{WGA}(f) = \min_{i=1,\dots,K}\Big(\frac{1}{|\mathcal{D}^i_{\text{test}}|}\sum_{(\mathbf{x}_j,y_j)\in\mathcal{D}^i_{\text{test}}}\mathbb{1}(\operatorname{argmax}\mathbf{f}(\mathbf{x}_j)=y_j)\Big). \tag{10}$$

On SpuCo Birds, the test groups are balanced so group and sample mean accuracies coincide. On BAR, the test data contains only one group per class which is disjoint from the groups in the training data such that we only evaluate the sample mean accuracy. On Spurious ImageNet, we consider two evaluation metrics. First, the validation accuracy on all images from the corresponding classes of the ImageNet validation split. Second, the mean spurious score introduced by [42], computed using the predicted probability for the corresponding class, which we denote by SpuScore. The SpuScore measures the class-wise separation of images from Spurious ImageNet [42], containing the spurious feature but not the class object, versus the validation set of images of that class. The SpuScore averages these scores over all 100 classes.

4.2 Group Shift Datasets

Quantitative Results. Tab. 2 summarizes results on CelebA-HQ, SpuCo Birds and BAR. We make two key observations: First, synthetic data augmentations

Table 2: Knowledge distillation using diffusion-based data augmentation under covariate shift: We report mean in % and standard deviation over six training runs. The student is a ViT-T distilled from a CLIP ViT-L teacher using EDRM with random resized crops as standard augmentations. Real Data refers to the student obtained using only real training images whereas the rows below correspond to the student models of different diffusion-based data augmentation methods. The highest score in each column is highlighted in **bold**, the second highest is underlined.

	CelebA-HQ			SpuCo Birds		BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**	**WGA**	**SMA**
Teacher	99.03	97.72	91.89	96.90	93.80	95.72
Real Data	93.98 ±1.68	86.89 ±2.48	53.44 ±8.12	56.53 ±2.40	12.97 ±6.65	37.69 ±3.58
GIF	96.70 ±0.87	90.78 ±1.00	68.92 ±7.60	63.15 ±1.20	25.90 ±2.53	47.91 ±3.30
DistDiff	96.33 ±0.58	90.41 ±1.11	70.25 ±5.46	63.94 ±2.33	24.07 ±4.62	45.91 ±3.36
DA-Fusion	95.39 ±1.52	90.09 ±0.78	67.96 ±6.79	71.45 ±1.54	30.20 ±5.18	53.03 ±3.34
Diff-Mix	97.36 ±0.16	93.07 ±0.73	70.63 ±3.76	69.49 ±2.33	32.13 ±2.22	**59.28** ±2.32
Diff-II	95.08 ±1.75	89.48 ±1.41	64.96 ±7.62	61.88 ±1.80	22.90 ±4.25	56.40 ±2.34
ActGen	97.04 ±0.48	92.61 ±0.67	75.68 ±4.22	64.54 ±2.07	25.97 ±1.80	55.71 ±2.70
ConfiG	**97.76** ±0.25	**95.67** ±0.55	**88.15** ±2.10	**73.38** ±1.28	**39.50** ±4.53	58.66 ±3.76

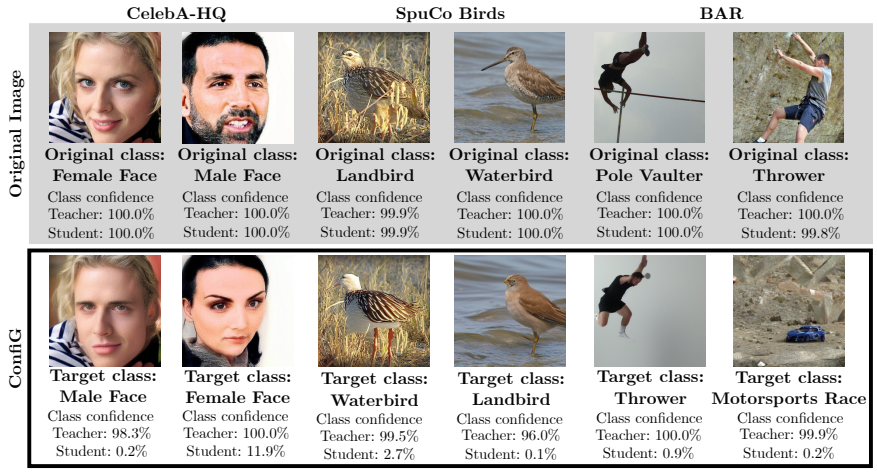

Fig. 5: Qualitative Examples of ConfiG on CelebA-HQ, SpuCo Birds and BAR. Note that the target class is chosen randomly to be different from the original class. The images generated by ConfiG have low confidence of the student and show that it relies on spurious features: a young male with blond hair is recognized as female (CelebA-HQ) or a landbird standing in water as waterbird.

generally improve the test performance in comparison to only training with real images. Second, ConfiG consistently yields the best or second-best SMA and best GMA and WGA results. Across all metrics, our method is at least on par with the best baseline despite using fewer synthetic augmentations. On CelebA-HQ, the performance of the student using ConfiG augmentations is generally the highest, with a 1% gap to the teacher in SMA and a 4% gap in WGA. In comparison to only real data, the absolute WGA improved by over 34%. On the SpuCo Birds and BAR datasets, a larger gap to the teacher remains of over 40% and 58% in SMA when training with real data only. ConfiG augmentations reduce this difference to 23% and 37% respectively. We demonstrate that combining stronger, non-synthetic data augmentations with ConfiG can further improve the student performance in Sec. 4.4.

Qualitative Results. We show six examples for ConfiG augmentations in Fig. 5. The ConfiG augmentations retain high teacher confidence while reducing the student confidence. Further examples can be found in App. O.

4.3 ImageNet-100

Tab. 3 presents results for fine-tuning ViT-T and ViT-S students on ImageNet-100. Synthetic data augmentations generally improve the validation further, with ConfiG achieving the best validation accuracies and spurious scores. For the ViT-T student, ConfiG improves the validation accuracy by more than 1.6% and the spurious score by more than 1.8% in comparison to all baselines. The improvements for the ViT-S student by 0.6% in validation accuracy and by 0.9% in spurious score are

slightly smaller. These results show that ConfiG consistently enhances knowledge distillation under covariate shift, not only in group-based settings but also for class-specific spurious features.

Table 3: Results on ImageNet-100. ConfiG improves the SpuScore in comparison to all baselines and yields the best validation accuracy. The reported values are in %. The highest score in each column is highlighted in **bold**, the second highest is underlined. The teacher has a SMA of 96.12% and a SpuScore of 85.45.

	ViT-T		ViT-S	
Method	**SMA**	**SpuScore**	**SMA**	**SpuScore**
Real data	71.46	54.05	86.46	68.62
GIF	76.22	58.55	89.36	72.64
DistDiff	76.84	59.13	90.66	74.59
DA-Fusion	74.92	56.99	89.84	73.40
Diff-Mix	76.88	59.48	89.98	72.32
Diff-II	74.76	58.43	89.00	74.53
ActGen	78.36	60.48	91.04	73.17
ConfiG	**80.02**	**62.36**	**91.68**	**75.49**

4.4 Ablations

We perform ablation studies on the group shift datasets to investigate the impact of inter-class mixing and latent optimization and analyze the effect of more synthetic and stronger non-synthetic augmentations. Furthermore, we investigate the influence of weaker teacher models and demonstrate that ConfiG improves the performance of model-centric bias mitigation methods.

Individual Components of ConfiG. As introduced in Sec. 3.2, ConfiG combines inter-class mixing through prompt-to-prompt editing with latent optimization. We assess the effect of these components individually in Tab. 4. Here only latent optimization refers to optimizing Problem (7) without inter-class mixing which means the target class is the same original class. While both components individually improve performance over training on purely real data, inter-class mixing alone is sufficient to be on par or even outperform the strongest baselines on CelebA-HQ. On BAR only performing inter-class mixing through prompt-to-prompt editing is less effective. Latent optimization directly after null-text

Table 4: Only Inter-class Mixing or Latent Optimization compared to ConfiG. Using only latent optimization after null-text inversion without inter-class mixing (using the original class as target class) or purely inter-class mixing with prompt-to-prompt (p2p) editing as data augmentation improves the performance in comparison to training on real data only. Combining both into ConfiG achieves the best results.

	CelebA-HQ			SpuCo Birds		BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**	**WGA**	**SMA**
Real Data	93.98 ±1.68	86.89 ±2.48	53.44 ±8.12	56.53 ±2.40	12.97 ±6.65	37.69 ±3.58
Latent Opt.	93.35 ±1.33	89.01 ±1.13	71.11 ±3.74	61.53 ±2.03	21.63 ±3.61	39.40 ±3.53
Only p2p	97.40 ±0.23	94.74 ±0.55	86.99 ±1.61	66.64 ±1.83	33.20 ±3.41	44.09 ±3.77
ConfiG	**97.76** ±0.25	**95.67** ±0.55	**88.15** ±2.10	**73.38** ±1.28	**39.50** ±4.53	**58.66** ±3.76

Table 5: Stronger Standard Data Augmentation. Combining stronger, non-synthetic augmentations with diffusion-based augmentations further improves the student performance both for Diff-Mix and ConfiG. ConfiG yields stronger gains than Diff-Mix.

	CelebA-HQ			SpuCo Birds		BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**	**WGA**	**SMA**
Stronger Random Resized Crops + Flips						
Real Data	97.17 ±0.66	91.81 ±0.77	69.97 ±7.19	60.58 ±0.79	17.63 ±2.77	52.17 ±3.98
Diff-Mix	97.69 ±0.38	93.58 ±1.02	72.62 ±6.69	70.28 ±3.28	33.40 ±6.39	64.91 ±1.16
ConfiG	**98.42** ±0.05	**96.72** ±0.30	**90.98** ±1.40	**73.78** ±1.60	**42.10** ±5.00	**65.39** ±1.91
Stronger Random Resized Crops + Flips + AutoAugment [7]						
Real Data	97.82 ±0.11	92.50 ±0.67	72.22 ±3.76	65.46 ±0.80	25.17 ±1.99	61.65 ±2.99
Diff-Mix	97.86 ±0.21	93.60 ±0.81	74.21 ±5.61	72.63 ±2.24	<u>36.50</u> ±7.86	67.89 ±2.27
ConfiG	**98.53** ±0.05	**96.55** ±0.36	**89.45** ±1.59	**78.62** ±1.61	**50.53** ±5.34	**69.78** ±0.70
Stronger Random Resized Crops + Flips + MixUp [67]						
Real Data	97.66 ±0.43	93.15 ±1.08	74.35 ±8.53	68.99 ±0.71	37.20 ±1.31	61.70 ±1.71
Diff-Mix	<u>98.06</u> ±0.08	94.38 ±0.60	77.65 ±3.60	73.19 ±1.28	<u>43.00</u> ±4.93	65.62 ±1.10
ConfiG	**98.42** ±0.11	**96.79** ±0.48	**89.95** ±1.38	**77.99** ±1.97	**51.57** ±5.73	**68.81** ±1.86
Stronger Random Resized Crops + Flips + CutMix [65]						
Real Data	97.74 ±0.26	93.17 ±0.77	70.24 ±4.59	70.05 ±2.15	36.63 ±4.29	58.21 ±2.61
Diff-Mix	97.77 ±0.07	93.50 ±0.40	71.43 ±2.41	71.48 ±2.37	<u>38.97</u> ±3.55	62.31 ±1.69
ConfiG	**98.33** ±0.21	**96.14** ±0.67	**87.90** ±2.80	**77.66** ±0.52	**53.10** ±3.42	**65.34** ±1.75

inversion performs worse than only inter-class mixing. However, one gets a substantial boost in performance in all evaluation metrics when combining both techniques as done in ConfiG.

Stronger Standard Augmentations. We report the results of distilling students with stronger, non-synthetic augmentations in Tab. 5. We compare ConfiG to Diff-Mix as the strongest baseline. The complete results for all baselines can be found in App. D. Compared to Tab. 2, we find that on BAR stronger standard augmentations with AutoAugment [7] and MixUp [67] considerably improve the student performance, outperforming synthetic data augmentations with only random resized crops. On CelebA-HQ and SpuCo Birds all standard augmentations with only real data remain worse in GMA and WGA than ConfiG with purely random resized crops. In all cases, the best overall performance can be achieved when combining stronger standard augmentations with ConfiG.

Larger Number of Synthetic Augmentations. Tab. 6 shows results with two augmentations per image for ConfiG and 56 instead of 28 augmentations per image for Diff-Mix as the strongest baseline matched by the approximate generation FLOPs. Results for the remaining baselines can be found in App. F. On CelebA-HQ the student performance remains close to the main results from Tab. 2 for both ConfiG and Diff-Mix. On SpuCo Birds and BAR the performance of ConfiG improves with twice as many augmentations and the gap between Diff-Mix and ConfiG increases. This indicates more saturated performances with

Table 6: Using two augmentations per image for ConfiG. The results with Diff-Mix are obtained with 56 augmentations per image to approximately match the generation costs in FLOPs (see App. C). The replacement probability α is kept at 0.5, δ refers to the difference to the results from Tab. 2.

	CelebA-HQ			SpuCo Birds		BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**	**WGA**	**SMA**
Diff-Mix	97.61 ±0.19	93.62 ±0.44	72.62 ±2.83	69.85 ±2.01	34.53 ±5.30	61.44 ±2.11
$\delta_{28\to56}$	+0.25	+0.55	+1.99	+0.36	+2.40	+2.16
ConfiG	**97.75** ±0.18	**95.67** ±0.34	**87.83** ±1.64	**74.64** ±1.72	**46.90** ±6.36	**63.12** ±1.71
$\delta_{1\to2}$	-0.01	+0.00	-0.32	+1.26	+7.40	+4.46

Diff-Mix while ConfiG demonstrates better scaling.

Weaker Teachers. Since the teacher influences both the generated augmentations and the downstream training of the student model when using ConfiG, we investigate the effect of using weaker teacher models. We exchange the CLIP ViT-L/14 used in Sec. 4.2 by a CLIP RN50 trained on YFCC15M for the CelebA-HQ dataset and a CLIP ViT-B/32 trained on DataComp M for BAR. The student model and all hyperparameters remain the same as in Sec. 4.2. Tab. 7 demonstrates that a weaker teacher can indeed decrease the student performance. However, ConfiG still improves the performance over real data and Diff-Mix. The complete results for all baselines can be found in App. J.

Table 7: Results with Weaker Teacher Models. We replace the CLIP ViT-L/14 used in Sec. 4.2 by weaker teachers (RN50 for CelebA-HQ and CLIP ViT-B/32 for BAR) which decreases the student performance but ConfiG still improves the performance over only real data and the Diff-Mix baseline.

	CelebA-HQ			BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**
Teacher	96.48	92.23	75.68	59.63
Real Data	94.37 ±0.47	83.37 ±1.48	38.36 ±5.37	36.65 ±1.68
Diff-Mix	94.48 ±0.73	86.62 ±1.33	50.79 ±4.99	40.67 ±2.82
ConfiG	**95.25** ±0.52	**90.16** ±0.63	**67.72** ±2.91	**44.55** ±1.76

Data-Centric vs. Model-Centric Bias Mitigation. The goal of ConfiG is to use synthetic data augmentations to mitigate the problem of limited coverage of the training data which can lead to biases and shortcut learning, particularly in scenarios where certain subgroups are entirely absent. This approach complements existing strategies that focus on learning robust models from a fixed dataset with potential biases [66]. We demonstrate this by exemplarily combining ConfiG with pure ERM, LfF [41] and uLA [55], three approaches for training classification models on biased datasets without bias information. Since LfF and uLA are based on ERM, we label the synthetic data augmentations using the teacher before training. In Tab. 8 we show that LfF, uLA, and pure ERM are substantially improved with synthetic data augmentations from ConfiG.

Table 8: Using ConfiG Augmentations Improves Model-Centric Methods for Bias Mitigation. We compare pure ERM, LfF [41] and uLA [55] with and without ConfiG augmentations. We labeled the ConfiG augmentations using the teacher prediction.

	CelebA-HQ			SpuCo Birds		BAR
Method	**SMA**	**GMA**	**WGA**	**SMA**	**WGA**	**SMA**
ERM	90.34 ±3.24	84.94 ±1.89	57.84 ±3.93	53.17 ±1.03	3.67 ±2.27	38.43 ±1.88
uLA	84.51 ±3.72	79.64 ±2.43	45.57 ±7.38	53.16 ±0.79	3.37 ±1.60	39.14 ±3.94
LfF	87.98 ±5.25	82.60 ±3.84	51.22 ±7.15	52.47 ±0.40	2.87 ±1.22	35.55 ±1.60
ERM+ConfiG	97.46 ±0.16	95.52 ±0.40	90.25 ±1.09	69.26 ±3.99	40.57 ±7.89	**58.33** ±2.42
uLA+ConfiG	97.33 ±0.33	95.30 ±0.90	89.12 ±4.97	**70.38** ±2.72	**40.87** ±4.14	57.75 ±1.92
LfF+ConfiG	**97.49** ±0.28	**95.82** ±0.47	**90.72** ±1.92	67.86 ±0.64	32.93 ±3.07	56.75 ±4.18

5 Conclusion

We introduce ConfiG, a novel confidence-guided data augmentation method designed to enhance knowledge distillation under unknown covariate shift. Our approach effectively addresses the challenge of group shift or class-level spurious features in training data by generating targeted augmentations that maximize the disagreement between a robust teacher and biased student models. Experimental results on datasets such as CelebA-HQ, SpuCo Birds, and BAR demonstrate that ConfiG improves both worst group and mean group accuracy in a group shift setting and is best or second-best in sample mean accuracy. On ImageNet-100, ConfiG improves the validation accuracy and SpuScore, indicating increased robustness of the student against class-wise spurious features.

Limitations and Future Work. We have demonstrated ConfiG in a classification setting where images are augmented globally. Location-aware tasks such as object detection require constraining the image augmentation process [10], which we leave for future work. Furthermore, our experimental setup requires pre-trained teacher models for knowledge distillation and data generation for both ConfiG and the baselines. As dataset and model sizes continue to grow, general purpose models such as CLIP and Stable Diffusion are suitable teachers for a wide range of domains. Future research could explore dedicated combinations of synthetic data augmentation and model-centric bias mitigation strategies. Finally, while we demonstrate that ConfiG works with Hyper-SD as a modern diffusion backbone in App. E, we believe that these initial results can be further improved. We leave the exploration of other model types, e.g. context-guided generation [30, 62], to future work.

Acknowledgements

We would like to thank Maximilian Augustin and Yannic Neuhaus for support on the DiG-IN implementation as well as Dan Zhang for helpful discussions on the manuscript. We also thank the European Laboratory for Learning and Intelligent Systems (ELLIS) for supporting Niclas Popp.

References

1. Ashurst, C., Weller, A.: Fairness without demographic data: A survey of approaches. In: EAAMO (2023)
2. Augustin, M., Neuhaus, Y., Hein, M.: Dig-in: Diffusion guidance for investigating networks - uncovering classifier differences, neuron visualisations, and visual counterfactual explanations. In: CVPR (2024)
3. Augustin, M., Neuhaus, Y., Hein, M.: Dash: Detection and assessment of systematic hallucinations of vlms. In: CVPR (2025)
4. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. TMLR (2023)
5. Bender, S., Delzer, O., Herrmann, J., Marxfeld, H.A., Müller, K.R., Montavon, G.: Mitigating clever hans strategies in image classifiers through generating counterexamples. Information Fusion **135**, 104406 (2026)
6. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: CVPR (2022)
7. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: CVPR (2019)
8. Dong, Y., Su, F.Y., Chiang, J.H.: Sgd-mix: Enhancing domain-specific image classification with label-preserving data augmentation. In: WACV (2026)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. Fang, H., Han, B., Zhang, S., Zhou, S., Hu, C., Ye, W.M.: Data augmentation for object detection via controllable diffusion models. In: WACV (2024)
11. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S.M., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A.J., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: Datacomp: In search of the next generation of multimodal datasets. In: NeurIPS (2023)
12. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R.S., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11) (2020)
13. Ghosh, S., Evuru, C.K.R., Kumar, S., Tyagi, U., Sakshi, S., Chowdhury, S., Manocha, D.: ASPIRE: language-guided data augmentation for improving robustness against spurious correlations. In: ACL Findings (2024)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. IJCV **129**(6) (2021)
15. Hamidi, S.M., Deng, X., Tan, R., Ye, L., Salamah, A.H.: How to train the teacher model for effective knowledge distillation. In: ECCV (2024)
16. Hao, Z., Guo, J., Han, K., Hu, H., Xu, C., Wang, Y.: Revisit the power of vanilla knowledge distillation: from small scale to large scale. In: NeurIPS (2023)
17. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? In: ICLR (2023)
18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint: https://arxiv.org/abs/2208.01626 (2022)

19. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint: https://arxiv.org/abs/1503.02531 (2015)
20. Huang, T., Liu, J., You, S., Xu, C.: Active generation for image classification. In: ECCV (2024)
21. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip. 10.5281/zenodo.5143773 (2021)
22. Joshi, S., Yang, Y., Xue, Y., Yang, W., Mirzasoleiman, B.: Challenges and opportunities in improving worst-group generalization in presence of spurious features. arXiv preprint: https://arxiv.org/abs/2306.11957 (2025)
23. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
24. Kenfack, P.J., Aïvodji, U., Kahou, S.E.: Adaptive group robust ensemble knowledge distillation. TMLR (2025)
25. Kenfack, P.J., Kahou, S.E., Aïvodji, U.: A survey on fairness without demographics. TMLR (2024)
26. Kim, J.M., Bader, J., Alaniz, S., Schmid, C., Akata, Z.: Datadream: Few-shot guided dataset generation. In: ECCV (2024)
27. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR (2014)
28. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. In: ICLR (2023)
29. LaBonte, T., Muthukumar, V., Kumar, A.: Towards last-layer retraining for group robustness with fewer annotations. In: NeurIPS (2023)
30. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint: https://arxiv.org/abs/2506.15742 (2025)
31. Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: CVPR (2022)
32. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Adversarial supervision makes layout-to-image diffusion models thrive. In: ICLR (2024)
33. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: ICML (2021)
34. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
35. Lu, Z., Xu, Q., Wen, P., Dai, S., Huang, Q.: Higfa: Hierarchical guidance for fine-grained data augmentation with diffusion models. In: AAAI (2026)
36. Mansourian, A.M., Ahmadi, R., Ghafouri, M., Babaei, A.M., Golezani, E.B., Ghamchi, Z.Y., Ramezanian, V., Taherian, A., Dinashi, K., Miri, A., Kasaei, S.: A comprehensive survey on knowledge distillation. TMLR (2025)
37. Menke, M., Wenzel, T., Schwung, A.: AWADA: attention-weighted adversarial domain adaptation for object detection. arXiv preprint: https://arxiv.org/abs/2208.14662 (2022)
38. Menon, A.K., Rawat, A.S., Reddi, S.J., Kim, S., Kumar, S.: A statistical perspective on distillation. In: ICML (2021)
39. Michaeli, E., Fried, O.: Advancing fine-grained classification by structure and subject preserving augmentation. In: NeurIPS (2024)
40. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR (2023)

41. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: Training debiased classifier from biased classifier. In: NeurIPS (2020)
42. Neuhaus, Y., Augustin, M., Boreiko, V., Hein, M.: Spurious features everywhere - large-scale detection of harmful spurious features in imagenet. In: CVPR (2023)
43. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint: https://arxiv.org/abs/2208.06366 (2022)
44. Phuong, M., Lampert, C.: Towards understanding knowledge distillation. In: ICML (2019)
45. Popp, N., Metzen, J.H., Hein, M.: Feature distillation improves zero-shot transfer from synthetic images. TMLR (2024)
46. Qiu, S., Potapczynski, A., Izmailov, P., Wilson, A.G.: Simple and fast group robustness by automatic feature reweighting. In: ICML (2023)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
48. Ren, Y., Xia, X., Lu, Y., Zhang, J., Wu, J., Xie, P., Wang, X., Xiao, X.: Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In: NeurIPS (2024)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV **115**(3) (2015)
51. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: ICLR (2020)
52. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? In: NeurIPS (2021)
53. Steiner, A.P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers. TMLR (2022)
54. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: ICLR (2024)
55. Tsirigotis, C., Monteiro, J., Rodríguez, P., Vázquez, D., Courville, A.C.: Group robust classification without any group information. In: NeurIPS (2023)
56. Wang, C., Yang, Q., Huang, R., Song, S., Huang, G.: Efficient knowledge distillation from model checkpoints. In: NeurIPS (2022)
57. Wang, H., Lohit, S., Jones, M.J., Fu, Y.: What makes a "good" data augmentation in knowledge distillation - A statistical perspective. In: NeurIPS (2022)
58. Wang, J., Laube, K.A., Li, Y., Metzen, J.H., Cheng, S., Borges, J., Khoreva, A.: Label-free neural semantic image synthesis. In: ECCV (2024)
59. Wang, Y., Chen, L.: Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In: CVPR (2025)
60. Wang, Z., Wei, L., Wang, T., Chen, H., Hao, Y., Wang, X., He, X., Tian, Q.: Enhance image classification via inter-class image mixup with diffusion model. In: CVPR (2024)
61. Wightman, R., Touvron, H., Jegou, H.: Resnet strikes back: An improved training procedure in timm. In: NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future (2021)

62. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint: https://arxiv.org/abs/2508.02324 (2025)
63. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. In: NeurIPS (2023)
64. Yang, W., Wu, X., Deng, Z., Tureci, E., Russakovsky, O.: Beyond objects: Contextual synthetic data generation for fine-grained classification. In: CVPR (2026)
65. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
66. Zarlenga, M.E., Sankaranarayanan, S., Andrews, J.T.A., Shams, Z., Jamnik, M., Xiang, A.: Efficient bias mitigation without privileged information. In: ECCV (2024)
67. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
68. Zhang, Y., Zhou, D., Hooi, B., Wang, K., Feng, J.: Expanding small-scale datasets with guided imagination. In: NeurIPS (2023)
69. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: CVPR (2021)
70. Zhao, T., Chen, X., Li, Z., Fang, J., An, D., Xu, X., Tu, Z., Xing, Y.: Salient concept-aware generative data augmentation. In: NeurIPS (2025)
71. Zhou, A., Wang, J., Wang, Y.X., Wang, H.: Distilling out-of-distribution robustness from vision-language foundation models. In: NeurIPS (2023)
72. Zhu, H., Yang, L., Yong, J., Yin, H., Jiang, J., Xiao, M., Zhang, W., Wang, B.: Distribution-aware data expansion with diffusion models. In: NeurIPS (2024)