Fine-Grained Generalization via Structuralizing Concept and Feature Space into Commonality, Specificity and Confounding

Zhen Wang[1], Jiaojiao Zhao[1], Qilong Wang[2], Yongfeng Dong[1], Wenlong Yu[2,3]*

[1]School of Artificial Intelligence, Hebei University of Technology, China
[2]Tianjin University, China
[3]Singapore Management University, Singapore
{wangzhen@, 202432803020@stu., dongyf@}hebut.edu.cn, {qlwang, ywl_95}@tju.edu.cn

Abstract

Fine-Grained Domain Generalization (FGDG) presents greater challenges than conventional domain generalization due to the subtle inter-class differences and relatively pronounced intra-class variations inherent in fine-grained recognition tasks. Under domain shifts, the model becomes overly sensitive to fine-grained cues, leading to the suppression of critical features and a significant drop in performance. Cognitive studies suggest that humans classify objects by leveraging both common and specific attributes, enabling accurate differentiation between fine-grained categories. However, current deep learning models have yet to incorporate this mechanism effectively. Inspired by this mechanism, we propose Concept-Feature Structuralized Generalization (CFSG). This model explicitly disentangles both the concept and feature spaces into three structured components: common, specific, and confounding segments. To mitigate the adverse effects of varying degrees of distribution shift, we introduce an adaptive mechanism that dynamically adjusts the proportions of common, specific, and confounding components. In the final prediction, explicit weights are assigned to each pair of components. Extensive experiments on three single-source benchmark datasets demonstrate that CFSG achieves an average performance improvement of 9.87% over baseline models and outperforms existing state-of-the-art methods by an average of 3.08%. Additionally, explainability analysis validates that CFSG effectively integrates multi-granularity structured knowledge and confirms that feature structuralization facilitates the emergence of concept structuralization.

Code — https://github.com/zhaozz-j/CFSG

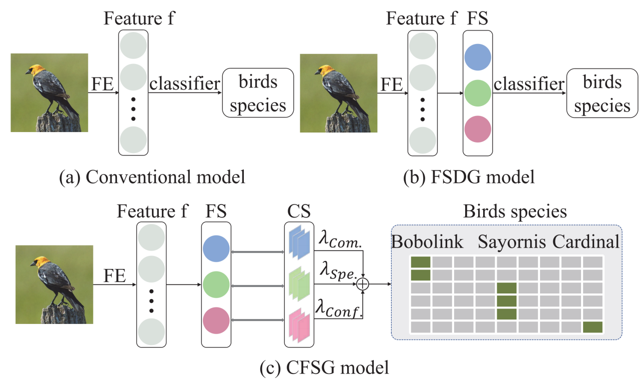

Figure 1: Decision mechanisms of different methods: conventional models predict directly from extracted features; FSDG structures the feature space, but classification still relies on features. In contrast, CFSG first achieves concept structuralization through structured features and then performs classification based on these structured concepts. FE stands for Feature Extractor, FS refers to Feature Structuralization, and CS represents Concept Structuralization.

Introduction

Deep learning has recently led to significant advances in computer vision (Krizhevsky, Sutskever, and Hinton 2017; Dong et al. 2025) and natural language processing (Devlin et al. 2019). However, these achievements rely primarily on the assumption that the training and testing data are drawn from the same distribution. In real-world scenarios, this assumption often fails due to distribution shifts across domains. For example, in image classification tasks, training data may consist of real-world bird images, while testing data could include artistic representations such as watercolor or oil paintings of birds. Due to these distributional differences across domains, traditional models often perform poorly on testing data. This issue is commonly referred to as the out-of-distribution (OOD) problem (Liu et al. 2021). Domain Generalization (DG) aims to train deep learning models solely on source domain data to generalize effectively to OOD samples and has gained increasing attention in recent years (Wang et al. 2022).

Although some DG methods have made significant progress in addressing OOD challenges, their performance on fine-grained domain generalization (FGDG) tasks remains unsatisfactory (Wei et al. 2021; Bi et al. 2025). Fine-grained recognition focuses on subtle inter-class differences, such as variations among breeds within the same species. In these tasks, intra-class variations are often more pronounced than inter-class differences and may manifest differently across domains (Du et al. 2020). Fine-grained tasks heavily rely on local discriminative regions; however, conventional techniques such as data augmentation damage these criti-

*Corresponding author.

cal areas, leading to substantial performance drops. In addition, FGDG is particularly sensitive to cross-domain distribution shifts, posing greater challenges to traditional DG methods and limiting their generalization capabilities (Zhou et al. 2021). Moreover, in real-world scenarios, conventional domain generalization methods, especially in single-source domain scenarios, tend to fail; the scarcity of multi-source fine-grained data makes single-source FGDG an even more demanding problem (Mao and Zhang 2024).

Humans significantly outperform machines on such tasks, primarily due to fundamental differences in decision-making processes. Cognitive studies indicate that humans do not rely solely on superficial similarity when distinguishing categories; instead, they leverage semantic concepts based on both commonality and specificity between categories, enabling robust generalization even when the target's appearance changes (Luppi et al. 2024; Yu et al. 2025b; Mervis and Rosch 1975; Mahner et al. 2025; Du et al. 2025; Yu, Yan, and Jin 2025). However, this decision-making strategy has yet to be effectively embedded in current deep learning models. Recently, Yu et al. (Yu et al. 2025a) proposed an FSDG model that incorporates human decision-making mechanisms into the training process. By structuring the feature space and applying constraints, the model is guided to focus on both common and specific features, resulting in significantly improved generalization performance. However, this approach exclusively emphasizes structuralization in the feature space, while overlooking the equally important structural organization of the concept space.

It is essential to integrate semantic concept learning into the decision-making process of deep learning models. Concept Bottleneck Models (CBMs) offer a representative approach by introducing a human-defined semantic concept bottleneck layer. The model first predicts these plain concepts under auxiliary supervision, and then makes the final decision based on them, leading to improvements in both performance and interpretability (Koh et al. 2020; He et al. 2025). Moreover, the Neural Collapse theory indicates that, under the influence of scaling factors, the classification layer weights in deep neural networks converge to the mean vectors of their respective classes, known as class concept prototypes (Papyan, Han, and Donoho 2020; Song et al. 2024). Inspired by these works and cognitive science, this paper argues that deep networks should be structurally organized in both the concept and feature spaces to embed multi-granularity structured knowledge (shown in Fig. 1).

We propose a novel domain generalization framework, Concept-Feature Structuralized Generalization (CFSG), that structurally disentangles the concept and feature spaces into three components: commonality, specificity, and confounding. Final classification is performed based on this structured learning and inference mechanism. Specifically, we first achieve concept space structuralization by disentangling the classification layer weights, and feature space structuralization by disentangling backbone-extracted features. Secondly, feature space structuralization is achieved by introducing multi-granularity structured constraints (Yu et al. 2025a). We further extends the Neural Collapse theory, proposing that structural constraints in the feature space also drive the concept space to exhibit similar structuralization, even in the absence of explicit semantic supervision. To achieve the third objective—classification based on structured features and concepts—this paper proposes adaptively mitigating the impact of varying degrees of distribution shifts on model performance by controlling the proportions of common, specific, and confounding components and explicitly assigning weights to each. Following FSDG, we also disentangle the confounding component in the concept space to account for redundancy and randomness among category concepts. The main contributions of this paper are summarized as follows:

- Drawing inspiration from human cognitive mechanisms, we propose the CFSG framework, which simultaneously and structurally decouples the concept and feature spaces into three components: commonality, specificity, and confounding.
- We argue that the impact of varying degrees of distribution shift on model performance can be adaptively mitigated by controlling the proportions of commonality, specificity, and confounding. To this end, we explicitly assign contribution weights to these three components during classification, thereby enhancing the model's generalization ability.
- Extensive experiments demonstrate that the proposed CFSG model achieves significant performance improvements on three single-source benchmark datasets, with an average gain of 9.87% over baseline models and 3.08% over state-of-the-art methods. Further explainability analysis reveals that CFSG effectively embeds multi-granularity structured knowledge and confirms that feature structuralization facilitates the emergence of concept structuralization.

Related Work

Domain Generalization. DG trains models solely on source domain data, aiming to improve generalization to OOD data. Deep learning has recently advanced under both independent and identically distributed (i.i.d.) and cross-domain settings (Wang, Zhao, and Dong 2025), with DG receiving growing attention. Many methods have been proposed to improve DG, including domain alignment (Rosenfeld, Ravikumar, and Risteski 2021), meta-learning (Finn, Abbeel, and Levine 2017), and data augmentation (Yan et al. 2020).

The core idea of domain alignment is to minimise discrepancies across multiple source domains to learn domain-invariant representations. DgCD (Kim et al. 2024) measured inter-source domain differences and reduced reliance on domain-specific features via stochastic channel dropout. Data augmentation generated multiple virtual domains from a single source (Qiao, Zhao, and Peng 2020), improving cross-domain generalization. Zhou et al. (Zhou et al. 2021) proposed the MixStyle method, which increased source domain diversity by mixing style information from training samples. Li et al. (Li et al. 2019) enhanced generalization by enabling the model to self-learn auxiliary loss functions.

However, these methods overlook the small inter-class and significant intra-class variations in fine-grained images.

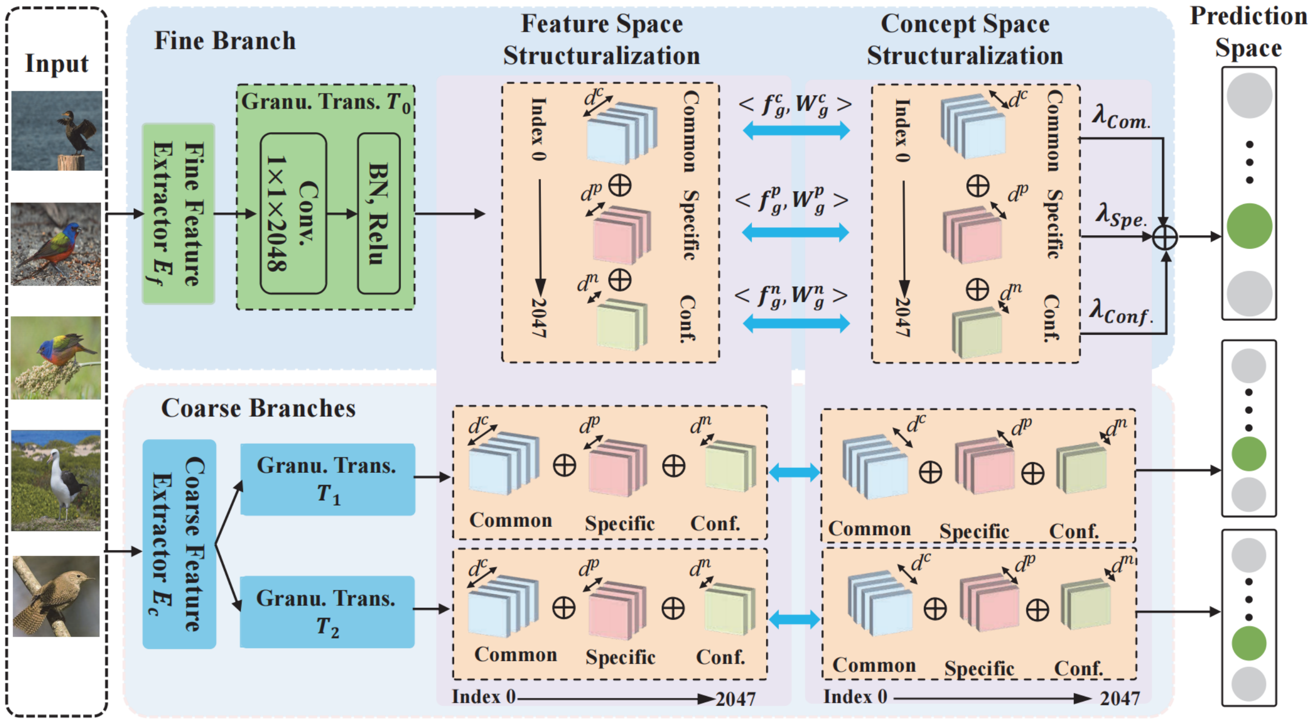

Figure 2: The model illustration presents an example based on a three-level granularity hierarchy. Granu. Trans. denotes the Granularity Transition Layer. Given an input image, CFSG performs structuralization and disentanglement in both the feature and concept spaces, and conducts classification by assigning different weights to the structured representations.

In addition, their reliance on multi-source data limits performance in single-source FGDG tasks.

Fine-Grained Visual Categorization. Fine-grained classification has garnered significant research attention in recent years (Xu et al. 2025). Unlike general image recognition tasks, the goal of Fine-Grained Visual Classification (FGVC) is to distinguish subtle differences among subordinate categories within the same parent class, differences that are often imperceptible to the human eye (Wang et al. 2020a; Wei et al. 2021). To address this issue, early studies primarily relied on additional annotations to assist classification through object localization (Berg and Belhumeur 2013). However, such methods suffer from limited scalability. Over time, researchers shifted their focus to more scalable representation-learning frameworks (Chen et al. 2024). Du et al. (Du et al. 2021) proposed a progressive training strategy to integrate multi-granularity features, thereby improving fine-grained visual classification performance.

These methods mainly focus on fine-grained recognition and do not adequately address domain shift. Achieving strong generalization in single-source FGDG remains a significant challenge, and existing approaches are still limited. Currently, few methods have been proposed for single-source FGDG. Yu et al. (Yu et al. 2025a) were the first to enhance cross-domain fine-grained semantic alignment by structuring the feature space. However, their approach overlooks the structuralization of the concept space, which is a crucial component. The HSSH framework incorporated a state-space model into the backbone network to enrich the style diversity of source domains, thereby improving generalization. However, it lacks a controllable mechanism during inference and applies only to the Vmamba architecture (Bi et al. 2025).

Unlike existing methods for addressing single-source FGDG, we draw inspiration from human cognitive mechanisms to simultaneously structure both the feature and concept spaces into common, specific, and confounding components. Furthermore, we believe that the impact of varying degrees of distribution shift on model performance can be adaptively mitigated by controlling the proportions of commonality, specificity, and confounding. Therefore, during classification, explicit weights are assigned separately to the three components.

Methodology

Preliminaries

Let $\mathcal{X} = \{(x)\}$ be the input space and $\mathcal{Y} = \{(y)\}$ the label space. In the context of FGDG, data is constructed at multiple granularities. We focus exclusively on the single-source setting, where the source domain is defined as $\mathcal{D}_S = \left(x, y_f, y_g|_{g=1}^{G-1}\right)$. Specifically, the source domain consists of input data x and corresponding labels $y = \left(y_f, y_g|_{g=1}^{G-1}\right)$, y_f represents the fine-grained labels, and $y_g|_{g=1}^{G-1}$ represents the coarse-grained labels. In our method, data from a single source domain can generalize to multiple target domains, denoted as $\mathcal{D}_{\mathcal{T}} = \{\mathcal{T}_k\}$. For different target domains, we assign distinct contribution weights to the structured common, specific, and confounding components during classification. Therefore, we aim to learn a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ on the source domain data, to ensure that f generalizes well to the target domain.

Fig. 2 illustrates the overall framework of the proposed CFSG model. We jointly structuralize both the feature and concept spaces into three components: commonality, specificity, and confounding, and assign distinct classification weights to each structured pair of components.

Concept-Feature Space Structuralization

To embed the human decision-making mechanism based on commonality and specificity into deep learning models, we structurally disentangle the concept and feature spaces into three components: common, specific, and confounding. This disentanglement is performed along the channel dimension, with a fixed partition determined by channel indices and the proportion of each component.

Neural collapse theory indicates that, following proper scaling, the last-layer classifier weights converge to the class means, effectively representing the concept prototypes for each category (Papyan, Han, and Donoho 2020). To go one step further, we propose that structurally disentangling features within deep learning models induces a corresponding disentanglement of the classifier weights, i.e., the concept prototypes of each class. In this work, we treat the classification layer weights as concept prototypes for each category and structurally disentangle the concept space into three components: common, specific, and confounding.

Specifically, in CFSG, data is organized in a multi-granularity format, where the number of categories varies across different granularities g. Consequently, the shape of the classification layer weight matrix W_g differs across granularities g. We disentangle the weight matrix W_g along the channel dimension into three components: common, specific, and confounding, with the partitioning fixed by channel indices and ratios. Formally defined as follows:

$$Disentangle(W_g) = \{W_g^c, W_g^p, W_g^n\}. \tag{1}$$

Here, $W_g^c = \{w_g^{c,i} \mid i = 1, 2, \ldots, d^c\}$ represents the d^c common parts, $W_g^p = \{w_g^{p,i} \mid i = 1, 2, \ldots, d^p\}$ reflects the specific parts, and $W_g^n = \{w_g^{n,i} \mid i = 1, 2, \ldots, d^n\}$ denotes the confounding parts. Through the above approach, the concept space is structured into common, specific, and confounding components.

After structurally disentangling the concept space, we further disentangle the feature space. Specifically, given a batch of source images $x \in R^{B \times 3 \times W \times H}$, we adopt both coarse-grained and fine-grained feature extractors to capture fine-grained semantic information better and ensure a fair comparison with FSDG. For comparison, we also conduct experiments using a single shared feature extractor. After passing through two feature extractors, E_c and E_f, the coarse-grained features F_c and fine-grained features F_f are obtained, respectively. Subsequently, these features pass through a Granularity Transition Layer (GTL), which selects the relevant features for each granularity. This yields a feature set F_g, composed of individual features $\{f_g^1, f_g^2, \ldots, f_g^d\}$, where d denotes the number of channels. We then disentangle these features along the channel dimension into three components, following the same approach as in the concept space. The components are formally defined as:

$$Disentangle(F_g) = \{F_g^c, F_g^p, F_g^n\}, \tag{2}$$

where $F_g^c = \{f_g^{c,i} \mid i = 1, 2, \ldots, d^c\}$ represents the d^c common features, $F_g^p = \{f_g^{p,i} \mid i = 1, 2, \ldots, d^p\}$ reflects the specific parts, and $F_g^n = \{f_g^{n,i} \mid i = 1, 2, \ldots, d^n\}$ denotes the confounding parts, satisfying $d = d^c + d^p + d^n$.

With the above design, we achieve structural disentanglement of both the concept space and the feature space simultaneously. The key to achieving structuralization lies in disentanglement followed by alignment. Inspired by neural collapse theory (Papyan, Han, and Donoho 2020), we propose that concept structuralization can be realized through constraints imposed by feature structuralization, such that the alignment constraints for concepts follow the same principles as those for features. Specifically, we compute the similarity between the common, specific, and confounding features and their corresponding concepts, respectively.

Concept Bottleneck-Based Classification

After achieving structuralization in both the concept and feature spaces, classification is performed based on the structured feature and concept representations. We propose that adaptively controlling the proportions of common, specific, and confounding components can mitigate the impact of varying degrees of distribution shifts on model performance by assigning explicit weights to each component.

Specifically, our classifier differs from conventional ones. Traditional deep learning models typically employ a parameterized softmax classifier, which is generally formulated as: $H = (W^T \times f_\theta(x)) + b$. Here, f_θ denotes the feature extracted from the input image x, W is the weight matrix of the fully connected layer, and b is the bias term. Both W and b are learnable parameters optimized during training.

In CFSG, our classifier is defined as follows:

$$\begin{aligned} H = {} & \lambda_c \langle f_g^c, W_{k,g}^c \rangle + \lambda_p \langle f_g^p, W_{k,g}^p \rangle \\ & + \lambda_n \langle f_g^n, W_{k,g}^n \rangle + b, \end{aligned} \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. f_g^c, f_g^p, and f_g^n represent the features of the sample after decoupling along the channel dimension through the GTL. $W_{k,g}^c$, $W_{k,g}^p$, and ${}_{k,g}^{n}$ denote the classifier weights corresponding to the commonality, specificity, and confounding components, respectively, for category k at granularity g. λ_c, λ_p, and λ_n represent the explicit weighting coefficients for the three parts, indicating their relative contributions in the weighted aggregation of distances between the sample features and the respective sub-weight components. During inference, the weights of the three components are normalized to sum to one and are manually adjusted to achieve optimal generalization.

Feature Alignment Constraint

During training, we adopt the standard $\mathcal{L}_{FSDG}$ loss from the FSDG (Yu et al. 2025a) method as our objective function, which is defined as follows:

$$\mathcal{L}_{FSDG} = \mathcal{L}_c + \mathcal{L}_{lf} + \mathcal{L}_{FS}. \quad (4)$$

The detailed explanation is as follows:

- $\mathcal{L}_c$: The coarse-grained cross-entropy loss, defined as $\mathcal{L}_c = \sum_{g=1}^{G-1} L_{CE}(\widehat{y}_g, y_g)$.
- $\mathcal{L}_{lf}$: The prediction alignment loss, where fine-grained ground-truth labels are progressively fused with the coarse-grained branch's prediction distribution. This fusion leverages coarse-grained information to enhance the performance of the fine-grained branch, thereby achieving effective alignment in the prediction space.
- $\mathcal{L}_{FS}$: The constraint loss from the feature structuralization module, encompassing feature disentanglement $\mathcal{L}_{dec}$ and the feature alignment loss $\mathcal{L}_{fa}$. Formally, $\mathcal{L}_{FS} = \mathcal{L}_{dec} + \mathcal{L}_{fa}$. The feature alignment loss $\mathcal{L}_{fa}$ enforces constraints on both common and specific features, expressed as $\lambda_p \mathcal{S}_p - \lambda_{cs}\mathcal{S}_{cs} - \lambda_{cd}\mathcal{S}_{cd}$.

We further optimize the loss function and training procedure of FSDG and adopt a more advanced inference strategy during model prediction.

For the feature extractor, we instantiate the backbone using networks from the ResNet (RN) series (He et al. 2016), Vision Transformer (ViT) series (Dosovitskiy et al. 2021), or ASMLP series (Lian et al. 2022). Additionally, to enable fair comparison with the HSSH method, we adopt the Vmamba (Liu et al. 2024) series as the backbone on the CUB-Paintings and Birds-31 datasets. Since the official HSSH code has not been released, we are unable to conduct comparisons on the CompCars dataset. We use fine-grained classification accuracy as the evaluation metric. All experiments are performed on an A30 GPU. At deployment, only the fine-grained branch is used, while the coarse-grained branch serves solely as an auxiliary module during training.

Experiments

We evaluate CFSG on three fine-grained datasets and compare it with various state-of-the-art DG and FGDG methods, demonstrating its robustness across multiple mainstream backbone networks.

Datasets

CUB-Paintings consists of two distinct domains: CUB-200-2011 (C) (Wah et al. 2011) and CUB-200-Paintings (P) (Wang et al. 2020b). Both datasets follow a four-level fine-grained classification hierarchy, encompassing 14 orders, 38 families, 122 genera, and 200 species. CUB-200-2011 contains 11,788 real bird images, whereas CUB-200-Paintings includes 3,047 cross-media artistic renditions.

CompCars (Yang et al. 2015) includes car images from two distinct sources: the web (**W**) and surveillance (**S**). It is organized into a hierarchical classification system with 68 coarse categories and 281 fine-grained categories.

Method	C → P	P → C	Avg	Params
ResNet-Based:				
PAN(DA)	67.40	50.92	59.16	103M
ERM	54.94	35.67	45.31	24M
ARM	47.98	31.53	39.76	24M
DANN	54.06	37.09	45.57	24M
MLDG	55.40	34.15	44.78	23M
GroupDRO	54.94	35.67	45.31	23M
CORAL	54.70	35.29	45.00	23M
SagNet	56.33	36.71	46.52	24M
MixStyle	52.97	28.44	40.71	23M
Mixup	54.58	34.66	44.62	23M
RIDG	36.41	24.11	30.26	24M
SAGM	57.83	37.16	47.50	23M
MIRO	56.29	41.28	48.79	47M
S-FSDG	63.42	44.87	54.14	26M
S-CFSG (Ours)	<u>64.06</u>	<u>54.06</u>	<u>59.06</u>	28M
FSDG	61.84	49.46	55.65	26M
CFSG (Ours)	**67.50**	**65.62**	**66.56**	28M
VMamba-Based:				
HSSH	<u>67.48</u>	<u>64.57</u>	<u>66.03</u>	35M
CFSG (Ours)	**70.94**	**73.75**	**72.35**	35M

Table 1: Classification accuracy (%) of the proposed CFSG method and existing methods on the Cub-Paintings dataset. The best results are highlighted in bold, and the second-best is underlined.

Method	W → S	S → W	Avg	Params
ResNet-Based:				
PAN (DA)	47.05	15.57	31.31	103M
ERM	44.15	7.54	25.85	24M
ARM	20.25	4.74	12.50	24M
DANN	35.10	6.80	20.95	24M
MLDG	44.94	7.56	26.25	23M
GroupDRO	43.60	7.75	25.68	23M
CORAL	43.05	7.97	25.51	23M
SagNet	45.33	8.89	27.11	24M
MixStyle	38.37	6.28	22.33	23M
Mixup	43.07	7.56	25.32	23M
RIDG	36.57	8.11	22.34	24M
SAGM	49.55	8.58	29.07	23M
MIRO	46.01	7.88	26.95	47M
S-FSDG	53.44	10.83	32.14	26M
S-CFSG (Ours)	<u>73.12</u>	<u>17.19</u>	<u>45.16</u>	28M
FSDG	51.78	11.30	31.54	26M
CFSG (Ours)	**74.69**	**20.62**	**47.66**	28M

Table 2: Classification accuracy (%) of the proposed CFSG method and existing methods on the ComCars dataset. The best results are highlighted in bold, and the second-best is underlined.

Birds-31 consists of three domains: CUB-200-2011 (**C**), NABirds (**N**) (Van Horn et al. 2015), and iNaturalist2017 (**I**) (Van Horn et al. 2018). In (Wang et al. 2020b), the categories from these three datasets were united, and 31 fine-grained categories were selected, with corresponding image

Method	C → I	C → N	I → C	I → N	N → C	N → I	Avg	Params
ResNet-Based:								
PAN (DA)	69.79	84.19	90.46	88.10	92.51	75.03	83.34	103M
ERM	54.64	72.93	85.01	74.97	86.10	62.51	72.69	24M
ARM	50.51	71.25	77.38	74.20	84.74	59.82	69.65	24M
DANN	52.75	71.82	80.79	73.59	85.55	61.53	71.01	24M
MLDG	53.55	72.19	80.74	74.83	85.61	61.95	71.48	23M
GroupDRO	52.61	70.78	81.87	74.40	86.26	61.32	71.21	23M
CORAL	54.64	72.93	81.01	74.97	86.10	62.51	72.03	23M
SagNet	53.66	71.75	81.39	74.13	85.66	62.06	71.44	24M
MixStyle	49.95	69.04	74.46	6834	83.60	57.12	67.09	23M
Mixup	52.36	71.65	82.36	75.17	85.61	62.34	71.58	23M
RIDG	47.15	66.71	82.47	73.63	85.77	60.98	69.45	24M
SAGM	54.04	73.63	82.96	77.01	87.88	63.49	73.17	23M
MIRO	54.39	74.87	82.36	75.34	86.42	62.48	72.64	47M
S-FSDG	63.66	82.43	89.56	85.80	92.03	72.91	81.06	26M
S-CFSG (Ours)	70.94	76.88	90.63	84.69	94.06	75.63	82.14	28M
FSDG	66.32	**83.71**	90.96	87.36	91.95	74.20	82.37	26M
CFSG (Ours)	**72.81**	80.94	**94.06**	**88.13**	**95.31**	**78.44**	**84.95**	28M
VMamba-Based:								
HSSH	80.43	**90.39**	95.25	**94.33**	96.17	**87.57**	**90.69**	35M
CFSG (Ours)	**81.25**	85.31	**98.12**	93.44	**99.06**	85.94	90.52	35M

Table 3: Classification accuracy (%) of the proposed CFSG method and existing methods on the Birds-31 dataset. The best results are highlighted in bold, and the second-best is underlined.

counts of 1,848, 2,988, and 2,857. A four-level granularity structure was subsequently constructed, including 4 orders, 16 families, 25 genera, and 31 species.

Experimental Results

We evaluate the effectiveness of the proposed method by comparing it with a variety of DG and FGDG approaches, including ERM (Vapnik 1999), ARM (Zhang et al. 2021), DANN (Ganin et al. 2016), MLDG (Li et al. 2018), GroupDRO (Sagawa et al. 2019), CORAL (Sun and Saenko 2016), SagNet (Nam et al. 2021), MixStyle (Zhou et al. 2021), Mixup (Yan et al. 2020), RIDG (Chen et al. 2023), SAGM (Wang et al. 2023), MIRO (Cha et al. 2022), FSDG (Yu et al. 2025a), and HSSH (Bi et al. 2025). All methods are implemented based on their official code, except for HSSH. Since the official HSSH code is not publicly available, we compare our method against the experimental results reported in its original paper. We evaluate the proposed CFSG method under both dual-backbone and single-backbone (prefixed with "S-") training settings.

On the CUB-Paintings dataset, as shown in Table 1, our method achieves the best performance among all DG competitors. The dual-backbone and single-backbone models outperform the baseline FSDG model by 10.91% and 4.92%, respectively. When replacing the backbone with the Vmamba series, CFSG outperforms HSSH by 6.32%. These results significantly confirm the effectiveness of our proposed CFSG model, which performs classification based on structured feature and concept representations while adaptively weighting the contributions of the three components.

On the ComCars dataset, as shown in Table 2, our method demonstrates significant improvement. The dual-backbone and single-backbone models achieve improvements of 16.12% and 13.02% over the baseline FSDG model. These results demonstrate that our proposed method effectively addresses distribution shifts between source and target domains and highlights the critical role of concept structuralization in enhancing generalization.

On the Birds-31 dataset, as shown in Table 3, our method achieves the highest average accuracy, exhibiting a clear performance improvement. Under the dual-backbone configuration, it surpasses FSDG by 2.58%, while in the single-backbone setting, it outperforms the baseline by 1.08%. When replacing the backbone network with the Vmamba series, the CFSG model achieves performance comparable to the HSSH method, demonstrating the robustness of our approach across different backbone architectures. In contrast, HSSH shows superior performance only on the Vmamba architecture.

Experiments on various backbone architectures with different depths. To verify the robustness of CFSG across different architectures, we conduct experiments on the feature extractors within the model pipeline using various backbone networks and deep architectures, including RN-50, RN-101, ViT-Tiny, ViT-Small, ASMLP-Tiny, and ASMLP-Small. The detailed experimental results are presented in Table 4. Experimental results show that our method consistently improves performance across various architectures, with average performance gains ranging from 2.05% to 11.19%. These results demonstrate the robustness of our method across different backbone networks.

Backbone	Method	C → P	P → C	Avg	Params
RN-101	FSDG	64.64	49.86	57.25	45M
	CFSG	65.62	71.25	68.44	47M
ViT-T	FSDG	60.71	48.82	54.76	11M
	CFSG	62.50	52.50	57.50	13M
ViT-S	FSDG	70.26	66.25	68.26	31M
	CFSG	70.31	70.31	70.31	33M
MLP-T	FSDG	60.90	50.51	55.70	31M
	CFSG	63.13	59.06	61.10	33M
MLP-S	FSDG	63.67	53.84	58.75	53M
	CFSG	67.50	66.87	67.19	55M

Table 4: Classification accuracy (%) of the proposed CFSG method on the Cub-Paintings dataset using various backbones with different depths. RN and MLP denote ResNet and ASMLP backbones, respectively. T and S stand for Tiny and Small, respectively.

Backbone	CS	FS	C → P	P → C	Avg
✓			56.76	46.53	51.65
✓		✓	61.84	49.46	55.65
✓	✓		57.50	60.00	58.75
✓	✓	✓	**67.50**	**65.62**	**66.56**

Table 5: Classification accuracy (%) of different modules in the CFSG model, where CS denotes Concept Structuralization and FS denotes Feature Structuralization.

Analysis

Parameter Impact Analysis. CFSG adaptively mitigates the impact of distribution shifts of varying degrees on model performance by controlling the proportions of common, specific, and confounding components. To validate the effectiveness of our method, we conduct extensive experiments to explore the existence of an optimal balance strategy. The results indicate the existence of such an optimum from domain P to domain C, with the best weights being 0.75, 0.2, and 0.05, respectively. We visualize different weight configurations and perform experiments treating the weights as learnable parameters. More implementation details are provided in the Appendix.

Concept Structuralization and Feature Structuralization Modules. We evaluate the effects of concept structuralization only, feature structuralization only, and their combination on top of the backbone network. When the feature structuralization module is not used, the model employs the standard cross-entropy loss. As shown in Table 5 shows that feature structuralization alone improves performance by 4.0%, concept structuralization by 7.1%, and their combination achieves the highest improvement of 14.91%. These results further validate the effectiveness of jointly structuralizing both the concept and feature spaces.

Compared to methods that perform classification by constructing sub-centroids. According to the Neural Collapse theory, classifier weights converge to class feature means up to a scaling factor (Papyan, Han, and Donoho 2020). Inspired by this, we assume that the structuralized classifier weights also converge to sub-centroids. Based on this assumption, we design a sub-centroid-based classification method and compare it with CFSG. During training, sub-centroids are constructed, and during inference, classification is performed by computing the weighted distances between samples and the sub-centroids. Detailed construction procedures are provided in the Appendix. Table 6 reports the experimental results. The sub-centroids method based on sample means achieves a 3.88% improvement over the FSDG baseline, demonstrating its effectiveness. However, its performance remains inferior to that of the CFSG method. Classification based on concept structuralization achieves a substantial performance boost.

Method	C → P	P → C	Avg
FSDG	61.84	49.46	55.65
CFSG(sub-centroids)	62.81	56.25	59.53
CFSG	**67.50**	**65.62**	**66.56**

Table 6: Classification accuracy (%) of the sub-centroid-based method on the CUB-Paintings dataset.

FGDG	All	Com.	Spe.	Conf.
Ground Truth	0.69	0.70	0.67	0.67
CFSG	All	Com.	Spe.	Conf.
Ground Truth	0.97	0.97	0.97	0.91

Table 7: Spearman's rank correlation between region-level concepts and ground-truth labels for each model.

Explainability analysis. We posit that, under multi-granularity structured knowledge, categories that are closer in distance exhibit higher similarity in the concept space. We measured both the ground-truth category similarity and the concept space similarity, then computed the Spearman rank correlation between them. As shown in Table 7, the CFSG model achieves a high correlation of 0.97, significantly outperforming the unstructured FGDG baseline with 0.69. This result confirms that CFSG effectively embeds multi-granularity knowledge into both concept and feature spaces, and that feature structuralization facilitates the emergence of concept structuralization. More implementation details are provided in the Appendix.

Conclusion

The core contribution of this work lies in interpreting classifier weights as an unsupervised concept space and proposing that structured patterns in the feature space can induce structural organization in the concept space. We jointly structuralize both the feature and concept spaces into three components: common, specific, and confounding. By controlling the proportions of common, specific, and confounding parts, the model adaptively mitigates the impact of varying degrees of distribution shifts on performance. During classification, the model computes the similarity between structured features and structured concept representations, assigning explicit weights to each component. Extensive experiments demonstrate that the proposed CFSG method significantly improves performance in FGDG scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62306103, 62276186, and 62376194, and in part by the Higher Education Science and Technology Research Project of Hebei Province under Grant QN2023262.

References

Berg, T.; and Belhumeur, P. N. 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 955–962.

Bi, Q.; Yi, J.; Zhan, H.; Ji, W.; and Xia, G.-S. 2025. Learning fine-grained domain generalization via hyperbolic state space hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1853–1861.

Cha, J.; Lee, K.; Park, S.; and Chun, S. 2022. Domain generalization by mutual-information regularization with pre-trained models. In *European conference on computer vision*, 440–457. Springer.

Chen, H.; Zhang, H.; Liu, C.; An, J.; Gao, Z.; and Qiu, J. 2024. FET-FGVC: Feature-enhanced transformer for fine-grained visual classification. *Pattern Recognition*, 149: 110265.

Chen, L.; Zhang, Y.; Song, Y.; Van Den Hengel, A.; and Liu, L. 2023. Domain generalization via rationale invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1751–1760.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dong, Y.; Wu, G.; Wang, J.; and Wang, Z. 2025. Sufficient Learning for Label Noise with Dual-Regularization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, C.; Fu, K.; Wen, B.; Sun, Y.; Peng, J.; Wei, W.; Gao, Y.; Wang, S.; Zhang, C.; Li, J.; et al. 2025. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 1–16.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European conference on computer vision*, 153–168. Springer.

Du, R.; Xie, J.; Ma, Z.; Chang, D.; Song, Y.-Z.; and Guo, J. 2021. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9521–9535.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

He, H.; Zhu, L.; Zhang, X.; Zeng, S.; Chen, Q.; and Lu, Y. 2025. V2C-CBM: Building Concept Bottlenecks with Vision-to-Concept Tokenizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3401–3409.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kim, S.; Park, B.; Lee, H.; and Chae, D.-K. 2024. Discrepancy-guided Channel Dropout for Domain Generalization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1110–1120.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li, Y.; Yang, Y.; Zhou, W.; and Hospedales, T. 2019. Feature-critic networks for heterogeneous domain generalization. In *International conference on machine learning*, 3915–3924. PMLR.

Lian, D.; Yu, Z.; Sun, X.; and Gao, S. 2022. AS-MLP: An Axial Shifted MLP Architecture for Vision. In *International Conference on Learning Representations*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Liu, Y.; Xiong, Z.; Li, Y.; Tian, X.; and Zha, Z.-J. 2021. Domain generalization via encoding and resampling in a unified latent space. *IEEE Transactions on Multimedia*, 25: 126–139.

Luppi, A. I.; Rosas, F. E.; Mediano, P. A.; Menon, D. K.; and Stamatakis, E. A. 2024. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences*, 28(4): 352–368.

Mahner, F. P.; Muttenthaler, L.; Güçlü, U.; and Hebart, M. N. 2025. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6): 848–859.

Mao, S.; and Zhang, S. 2024. Robust Fine-Grained Visual Recognition With Neighbor-Attention Label Correction. *IEEE Transactions on Image Processing*, 33: 2614–2626.

Mervis, C. B.; and Rosch, E. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7: 573–605.

Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8690–8699.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12556–12565.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Song, K.; Tan, Z.; Zou, B.; Chen, J.; Ma, H.; and Huang, W. 2024. Exploring Information-Theoretic Metrics Associated with Neural Collapse in Supervised Training. *arXiv preprint arXiv:2409.16767*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, 443–450. Springer.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 595–604.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Vapnik, V. 1999. *The nature of statistical learning theory*. Springer science & business media.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. S. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8): 8052–8072.

Wang, L.; Li, D.; Zhu, Y.; Tian, L.; and Shan, Y. 2020a. Dual super-resolution learning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3774–3783.

Wang, P.; Zhang, Z.; Lei, Z.; and Zhang, L. 2023. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3769–3778.

Wang, S.; Chen, X.; Wang, Y.; Long, M.; and Wang, J. 2020b. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9213–9222.

Wang, Z.; Zhao, W.; and Dong, Y. 2025. Learning with Coupled Noisy Labels for Visible-Infrared Person Re-identification via Graph Consistency. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8927–8948.

Xu, Q.; Li, S.; Wang, J.; Jiang, B.; Luo, B.; and Tang, J. 2025. Context-semantic quality awareness network for fine-grained visual categorization. *Pattern Recognition*, 112033.

Yan, S.; Song, H.; Li, N.; Zou, L.; and Ren, L. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*.

Yang, L.; Luo, P.; Change Loy, C.; and Tang, X. 2015. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3973–3981.

Yu, W.; Chen, D.; Wang, Q.; and Hu, Q. 2025a. Fine-Grained Domain Generalization With Feature Structuralization. *IEEE Transactions on Multimedia*, 27: 8510–8524.

Yu, W.; Wang, Q.; Liu, C.; Li, D.; and Hu, Q. 2025b. CoE: Chain-of-Explanation via Automatic Visual Concept Circuit Description and Polysemanticity Quantification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4364–4374.

Yu, Y.; Yan, Y.; and Jin, Y. 2025. Structural knowledge: from brain to artificial intelligence. *Artificial Intelligence Review*, 58(9): 264.

Zhang, M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; and Finn, C. 2021. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34: 23664–23678.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.