This appendix is organized as follows:

- Section A provides the detailed dataset information.
- Section B provides the algorithm details.
- Section C provides the additional training implementation details.
- Section D gives additional experiment results, including additional comparisons of domain-wise results of multi-source domain generalization and single-source domain generalization.

A Dataset Details

PACS [21] is a commonly used small-scaled dataset in the field of domain adaptation and domain generalization. It consists of 4 domains, a total of 9991 images, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images), and Sketch (3,929 images). Each domain contains 7 categories.

VLCS [11] is also a small-scaled benchmark dataset, a total of 7,510 images, including 4 domains: Caltech (991 images), LabelMe (1,859 images), Pascal (2,363 images) and Sun (2,297 images). Each domain contains 5 categories.

Office-Home [33] is a medium-scaled benchmark for domain adaptation and domain generalization. It contains a total of around 15,500 images from 4 distinct domains: Art (2,427 images), Clip Art (4,365 images), Product (4,439 images), and Real World (4,357 images). Each domain contains objects from 65 categories commonly found in office and home environments.

TerraIncognita [3] is a large-scaled benchmark for visual recognition. It contains 243,187 images from 140 camera locations. For DG, a subset is selected that includes 4 domains: Location38 (9,736 images), Location43 (3,970 images), Location46 (5,883 images) and Location100 (4,741 images). Each domain contains animals from 10 categories found in the wild.

DomainNet [30] is a large-scaled benchmark for domain adaptation and domain generalization. It contains a total of around 586,575 images from 6 distinct domains: Clipart (48,129 images), Infograph (51,605 images), Painting (72,266 images), Quickdraw (172,500 images), Real (172,947 images), Sketch (69,128 images). Each domain includes 345 categories of objects. We sample around 20 thousand data from all domains as the subset.

B Algorithm Details

The overall framework of the pseudo-code of SPG is described in Algorithm 1 and Algorithm 2. Algorithm 1 demonstrates the process of stage I: Domain Prompt Labels Learning, and Algorithm 2 shows the process of stage II: Generative Model Pre-training.

Algorithm 1 Soft Prompt Generation - Domain Prompt Labels Learning

Requirement: pre-defined N_c class names in the target task
Input: images and labels of training samples $(\mathbf{x}_j^{d_i}, y_j^{d_i})$, number of training iterations L
Output: N_d domain prompt labels

# learn prompt label on each domain separately
for $i = 1, 2, \ldots, N_d$ **do**
  # initialize i-th domain prompt label with prompt prefix $\mathbf{v}^p$ and learnable vector $\mathbf{v}^i$.
  $\mathbf{v}^{d_i} \leftarrow \texttt{initialize}(\mathbf{v}^p, \mathbf{v}^i)$
  # L training iterations for learning each domain prompt label
  for iteration $= 1, 2, \ldots, L$ **do**
    # update learnable vector v^{d_i}
    $\mathbf{v}^{d_i}{}^* = \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{x}_j^{d_i}, y_j^{d_i}}[-\log p(y_j^{d_i} \mid \mathbf{x}_j^{d_i}, \mathbf{v}^{d_i})]$
    Update $\mathbf{v}^{d_i}$ by gradient descent
  end for
end for
Store optimal domain prompt label $\mathbf{v}^{d_i}$ for each domain

Algorithm 2 Soft Prompt Generation - Generative Model Pre-training

Requirement: A CGAN with a generator G and a discriminator D, real vector $\mathbf{v}_{\text{real}}$ and fake vector $\mathbf{v}_{\text{fake}}$, and domain prompt labels $\mathbf{v}^{d_i}$
Input: image embeddings $f(\mathbf{x})$, number of training iterations L
Output: optimal prompt for each image

# L training iterations
for iteration $= 1, 2, \ldots, L$ **do**
  # define input noise variable $\mathbf{z}$ and combines these noise variables with image embeddings $f(\mathbf{x})$ as input of generator G and output $\mathbf{v}_g$
  $\mathbf{v}_g = G(input) \leftarrow input = \texttt{concat}([\mathbf{z}, f(\mathbf{x})])$
  # discriminator determines the authenticity of domain prompt labels $\mathbf{v}^{d_i}$ and generated prompt $\mathbf{v}_g$
  $\mathbf{v}_{d,\text{real}} = D(\mathbf{v}^{d_i})$, $\mathbf{v}_{d,\text{fake}} = D(\mathbf{v}_g)$
  # compute $\mathcal{L}_{\text{real}}$ with discriminator_real output $\mathbf{v}_{d,\text{real}}$ and pre-defined real vector $\mathbf{v}_{\text{real}}$
  $\mathcal{L}_{\text{real}} \leftarrow \texttt{mse\_loss}(\mathbf{v}_{d,\text{real}}, \mathbf{v}_{\text{real}})$
  # compute $\mathcal{L}_{\text{fake}}$ with discriminator_fake output $\mathbf{v}_{d,\text{fake}}$ and pre-defined fake vector $\mathbf{v}_{\text{fake}}$
  $\mathcal{L}_{\text{fake}} \leftarrow \texttt{mse\_loss}(\mathbf{v}_{d,\text{fake}}, \mathbf{v}_{\text{fake}})$
  $\mathcal{L}_{\text{discriminator}} \leftarrow \mathcal{L}_{\text{real}} + \mathcal{L}_{\text{fake}}$
  Update parameters of D using $\mathcal{L}_{\text{discriminator}}$ by gradient descent
  # compute $\mathcal{L}_{\text{generator}}$ with discriminator_fake output $\mathbf{v}_{d,\text{fake}}$ and pre-defined real vector $\mathbf{v}_{\text{real}}$
  $\mathcal{L}_{\text{generator}} \leftarrow \texttt{mse\_loss}(\mathbf{v}_{d,\text{fake}}, \mathbf{v}_{\text{real}})$
  Update parameters of G using $\mathcal{L}_{\text{generator}}$ by gradient descent
end for
Generate the optimal prompt for each input image

C Implementation Details

For our proposed SPG, in the first phase of the training stage, we employ the text prompt [40] as our prompt design, which is also the prototype for the domain prompt labels and generated prompts. We initialize the context with the phrase "a photo of a" and set the prompt's context length to 4. In the second phase of the training stage, we train the CGAN model on different domains of various datasets, employing a tailored set of training parameters. Specifically, we set the batch size to around 32 and adjust the initial learning rate between 1e-4 and 2e-3 on different datasets, We employ a cosine learning rate scheduler and conduct training for 70 to 100 epochs, incorporating a linear learning rate warm-up phase at 1e-5 over the first 4 epochs. For optimization, we use the AdamW optimizer, configuring it with a weight decay of 1e-4 and beta values set to (0.9, 0.999).

Meanwhile, given the observed instability in CGAN training, we implement the gradient clipping strategy to control the magnitude of gradients within the generator and discriminator networks. Specifically, for the discriminator, we establish norm upper limits for general weights and biases in the range of 5e-2 to 5e-1, while setting those for particular weights and biases at a ceiling of 5. For the generator, we set the norm upper limit for universal weights between 5e-3 and 5e-2, the norm upper limit for universal biases between 5e-8 and 5e-7, and the norm upper limit for special biases between 0.5 and 5. This strategy aims to enhance training stability by preventing excessive gradient values.

D Supplement Experiments

The supplement experiments mainly demonstrate the domain-wise results of multi-source DG and single-source DG.

D.1 Multi-source DG Comparisons

Table 7∼11 demonstrate the per-domain multi-source DG top-1 classification accuracy on PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet, respectively. We observe that although SPG may not achieve the best performance in every individual domain, it consistently reaches state-of-the-art levels across five datasets under two different backbones on average. For some tasks, such as the target domain is Sketch in PACS and Sun in VLCS, SPG outperforms the previous SOTA method with a large margin of 4.3% and 3.6%, respectively.

D.2 Single-source DG Comparisons

Table 12∼16 demonstrate the per-domain single-source DG top-1 classification accuracy on PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet, respectively. Relying on a single domain for generalization can significantly degrade performance. While our approach may not be optimal in certain cases, it still achieves state-of-the-art performance on average.

Table 7: Comparisons with SOTA methods on PACS for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). Average accuracies are reported from three trials. Bold denotes the best scores.

Method	B.	Art	Cartoon	Photo	Sketch	Avg
ZS-CLIP [32]	RN50	90.9	93.3	99.2	79.5	90.7
Lin. Prob. [32]		90.8	92.7	99.1	79.8	90.6
CoOp [40]		92.0	93.8	98.6	80.7	91.3
CoCoOp [39]		93.1	**94.3**	99.3	80.8	91.9
DPL [36]		**93.6**	93.8	99.0	80.7	91.8
VP [1]		90.6	92.7	99.3	78.0	90.2
SPG (ours)		92.8	93.8	**99.5**	**85.1**	**92.8**
ZS-CLIP [32]	ViT-B/16	97.2	**99.1**	**99.9**	88.2	96.1
Lin. Prob. [32]		96.2	94.7	98.7	90.1	94.9
CoOp [40]		97.7	98.4	99.6	90.0	96.4
CoCoOp [39]		97.7	99.0	99.8	90.4	96.7
DPL [36]		97.8	98.5	**99.9**	89.5	96.4
VP [1]		96.9	98.9	**99.9**	87.3	95.8
VPT [17]		**97.9**	98.9	**99.9**	91.0	96.9
MaPLe [18]		**97.9**	98.7	99.7	89.8	96.5
SPG (ours)		97.7	99.0	**99.9**	**91.3**	**97.0**

Table 8: Comparisons with SOTA methods on VLCS for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). Average accuracies are reported from three trials. Bold denotes the best scores.

Method	B.	Caltech	LableMe	Pascal	Sun	Avg
ZS-CLIP [32]	RN50	99.4	64.9	84.1	71.6	80.0
Lin. Prob. [32]		99.3	61.1	81.8	76.9	79.8
CoOp [40]		99.7	64.0	84.7	77.3	81.4
CoCoOp [39]		99.7	63.7	84.8	78.8	81.8
DPL [36]		**99.8**	62.5	84.5	76.3	80.8
VP [1]		99.6	66.3	84.6	71.5	80.5
SPG (ours)		99.5	**68.7**	**85.4**	**82.4**	**84.0**
zero-shot CLIP	ViT-B/16	99.9	68.6	85.9	74.8	82.3
Lin. Prob. [32]		95.9	63.7	76.3	74.2	77.5
CoOp [40]		99.6	61.4	84.6	77.5	80.8
CoCoOp [39]		99.9	59.7	85.9	75.5	80.3
DPL [36]		99.8	61.5	84.6	77.8	80.9
VP [1]		**100.0**	**68.5**	**86.2**	73.9	82.2
VPT [17]		99.9	64.8	85.2	78.2	82.0
MaPLe [18]		98.3	64.8	85.1	80.6	82.2
SPG (ours)		99.7	64.7	84.4	**80.7**	**82.4**

Table 9: Comparisons with SOTA methods on OfficeHome for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). Average accuracies are reported from three trials. Bold denotes the best scores.

Method	B.	Art	Clipart	Product	Real	Avg
ZS-CLIP [32]	RN50	69.0	53.5	80.1	80.5	70.8
Lin. Prob. [32]		62.0	49.0	73.6	77.4	65.5
CoOp [40]		71.3	56.1	83.2	83.2	73.5
CoCoOp [39]		70.3	**56.7**	83.4	83.3	73.4
DPL [36]		**71.5**	56.2	83.5	83.1	73.6
VP [1]		67.7	52.5	80.0	80.4	70.2
SPG (ours)		71.3	55.6	**84.8**	**83.4**	**73.8**
ZS-CLIP [32]	ViT-B/16	80.1	70.0	88.2	89.0	81.8
Lin. Prob. [32]		73.5	69.9	87.4	86.4	79.3
CoOp [40]		81.2	72.0	89.7	89.2	83.0
CoCoOp [39]		81.8	71.7	**90.3**	89.7	83.4
DPL [36]		81.0	71.2	90.0	89.6	83.0
VP [1]		79.8	69.1	87.4	88.6	81.2
VPT [17]		80.9	72.5	89.0	90.4	83.2
MaPLe [18]		**81.6**	72.6	90.2	89.0	83.4
SPG (ours)		**81.6**	**72.7**	90.2	**89.9**	**83.6**

Table 10: Comparisons with SOTA methods on TerraIncognita for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). Average accuracies are reported from three trials. Bold denotes the best scores.

Method	B.	Location38	Location43	Location46	Location100	Avg
ZS-CLIP [32]	RN50	28.4	32.8	24.0	10.1	23.8
Lin. Prob. [32]		33.0	42.7	31.9	24.4	33.0
CoOp [40]		25.6	**43.5**	**34.5**	29.2	33.2
CoCoOp [39]		35.9	42.1	32.5	25.8	34.1
DPL [36]		36.0	41.1	32.9	27.6	34.4
VP [1]		28.8	34.0	26.8	12.6	25.6
SPG (ours)		**42.1**	38.9	32.1	**36.8**	**37.5**
ZS-CLIP [32]	ViT-B/16	20.5	32.8	29.6	**52.4**	33.8
Lin. Prob. [32]		48.0	50.5	43.8	44.0	46.6
CoOp [40]		**53.3**	47.4	41.1	45.5	46.8
CoCoOp [39]		51.6	46.9	39.3	43.2	45.3
DPL [36]		54.3	49.0	41.6	41.6	46.6
VP [1]		20.2	34.3	32.8	52.3	34.9
VPT [17]		46.8	52.8	41.8	45.5	46.7
MaPLe [18]		52.4	**53.0**	43.1	**52.4**	**50.2**
SPG (ours)		51.0	49.2	**50.7**	49.8	**50.2**

Table 11: Comparisons with SOTA methods on DomainNet for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 and ViT-B/16 as the backbone (B.). Average accuracies are reported from three trials. Bold denotes the best scores.

Method	B.	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
ZS-CLIP [32]	RN50	52.7	40.5	53.2	5.7	77.1	49.3	46.4
Lin. Prob. [32]		34.6	24.7	35.3	4.1	28.2	35.9	27.1
CoOp [40]		57.0	43.9	58.1	7.8	78.8	52.6	49.7
CoCoOp [39]		57.0	**44.0**	58.3	7.8	78.9	52.0	49.7
DPL [36]		56.7	43.9	57.9	**7.9**	78.2	53.0	49.6
VP [1]		52.4	40.3	52.7	5.3	76.8	47.1	45.8
SPG (ours)		**57.3**	41.7	**58.3**	**7.9**	**80.0**	**55.5**	**50.1**
ZS-CLIP [32]	ViT-B/16	70.2	46.3	65.0	13.0	83.0	62.0	56.6
Lin. Prob. [32]		62.9	35.4	56.8	11.3	65.8	56.7	48.2
CoOp [40]		72.7	50.2	68.5	15.6	84.2	65.9	59.5
CoCoOp [39]		72.1	50.4	67.9	15.8	**84.4**	65.5	59.4
DPL [36]		72.5	50.4	68.3	15.8	83.9	66.0	59.5
VP [1]		70.1	45.5	64.6	14.1	82.7	62.0	56.5
VPT [17]		71.0	48.5	66.2	16.3	83.6	65.2	58.5
MaPLe [18]		**73.1**	49.9	67.8	**16.6**	83.5	65.9	59.5
SPG (ours)		68.7	**50.2**	**73.2**	**16.6**	83.3	**68.5**	**60.1**

Table 12: Comparisons with CLIP-base fine-tuning methods on PACS for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	Art	Cartoon	Photo	Sketch	Avg
Lin. Prob. [32]	79.2	82.2	76.7	71.2	77.3
CoOp [40]	**91.4**	84.5	82.5	85.4	86.0
CoCoOp [39]	91.1	85.7	88.2	87.5	88.1
DPL [36]	90.4	83.7	84.5	88.5	86.8
SPG (ours)	90.5	**87.4**	**88.4**	**88.7**	**88.8**

Table 13: Comparisons with CLIP-base fine-tuning methods on VLCS for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	Caltech	LableMe	Pascal	Sun	Avg
Lin. Prob. [32]	58.9	62.6	76.7	63.8	65.5
CoOp [40]	**74.6**	76.9	78.7	71.0	75.3
CoCoOp [39]	70.0	58.7	80.3	63.8	68.2
DPL [36]	64.8	77.0	**80.7**	78.3	75.2
SPG (ours)	70.2	**79.3**	76.3	**80.2**	**76.5**

Table 14: Comparisons with CLIP-base fine-tuning methods on OfficeHome for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	Art	Clipart	Product	Real	Avg
Lin. Prob. [32]	36.8	42.1	50.8	55.8	46.4
CoOp [40]	71.3	75.6	65.8	**70.2**	70.7
CoCoOp [39]	72.0	**75.7**	65.1	69.7	70.6
DPL [36]	72.0	75.3	65.7	70.1	70.8
SPG (ours)	**72.3**	74.8	**66.1**	**70.2**	**70.9**

Table 15: Comparisons with CLIP-base fine-tuning methods on TerraIncognita for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	Location38	Location43	Location46	Location100	Avg
Lin. Prob. [32]	**29.6**	16.0	18.3	29.1	23.3
CoOp [40]	23.7	**39.2**	40.8	19.2	30.7
CoCoOp [39]	22.8	27.5	34.0	18.1	25.6
DPL [36]	23.2	32.9	28.5	29.1	28.4
SPG (ours)	21.5	30.3	**40.8**	**36.5**	**32.3**

Table 16: Comparisons with CLIP-base fine-tuning methods on DomainNet for single-source DG in terms of leave-all-but-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
Lin. Prob. [32]	4.2	3.3	6.6	2.7	16.4	4.4	6.3
CoOp [40]	**47.2**	46.6	45.3	48.0	35.0	47.1	44.9
CoCoOp [39]	46.8	**48.0**	**46.3**	50.4	34.4	47.1	45.5
DPL [36]	46.6	47.5	45.2	40.6	35.6	**47.3**	43.8
SPG (ours)	44.0	45.4	45.2	**55.6**	**36.6**	46.8	**45.6**

D.3 Ablation Study

Component ablation. The domain prompt label and generative model are indispensable components of our SPG method and cannot be directly removed, but can be replaced. For domain prompt label, We replace the text prompt used in our work with VP [1]. For backbone, we replace the CGAN [27] with an MLP. Additional domain label and backbone ablations are as follows.

Table 17: Component ablation for multi-source DG in terms of mean leave-one-domain-out performance with ResNet50 as the backbone. Bold denotes the best scores.

Examples	PACS	VLCS	O.H.	TerraInc.	Do.Net	Avg
w/ VP	89.4	79.8	67.8	19.1	44.7	60.2
w/ MLP	90.4	79.6	72.3	31.7	47.8	64.4
SPG (Ours)	**92.8**	**84.0**	**73.8**	**37.5**	**50.1**	**67.6**