Supplementary Material for Learning a Cross-Modal Schrödinger Bridge for Visual Domain Generalization

Anonymous Author(s)
Affiliation
Address
email

Contents

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

A Theoretical Analysis: Generalization Error Bound

In this section, we derive a generalization error bound on the unseen target domains of the proposed SBGen, and demonstrate its superiority over the generalization error bound over the VLM baseline.

We start from some key definitions. Let P_0 and P_1 denote the source and the target feature distributions in $\mathbb{R}^C$. Let $\mathbb{Q}$ denote the law of our learned stochastic evolution (Schrödinger Bridge) from P_0 to P_1. The risk of a classification or segmentation model h w.r.t. distribution P can be defined as

$$R_P(h) \;=\; \mathbb{E}_{z\sim P}\big[\ell\big(h(z), y\big)\big], \tag{1}$$

where the task loss function ℓ is bounded in $[0, 1]$, and y denotes the ground truth.

The analysis will be based on the deduction of the empirical error on source domain and the expected error on target domain, defined as

$$R_{P_1}(h) \quad \text{(target risk)} \quad \text{to} \quad R_{P_0}(h) \quad \text{(source risk)}. \tag{2}$$

Lemma 1. Ben-David Transfer Bound. *Let P_0 and P_1 be two distributions over a common feature space $\mathcal{Z} \subseteq \mathbb{R}^C$, corresponding to the source and target domains, respectively. Let $h : \mathcal{Z} \to \mathcal{Y}$ be a hypothesis, and let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a bounded loss function. Then, the target risk of h satisfies:*

$$R_{P_1}(h) \;\le\; R_{P_0}(h) \;+\; \mathrm{Distance}_{\mathrm{TV}}(P_0, P_1) \;+\; \epsilon_{\mathrm{joint}}, \tag{3}$$

where $R_P(h) := \mathbb{E}_{(z,y)\sim P}[\ell(h(z), y)]$ denotes the expected risk under distribution P, $\mathrm{Distance}_{\mathrm{TV}}(P_0, P_1) := \frac{1}{2}\int |dP_0 - dP_1|$ denotes the total variation distance between the distributions, and $\epsilon_{\mathrm{joint}} := \min_{h'\in\mathcal{H}} [R_{P_0}(h') + R_{P_1}(h')]$ denotes the joint risk of the optimal shared hypothesis.

Proof. Please refer to [1] for the detailed proof.

Theorem 1. Variation Distance via Schrödinger Bridge. *Let $\mathbb{Q}$ be the solution to the Schrödinger Bridge problem between distributions P_0 and P_1 over $\mathbb{R}^C$, i.e., a path measure such that $\mathbb{Q}_{t=0} = P_0$ and $\mathbb{Q}_{t=1} = P_1$. Let $\mathbb{P}$ denote the reference Brownian motion with the same marginals at $t = 0$ and $t = 1$. Then the total variation distance between P_0 and P_1 is bounded by the KL divergence between $\mathbb{Q}$ and $\mathbb{P}$ as:*

$$\mathrm{Distance}_{\mathrm{TV}}(P_0, P_1) = \mathrm{Distance}_{\mathrm{TV}}(\mathbb{Q}_0, \mathbb{Q}_1) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{Q} \,\|\, \mathbb{P})}. \tag{4}$$

Proof: We apply Pinsker's inequality to the marginals of the SB process:

$$\mathrm{Distance}_{\mathrm{TV}}(\mu, \nu) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mu \,\|\, \nu)} \quad \text{for all probability measures } \mu, \nu. \tag{5}$$

Since the Schrödinger Bridge process $\mathbb{Q}$ interpolates from P_0 to P_1 over time $t \in [0, 1]$, and $\mathbb{Q}_0 = P_0$, $\mathbb{Q}_1 = P_1$, we apply Pinsker's inequality to the terminal marginal distributions of $\mathbb{Q}$ and $\mathbb{P}$.

Because $\mathbb{Q}$ and $\mathbb{P}$ are path measures with the same support, we have:

$$\mathrm{Distance}_{\mathrm{TV}}(\mathbb{Q}_0, \mathbb{Q}_1) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{Q} \,\|\, \mathbb{P})}, \tag{6}$$

and by definition $\mathbb{Q}_0 = P_0$, $\mathbb{Q}_1 = P_1$, so:

$$\mathrm{Distance}_{\mathrm{TV}}(P_0, P_1) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{Q} \,\|\, \mathbb{P})}. \tag{7}$$

Theorem 2. Generalization Error Bound of Schrödinger Bridge. *Let P_0 and P_1 be the source and target feature distributions over $\mathbb{R}^C$. Let $\mathbb{Q}_\theta$ be the path distribution induced by the Schrödinger Bridge model trained to transport $z_0 \sim P_0$ to $z_T \sim P_1$, and let $\mathbb{P}$ be the Brownian reference process. Let h_θ be the hypothesis (e.g., classifier or segmenter) composed with the SB mapping. Then, the expected target-domain risk is bounded as:*

$$R_{P_1}(h_\theta) \;\le\; R_{P_0}(h_\theta) \;+\; \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{Q}_\theta \,\|\, \mathbb{P})} \;+\; \epsilon_{\mathrm{joint}}, \tag{8}$$

where $R_P(h) := \mathbb{E}_{(z,y)\sim P}[\ell(h(z), y)]$ is the expected risk under distribution P, and $\epsilon_{\text{joint}} := \min_{h' \in \mathcal{H}} [R_{P_0}(h') + R_{P_1}(h')]$ is the optimal joint risk over the hypothesis class.

Proof: From Lemma 1, the basic transfer bound gives:

$$R_{P_1}(h_\theta) \leq R_{P_0}(h_\theta) + \text{Distance}_{\text{TV}}(P_0, P_1) + \epsilon_{\text{joint}}. \tag{9}$$

From Theorem 1, we apply Pinsker's inequality to the SB marginals:

$$\text{Distance}_{\text{TV}}(P_0, P_1) \leq \sqrt{\tfrac{1}{2}\text{KL}(\mathbb{Q}_\theta \| \mathbb{P})}. \tag{10}$$

Substituting yields:

$$R_{P_1}(h_\theta) \leq R_{P_0}(h_\theta) + \sqrt{\tfrac{1}{2}\text{KL}(\mathbb{Q}_\theta \| \mathbb{P})} + \epsilon_{\text{joint}}. \tag{11}$$

Theorem 3. Tighter Generalization Bound for Schrödinger Bridge Model. *Let P_0 and P_1 be the source and target distributions over $\mathbb{R}^C$. Let $\mathbb{Q}_\theta$ denote the Schrödinger Bridge process that evolves samples from P_0 to P_1 with reference prior $\mathbb{P}$. Let $T_\phi : \mathbb{R}^C \to \mathbb{R}^C$ be a deterministic baseline transport (e.g., cosine projection or prompt-aligned mapping), and let $P_1^\phi := T_{\phi\#}P_0$ denote the induced pushforward distribution. Let ℓ be a bounded loss function and h_θ, h_ϕ the hypotheses composed with the SB and baseline mappings, respectively. Then the generalization error of the SB model satisfies a strictly tighter upper bound:*

$$R_{P_1}(h_\theta) \leq R_{P_0}(h_\theta) + \sqrt{\tfrac{1}{2}\text{KL}(\mathbb{Q}_\theta \| \mathbb{P})} + \epsilon_{\text{joint}}, \tag{12}$$

$$R_{P_1}(h_\phi) \leq R_{P_0}(h_\phi) + \text{Distance}_{\text{TV}}(P_0, P_1^\phi) + \epsilon_{\text{joint}}. \tag{13}$$

Moreover, since $\mathbb{Q}_\theta$ minimizes the entropy-regularized transport cost from P_0 to P_1, and T_ϕ induces a deterministic coupling,

$$\sqrt{\tfrac{1}{2}\text{KL}(\mathbb{Q}_\theta \| \mathbb{P})} < \text{Distance}_{\text{TV}}(P_0, P_1^\phi) \tag{14}$$

unless T_ϕ itself induces the SB-optimal coupling.

Proof: The bound for the SB model is established in Theorem 2. For the deterministic baseline, we consider the mapping $z_1 = T_\phi(z_0)$ and define $P_1^\phi := T_{\phi\#}P_0$ as the transformed distribution.

Using the basic transfer bound (Lemma 1) again:

$$R_{P_1}(h_\phi) \leq R_{P_0}(h_\phi) + \text{Distance}_{\text{TV}}(P_0, P_1^\phi) + \epsilon_{\text{joint}}. \tag{15}$$

In contrast, the SB model produces a path distribution $\mathbb{Q}_\theta$ over z_t such that $\mathbb{Q}_{t=0} = P_0$, $\mathbb{Q}_{t=1} = P_1$. Applying Pinsker's inequality as in Theorem 2, we have:

$$\text{Distance}_{\text{TV}}(P_0, P_1) \leq \sqrt{\tfrac{1}{2}\text{KL}(\mathbb{Q}_\theta \| \mathbb{P})}. \tag{16}$$

Since the Schrödinger Bridge is known to minimize the KL divergence over all couplings between P_0 and P_1, and the deterministic map T_ϕ induces a coupling $\pi^\phi(z_0, z_1) = \delta(z_1 - T_\phi(z_0))$, we have:

$$\text{KL}(\mathbb{Q}_\theta \| \mathbb{P}) < \text{KL}(\pi^\phi \| \mathcal{R}), \tag{17}$$

for any reference coupling $\mathcal{R}$, unless π^ϕ itself is the SB-optimal coupling.

Therefore, the divergence and the TV-based generalization bound is strictly tighter under the SB transport.

Corollary 1. Match of the generalization bound between the SB model and the Deterministic Baseline. *Under the assumptions of Theorem 3, the generalization bounds of the Schrödinger Bridge model and the deterministic baseline coincide if and only if the SB-induced coupling $\mathbb{Q}_\theta$ corresponds to a deterministic map T^* satisfying:*

$$\mathbb{Q}_\theta(z_0, z_1) = \delta(z_1 - T^*(z_0)) \cdot P_0(z_0),$$

and this map T^* pushes P_0 exactly onto P_1, i.e.,

$$T^*_\# P_0 = P_1.$$

Algorithm 1 Schrödinger Bridge-Guided Domain Generalization

Require: Source images $\{x_i\}_{i=1}^{N}$, class text queries $\{\mathcal{Q}_c\}_{c=1}^{C}$, vision encoder $\mathcal{E}$, text encoder $\mathcal{T}$, time horizon T, noise scale ε, number of steps L
Ensure: Learned drift model $\mathcal{U}_\theta$, prediction decoder $\mathcal{D}$

Initialize $\mathcal{U}_\theta$, $\mathcal{D}$
for each training iteration **do**
  Sample mini-batch $\{x_i, y_i\}_{i=1}^{B}$ from source domain
  ### Domain-aware Visual Feature Selection ###
  Extract dense visual features: $\mathcal{F}_i = \mathcal{E}(x_i)$
  Encode class queries: $q_c = \mathcal{T}(\mathcal{Q}_c)$
  Compute similarity scores $S_{h,w,c} = \langle \mathcal{F}_{h,w}, q_c \rangle$
  Select top-k features: $\mathcal{F}_s \leftarrow$ query-guided selection from $\mathcal{F}$
  for each feature vector $z_0 \in \mathcal{F}_s$ **do**
    Initialize $z_t \leftarrow z_0$
    for $l = 1$ to L **do**
      $t \leftarrow \frac{l}{L}$
      Sample noise $\xi \sim \mathcal{N}(0, I)$
      ### Stochastic Cross-Domain Evolution & Domain-Agnostic Interpolation ###
      Update: $z_t \leftarrow z_t + \mathcal{U}_\theta(z_t, t)\,\Delta t + \sqrt{2\varepsilon\,\Delta t}\,\xi$
    end for
    Store final evolved feature z_T
  end for
  ### Prediction Head ###
  Predict: $\hat{y}_{\text{cls}}, \hat{y}_{\text{seg}} \leftarrow \mathcal{D}(\{z_T\}, \{q_c\})$
  Compute task losses $\mathcal{L}_{\text{sup}}$
  Estimate SB divergence (e.g., via score matching or IPFP): $\mathcal{L}_{\text{SB}}$
  Update parameters via $\nabla_\theta(\mathcal{L}_{\text{sup}} + \lambda\mathcal{L}_{\text{SB}})$
end for

In this case, the KL divergence collapses to:

$$\mathrm{KL}(\mathbb{Q}_\theta \| \mathbb{P}) = 2 \cdot \mathrm{Distance}_{\mathrm{TV}}{}^2(P_0, P_1),$$

and the generalization bounds for both models are equal:

$$R_{P_1}(h_\theta) = R_{P_1}(h_\phi).$$

We conclude this section by the following remark. The proposed `SBGen`, a Schrödinger Bridge guided framework, not only provides a principled dynamic interpolation between source and target distributions but also enjoys a strictly tighter generalization error upper bound compared to the deterministic baseline.

B Pseudo-code: Schrödinger Bridge-Guided Domain Generalization

A pesudo-code implementation of the proposed `SBGen` is given in Algorithm 1.

C More Implementation Details

Following prior work [7], the same training configuration is set for all types of pre-trained foundation models (*e.g.*, CLIP, DINOv2, and EVA02), and for both domain generalization in classification and semantic segmentation.

In all the experiments, the images are cropped and resized into 512×512 pixels. The batch size is set 16, with an AdamW optimizer. The initial learning rate is set to be 1×10^{-5} for all the synthetic-to-real settings, and is set to be 1×10^{-4} for all the real-to-real settings. The learning rate of the backbone is further scaled by 0.1. The training does not terminate after 20,000 iterations. Following [7], a linear warm-up is applied after 1500 iterations, followed by a linear decay. some common data augmentation techniques, namely, random scaling, random cropping, random flipping, color jittering, and rare class sampling, are also used.

Table 1: Impact of hyper-parameter λ. Evaluation metric is mIoU in %.

λ	G $\rightarrow$ C	G $\rightarrow$ B	G $\rightarrow$ M	Avg.
0.01	69.83	60.67	70.68	67.06
0.1	70.65	61.72	71.34	67.90
1	**71.24**	**62.26**	**71.91**	**68.74**
10	71.01	61.17	71.27	67.82
100	70.37	61.26	71.28	67.64

Table 2: Impact of hyper-parameter K. Evaluation metric is mIoU in %.

K	G$\rightarrow$C	G$\rightarrow$B	G$\rightarrow$G	Avg.
0	70.39	60.57	70.54	67.17
0.1	70.57	60.81	70.90	67.43
0.2	71.04	61.58	71.59	68.07
0.3	**71.24**	**62.26**	**71.91**	**68.74**
0.4	71.15	62.04	71.37	68.19
0.5	70.92	61.75	71.16	67.94

Domain generalization in classification. For the classification task, the image encoder $\mathcal{E}$ and the text encoder $\mathcal{T}$ use the pre-trained CLIP in align with the prior DG methods. The task-specific decoder $\mathcal{D}$ is a linear layer followed by a Softmax layer.

Domain generalization in segmentation. Following prior domain generalized semantic segmentation methods [7, 8], the default image encoder $\mathcal{E}$ and the text encoder $\mathcal{T}$ use the pre-trained EVA-02 [5]. The image encoder $\mathcal{E}$ can also be switched to CLIP, SAM and DINOv2 in our experiments. The task-specific decoder $\mathcal{D}$ integrates the pixel decoder of the Mask2Former model [4].

D More Ablation Studies

D.1 On Hyper-parameter λ

The hyper-parameter λ in Eq.9 balances the impact of the task-specific loss and the cross-modal Schrödinger Bridge loss. By default, it is set to be 1 in all of our experiments. To observe its impact on domain generalization, we further conduct the experiments when it is set to be 0.01, 0.1, 10 and 100, respectively.

The results are reported in Table 1. We observe that when λ is set to 1, the segmentation performance on unseen target domains achieves the optimal performance, yielding an average of 68.74% mIoU. A too small λ (*e.g.*, 0.01) may not be able to impose an effective and sufficient alignment between the domain-agnostic class text and the domain-specific visual features, thereby degrading the generalization. A too large λ (*e.g.*, 100) may overwhelm the task loss, also leading to a performance drop.

D.2 On Feature Selection Ratio K

The Domain-aware Visual Feature Selection module selects the top-K visual features that are domain-specific to align the domain-agnostic class-wise text embeddings. By default, K is set to be 0.3 under all of our experiments, for domain generalization in both classification and segmentation. To observe its impact on domain generalization, we further conduct the experiments when it is set to be 0, 0.1, 0.2, 0.4 and 0.5, respectively.

The results are reported in Table 2. We observe that when K is set to 0.3, the segmentation performance on unseen target domains achieves the optimal performance, yielding an average of 68.74% mIoU. A too small K (*e.g.*, 0 and 0.1) may not be able to select sufficient visual features that are domain-specific for the alignment between the domain-agnostic class text and the domain-specific visual features, which may under-fit the generalization representation. A too large K (*e.g.*, 0.4 and 0.5) may introduce more visual features that are not domain-specific in the representation learning, which may lead to over-fit and result in a slight performance drop.

E More Feature Space Analysis

Fig.4 in the main text inspects whether the proposed `SBGen` can improve the generalization ability over the baseline, on the task of domain generalized semantic segmentation (DGSS). In the supplementary material, we further inspect whether the proposed `SBGen` can improve the generalization ability over the baseline on domain generalization in the classification task.

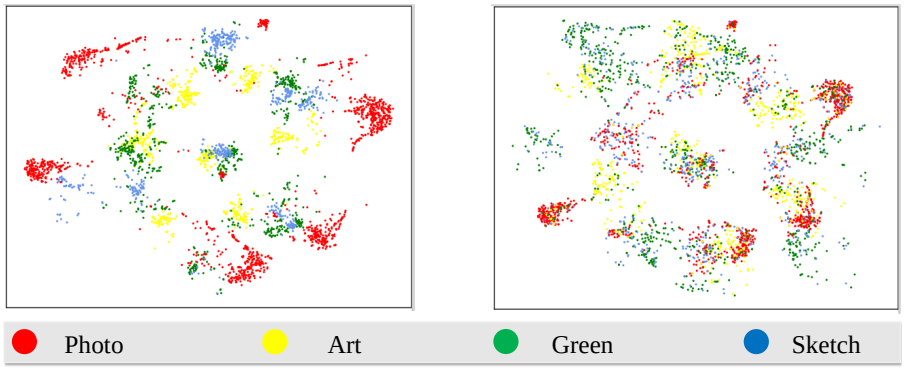

Figure 1: t-SNE visualization. Feature embedding is extracted before the decoder. Left: EVA02 baseline; Right: ours.

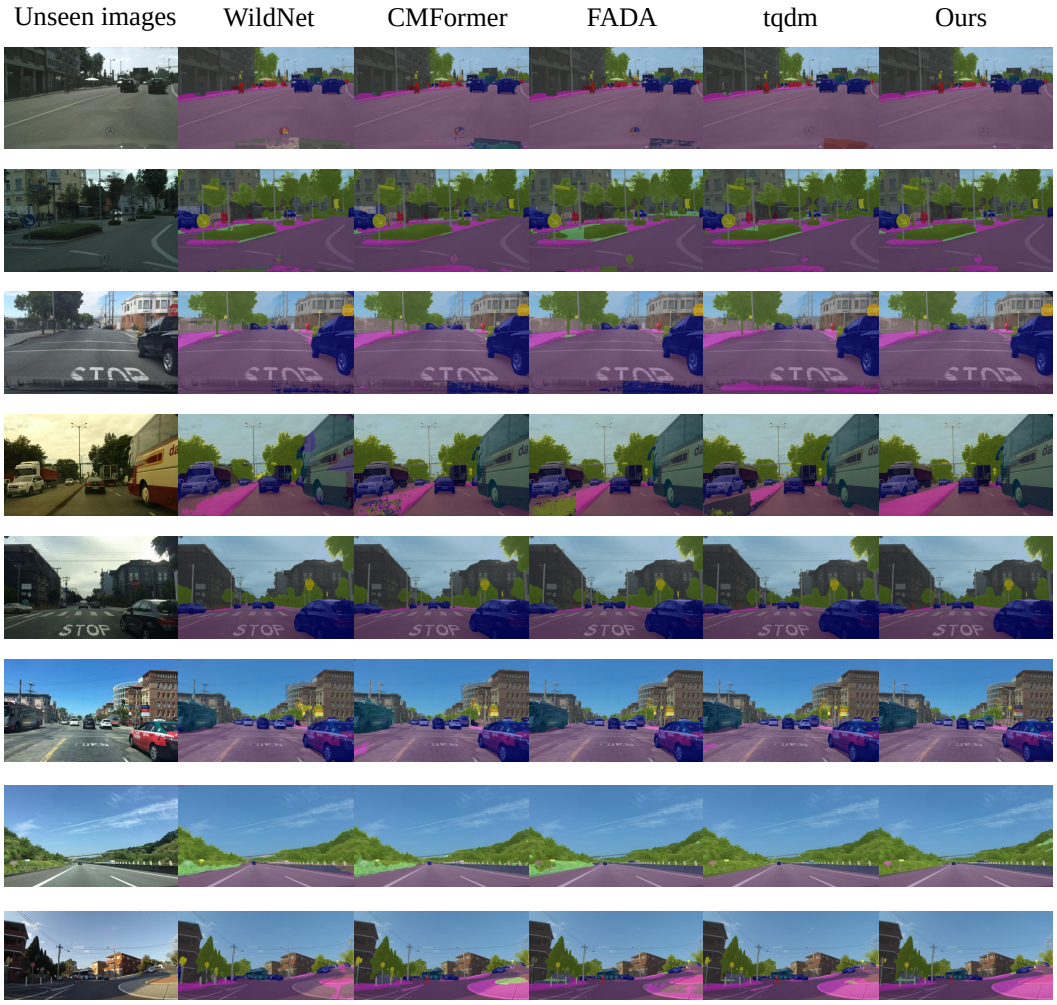

Figure 2: Visual segmentation results on unseen target domains under the G $\rightarrow$ B, M, C setting. The proposed `SBGen` is compared with WildNet [6], CMFormer [3], FADA [2], and tqdm [7].

Specifically, we extract the feature of each sample from the PACS dataset before the decoder and concatenate it into a feature vector. Then, we display the feature vector of each sample regardless of the domain identity by t-SNE visualization. Feature vectors from the Photo, Art Painting, Cartoon and Sketch domains are colored in red, yellow, green and blue, respectively.

The feature space of the original baseline and the proposed SBGen is visualized in the left and right of Fig. 1, respectively. In each cluster that shares the same semantic category, the samples from different domains are more uniformly distributed by the proposed SBGen, indicating its effectiveness to mitigate the domain gap.

F More Visual Prediction Results

Fig. 2 shows more results under G $\rightarrow$ B, M, C setting. The segmentation results show that the proposed SBGen shows better pixel-wise prediction than the compared DGSS methods, especially in terms of the completeness of objects.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[2] Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37:94047–94072, 2024.

[3] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 819–827, 2024.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[5] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

[6] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022.

[7] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, pages 37–54, 2024.

[8] PeiYuan Tang, Xiaodong Zhang, Chunze Yang, Haoran Yuan, Jun Sun, Danfeng Shan, and Zijiang James Yang. Unleashing the power of visual foundation models for generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20823–20831, 2025.