# A Simple Recipe for Language-guided Domain Generalized Segmentation

## Supplementary Material

We provide details about Table 1 experiments in Appendix A. Further, we report class-wise performance for FAMix in Appendix B, and detail the prompts used in our experiments in Appendix C. Finally, we discuss the limitations and perspectives in Appendix D.

We refer to the **supplementary video** for further demonstration of FAMix qualitative performance: https://youtu.be/vyjtvx2El9Q.

## A. CLIP *vs*. ImageNet initialization

In Table 1 of the main paper, we introduce a comparison of ImageNet and CLIP pretraining for out-of-distribution semantic segmentation. We clarify here its implementation.

To produce Table 1 we employ the public code[1] and fine-tuned with SGD using a learning rate of $10^{-1}$ for the segmenter and $10^{-2}$ for the backbone. We note that we freeze the *stem layers* and *Layer1* for both backbones, *i.e*., ImageNet and CLIP initialized ResNet-50, after observing that full fine-tuning leads to subpar in-domain results for CLIP. Crucially, in this setting, *both* ImageNet and CLIP initialized networks converge and achieve the same performance in-domain (on GTA5 validation set). Hence, we argue that the poor OOD performance of CLIP initialization in Table 1 may originate from the distortion of the robust CLIP representation towards the source domain distribution as advocated in [4]. To alleviate such distortion, we freeze most of the backbone layers in FAMix.

However, we highlight that different hyper-parameter choices could boost the performance to some extent. For example, Rao *et al.* [8] observed that fine-tuning CLIP for semantic segmentation with the default configuration in MMSegmentation[2] leads to 15.6% mIoU lower performance than its ImageNet pre-trained counterpart on ADE20K [9]. Consequently, they propose using AdamW [6] for optimization.

In Tab. 11 we reproduce the experiment of Table 1 experiment but using AdamW as optimizer. $O_1$ refers to the optimization configuration adopted in our paper, *i.e*., SGD optimizer with a learning rate of $10^{-1}$ for the segmenter and $10^{-2}$ for the backbone. $O_2$ and $O_3$ both refer to the use of AdamW with a learning rate of $10^{-4}$ for the segmenter and $10^{-5}$ for the backbone. In $O_2$ all the backbone is fine-tuned while in $O_3$ the *stem layers* and *Layer1* are frozen, similar to $O_1$. Results show that using AdamW improves performance across out-of-distributions (OOD) domains for both CLIP and ImageNet initialized networks, but still largely lags behind FAMix (mean mIoU=38.88%) and even our variant (Freeze ✓, Augment ✗, Mix ✗) (mean mIoU=32.44%) in Table 6.

| Optim. | Pretraining | C | B | M | S | AN | AS | AR | AF | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| $O_1$ | ImageNet | **29.04** | **32.17** | **34.26** | **29.87** | **4.36** | **22.38** | **28.34** | **26.76** | **25.90** |
| | CLIP | 16.81 | 16.31 | 17.80 | 27.10 | 2.95 | 8.58 | 14.35 | 13.61 | 14.69 |
| $O_2$ | ImageNet | 28.00 | **36.82** | **37.00** | 30.60 | **3.56** | **24.14** | **29.51** | **26.23** | **26.98** |
| | CLIP | **31.73** | 25.89 | 30.68 | **33.32** | 2.56 | 19.17 | 21.42 | 17.58 | 22.79 |
| $O_3$ | ImageNet | **28.74** | **36.91** | **37.86** | 30.32 | **4.46** | **22.48** | **28.49** | **25.38** | **26.83** |
| | CLIP | 26.81 | 23.11 | 29.82 | **32.38** | 4.20 | 18.50 | 22.59 | 20.31 | 22.22 |

Table 11. **Effect of optimization configurations on OOD performance.** Performance (mIoU %) of CLIP *vs*. ImageNet initialized networks for different optimization configurations.

These results hint that using AdamW with relatively low learning rates might reduce the feature distortion of CLIP. Motivated by this observation, one could question the necessity of the minimal fine-tuning part of FAMix, and whether similar results could be achieved only by augmenting, mixing, and fine-tuning with AdamW and low learning rate. We call this variant AMix (Augment and Mix) and show the results in Tab. 12, which support the necessity of our full recipe.

| Method | C | B | M | S | AN | AS | AR | AF | Mean |
|---|---|---|---|---|---|---|---|---|---|
| AMix | 40.50 | 38.69 | 36.05 | 33.61 | 4.03 | 23.03 | 30.01 | 26.89 | 29.10 |
| FAMix | **48.15** | **45.61** | **52.11** | **34.23** | **14.96** | **37.09** | **38.66** | **40.25** | **38.88** |

Table 12. **AMix with AdamW optimizer *vs*. FAMix.** Performance (mIoU %) of FAMix in our default configuration compared to a variant with no minimal fine-tuning, replacing SGD with AdamW optimizer.

## B. Class-wise performance

We report class-wise IoUs in Tab. 13 and Tab. 14. The standard deviations of the mIoU (%) over three runs are also reported.

## C. Prompts used for mining

The <random style prompt> used for training FAMix: $\mathcal{R}_1$ = <random style prompt> = { *Ethereal Mist, Cyberpunk Cityscape, Rustic Charm, Galactic Fantasy, Pastel Dreams, Dystopian Noir, Whimsical Wonderland, Urban Grit, Enchanted Forest, Retro Futurism, Monochrome Elegance, Vibrant Graffiti, Haunting Shadows, Steampunk Adventures, Watercolor Serenity, Industrial Chic, Cosmic*

[1] https://github.com/VainF/DeepLabV3Plus-Pytorch
[2] https://github.com/open-mmlab/mmsegmentation

| Target eval. | road | sidewalk | building | wall | fence | pole | traffic light | traffic sign | vegetation | terrain | sky | person | rider | car | truck | bus | train | motorcycle | bicycle | mIoU% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C** | 88.97 | 39.48 | 83.12 | 28.52 | 29.06 | 38.64 | 42.67 | 36.26 | 86.48 | 24.70 | 78.26 | 69.08 | 23.88 | 85.35 | 29.63 | 38.82 | 9.28 | 38.02 | 44.68 | 48.15 $\pm$0.38 |
| **B** | 87.83 | 40.33 | 78.44 | 15.88 | 35.20 | 38.13 | 40.63 | 29.49 | 77.52 | 31.19 | 90.80 | 60.28 | 23.23 | 82.99 | 26.73 | 34.53 | 0.00 | 43.55 | 29.92 | 45.61 $\pm$0.84 |
| **M** | 86.65 | 41.65 | 78.67 | 26.91 | 30.88 | 45.91 | 46.50 | 61.48 | 81.84 | 38.79 | 94.09 | 68.65 | 33.59 | 84.52 | 40.90 | 42.40 | 10.15 | 41.20 | 35.24 | 52.11 $\pm$0.17 |
| **S** | 60.55 | 49.61 | 82.63 | 7.80 | 5.42 | 29.23 | 15.71 | 15.26 | 68.18 | 0.00 | 90.83 | 61.59 | 12.09 | 61.29 | 0.00 | 35.23 | 0.00 | 32.75 | 22.19 | 34.23 $\pm$0.53 |
| **AN** | 47.44 | 7.01 | 38.73 | 8.42 | 3.59 | 23.04 | 18.42 | 5.75 | 19.33 | 5.82 | 5.65 | 26.61 | 10.39 | 50.46 | 4.10 | 0.00 | 0.79 | 8.42 | 0.28 | 14.96 $\pm$0.09 |
| **AS** | 66.93 | 10.15 | 62.17 | 33.95 | 22.10 | 35.26 | 51.20 | 35.57 | 73.12 | 20.72 | 77.55 | 52.02 | 0.63 | 71.62 | 21.20 | 1.14 | 12.36 | 47.32 | 9.71 | 37.09 $\pm$0.83 |
| **AR** | 73.41 | 21.42 | 77.58 | 19.41 | 16.96 | 33.63 | 44.90 | 38.53 | 80.96 | 29.31 | 94.98 | 56.89 | 17.04 | 76.15 | 16.15 | 7.07 | 5.11 | 23.74 | 1.24 | 38.66 $\pm$1.12 |
| **AF** | 77.61 | 31.99 | 76.42 | 28.84 | 10.30 | 31.42 | 52.92 | 29.99 | 68.09 | 24.55 | 92.03 | 54.52 | 34.91 | 68.21 | 26.07 | 11.02 | 1.44 | 7.71 | 36.78 | 40.25 $\pm$0.71 |

Table 13. **ResNet-50 class-wise performance.** We report the performance of FAMix (IoU %) trained on **G** with ResNet-50 as backbone.

| Target eval. | road | sidewalk | building | wall | fence | pole | traffic light | traffic sign | vegetation | terrain | sky | person | rider | car | truck | bus | train | motorcycle | bicycle | mIoU% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C** | 90.25 | 44.78 | 84.54 | 31.71 | 31.48 | 44.17 | 45.45 | 35.78 | 87.17 | 35.58 | 84.30 | 69.62 | 20.48 | 86.87 | 31.11 | 44.38 | 7.73 | 34.06 | 30.46 | 49.47 $\pm$0.36 |
| **B** | 86.79 | 41.47 | 79.53 | 16.67 | 41.27 | 39.41 | 42.31 | 33.35 | 78.64 | 36.86 | 91.03 | 60.32 | 23.73 | 81.51 | 31.13 | 25.99 | 0.00 | 45.78 | 25.73 | 46.40 $\pm$0.50 |
| **M** | 78.71 | 38.89 | 81.85 | 26.56 | 40.22 | 47.32 | 49.27 | 62.19 | 82.68 | 41.54 | 95.63 | 67.60 | 25.87 | 85.50 | 41.62 | 35.87 | 12.55 | 43.69 | 29.92 | 51.97 $\pm$1.30 |
| **S** | 63.72 | 56.80 | 85.05 | 9.30 | 21.62 | 33.26 | 16.44 | 18.96 | 69.42 | 0.00 | 92.10 | 63.52 | 10.95 | 64.86 | 0.00 | 29.84 | 0.00 | 39.67 | 22.22 | 36.72 $\pm$0.71 |
| **AN** | 65.87 | 23.23 | 37.83 | 13.72 | 4.60 | 30.34 | 16.49 | 7.48 | 27.37 | 7.83 | 17.61 | 35.16 | 18.48 | 53.71 | 5.67 | 0.00 | 0.84 | 10.25 | 1.44 | 19.89 $\pm$1.22 |
| **AS** | 75.42 | 31.90 | 72.15 | 36.75 | 27.85 | 38.89 | 49.63 | 33.09 | 72.45 | 22.98 | 84.21 | 56.75 | 1.91 | 75.84 | 34.61 | 4.44 | 4.17 | 48.91 | 14.26 | 41.38 $\pm$0.34 |
| **AR** | 57.58 | 26.76 | 79.76 | 19.79 | 21.06 | 37.70 | 46.34 | 37.66 | 83.85 | 36.96 | 94.80 | 55.40 | 31.61 | 79.53 | 14.84 | 14.01 | 6.97 | 29.36 | 3.34 | 40.91 $\pm$1.28 |
| **AF** | 77.96 | 41.73 | 77.99 | 34.73 | 6.85 | 36.80 | 49.49 | 34.51 | 72.00 | 32.60 | 91.52 | 46.28 | 27.28 | 70.77 | 31.17 | 19.08 | 4.87 | 10.75 | 34.55 | 42.15 $\pm$1.87 |

Table 14. **ResNet-101 class-wise performance.** We report the performance of FAMix (IoU %) trained on **G** with ResNet-101 as backbone.

*Voyage, Pop Art Popularity, Abstract Symphony, Magical Realism, Abstract Geometric Patterns, Vintage Film Grain, Neon Cityscape Vibes, Surreal Watercolor Dreams, Minimalist Nature Scenes, Cyberpunk Urban Chaos, Impressionist Sunset Hues, Pop Art Explosion, Fantasy Forest Adventures, Pixelated Digital Chaos, Monochromatic Street Photography, Vibrant Graffiti Expressions, Steampunk Industrial Charm, Ethereal Cloudscapes, Retro Futurism Flare, Dark and Moody Landscapes, Pastel Dreamworlds, Galactic Space Odyssey, Abstract Brush Strokes, Noir Cinematic Moments, Whimsical Fairy Tale Realms, Modernist Architectural Wonders, Macro Botanical Elegance, Dystopian Sci-Fi Realities, High Contrast Street Art, Impressionist City Reflections, Pixel Art Nostalgia, Dynamic Action Sequences, Soft Focus Pastels, Abstract 3D Renderings, Mystical Moonlit Landscapes, Urban Decay Aesthetics, Holographic Futuristic Visions, Vintage Polaroid Snapshots, Digital Glitch Anomalies, Japanese Zen Gardens, Psychedelic Kaleidoscopes, Cosmic Abstract Portraits, Subtle Earthy Textures, Hyperrealistic Wildlife Portraits, Cybernetic Neon Lights, Warped Reality Illusions, Whimsical Watercolor Animals, Industrial Grunge Textures, Tropical Paradise Escapes, Dynamic Street Performances, Abstract Architectural Wonders, Comic Book Panel Vibes, Soft Glow Sunsets, 8-Bit Pixel Adventures, Galactic Nebula Explosions, Doodle Sketchbook Pages, High-Tech Futuristic Landscapes, Cinematic Noir Shadows, Vibrant Desert Landscapes, Abstract Collage Chaos, Nature in Infrared, Surreal Dream Sequences, Abstract Light Painting, Whimsical Fantasy Creatures, Cybernetic Augmented Reality, Impressionist Rainy Days, Vintage Aged Photographs, Neon Anime Cityscapes, Pastel Sunset Palette, Surreal Floating Islands, Abstract Mosaic Patterns, Retro Sci-Fi Spaceships, Futuristic Cyber Landscapes, Steampunk Clockwork Contraptions, Monochromatic Urban Decay, Glitch Art*

*Distortions, Magical Forest Enchantments, Digital Oil Painting, Pop Surrealist Dreams, Dynamic Graffiti Murals, Vintage Pin-up Glamour, Abstract Kinetic Sculptures, Neon Jungle Adventures, Minimalist Futuristic Interfaces* }

The <`random character prompt`> used in Tab. 6 experiments (*i.e.*, 'RCP') are:
$\mathcal{R}_2$ = <`random character prompt`> = { *ioscjspa, cjosae, wqvsecpas, csavwggw, csanoiaj, zfaspf, atpwqkmfc, mdmfejh, casjicjai, cnoacpoaj, noiasvnai, kcsakofnaoi, cjncioasn, wkqgmdc, jqblhyu, pqwfkgr, mzxanqnw, wnzsalml, sdqlhkjr, odfeqfit* }

Both $\mathcal{R}_1$ and $\mathcal{R}_2$ are concatenated with the word `style`.

# D. Limitations and perspectives

## D.1. Limitations

**Failures conditions.** We show in Fig. 5 failure cases in rare conditions, which include extreme illumination or darkness, low visibility due to rain drops on the windshield, and other adverse conditions (*e.g.*, snowy road). While FAMix improves over the baseline, the results remain unsatisfactory for safety-critical applications as it fails to segment critical objects in the scenes (*e.g.*, car, road, sidewalk, person etc). We leave for future research the generalization to the above mentioned conditioned. One possible direction could be to design specific methods for specific corner conditions (*e.g.*, [3, 5]), although we highlight this is orthogonal to generalization.

**Stylization.** At the heart of FAMix lies the assumption that unseen target distributions could be covered by augmenting the mean and standard deviation of the low-level features. While the correlation between "style" and these parameters has been shown in previous research [7, 10], we believe that the hypothesis stating that the domain shift could be described only by these parameters over-simplifies generalization. Moreover, FAMix does not handle or provide an estimation of uncertainty, which is crucial for both classes in and outside the label set for the application at hand.

## D.2. Perspectives

Vision transformers (ViTs) [1] have recently emerged as an alternative to CNNs. We leave a ViT implementation of FAMix for future work. Applying prompt-driven instance normalization (PIN) [2] to ViTs appears non-trivial as the relation between statistics of low-level feature maps and style is established only for CNNs so far, and could raise some technical challenges. Exploring this direction might first involve a study of the correlation between style and statistics of patches. If such correlation is demonstrated, a naive way to apply FAMix could be by applying PIN with tied parameters across the patches.

While some modern architectures are inherently more robust than older ones, the problem of DGSS with ResNets (*e.g.* ResNet-50 and ResNet-101) is still not solved. As long as the gap exists between in-domain and out-of-distribution performances, we believe that this setting remains interesting, and that a general understanding of domain generalization could emerge from the algorithms proposed to address it.

# References


[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[2] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Poda: Prompt-driven zero-shot domain adaptation. In *ICCV*, 2023. 3

[3] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *ICCV*, 2019. 3

[4] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 1

[5] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *ICCV*, 2021. 3

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[7] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 3

[8] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 1

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1

[10] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 3

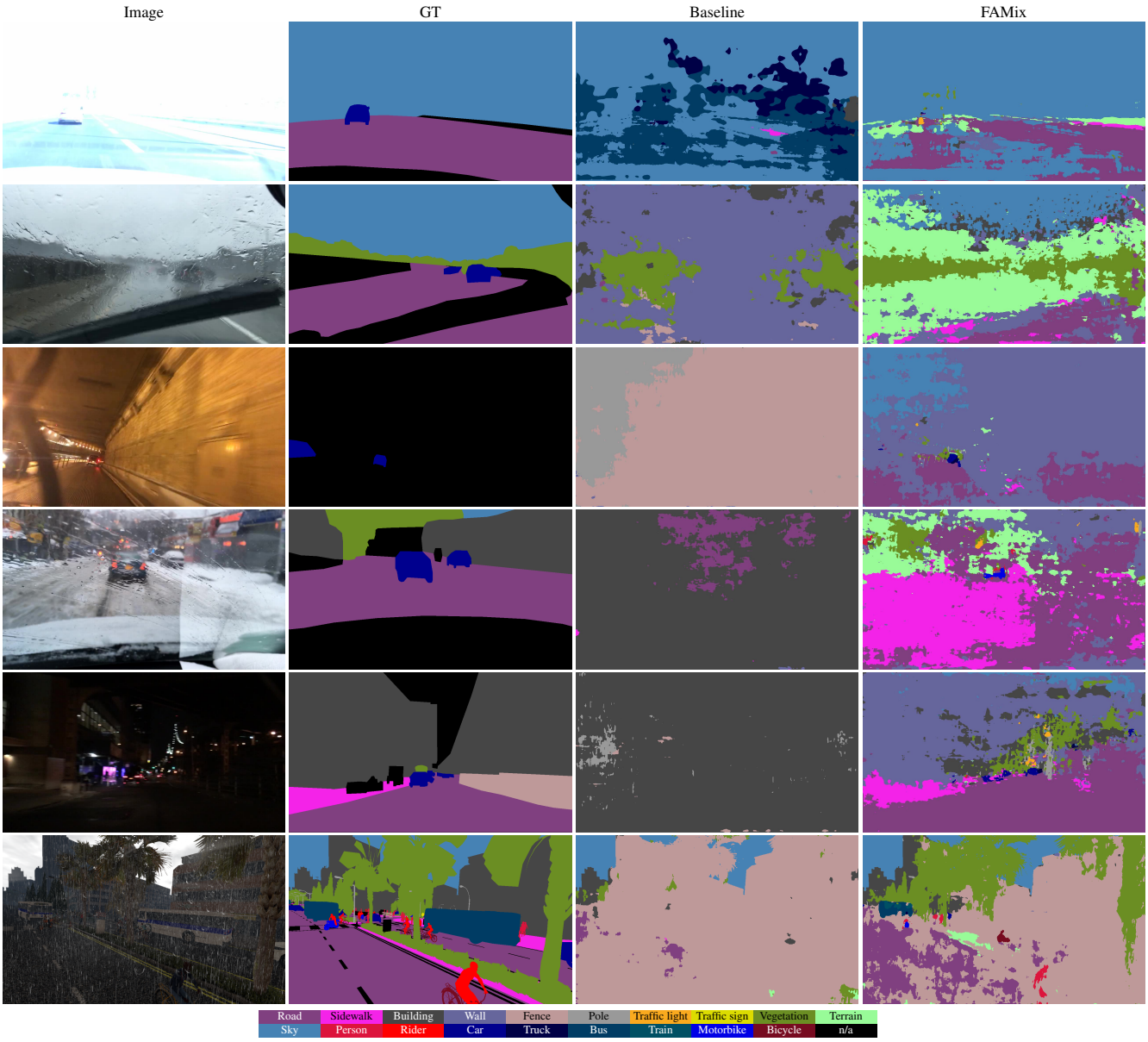


Figure 5. **Examples of failure cases.** *Columns 1-2*: Image and Ground Truth (GT), *Column 3*: Baseline (Freeze ✗, Augment ✗, Mix ✗), *Column 4*: FAMix results. The models are trained on G with ResNet-50 backbone.