CFRNET: COUNTERFACTUAL FREQUENCY-GUIDED SELECTIVE RETENTION FOR SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-frequency information is important for object boundaries and structures in semantic segmentation, but can also arise from background artifacts and object-like patterns, making preservation or attenuation unreliable. We propose CFRNet, a counterfactual frequency-guided framework that adaptively modulates high-frequency retention by combining probe-output differences from controlled removal and perturbation with multi-scale structural support. The resulting spatial and channel-wise retention signals selectively reduce or preserve reliable high-frequency information. On PASCAL VOC 2012, COCO-Stuff 164K, and ADE20K, CFRNet achieves mIoUs of 0.701, 0.468, and 0.390, respectively. CFRNet outperforms the strongest counterparts by 0.058 and 0.016 mIoU on COCO-Stuff and ADE20K, while remaining within 0.002 mIoU of the highest VOC result. Mathematical analysis and ablations further examine the component contributions and the conditional retention directions induced by the CFR objective.

1 INTRODUCTION

Semantic segmentation is a task that interprets an image by assigning a class label to each pixel (Long et al., 2015; Minaee et al., 2022). Such pixel-level scene understanding is widely used in various fields, including autonomous driving, robotic perception, and remote sensing (Minaee et al., 2022). Accurate segmentation requires information that captures object structure and semantic consistency, together with fine-grained cues for distinguishing local structures and object boundaries (Takikawa et al., 2019; Yuan et al., 2020; Kirillov et al., 2020; Wang et al., 2021). Regarding this point, conventional methods focused on frequency-based representation created by extracting the information that lies in a specific frequency band of input image (Chen et al., 2024b;c;a; 2025a).

Frequency decomposition can be used to separate and process information with different characteristics (Chen et al., 2019; Qin et al., 2021). The low-frequency component represents global structures with relatively small variations in pixel values, such as object regions, whereas the high-frequency component contains local and fine-grained details, such as object boundaries(Chen et al., 2019; 2024a; 2025a). One well-known tricky point is that, while the high-frequency component is important for representing accurate boundary information, it can also arise from background artifacts or object-like local structures (Chen et al., 2024c;a).

Although existing frequency-aware methods have improved semantic segmentation performance (Chen et al., 2024b;c;a; 2025a; Luo et al., 2026; Chen et al., 2025b), these methods provide effective mechanisms for modulating frequency components while putting relatively little focus on whether local high-frequency information provides reliable evidence for semantic prediction (Chen et al., 2024b;c;a; 2025a; Luo et al., 2026; Chen et al., 2024c;a). In this study, we propose CFRNet, a frequency-guided semantic segmentation framework that selectively modulates high-frequency information according to a learned retention score. Specifically, CFRNet constructs a high-frequency retention score by combining differences in probe outputs with multi-scale structural support and modulates high-frequency retention in decoder features using spatial and channel-wise retention signals. In addition, our mathematical derivation shows that, under the specified active-region conditions, the CFR objective induces directions that decrease high-frequency retention at prediction-error locations receiving nonzero error-suppression weights and increase high-frequency retention

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

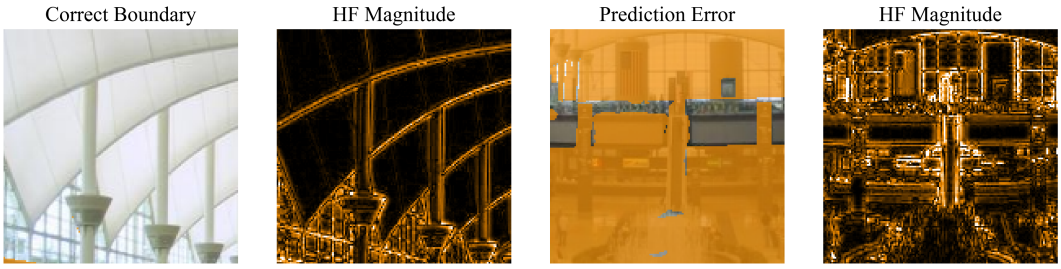


Figure 1: Motivating example of high-frequency magnitude in semantic segmentation. Strong high-frequency responses appear at correctly predicted object boundaries and misclassified regions, indicating that high-frequency magnitude alone is insufficient to determine whether local high-frequency information should be retained.

at correctly predicted locations receiving nonzero preservation weights whose retention score falls below a predefined lower bound.

2 Related Work

Existing semantic segmentation methods have improved pixel-level prediction through multi-scale spatial modeling, boundary-aware representations, and enhanced semantic context. DeepLabv3+ combines atrous spatial pyramid pooling with encoder–decoder refinement, while Gated-SCNN employs a dedicated shape stream for boundary prediction (Chen et al., 2018; Takikawa et al., 2019). Context-oriented methods such as OCRNet, SegFormer, and Mask2Former enhance semantic representations through object-level context, hierarchical Transformer features, and masked attention (Yuan et al., 2020; Xie et al., 2021; Cheng et al., 2022), while related architectures preserve spatial details and contextual information (Zhao et al., 2017; Xiao et al., 2018; Wang et al., 2021; Yu et al., 2018; Zheng et al., 2021; Kirillov et al., 2020).

Other studies have exploited frequency information to alleviate frequency distortion and preserve fine-grained structures (Chen et al., 2019; Qin et al., 2021; Zhang, 2019; Zou et al., 2020; Yang et al., 2025; Chen et al., 2026). When Semantic Segmentation Meets Frequency Aliasing related segmentation errors to aliasing introduced during downsampling and proposed DAF and FreqMix for suppressing aliased components and dynamically selecting high-frequency information (Chen et al., 2024b). FADC adapts dilation and frequency composition according to local frequency characteristics (Chen et al., 2024c), while FreqFusion performs frequency-aware multi-level feature fusion using adaptive filtering and feature resampling (Chen et al., 2024a). SFM modulates high-frequency information across downsampling and upsampling (Chen et al., 2025a), and more recent work employs dynamic Gaussian filtering and learnable Gabor-based frequency selection to reduce aliasing and phase distortion (Luo et al., 2026). Collectively, these methods exploit frequency information through aliasing suppression, spatial adaptation, feature consistency, and frequency recovery (Chen et al., 2024b;c;a; 2025a; Luo et al., 2026; Chen et al., 2026).

Existing frequency-aware segmentation methods primarily modulate high-frequency information based on aliasing behavior, spatial adaptation, or local feature consistency (Chen et al., 2024b;c;a; 2025a; Luo et al., 2026). In contrast, they generally do not explicitly combine probe-output differences under controlled high-frequency interventions with multi-scale structural support to construct an adaptive high-frequency retention signal (Chen et al., 2024b;c;a; 2025a; Luo et al., 2026).

3 Proposed Method

We present the CFR formulation and local retention analysis below, with complete proofs provided in Appendix A.

Table 1: Notation used in the theoretical analysis.

Notation	Description
s	CFC stage index, $s \in \{1, \ldots, 5\}$
$\mathbf{F}_s, \widetilde{\mathbf{F}}_s, \mathbf{S}_s$	Input feature, calibrated feature, and structural support at stage s, respectively
$\mathbf{L}_s, \mathbf{H}_s$	Low- and high-frequency components $\mathbf{F}_s = \mathbf{L}_s + \mathbf{H}_s$, respectively
$\mathbf{F}_s^{\mathrm{rem}}, \mathbf{F}_s^{\mathrm{pert}}, \mathbf{U}_s^b$	Counterfactual features and displacement from $b \in \{\mathrm{rem}, \mathrm{pert}\}$, respectively
α	High-frequency perturbation strength, $\alpha \in [0, 1]$
P_s, D_s	Shared probe and intervention-induced probe-output difference map, respectively
$\mathcal{C}_s$	CFC mapping at stage s
$\mathbf{R}_s, \overline{R}_s$	High-frequency retention-score map and its channel average, respectively
$Q_s^{\mathrm{err}}, Q_s^{\mathrm{cor}}, Q_s^{\mathrm{bd}}, \mathcal{L}_{\mathrm{CFR},s}$	Detached CFR weighting maps and per-stage CFR objective, respectively
$\mathcal{V}_s$	Valid pixel region at stage s, excluding ignored labels
$\Omega_s^{\mathrm{err}}, \Omega_s^{\mathrm{cor}}$	Prediction-error and correct-prediction regions within $\mathcal{V}_s$, respectively
Ω_s^{bd}	Multiclass semantic-boundary region derived from valid ground-truth labels
$\rho_{\mathrm{cor}}, \rho_{\mathrm{bd}}, w_{\mathrm{err}}, w_{\mathrm{cor}}, w_{\mathrm{bd}}$	Retention-score floors and corresponding CFR weights, respectively
$\xi_s, \gamma_s, \bar{\beta}_s(x)$	Calibration logit, strength $\gamma_s = \sigma(\xi_s)$, and retention coefficient, respectively

3.1 MOTIVATION

High-frequency information can appear at object boundaries as well as in background textures and object-like local structures, complicating direct high-frequency processing. As shown in Figure 1, strong high-frequency magnitude is observed in both correctly predicted object boundaries and misclassified regions, indicating that magnitude alone is insufficient to determine how strongly local high-frequency information should be retained. Thus, an additional control signal is required to adaptively modulate high-frequency retention at each spatial location according to both intervention-induced feature responses and surrounding structural context. In this study, we define R_s as a learned high-frequency retention score rather than a calibrated measure of prediction confidence or local high-frequency utility. The retention score is constructed during decoding from two complementary signals: probe-output differences induced by controlled high-frequency interventions and consistency with multi-scale structural support for stage-wise feature calibration.

3.2 COUNTERFACTUAL HIGH-FREQUENCY GUIDED RETENTION ANALYSIS

Table 1 summarizes the notation used throughout the following theoretical analysis. We analyze the probe-output responses induced by controlled high-frequency interventions and the corresponding local retention directions imposed by the CFR objective under the specified active-region conditions.

Definition 1 (CFC Module). At each calibration stage $s \in \{1, \ldots, 5\}$, let $F_s \in \mathbb{R}^{c_s \times h_s \times w_s}$ and $S_s \in \mathbb{R}^{d \times h_s \times w_s}$ denote the corresponding stage-specific input feature and learned multi-scale structural support, respectively. The CFC module then produces the retention-calibrated feature as

$$\tilde{F}_s = C_s(F_s, S_s). \tag{1}$$

C_s denotes the stage-s CFC mapping, and S_s is a learned, non-ground-truth reference.

Proposition 1 (Counterfactual Features under High-Frequency Interventions). We decompose the stage feature F_s into low- and high-frequency components as $F_s = L_s + H_s$, where L_s is obtained by Gaussian low-pass filtering of F_s and $H_s = F_s - L_s$. The removal intervention eliminates H_s while preserving L_s, yielding $F_s^{\mathrm{rem}} = L_s$. For perturbation, let $H_s^{\mathrm{ch}} = \mathrm{Roll}_c(H_s; 1)$ and $H_s^{\mathrm{sp}} = \mathrm{Roll}_{h,w}(H_s; 1, -1)$. With $\alpha \in [0, 1]$, the perturbed counterfactual feature is defined as

$$F_s^{\mathrm{pert}} = L_s + H_s^{\mathrm{pert}} = L_s + (1 - \alpha)H_s + \frac{\alpha}{2}\left(H_s^{\mathrm{ch}} + H_s^{\mathrm{sp}}\right). \tag{2}$$

Both interventions preserve L_s while modifying only H_s by construction; here, *counterfactual* refers specifically to these controlled feature-level interventions and does not imply causal identification.

Proposition 2 (First-Order Approximation of Counterfactual Probe Output Differences). Let P_s denote the shared probe at stage s, with identical parameters for original and counterfactual features, and let $U_s^b := F_s - F_s^b$ for $b \in \{\mathrm{rem}, \mathrm{pert}\}$. Assuming that P_s is continuously differentiable along the segment between F_s and F_s^b, the signed probe difference has the first-order expansion as

$$P_s(F_s) - P_s(F_s^b) = J_{P_s}(F_s)[U_s^b] + \epsilon_s^b. \tag{3}$$

$J_{P_s}(F_s)$ denotes the shared-probe Jacobian, and ϵ_s^b denotes the corresponding first-order remainder.

Lemma 1 (Quadratic Bound on the First-Order Remainder). If P_s is twice continuously differentiable along the corresponding line segment in feature space, then there exists a finite constant $M_s \geq 0$ such that the first-order remainder satisfies the following quadratic norm bound.

$$|\epsilon_s^b|_p \leq \frac{M_s}{2}|U_s^b|_p^2. \tag{4}$$

We define the intervention-response map D_s by averaging absolute probe-output differences induced by removal and perturbation interventions, summarizing sensitivity to high-frequency changes as

$$D_s = \frac{1}{2}\sum_{b\in\{\mathrm{rem},\mathrm{pert}\}} \left|P_s(F_s) - P_s(F_s^b)\right|. \tag{5}$$

D_s captures the response to controlled high-frequency changes at each spatial feature location and is combined with S_s to construct the retention score. For the CFR objective, prediction errors are candidate suppression regions weighted by detached high-frequency, intervention-response, and structural-support evidence W_s^{hf}, W_s^{int}, and W_s^{sup}, yielding the error-suppression weight as

$$Q_s^{\mathrm{err}} = \mathrm{sg}\big(\mathbf{1}_{\Omega_s^{\mathrm{err}}} W_s^{\mathrm{hf}} W_s^{\mathrm{int}} (1 - W_s^{\mathrm{sup}}) M_s^{\mathrm{err}}\big). \tag{6}$$

The wrong-class confidence margin is $M_s^{\mathrm{err}}(x) = [p_s(\hat{y}_s(x) \mid x) - p_s(y_s(x) \mid x)]_+$, where p_s is the stage-s prediction probability and $\hat{y}_s(x)$ and $y_s(x)$ are the predicted and ground-truth classes. W_s^{hf}, W_s^{int}, and W_s^{sup} are normalized evidence maps from high-frequency magnitude, the intervention-response map D_s, and structural support S_s, respectively. Thus, prediction error alone does not activate suppression, which depends on high-frequency and intervention evidence, structural support, and prediction confidence. The preservation weight for correct predictions is defined as

$$Q_s^{\mathrm{cor}} = \mathrm{sg}\big(\mathbf{1}_{\Omega_s^{\mathrm{cor}}} W_s^{\mathrm{hf}} W_s^{\mathrm{int}} W_s^{\mathrm{sup}}\big). \tag{7}$$

The boundary weight restricts Q_s^{cor} specifically to valid semantic-boundary locations as $Q_s^{\mathrm{bd}} = \mathbf{1}_{\Omega_s^{\mathrm{bd}}} Q_s^{\mathrm{cor}}$. For any non-negative weighting map Q, the weighted spatial mean is defined as

$$\langle A\rangle_Q = \frac{\sum_x Q(x)A(x)}{\max(1, \sum_x Q(x))}. \tag{8}$$

With $\bar{R}_s(x) = c_s^{-1}\sum_{c=1}^{c_s} R_s(c,x)$, the corresponding per-stage CFR training objective is defined as

$$\mathcal{L}_{\mathrm{CFR},s} = w_{\mathrm{err}}\langle \bar{R}_s\rangle_{Q_s^{\mathrm{err}}} + w_{\mathrm{cor}}\langle[\rho_{\mathrm{cor}} - \bar{R}_s]_+\rangle_{Q_s^{\mathrm{cor}}} + w_{\mathrm{bd}}\langle[\rho_{\mathrm{bd}} - \bar{R}_s]_+\rangle_{Q_s^{\mathrm{bd}}}. \tag{9}$$

The CFR objective suppresses retention at weighted errors while imposing floors on structurally supported correct and boundary regions. Without CFC, local high-frequency derivative is given as

$$\Psi'_{s,x}(0) = -\left\langle \nabla_{F_s(:,x)}\mathcal{L}_{\mathrm{CE},s}, H_s(:,x)\right\rangle. \tag{10}$$

Prediction error at each spatial location does not constrain the alignment between $H_s(:,x)$ and the cross-entropy feature gradient, so it alone cannot determine the sign of $\Psi'_{s,x}(0)$. Using $F_s = L_s + H_s$, the retention-calibrated feature under the CFC mapping is rewritten as

$$\tilde{F}_s = F_s + \gamma_s(L_s + R_s \odot H_s - F_s) = L_s + [1 - \gamma_s(1 - R_s)] \odot H_s. \tag{11}$$

The channel-averaged high-frequency retention coefficient at each location is defined as

$$\bar{\beta}_s(x) = 1 - \gamma_s\left(1 - \bar{R}_s(x)\right). \tag{12}$$

For $\bar{R}_s(x) \in [0,1]$, $\bar{\beta}_s(x) \in [1-\gamma_s, 1]$ and increases with $\bar{R}_s(x)$ since $\partial\bar{\beta}_s(x)/\partial\bar{R}_s(x) = \gamma_s > 0$. We analyze the CFR-induced local direction under the stated active-region conditions using $\bar{R}_s^+(x) = \bar{R}_s(x) - \eta_s g_s(x)$, where $g_s(x)$ is a pointwise subgradient and $\eta_s > 0$ is sufficiently small.

Lemma 2 (Error-Suppression Retention Direction). Let $Z_s^{\mathrm{err}} = \max(1, \sum_x Q_s^{\mathrm{err}}(x))$. For $Q_s^{\mathrm{err}}(x) > 0$, the error-suppression term yields gradient $g_s(x) = w_{\mathrm{err}} Q_s^{\mathrm{err}}(x)/Z_s^{\mathrm{err}} > 0$ for $\bar{R}_s(x)$. Under the output-space step, this gradient decreases the high-frequency retention coefficient as

$$\Delta\bar{\beta}_s(x) = -\eta_s\gamma_s w_{\mathrm{err}}\frac{Q_s^{\mathrm{err}}(x)}{Z_s^{\mathrm{err}}} < 0. \tag{13}$$

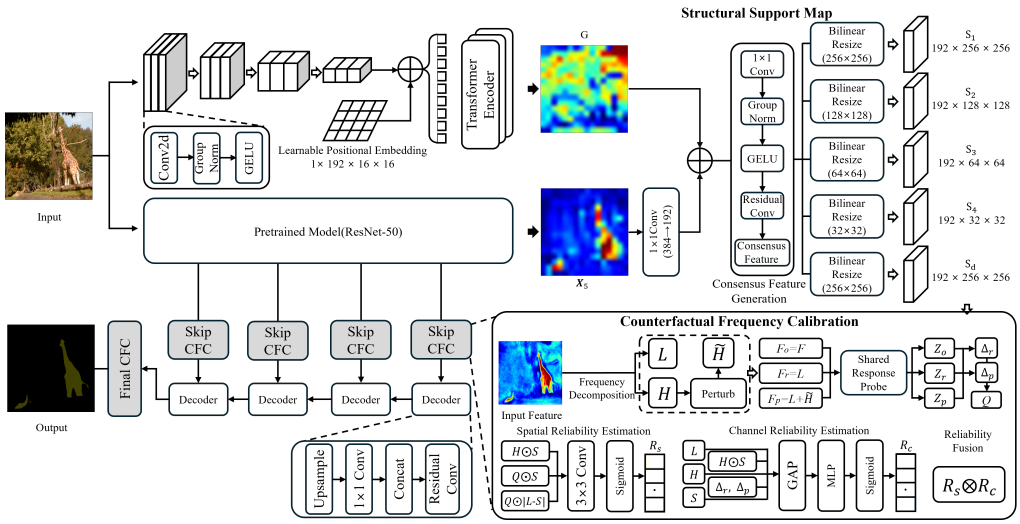


Figure 2: Overall architecture of CFRNet.

Lemma 3 (Correct-Prediction Retention Direction Below the Retention-Score Floor). Let $Z_s^{\mathrm{cor}} = \max(1, \sum_x Q_s^{\mathrm{cor}}(x))$. For $Q_s^{\mathrm{cor}}(x) > 0$ and $\bar{R}_s(x) < \rho_{\mathrm{cor}}$, the correct-prediction term yields $g_s(x) = -w_{\mathrm{cor}} Q_s^{\mathrm{cor}}(x)/Z_s^{\mathrm{cor}} < 0$. An active boundary term at the location also adds a nonpositive gradient, so the output-space step increases the high-frequency retention coefficient as

$$\Delta\bar{\beta}_s(x) \geq \eta_s \gamma_s w_{\mathrm{cor}} \frac{Q_s^{\mathrm{cor}}(x)}{Z_s^{\mathrm{cor}}} > 0. \tag{14}$$

Theorem 1 (Local Retention Directions in the CFC-Free and CFR-Regularized Cases). For x_{err}, prediction error alone does not determine the sign of $\Psi'_{s,x_{\mathrm{err}}}(0)$. Under Lemmas 2 and 3, if $Q_s^{\mathrm{err}}(x_{\mathrm{err}}) > 0$, $Q_s^{\mathrm{cor}}(x_{\mathrm{cor}}) > 0$, and $\bar{R}_s(x_{\mathrm{cor}}) < \rho_{\mathrm{cor}}$, the CFR-induced retention directions are

$$\Delta\bar{\beta}_s(x_{\mathrm{err}}) = -\eta_s \gamma_s w_{\mathrm{err}} \frac{Q_s^{\mathrm{err}}(x_{\mathrm{err}})}{Z_s^{\mathrm{err}}} < 0, \qquad \Delta\bar{\beta}_s(x_{\mathrm{cor}}) \geq \eta_s \gamma_s w_{\mathrm{cor}} \frac{Q_s^{\mathrm{cor}}(x_{\mathrm{cor}})}{Z_s^{\mathrm{cor}}} > 0. \tag{15}$$

CFR imposes conditional local retention directions under the stated active-region conditions, absent from the CFC-free error condition. These output-space directions do not imply calibrated high-frequency utility, exact optimizer-step behavior, prediction-error reduction, or global convergence.

3.3 Architecture

Figure 2 illustrates the overall CFRNet architecture. Multi-scale CNN and global Transformer features jointly form stage-specific structural support across the five calibration stages, while the CFC module combines this support with intervention-induced probe differences to selectively modulate local high-frequency information in the skip and final decoder features during feature decoding.

A CNN encoder and lightweight Transformer branch extract multi-scale local features and global structural context across progressively coarser spatial resolutions. The Transformer feature is fused with the deepest CNN feature, then resized and refined across calibration scales throughout the decoder to form the stage-specific structural support S_s. Rather than directly predicting segmentation, S_s provides structural context at each stage for evaluating local high-frequency responses.

At each stage, the CFC module receives F_s and S_s at the same spatial resolution. Gaussian low-pass decomposition separates F_s into L_s and H_s, after which the removal and perturbation interventions modify only H_s. A shared probe compares the original and intervened features for both interventions to obtain the intervention-induced probe-output differences for retention estimation.

The CFC module combines probe-output differences with S_s to construct spatial and channel-wise retention signals for each calibration stage. The resulting score R_s modulates the high-frequency retention coefficient at each location to produce $\tilde{F}_s$, assigning adaptive retention strengths rather than preserving or attenuating high-frequency responses at individual feature locations.

Table 2: Overall comparison on PASCAL VOC 2012, COCO-Stuff 164K, and ADE20K.

DATASET	METHOD	OVERALL		Region-wise Segmentation		Region-wise Error	
		mIoU	mAcc	B-IoU	I-IoU	$\mathrm{Err}_{\mathrm{bd}}$	$\mathrm{Err}_{\mathrm{nonbd}}$
VOC	FADC	$\mathbf{0.703 \pm 0.005}$	0.792	0.605	**0.720**	0.188	0.073
	DAF+FreqMix	0.236 ± 0.013	0.320	0.159	0.247	0.474	0.184
	FDConv	0.522 ± 0.011	0.582	0.328	0.410	0.394	0.155
	SFM	0.689 ± 0.018	**0.797**	**0.616**	0.653	0.198	0.089
	FreqFusion	0.615 ± 0.008	0.695	0.532	0.621	0.228	0.075
	CFRNet	0.701 ± 0.009	**0.797**	0.607	0.686	**0.185**	**0.065**
COCO-Stuff	FADC	0.314 ± 0.007	0.453	0.238	0.385	0.446	0.285
	DAF+FreqMix	0.249 ± 0.005	0.370	0.147	0.264	0.612	0.381
	FDConv	0.339 ± 0.015	0.411	0.333	0.367	0.450	0.280
	SFM	0.388 ± 0.006	0.537	0.286	0.404	0.491	0.256
	FreqFusion	0.410 ± 0.022	0.519	0.239	0.425	0.539	0.292
	CFRNet	$\mathbf{0.468 \pm 0.018}$	**0.623**	**0.343**	**0.491**	**0.396**	**0.103**
ADE20K	FADC	0.362 ± 0.004	0.510	0.267	0.392	0.436	0.186
	DAF+FreqMix	0.320 ± 0.031	0.435	0.248	0.325	0.438	0.191
	FDConv	0.332 ± 0.053	0.448	0.254	0.344	0.421	0.189
	SFM	0.315 ± 0.007	0.425	0.242	0.324	0.423	0.208
	FreqFusion	0.374 ± 0.002	0.493	0.275	0.414	0.445	0.182
	CFRNet	$\mathbf{0.390 \pm 0.008}$	**0.518**	**0.299**	**0.415**	**0.391**	**0.162**

Table 3: Model complexity comparison on PASCAL VOC 2012.

Method	Params (M)	Backbone MACs (G)	Head MACs (G)	Total MACs (G)	Head MAC Share (%)
DAF+FreqMix	31.541	23.512	14.604	38.115	38.3
FADC	67.343	24.289	214.235	238.524	89.8
FDConv	70.045	24.171	214.235	238.406	89.9
FreqFusion	31.343	22.730	13.811	36.541	37.8
SFM	44.452	21.527	38.460	59.988	64.1
CFRNet	35.577	23.806	17.865	41.671	42.9

4 Experimental results

4.1 Experimental settings

We evaluate CFRNet on PASCAL VOC 2012 (Everingham et al., 2010), COCO-Stuff 164K (Caesar et al., 2018), and ADE20K (Zhou et al., 2017). DAF+FreqMix (Chen et al., 2024b), FDConv (Chen et al., 2025c), FADC (Chen et al., 2024c), SFM (Chen et al., 2025a), and FreqFusion (Chen et al., 2024a) serve as frequency-aware counterparts under a unified protocol. All methods use a ResNet-50 encoder, an 80K-iteration schedule, the same input resolution per dataset, and method-specific decoders and frequency-aware components. On PASCAL VOC 2012, all methods are retrained on the original 1,464-image split without SBD augmentation. Since the original DAF+FreqMix setting uses 10,582 augmented images, its reported VOC result is not directly comparable to our unified-protocol result.

We report mIoU and mAcc for overall performance and B-IoU, I-IoU, and boundary/non-boundary error rates for region-wise evaluation. Boundary regions are defined by a 3-pixel dilation of ground-truth semantic boundaries, with the remaining valid pixels treated as interior regions. B-IoU evaluates segmentation overlap near semantic boundaries, whereas I-IoU evaluates overlap within object interiors; the corresponding error rates measure pixel-wise errors in boundary and non-boundary regions. mIoU is reported as mean $\pm$ standard deviation over three independent runs. Additional settings are provided in Appendix C.

4.2 Comparison Results and Analysis

Table 2 compares CFRNet with existing frequency-aware segmentation methods across the three benchmarks under the unified ResNet-50 setting. On PASCAL VOC 2012, CFRNet achieves 0.701

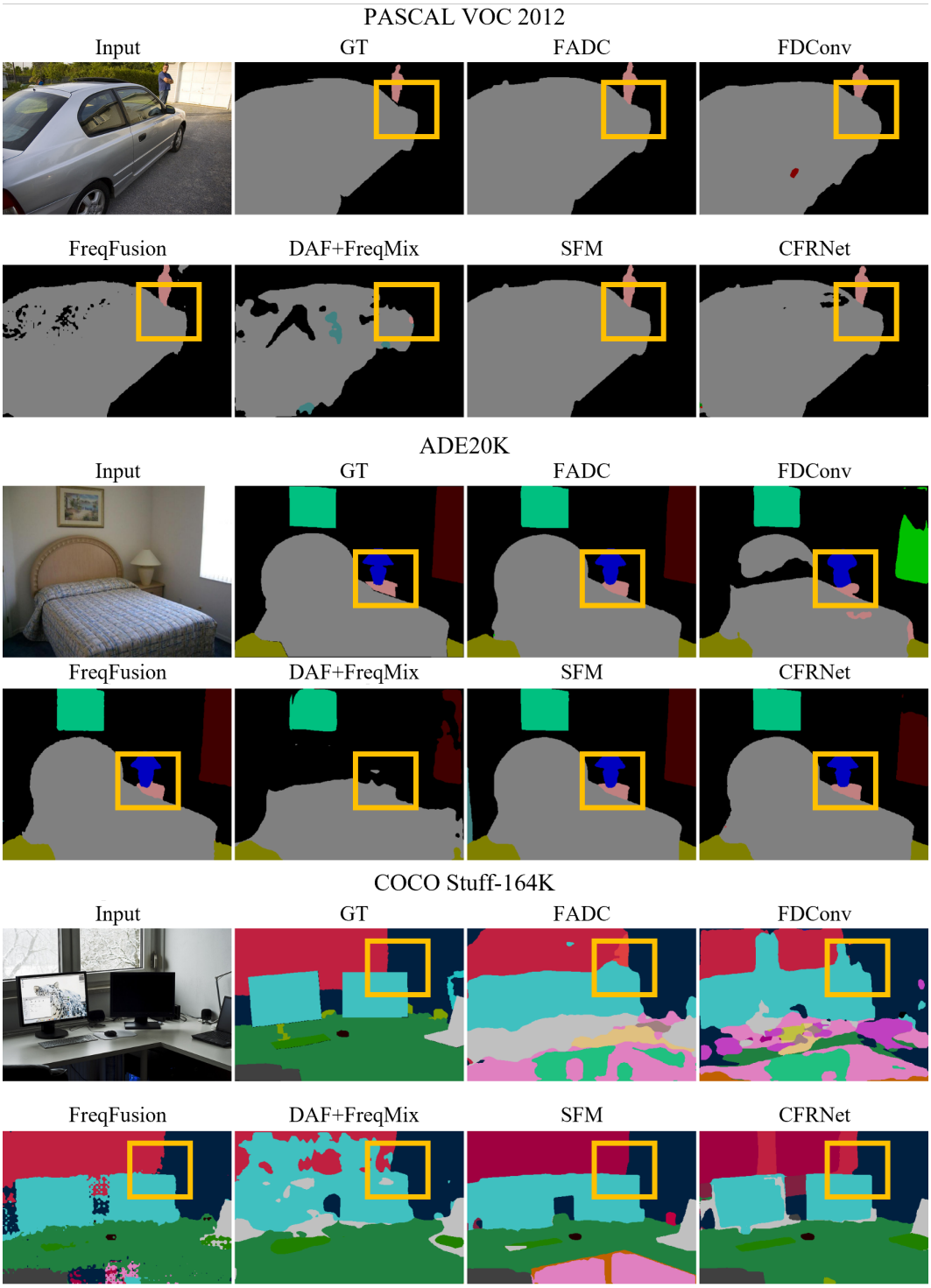
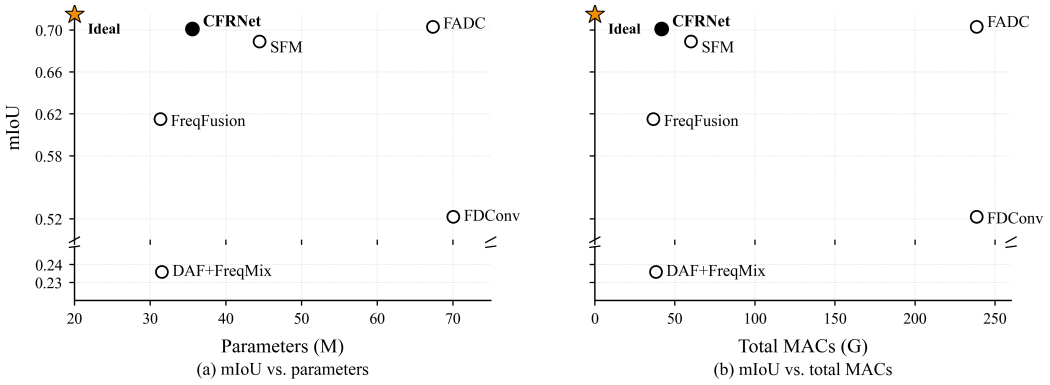


Figure 3: Qualitative comparison on PASCAL VOC 2012, ADE20K, and COCO-Stuff 164K. Orange boxes highlight representative regions where CFRNet selectively retains structurally supported high-frequency cues, such as semantic boundaries, while attenuating unsupported responses associated with spurious predictions.



(a) mIoU vs. parameters (b) mIoU vs. total MACs

Figure 4: Performance–complexity trade-off on PASCAL VOC 2012.

Table 4: Diagnostic analysis of the CFR mechanism on VOC 2012

Analysis	Region / Stage	Metric	Value
No-CFC direction	Error / Stage 1	$\Psi'_{s,x}(0) < 0 \,/\, > 0$	50.37 / 49.63
	Error / Stage 2	$\Psi'_{s,x}(0) < 0 \,/\, > 0$	49.14 / 50.86
	Error / Stage 3	$\Psi'_{s,x}(0) < 0 \,/\, > 0$	51.65 / 48.35
	Error / Stage 4	$\Psi'_{s,x}(0) < 0 \,/\, > 0$	53.70 / 46.30
	Error / Average	$\Psi'_{s,x}(0) < 0 \,/\, > 0$	51.21 / 48.79
CFR direction	Prediction error	$\Delta\bar{\beta}$ (after − before)	-0.0195
	Correct prediction below floor	$\Delta\bar{\beta}$ (after − before)	$+0.0059$
	Correct boundary below floor	$\Delta\bar{\beta}$ (after − before)	$+0.0348$
Retention allocation	Prediction error	Mean $\bar{R}$	0.1041
	Correct non-boundary	Mean $\bar{R}$	0.0895
	Correct boundary	Mean $\bar{R}$	0.1890

mIoU, within 0.002 of the highest result, while matching the highest mAcc and yielding the lowest boundary and non-boundary error rates. On COCO-Stuff 164K and ADE20K, CFRNet achieves the highest mIoUs of 0.468 and 0.390, outperforming the strongest counterparts by 0.058 and 0.016, respectively, while attaining the best region-wise metrics. Figure 3 complements these results with representative predictions illustrating object consistency, boundary delineation, and spurious errors. Under the same COCO-Stuff 164K evaluation protocol, replacing ResNet-50 with ConvNeXt-T increases mIoU from 0.468 to 0.568 and reduces the boundary and non-boundary error rates from 0.396/0.103 to 0.360/0.081, providing additional evidence that CFRNet can operate with a different backbone architecture.

Table 3 reports parameter counts and inference-path MACs, separating backbone and segmentation-head computation while including method-specific operations. Figure 4 further illustrates the accuracy–complexity trade-off on PASCAL VOC 2012, where points closer to the upper-left correspond to higher mIoU with lower model complexity. Although FADC achieves a marginally higher mIoU than CFRNet (0.703 vs. 0.701), it requires substantially greater computation (238.524G vs. 41.671G MACs), while CFRNet maintains competitive segmentation accuracy in a more favorable efficiency region.

Table 4 is organized into three diagnostics. The No-CFC block reports the percentages of prediction-error locations with negative and positive $\Psi'_{s,x}(0)$ across the four skip-CFC stages. The CFR block reports $\Delta\bar{\beta}$ after the analytical CFR step, where negative and positive values indicate decreased and

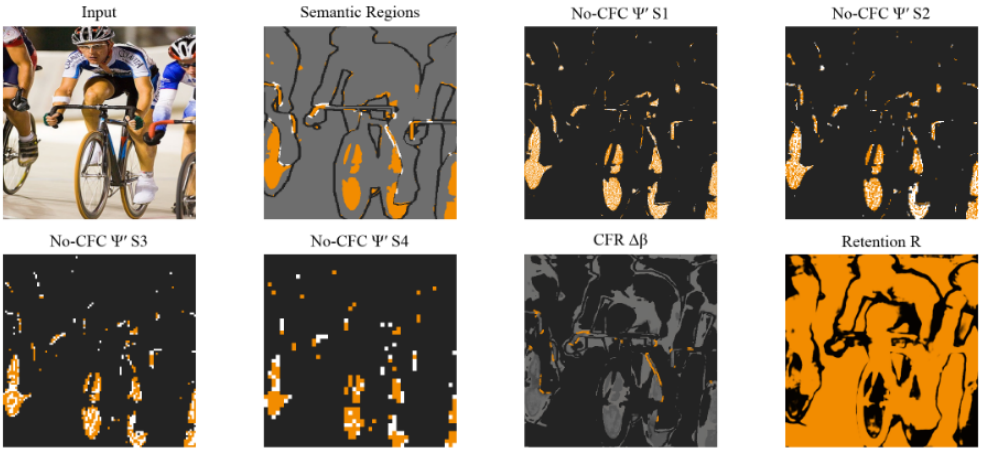


Figure 5: Spatial diagnostics of CFR on PASCAL VOC 2012 validation images.

Table 5: Ablation study of the main components of CFRNet on COCO-Stuff 164K

SETTING	SS	REM	PERT	S+C	CFR OBJ	mIoU
Baseline	–	–	–	–	–	0.398
w/o structural support	–	✓	✓	✓	✓	0.418
Removal only	✓	✓	–	✓	✓	0.438
Perturbation only	✓	–	✓	✓	✓	0.434
Spatial only	✓	✓	✓	S	✓	0.430
Channel only	✓	✓	✓	C	✓	0.436
w/o CFR objective	✓	✓	✓	S+C	–	0.443
Full CFRNet	✓	✓	✓	S+C	✓	0.468

increased high-frequency retention, respectively. The final block reports the mean learned retention score $\bar{R}$, with larger values indicating stronger retention. Figure 5 provides the spatial counterpart: orange/white denote negative/positive no-CFC directions, orange/dark responses in $\Delta\bar{\beta}$ indicate increased/decreased retention, and stronger orange in the retention map indicates larger $\bar{R}$. Thus, Table 4 summarizes the aggregate behavior, while Figure 5 shows its spatial distribution.

4.3 Ablation Study

Table 5 shows that removing any major component reduces performance. The CFC architecture improves the CFR-free baseline from 0.398 to 0.443 mIoU, and the CFR objective further increases it to 0.468. As all variants use a fixed random seed, these results indicate relative trends rather than statistical significance. An additional architectural control is provided in Appendix B to examine whether the performance gain can be attributed solely to the Transformer-derived structural context.

5 Conclusion

In this study, we proposed CFRNet, a counterfactual frequency-guided framework that selectively modulates high-frequency retention using intervention responses and multi-scale structural support. Theoretical analysis showed that prediction errors alone do not determine a consistent attenuation direction, whereas the CFR objective induces conditional retention changes. CFRNet achieves mIoUs of 0.701, 0.468, and 0.390 on PASCAL VOC 2012, COCO-Stuff 164K, and ADE20K, respectively, remaining within 0.002 of the highest VOC result and outperforming the strongest counterparts by 0.058 and 0.016 on COCO-Stuff and ADE20K. Ablation and diagnostic results support the proposed components and retention behavior. The analysis remains limited to local directional properties and does not establish calibrated high-frequency utility or global optimization guarantees.

REPRODUCIBILITY STATEMENT

Source code, configuration files, and training instructions are available at `https://anonymous.4open.science/r/CFRNet__Counterfactual_Frequency_Guided_Selective_Retention_for_Semantic_Segmentation-13C4/`. Additional implementation details, experimental settings, and theoretical derivations are provided in the appendix.

AI USE STATEMENT

Generative AI tools were used to assist with feedback on research methodology and experimental design, implementation and code debugging, mathematical derivations and proofs, interpretation of experimental results, and manuscript drafting and editing. All AI-assisted outputs were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including all claims, results, proofs, code, and artifacts produced with the aid of generative AI.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1209–1218, 2018.

Hengzhi Chen, Liqian Feng, Wenhua Wu, Xiaogang Zhu, Qiuxia Wu, Lianlei Shan, and Kun Hu. F2Net: A frequency-fused network for ultra-high resolution remote sensing segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13275–13284, 2026.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10763–10780, 2024a. doi: 10.1109/TPAMI.2024.3449959.

Linwei Chen, Lin Gu, and Ying Fu. When semantic segmentation meets frequency aliasing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Linwei Chen, Lin Gu, Dezhi Zheng, and Ying Fu. Frequency-adaptive dilated convolution for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3414–3425, 2024c. doi: 10.1109/CVPR52733.2024.00328.

Linwei Chen, Ying Fu, Lin Gu, Dezhi Zheng, and Jifeng Dai. Spatial frequency modulation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (11):9767–9784, 2025a. doi: 10.1109/TPAMI.2025.3592621.

Linwei Chen, Lin Gu, and Ying Fu. Frequency-dynamic attention modulation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22620–22632, 2025b.

Linwei Chen, Lin Gu, Liang Li, Chenggang Yan, and Ying Fu. Frequency dynamic convolution for dense image prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30178–30188, 2025c.

Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3435–3444, 2019.

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 (CVPR)*, pp. 1290–1299, 2022.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9799–9808, 2020.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations (ICLR)*, 2020.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

Yu Bing Luo, Nian Shi, Jia Qin, Zekai Ji, Pinle Qin, Jianchao Zeng, and Jianghui Cai. Revisiting downsampling in semantic segmentation: Fighting aliasing with dynamic gaussian and gabor frequency filters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7707–7715, 2026. doi: 10.1609/aaai.v40i9.37713.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. doi: 10.1109/TPAMI.2021.3059968.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 783–792, 2021. doi: 10.1109/ICCV48922.2021.00082.

Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5229–5238, 2019. doi: 10.1109/ICCV.2019.00533.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. doi: 10.1109/TPAMI.2020.2983686.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Ziqian Yang, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, and Jimin Xiao. FFR: Frequency feature rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30261–30270, 2025. doi: 10.1109/CVPR52734.2025.02817.

Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient centralization: A new optimization technique for deep neural networks. In *Computer Vision – ECCV 2020*, pp. 635–652, 2020. doi: 10.1007/978-3-030-58452-8_37.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 325–341, 2018.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision – ECCV 2020*, pp. 173–190, 2020. doi: 10.1007/978-3-030-58539-6_11.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E. Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7324–7334. PMLR, 2019.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6881–6890, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. doi: 10.5244/C.34.134.

A Proofs

A.1 Proof of Proposition 1: Counterfactual Features under High-Frequency Interventions

Removing the high-frequency component from $\mathbf{F}_s$ gives

$$\mathbf{F}_s^{\mathrm{rem}} = \mathbf{L}_s. \tag{16}$$

The channel-rolled high-frequency component is

$$\mathbf{H}_s^{\mathrm{ch}} = \mathrm{Roll}_c(\mathbf{H}_s; 1). \tag{17}$$

The spatially rolled high-frequency component is

$$\mathbf{H}_s^{\mathrm{sp}} = \mathrm{Roll}_{h,w}(\mathbf{H}_s; 1, -1). \tag{18}$$

Averaging the rolled components and mixing them with the original high-frequency component gives

$$\mathbf{H}_s^{\mathrm{pert}} = (1-\alpha)\mathbf{H}_s + \frac{\alpha}{2}\left(\mathbf{H}_s^{\mathrm{ch}} + \mathbf{H}_s^{\mathrm{sp}}\right). \tag{19}$$

Adding the preserved low-frequency component $\mathbf{L}_s$ yields Equation 2. □

A.2 Proof of Proposition 2: First-Order Approximation of Counterfactual Probe Output Differences

By Proposition 1, the removal displacement is

$$\mathbf{U}_s^{\mathrm{rem}} = \mathbf{H}_s. \tag{20}$$

The perturbation displacement is

$$\mathbf{U}_s^{\mathrm{pert}} = \alpha\left[\mathbf{H}_s - \frac{1}{2}\left(\mathbf{H}_s^{\mathrm{ch}} + \mathbf{H}_s^{\mathrm{sp}}\right)\right]. \tag{21}$$

For either intervention b, applying the fundamental theorem of calculus along the path from $\mathbf{F}_s$ to $\mathbf{F}_s^b$ gives

$$\mathcal{P}_s(\mathbf{F}_s) - \mathcal{P}_s(\mathbf{F}_s^b) = \int_0^1 J_{\mathcal{P}_s}\left(\mathbf{F}_s - t\mathbf{U}_s^b\right)\left[\mathbf{U}_s^b\right] dt. \tag{22}$$

Define the remainder as

$$\boldsymbol{\epsilon}_s^b := \int_0^1 \left(J_{\mathcal{P}_s}\left(\mathbf{F}_s - t\mathbf{U}_s^b\right) - J_{\mathcal{P}_s}(\mathbf{F}_s)\right)[\mathbf{U}_s^b]\, dt. \tag{23}$$

It follows that

$$\mathcal{P}_s(\mathbf{F}_s) - \mathcal{P}_s(\mathbf{F}_s^b) = J_{\mathcal{P}_s}(\mathbf{F}_s)[\mathbf{U}_s^b] + \boldsymbol{\epsilon}_s^b, \tag{24}$$

which is Equation 3.

The implementation uses the element-wise absolute probe-output difference. Because

$$\| |\mathbf{A}| \|_p = \|\mathbf{A}\|_p, \tag{25}$$

the same norm bounds apply to the implemented probe-output difference. □

A.3 Proof of Lemma 1: Quadratic Bound on the First-Order Remainder

For fixed probe-output difference parameters, the line segment connecting $\mathbf{F}_s$ and $\mathbf{F}_s - \mathbf{U}_s^b$ is compact. Because $\mathcal{P}_s$ is twice continuously differentiable, its second derivative is bounded on this segment.

Define

$$M_s := \sup_{t\in[0,1]} \left\| D^2\mathcal{P}_s\left(\mathbf{F}_s - t\mathbf{U}_s^b\right)\right\|_{\mathrm{op},p}. \tag{26}$$

The constant M_s is finite.

The integral form of the first-order remainder gives

$$\epsilon_s^b = -\int_0^1 (1-t) D^2 P_s\big(F_s - tU_s^b\big)\,[U_s^b, U_s^b]\, dt. \tag{27}$$

Thus,

$$\|\epsilon_s^b\|_p \le \int_0^1 (1-t) M_s \|U_s^b\|_p^2\, dt = \frac{M_s}{2}\|U_s^b\|_p^2. \tag{28}$$

This is Equation 4. □

A.4 Proof of Lemma 2: Error-Suppression Retention Direction

For $Q_s^{\mathrm{err}}(x) > 0$, the correct-prediction and boundary terms vanish. Thus, the pointwise CFR gradient is

$$g_s(x) = w_{\mathrm{err}} \frac{Q_s^{\mathrm{err}}(x)}{Z_s^{\mathrm{err}}}. \tag{29}$$

Because $w_{\mathrm{err}} > 0$ and $Q_s^{\mathrm{err}}(x) > 0$, this gradient is strictly positive. Substituting it into the output-space update gives

$$\overline{R}_s^{+}(x) - \overline{R}_s(x) = -\eta_s w_{\mathrm{err}} \frac{Q_s^{\mathrm{err}}(x)}{Z_s^{\mathrm{err}}}. \tag{30}$$

Using the definition of $\Delta\overline{\beta}_s(x)$ yields Equation 13. □

A.5 Proof of Lemma 3: Correct-Prediction Retention Direction Below the Retention-Score Floor

For $Q_s^{\mathrm{cor}}(x) > 0$ and $\overline{R}_s(x) < \rho_{\mathrm{cor}}$, the error-suppression term vanishes and the correct-prediction hinge term is active. Its pointwise gradient is

$$-w_{\mathrm{cor}} \frac{Q_s^{\mathrm{cor}}(x)}{Z_s^{\mathrm{cor}}}. \tag{31}$$

Table 6: Architectural context control on COCO-Stuff 164K

SETTING	SC	CFC	CFR OBJ.	mIoU
No-CFR baseline	–	–	–	0.398
Structural-context-only	✓	–	–	0.403
w/o CFR objective	✓	✓	–	0.443
Full CFRNet	✓	✓	✓	0.468

Any active boundary term contributes an additional nonpositive gradient. Thus,

$$g_s(x) \le -w_{\mathrm{cor}} \frac{Q_s^{\mathrm{cor}}(x)}{Z_s^{\mathrm{cor}}} < 0. \tag{32}$$

The output-space update consequently satisfies

$$\overline{R}_s^{+}(x) - \overline{R}_s(x) \ge \eta_s w_{\mathrm{cor}} \frac{Q_s^{\mathrm{cor}}(x)}{Z_s^{\mathrm{cor}}} > 0. \tag{33}$$

Using the definition of $\Delta\overline{\beta}_s(x)$ yields Equation 14. □

A.6 Proof of Theorem 1: Local Retention Directions in the CFC-Free and CFR-Regularized Cases

For the CFC-free case, the preceding analysis gives

$$\Psi'_{s,x}(0) = -\left\langle \nabla_{F_s(:,x)} \mathcal{L}_{\mathrm{CE},s}, H_s(:,x) \right\rangle. \tag{34}$$

The prediction-error condition constrains the output prediction relative to the ground-truth class but does not constrain the alignment between $\nabla_{F_s(:,x)}\mathcal{L}_{\mathrm{CE},s}$ and $H_s(:,x)$. Thus, no sign of $\Psi'_{s,x}(0)$ follows from the prediction-error condition alone. For the CFR-regularized case, the negative retention-coefficient change at x_{err} follows from Lemma 2, whereas the positive change at x_{cor} follows from Lemma 3. □

Thus, the two cases differ in the available directional property: the CFC-free cross-entropy objective has no prediction-error-induced sign determination for local high-frequency attenuation, whereas the CFR objective provides explicit retention directions under the stated active conditions.

B Additional Architectural Control

To examine whether the performance improvement of CFRNet can be explained solely by the additional global structural context provided by the Transformer branch, we introduce a structural-context-only control. This variant retains the Transformer-based structural branch and incorporates its representation through a simple non-frequency-aware fusion, while disabling the controlled high-frequency interventions, probe-output difference computation, retention-score estimation, high-frequency calibration, and the CFR objective.

This control differs from the *w/o structural support* variant in Table 5. The latter evaluates the role of structural evidence within the CFR mechanism, whereas the structural-context-only control evaluates whether the additional Transformer-derived context alone accounts for the performance gain.

The structural-context-only variant achieves 0.403 mIoU, compared with 0.443 for the CFC architecture without the CFR objective and 0.468 for full CFRNet. This gap indicates that the improvement cannot be explained solely by the additional Transformer-derived structural context.

C Implementation Details

C.1 Counterpart Configurations

For all comparative experiments, the CNN encoder was standardized to ResNet-50, the input resolution was fixed to 512×512, and the training schedule was standardized to 80K iterations. The

Table 7: Implementation settings of CFRNet

SETTING	VALUE
Transformer dimension / depth / heads	192 / 3 / 6
Optimizer	Ranger
Initial learning rate	0.003
Weight decay	5×10^{-4}
Frequency cutoff initialization	0.18
Calibration coefficient γ_s initialization	0.1
Intervention strength α	0.5
CFR loss weight	0.1
Retention-score floors $(\rho_{\mathrm{cor}}, \rho_{\mathrm{bd}})$	0.50 / 0.55
CFR weights $(w_{\mathrm{err}}, w_{\mathrm{cor}}, w_{\mathrm{bd}})$	0.003 / 0.001 / 0.005

original decoder structure and method-specific frequency-aware components of each counterpart were retained. The remaining method-specific training configurations followed the corresponding implementations.

C.2 CFRNet Configuration

For CFRNet, the Transformer branch was configured with an embedding dimension of 192, a depth of 3, and 6 attention heads. The model was trained using the Ranger optimizer (Liu et al., 2020; Zhang et al., 2019; Yong et al., 2020) with an initial learning rate of 0.003 and a weight decay of 5×10^{-4}. Random scaling was applied in the range of 0.5 to 2.0. The frequency cutoff and the stage-wise calibration coefficient γ_s were initialized to 0.18 and 0.1, respectively, and optimized jointly with the network parameters. The high-frequency intervention strength α was set to 0.5, and the CFR loss weight was set to 0.1. The retention-score floors ρ_{cor} and ρ_{bd} were set to 0.50 and 0.55, respectively.