

# Supplementary Material

## A Notations

The notations used throughout this article are summarized in Table 1.

## B Theory and Proofs

### B.1 Proof of Lemma 1

**Lemma 1.** Consider two sub-sets  $S_a$  and  $S_b$  in set  $V$ , where  $S_a \subseteq S_b \subseteq V$ . Given an element  $\beta$ , where  $\beta = V \setminus S_b$ . The necessary and sufficient conditions for the function  $\mathcal{F}(\cdot)$  to satisfy the sub-modular property are:

$$\mathcal{F}(S_a \cup \{\beta\}) - \mathcal{F}(S_a) \geq \mathcal{F}(S_b \cup \{\beta\}) - \mathcal{F}(S_b). \quad (1)$$

**Proof.** Consider two sub-sets  $S_a$  and  $S_b$  in set  $V$ , where  $S_a \subseteq S_b \subseteq V$ . Given an element  $\beta$ , where  $\beta = V \setminus S_b$ . The necessary and sufficient conditions for the function  $\mathcal{F}(\cdot)$  to satisfy the sub-modular property are:

$$\mathcal{F}(S_a \cup \{\beta\}) - \mathcal{F}(S_a) \geq \mathcal{F}(S_b \cup \{\beta\}) - \mathcal{F}(S_b). \quad (11)$$

For Eq. 4, assuming that the individual element  $\beta$  of the collection partition is relatively small, according to the Taylor decomposition (Chen et al., 2018), we can locally approximate  $F_u(S_a + \beta) = F_u(S_a) + \nabla F_u(S_a) \cdot \beta$ . Thus:

$$\begin{aligned} s_{\text{conf}}(S_a + \beta) - s_{\text{conf}}(S_a) & \simeq \frac{K}{\exp(F_u(S_a)) + K} - \frac{K}{\exp(F_u(S_a + \beta)) + K} \end{aligned} \quad (2)$$

$$= \frac{K}{\exp(F_u(S_a)) + K} - \frac{K}{\exp(F_u(S_a) + \nabla F_u(S_a) \cdot \beta) + K} \quad (3)$$

$$= \frac{K}{\exp(F_u(S_a)) + K} - \frac{K}{\exp(F_u(S_a)) \exp(\nabla F_u(S_a) \cdot \beta) + K}, \quad (4)$$

since  $S_a \cap \beta = \emptyset$ ,  $S_b$  and  $\beta$  do not overlap in the image space, and  $\beta$  is small. Therefore, we can regard  $\nabla F_u(S_a) \cdot \beta \approx 0$ . Follow up:

$$\begin{aligned} s_{\text{conf}}(S_a + \beta) - s_{\text{conf}}(S_a) & \simeq \frac{K}{\exp(F_u(S_a)) + K} - \frac{K}{\exp(F_u(S_a)) \exp(0) + K} \\ & = 0, \end{aligned} \quad (5)$$

$$= 0, \quad (6)$$

and in the same way,  $s_{\text{conf}}(S_b + \beta) - s_{\text{conf}}(S_b) \simeq 0$ . We have:

$$s_{\text{conf}}(S_a + \beta) - s_{\text{conf}}(S_a) - (s_{\text{conf}}(S_b + \beta) - s_{\text{conf}}(S_b)) \approx 0. \quad (14)$$

For Eq. 6, when a new element  $\beta$  is added to the set  $S_a$ , the minimum distance between elements in  $S_a$  and other elements may be further reduced, i.e., for any element  $s_i \in S_a$ , we have:

$$\begin{aligned} \min_{s_j \in S_a \cup \{\beta\}, s_j \neq s_i} \text{dist}(F(s_i), F(s_j)) \\ \leq \min_{s_j \in S_a, s_j \neq s_i} \text{dist}(F(s_i), F(s_j)). \end{aligned} \quad (7)$$

Thus:

$$\begin{aligned} s_{\text{att}}(S_a \cup \{\beta\}) \\ = \min_{s_i \in S_a} \text{dist}(F(\beta), F(s_i)) \\ + \sum_{s_i \in S_a \cup \{\beta\}, s_j \neq s_i} \min \text{dist}(F(s_i), F(s_j)) \end{aligned} \quad (8)$$

$$\begin{aligned} = \min_{s_i \in S_a} \text{dist}(F(\beta), F(s_i)) \\ + \sum_{s_i \in S_a, s_j \in S_a \cup \{\beta\}, s_j \neq s_i} \min \text{dist}(F(s_i), F(s_j)) - \varepsilon_a, \end{aligned} \quad (9)$$

where  $\varepsilon_a$  is a constant, which is the sum of the minimum distance reductions of the elements in the original  $S_a$  after  $\beta$  is added. Then, we have:

$$s_{\text{att}}(S_a \cup \{\beta\}) - s_{\text{att}}(S_a) = \min_{s_i \in S_a} \text{dist}(F(\beta), F(s_i)) - \varepsilon_a, \quad (16)$$

and in the same way,

$$s_{\text{att}}(S_b \cup \{\beta\}) - s_{\text{att}}(S_b) = \min_{s_i \in S_b} \text{dist}(F(\beta), F(s_i)) - \varepsilon_b, \quad (17)$$

since  $S_a \subseteq S_b$ , the minimum distance between beta and elements in  $S_a \setminus S_b$  may be smaller than the minimum distance between beta and elements in  $S_b$ , thus,

$$\min_{s_i \in S_a} \text{dist}(F(\beta), F(s_i)) \geq \min_{s_i \in S_b} \text{dist}(F(\beta), F(s_i)), \quad (10)$$

since there are more elements in  $S_b$  than in  $S_a$ , more elements in  $S_b$  have the shortest distance from  $\beta$ , i.e.,  $\varepsilon_b \geq \varepsilon_a$ . Therefore, we have:

$$s_{\text{att}}(S_a \cup \{\beta\}) - s_{\text{att}}(S_a) \geq s_{\text{att}}(S_b \cup \{\beta\}) - s_{\text{att}}(S_b). \quad (18)$$

For Eq. 7, let  $G(S_a) = F(S_a) \cdot f_s$ . Assuming that the individual element  $\beta$  of the collection partition is relatively small, according to the Taylor decomposition, we can locally approximate  $G(S_a + \beta) = G(S_a) + \nabla G(S_a) \cdot \beta$ . Assuming that the searched  $\beta$  is valid, i.e.,  $\nabla G(S_a) > 0$ . Thus:

$$\begin{aligned} s_{\text{cons}}(S_a + \beta, f_s) - s_{\text{cons}}(S_a, f_s) \\ = \frac{G(S_a) + \nabla G(S_a) \cdot \beta}{\|F(S_a) + \nabla F(S_a) \cdot \beta\| \|f_s\|} - \frac{G(S_a)}{\|F(S_a)\| \|f_s\|} \end{aligned} \quad (11)$$

$$\simeq \frac{\nabla G(S_a) \cdot \beta}{\|F(S_a)\| \|f_s\|}, \quad (12)$$

since  $S_a \cap \beta = \emptyset$ ,  $S_b$  and  $\beta$  do not overlap in the image space, and  $\beta$  is small.  $\nabla G(S_a) \cdot \beta$  is small. Then, we have:

$$s_{\text{cons}}(S_a + \beta, f_s) - s_{\text{cons}}(S_a, f_s) - (s_{\text{cons}}(S_b + \beta, f_s) - s_{\text{cons}}(S_b, f_s)) \approx 0. \quad (20)$$

For Eq. 8, let  $G(I - S_a) = F(I - S_a) \cdot f_s$ . Assuming that the individual element  $\beta$  of the collection partition is relatively small, according to the Taylor decomposition, we can locally approximate  $G(I - S_a - \beta) = G(I - S_a) - \nabla G(I - S_a) \cdot \beta$ . Assuming that the searched alpha is valid, i.e.,  $\nabla G(I - S_a) > 0$ . Thus:

$$\begin{aligned} & s_{\text{contra}}(S_a + \beta, I, f_s) - s_{\text{contra}}(S_a, I, f_s) \\ &= \frac{G(I - S_a)}{\|F(I - S_a)\| \|f_s\|} - \frac{G(I - S_a) - \nabla G(I - S_a) \cdot \beta}{\|F(I - S_a - \beta)\| \|f_s\|} \end{aligned} \quad (13)$$

$$\simeq \frac{\nabla G(I - S_a) \cdot \beta}{\|F(I - S_a)\| \|f_s\|}, \quad (14)$$

since  $\beta$  is small,  $\nabla G(I - S_a) \cdot \beta$  is small. Then, we have:

$$s_{\text{contra}}(S_a + \beta, I, f_s) - s_{\text{contra}}(S_a, I, f_s) - (s_{\text{contra}}(S_b + \beta, I, f_s) - s_{\text{contra}}(S_b, I, f_s)) \approx 0. \quad (22)$$

Combining Eq. 14, 18, 20, and 22 we can get:

$$\mathcal{F}(S_a \cup \{\beta\}) - \mathcal{F}(S_a) \geq \mathcal{F}(S_b \cup \{\beta\}) - \mathcal{F}(S_b), \quad (23)$$

hence, we can prove that Eq. 9 is a sub-modular function.  $\square$

## B.2 Proof of Lemma 2

**Lemma 2.** Consider a subset  $S$ , given an element  $\gamma$ , assuming that  $\gamma$  is contributing to interpretation. The necessary and sufficient conditions for the function  $\mathcal{F}(\cdot)$  to satisfy the property of monotonically non-decreasing is:

$$\mathcal{F}(S \cup \{\gamma\}) - \mathcal{F}(S) > 0, \quad (15)$$

**Proof.** Consider a subset  $S$ , given an element  $\gamma$ , assuming that  $\gamma$  is contributing to interpretation. The necessary and sufficient conditions for the function  $\mathcal{F}(\cdot)$  to satisfy the property of monotonically non-decreasing is:

$$\mathcal{F}(S \cup \{\gamma\}) - \mathcal{F}(S) > 0, \quad (24)$$

where, for Eq. 4:

$$\begin{aligned} & s_{\text{conf}}(S + \gamma) - s_{\text{conf}}(S) \\ &= \frac{K}{\exp(F_u(S)) + K} - \frac{K}{\exp(F_u(S)) \exp(\nabla F_u(S) \cdot \gamma) + K}, \end{aligned} \quad (25)$$

since  $\gamma$  is contributing to interpretation,  $\nabla F_u(S) > 0$ , and  $\exp(\nabla F_u(S) \cdot \gamma) > 1$ , thus:

$$s_{\text{conf}}(S + \gamma) - s_{\text{conf}}(S) > 0. \quad (26)$$

For Eq. 6,

$$s_{\text{eff}}(S \cup \{\gamma\}) - s_{\text{eff}}(S) = \min_{s_i \in S} \text{dist}(F(\gamma), F(s_i)) - \epsilon, \quad (16)$$

since effective element  $\gamma$  are selected as much as possible, the value  $\epsilon$  will be small,

$$s_{\text{eff}}(S \cup \{\gamma\}) - s_{\text{eff}}(S) \simeq \min_{s_i \in S} \text{dist}(F(\gamma), F(s_i)) > 0. \quad (27)$$

For Eq. 7, assuming that the searched  $\gamma$  is valid,

$$s_{\text{cons}}(S + \gamma, f_s) - s_{\text{cons}}(S, f_s) \simeq \frac{\nabla G(S) \cdot \gamma}{\|F(S)\| \|f_s\|} > 0, \quad (28)$$

likewise, for Eq. 8,

$$s_{\text{colla}}(S + \gamma, I, f_s) - s_{\text{colla}}(S, I, f_s) \simeq \frac{\nabla G(I - S) \cdot \gamma}{\|F(I - S)\| \|f_s\|} > 0. \quad (29)$$

Combining Eq. 26, 27, 28, and 29 we can get:

$$\mathcal{F}(S \cup \{\gamma\}) - \mathcal{F}(S) > 0, \quad (30)$$

hence, we can prove that Eq. 9 is monotonically non-decreasing.  $\square$

## C Additional Results

The following table presents the results for a second experimental setting designed to evaluate attribution robustness on a face recognition task. In this setup, the CelebA dataset serves as the in-distribution (ID) data. To assess performance under distribution shifts, we use three distinct out-of-distribution (OOD) datasets: CelebA Transformed (a transformed-distribution OOD), VggFace2 (a related-distribution OOD), and CIFAR-100 (a complementary-distribution OOD). The table compares our proposed method against the original baseline across different partitioning strategies, using Insertion AUC ( $\uparrow$ ) and Deletion AUC ( $\downarrow$ ) as the primary evaluation metrics.

## D Ablation Study

To arrive at our proposed uncertainty-aware framework, we conducted a series of ablation studies to identify the most effective method for estimating model uncertainty under distribution shifts. This process involved a systematic evaluation of several candidate techniques, beginning with established baselines and progressively moving toward more sophisticated approaches. Our goal was to find a method that was not only effective but also lightweight, avoiding the need for extensive retraining or complex architectural changes.

### D.1 Benchmark Out-of-Distribution (OOD) Datasets

We selected a diverse suite of OOD datasets to rigorously test the limits of each uncertainty estimation technique. These datasets were chosen to represent a wide range of distribution shifts, from subtle variations to significant semantic changes:

- **Cars 196:** A fine-grained dataset of car models that tests the ability to handle subtle intra-class variations when the model is trained on a different domain (e.g., birds).
- **CUB Transformed:** An evaluation on corrupted in-distribution data, this dataset tests robustness to common image perturbations like noise and blur.
- **CIFAR-100 (50 Classes):** A complementary OOD dataset with classes semantically distant from the primary training domain, testing how methods handle completely unseen objects.
- **NA Birds:** A related-distribution OOD dataset, which is similar to CUB but contains different species and imaging conditions, testing for domain generalization.
- **Synthetic Dataset:** A controlled synthetic environment designed to isolate specific variables and provide a clean testbed for attribution faithfulness.

## D.2 Exploratory Uncertainty Estimation Techniques

Our investigation proceeded through several stages, with each technique building upon the insights gained from the last. We started with the HSIC+SMDL baseline and augmented it with the following uncertainty scores.

### D.2.1 Laplace Weight Sampling

As our initial exploration into Bayesian-inspired methods, we implemented Laplace Weight Sampling. This technique approximates the posterior distribution of the model’s weights with a Laplace distribution. At inference time, we sampled from this distribution to create a small ensemble of models. The variance in their predictions served as our uncertainty estimate, providing a principled yet computationally intensive way to capture model uncertainty.

### D.2.2 Cross-Entropy and Patch InfoNCE Uncertainty

To establish stronger baselines, we evaluated two alternative approaches. The first, **Cross-Entropy Uncertainty**, is a straightforward technique where uncertainty is derived directly from the model’s predictive entropy. The second, **Patch InfoNCE Loss**, was a more advanced attempt to learn discriminative patch-level representations by adding a contrastive loss during a fine-tuning stage. The goal was to see if better feature representations alone could improve OOD attribution, even without explicit uncertainty modeling at inference.

### D.2.3 SWAG Perturbation-Based Uncertainty

Moving towards more powerful weight-space approximations, we experimented with Stochastic Weight Averaging Gaussian (SWAG). This method captures the geometry of the loss landscape during training to form a Gaussian distribution over the weights. Perturbations are then drawn from this distribution to estimate uncertainty. While effective, this approach demonstrated that weight-space perturbations were a promising direction, ultimately leading us to develop our final, more adaptive technique.

## D.3 Ablation Results

The following tables 3, 4, 5, 6, 7 summarize the performance of each exploratory technique across our suite of OOD datasets. The results consistently show that while methods like Laplace and SWAG offered improvements over the baseline in some datasets, they were outperformed by our final adaptive perturbation approach. All results are based on Grad partitioning strategy

Table 1: Key notations used in this paper.

Notation	Description
$I$	an input image
$I^M$	a sub-region into which $I$ is partitioned
$V$	the set of all partitioned sub-regions
$S$	a subset of $V$
$\beta$	an element from $V \setminus S$
$k$	the size of the set $S$
$\mathcal{F}(\cdot)$	a function that maps a set to a value
$F_{\text{attr}}(S)$	objective function for visual attribution
$F_{\text{obj}}(S)$	objective function for object-level model interpret.
$s_{\text{conf}}(S)$	gradient-based confidence score
$s_{\text{eff}}(S)$	effectiveness score promoting diversity
$s_{\text{cons}}(S, f_s)$	consistency score ensuring relevance
$s_{\text{colla}}(S, I, f_s)$	collaboration score
$s_{\text{clue}}(S, b_{\text{target}}, c)$	clue score for object localization
$s_{\text{colla-obj}}(S, b_{\text{target}}, c)$	object-level collaboration score
$\mu_1, \mu_2, \mu_3, \mu_4$	weighting parameters
$X = \{x_i\}_{i=1}^B$	input batch with $B$ samples
$\theta_\ell$	trainable weights at layer $\ell$
$\tilde{\theta}_\ell^{(t)}$	perturbed weights at layer $\ell$ for pass $t$
$\epsilon_\ell$	Gaussian noise sampled from $\mathcal{N}(0, 1)$
$\sigma_\ell$	standard deviation of weights at layer $\ell$
$u(x)$	input-aware adaptive scaling factor
$\alpha, \beta, \gamma$	perturbation scaling parameters
$\phi_\ell(x)$	penultimate features at layer $\ell$
$\bar{\phi}$	centroid of training penultimate features
$\rho_0$	median distance threshold
$\eta_{\ell k}^{(t)}$	layer-wise grad. norm at layer $\ell$ for class $k$ at pass $t$
$d_i$	aggregated descriptor for sample $i$
$D_{\text{train}}$	training descriptors
$T$	number of stochastic forward passes
$s_i$	Mahalanobis distance score for sample $i$
$u_i$	normalized uncertainty score for sample $i$
$\Sigma$	covariance matrix of training descriptors
$\lambda$	ridge regularization parameter
$b_{\text{target}}$	target bounding box
$c$	class label
$N \times N$	grid size for image partitioning
$A$	saliency prior map

Table 2: Comparison of Insertion AUC( $\uparrow$ ) and Deletion AUC( $\downarrow$ ) scores on the CelebA (ID) dataset and its various OOD counterparts. The table compares our proposed method against the original baseline across different partitioning strategies.

Partition	Method	CelebA (ID)		CelebA Transformed		VggFace2 (OOD)		CIFAR-100 (OOD)	
		Del. ( $\downarrow$ )	Ins. ( $\uparrow$ )	Del. ( $\downarrow$ )	Ins. ( $\uparrow$ )	Del. ( $\downarrow$ )	Ins. ( $\uparrow$ )	Del. ( $\downarrow$ )	Ins. ( $\uparrow$ )
Grad	HSIC	0.0306	0.2524	0.0411	0.1530	0.0737	0.1289	0.0420	0.1299
SLICO	HSIC + SMDL	0.0099	0.3384	0.0155	0.2133	0.0015	0.2615	0.0256	0.3109
SLICO	HSIC + Ours	0.0101	<b>0.3512</b>	<b>0.0148</b>	0.1845	<b>0.0012</b>	<b>0.2833</b>	0.0256	<b>0.3493</b>
SEEDS	HSIC + SMDL	0.0060	0.3332	0.0121	0.2015	0.0085	0.2523	0.0297	0.3223
SEEDS	HSIC + Ours	<b>0.0055</b>	<b>0.3478</b>	<b>0.0115</b>	<b>0.2250</b>	<b>0.0079</b>	<b>0.2719</b>	<b>0.0286</b>	<b>0.3549</b>

Table 3: Ablation results on the **Cars 196** OOD dataset.

Metric	HSIC+SMDL	Laplace	Cross Entropy	Patch InfoNCE	SWAG
Insertion AUC	0.3301	0.3428	0.3354	0.3289	<b>0.3304</b>
Deletion AUC	0.0467	0.0446	0.0462	0.0447	<b>0.0471</b>
Insertion Score	0.0095	0.0096	0.0095	0.0096	<b>0.0089</b>
Deletion Score	0.0074	0.0079	0.0075	0.0073	<b>0.0142</b>

Table 4: Ablation results on the **CUB Transformed** OOD dataset.

Metric	HSIC+SMDL	Laplace	Cross Entropy	Patch InfoNCE	SWAG
Insertion AUC	0.1612	0.1465	0.1550	0.1495	<b>0.1687</b>
Deletion AUC	0.0191	0.0186	0.0232	0.0373	<b>0.021</b>
Insertion Score	0.0058	0.0059	0.0049	0.0051	<b>0.0054</b>
Deletion Score	0.0054	0.0056	0.0058	0.0059	<b>0.005</b>

Table 5: Ablation results on the **CIFAR-100 (50 Classes)** OOD dataset.

Metric	HSIC+SMDL	Laplace	Cross Entropy	Patch InfoNCE	SWAG
Insertion AUC	0.1299	0.1304	0.1280	0.1255	<b>0.1389</b>
Deletion AUC	0.0420	0.0415	0.0435	0.0440	<b>0.0262</b>
Insertion Score	0.0188	0.0192	0.0185	0.0181	<b>0.0196</b>
Deletion Score	0.0140	0.0137	0.0142	0.0145	<b>0.0135</b>

Table 6: Ablation results on the **NA Birds** OOD dataset.

Metric	HSIC+SMDL	Laplace	Cross Entropy	Patch InfoNCE	SWAG
Insertion AUC	0.4103	0.4084	0.3955	0.3901	<b>0.4215</b>
Deletion AUC	0.0239	0.0239	0.0245	0.0251	<b>0.0228</b>
Insertion Score	0.0028	0.0028	0.0026	0.0025	<b>0.0031</b>
Deletion Score	0.0018	0.0019	0.0020	0.0021	<b>0.0019</b>

Table 7: Ablation results on the **Synthetic Dataset**.

<b>Metric</b>	<b>HSIC+SMDL</b>	<b>Laplace</b>	<b>Cross Entropy</b>	<b>Patch InfoNCE</b>	<b>SWAG</b>
Insertion AUC	0.7105	0.7088	0.6950	0.6875	<b>0.7176</b>
Deletion AUC	0.2422	0.2440	0.2510	0.2550	<b>0.3162</b>
Insertion Score	0.0470	0.0468	0.0450	0.0445	<b>0.0487</b>
Deletion Score	0.0248	0.0250	0.0255	0.0260	<b>0.0252</b>