

Holistic-CAM: Ultra-lucid and Sanity Preserving Visual Interpretation in Holistic Stage of CNNs

Appendix

A Ablation Study on fundamental scale denoising

During *fundamental scale denoising* phase, we utilize two filters to identify the fundamental location region. Now, we conduct experiments to evaluate the contribution of the inner component *median_blur2d* and *mean_blur2d* of *low-pass wrap* (LPW). In addition, we also conduct experiments on the hyper-parameters *blurSize* and *ksize*.

Table 1: Ablation study on denoising components. “D” represents median filter and “I” represents mean filter. The best result for each metric are shown in underline bold, and the second one is shown with underline.

	D	I	Ins↑	Del↓	Over-all↑	EPG↑
Layer4	X	X	53.399	8.881	44.518	<u>57.144</u>
	X	X	54.728	9.272	45.456	53.811
	X	X	53.677	8.841	<u>44.796</u>	56.748
	X	X	55.099	9.067	<u>46.032</u>	57.102
	X	X	49.067	8.841	40.226	54.957
Layer2	X	X	52.942	10.528	42.414	54.873
	X	X	53.689	10.944	<u>42.716</u>	54.603
	X	X	53.878	10.904	<u>42.975</u>	<u>55.577</u>
	X	X	53.878	10.904	<u>42.975</u>	<u>55.577</u>

Study on the components. As presented in Table 1, there is a evident drop in *Ins* and *Over-all* while removing the median filter and mean filter. This is mainly attributed to the narrow positioning region of basic scale mask with the damage of boundary details. In regard to its high performance in positioning ability (i. e. *EPG*), we believe that this mainly associated to the concentrated prominent information which coincides with the flaw of the *EPG* metric. It solely evaluates the first 100 points but ignores other useful information at the edge. In addition, the introduction of *median_blur2d* aims to prevent the omission of boundary information, and the corresponding performance in *Over-all* certainly confirms this contribution. However, its processing of expanding the edge region is very rough, sometimes even damaging the original edge information, finally leads to a slight decrease in localization ability.

Following the introduction of the *mean_filter* and *median_filter* simultaneously, LPW generates a smoother mask that is capable of seeking a balance between preserving the internal information and edge contour details after integrating it with the fused mask. This aligns with the improvement in *Over-all* and *EPG* scores. The marginal increase in *Del* is mainly attributed to the ability to emphasize more significant areas of LPW, it sometimes leads the *Del* metric to remove superior but not the best features. Nevertheless, the performance of *Ins* still reached the peak.

Study on the hyper-parameters. We conduct experiment on the hyper-parameters of LPW. Detailed result on Layer4 and Layer2 of ResNet-50 can be indicated in the Table 2, Table 3, Table 4 and Table 5. In our analysis, we thoroughly evaluate the holistic fidelity of each interpretation and localization ability, and opt for the *blurSize* of 51, the *ksize* of (91, 91) based on these considerations.

Table 2: Study on the hyper-parameters of *low-pass wrap*. Over-all of Layer4, higher is better. The best result for each metric are shown in underline bold, and the second one is shown with underline.

blurSize	31	51	71	91	111	131
ksize						
(31, 31)	44.695	45.048	45.407	45.628	44.021	44.003
(51, 51)	45.378	45.576	45.791	45.891	43.993	43.852
(71, 71)	45.879	46.014	<u>46.092</u>	46.084	43.853	43.723
(91, 91)	46.278	<u>46.290</u>	46.284	46.234	43.993	44.723
(111, 111)	44.021	43.993	43.778	43.772	43.385	43.778
(131, 131)	43.993	43.852	44.034	43.908	43.766	44.037

Table 3: Study on the hyper-parameters of *low-pass wrap*. Over-all of Layer2, higher is better.

blurSize	31	51	71	91	111	131
ksize						
(31, 31)	43.201	43.486	43.497	43.462	43.231	42.978
(51, 51)	43.637	43.701	43.629	43.351	43.041	43.013
(71, 71)	43.812	<u>43.833</u>	43.678	43.554	42.873	42.931
(91, 91)	43.822	<u>43.842</u>	43.745	43.223	42.764	42.891
(111, 111)	43.722	43.721	43.367	43.123	42.794	42.988
(131, 131)	43.612	43.632	43.382	43.116	42.668	42.762

Table 4: Study on the hyper-parameters of *low-pass wrap*. EPG of Layer4, higher is better.

blurSize	31	51	71	91	111	131
ksize						
(31, 31)	56.805	<u>56.906</u>	56.888	56.453	56.491	56.411
(51, 51)	56.755	56.832	56.876	56.871	56.698	56.637
(71, 71)	56.589	56.664	56.753	56.698	56.696	56.851
(91, 91)	56.690	<u>56.934</u>	56.417	56.583	56.587	56.857
(111, 111)	55.876	55.911	55.978	56.094	56.081	56.458
(131, 131)	55.381	55.406	55.455	55.553	55.514	55.961

Table 5: Study on the hyper-parameters of *low-pass wrap*. EPG of Layer2, higher is better.

blurSize	31	51	71	91	111	131
ksize						
(31, 31)	56.906	<u>56.957</u>	55.739	55.294	54.605	53.682
(51, 51)	56.832	<u>56.836</u>	56.571	55.112	54.443	53.546
(71, 71)	56.645	56.519	55.254	54.810	54.159	53.314
(91, 91)	56.334	55.004	54.757	54.451	55.327	52.989
(111, 111)	55.911	55.357	54.127	53.766	53.257	52.576
(131, 131)	55.406	54.631	53.425	53.105	53.666	52.961

B Detailed Quantitative Experiment Results.

Detailed quantitative results can be referenced at Table 6.

Table 6: Detailed Quantitative Experiment Results on ResNet-50 and VGG-16. Layer4 to Layer1 represent the four bottleneck layers of ResNet-50 from deep to shallow, respectively. Layer43 to Layer13 represent the four max-pooling layers of VGG-16 from deep to shallow. The best results for each metric are shown in underline bold as well as dyed in **red, and the second one is shown with underline and colored in **blue**.**

ResNet-50							VGG-16						
Method		Ins \uparrow	Del \downarrow	Over-all \uparrow	ROAD \uparrow	EPG \uparrow	Method		Ins \uparrow	Del \downarrow	Over-all \uparrow	ROAD \uparrow	EPG \uparrow
Layer4	Grad-CAM	54.887	11.622	43.265	28.095	53.673	Layer43	Grad-CAM	48.390	10.894	37.496	25.851	49.730
	Grad-CAM++	51.165	14.172	36.993	22.470	51.424		Grad-CAM++	45.044	12.528	32.516	22.331	52.428
	XGrad-CAM	54.887	11.622	43.265	28.099	53.673		XGrad-CAM	49.029	10.761	38.268	26.223	49.288
	Eigen-CAM	53.249	12.705	40.544	25.595	53.167		Eigen-CAM	48.925	10.547	38.378	25.765	52.305
	Layer-CAM	54.018	11.882	42.136	26.799	52.963		Layer-CAM	48.125	10.444	37.681	26.002	51.433
	CAMERAS	54.439	8.698	45.741	28.606	52.931		CAMERAS	44.548	9.091	35.457	26.153	50.008
	Relevance-CAM	54.663	11.622	43.041	27.981	52.989		Relevance-CAM	49.296	10.043	39.253	25.980	50.894
	IAA	54.872	11.901	42.971	28.871	54.696		IAA	43.733	10.116	33.617	24.876	46.071
Holistic-CAM	55.056	8.947	46.109	29.047	57.635	Holistic-CAM	49.569	8.792	40.777	26.345	55.090		
Layer3	Grad-CAM	30.622	15.752	14.870	12.312	44.078	Layer33	Grad-CAM	18.575	20.033	-1.458	-0.835	38.102
	Grad-CAM++	32.827	15.011	17.816	13.691	45.752		Grad-CAM++	31.451	12.769	18.682	13.815	44.281
	XGrad-CAM	28.775	14.832	13.943	13.676	46.972		XGrad-CAM	32.523	9.162	23.361	19.099	43.814
	Eigen-CAM	50.731	9.132	41.599	27.526	55.215		Eigen-CAM	46.541	8.373	38.168	26.204	52.134
	Layer-CAM	49.364	8.338	41.026	27.707	54.271		Layer-CAM	45.242	7.699	37.543	26.209	51.501
	CAMERAS	49.483	7.569	41.914	27.879	53.127		CAMERAS	43.684	8.331	35.353	26.383	51.882
	Relevance-CAM	53.419	9.284	44.135	28.631	48.664		Relevance-CAM	42.494	8.685	33.809	23.464	47.296
	IAA	18.597	11.164	7.433	11.205	43.876		IAA	31.841	8.969	22.872	23.138	44.452
Holistic-CAM	54.823	9.923	44.900	28.268	58.311	Holistic-CAM	47.231	8.292	38.939	26.504	57.616		
Layer2	Grad-CAM	18.876	15.207	3.669	4.673	45.171	Layer23	Grad-CAM	11.226	14.503	-3.277	-3.696	37.526
	Grad-CAM++	19.779	14.901	4.879	7.165	44.16		Grad-CAM++	20.927	10.394	10.533	12.468	45.072
	XGrad-CAM	19.804	13.162	6.642	7.649	46.257		XGrad-CAM	18.142	9.216	8.926	12.036	40.761
	Eigen-CAM	47.725	8.716	39.009	26.380	54.459		Eigen-CAM	42.230	7.276	34.954	25.417	53.162
	Layer-CAM	45.826	7.660	38.166	26.864	52.579		Layer-CAM	39.160	5.882	33.278	25.656	51.396
	CAMERAS	47.048	7.314	38.332	27.236	51.852		CAMERAS	39.994	6.972	33.022	26.008	51.398
	Relevance-CAM	48.854	8.949	39.909	26.734	47.163		Relevance-CAM	32.180	8.487	23.693	19.995	46.068
	IAA	12.413	7.889	4.524	10.964	43.905		IAA	14.301	7.552	6.749	16.236	43.924
Holistic-CAM	53.825	10.912	42.913	27.650	55.745	Holistic-CAM	45.001	6.969	38.032	26.252	58.155		
Layer1	Grad-CAM	18.448	17.177	1.271	2.801	45.095	Layer13	Grad-CAM	10.181	10.523	-0.342	0.581	42.653
	Grad-CAM++	16.903	18.089	-1.186	1.513	43.631		Grad-CAM++	18.429	10.293	8.136	12.764	45.531
	XGrad-CAM	19.621	16.642	2.979	3.539	45.158		XGrad-CAM	14.480	8.690	5.79	11.918	43.796
	Eigen-CAM	38.471	7.277	31.194	22.692	51.651		Eigen-CAM	40.353	8.596	31.757	23.911	53.966
	Layer-CAM	35.491	6.803	28.688	23.347	50.925		Layer-CAM	35.638	6.032	29.606	26.597	51.759
	CAMERAS	36.306	6.373	29.933	22.628	50.101		CAMERAS	36.561	5.746	30.815	27.724	51.336
	Relevance-CAM	46.255	10.291	35.964	24.297	47.591		Relevance-CAM	27.971	10.883	17.088	18.015	45.987
	IAA	8.475	6.841	1.634	3.368	43.828		IAA	8.475	6.741	1.734	3.368	43.828
Holistic-CAM	52.639	10.899	41.740	25.439	52.278	Holistic-CAM	45.776	8.604	37.172	26.771	56.467		

C Comparisons of Various Attribution Methods.

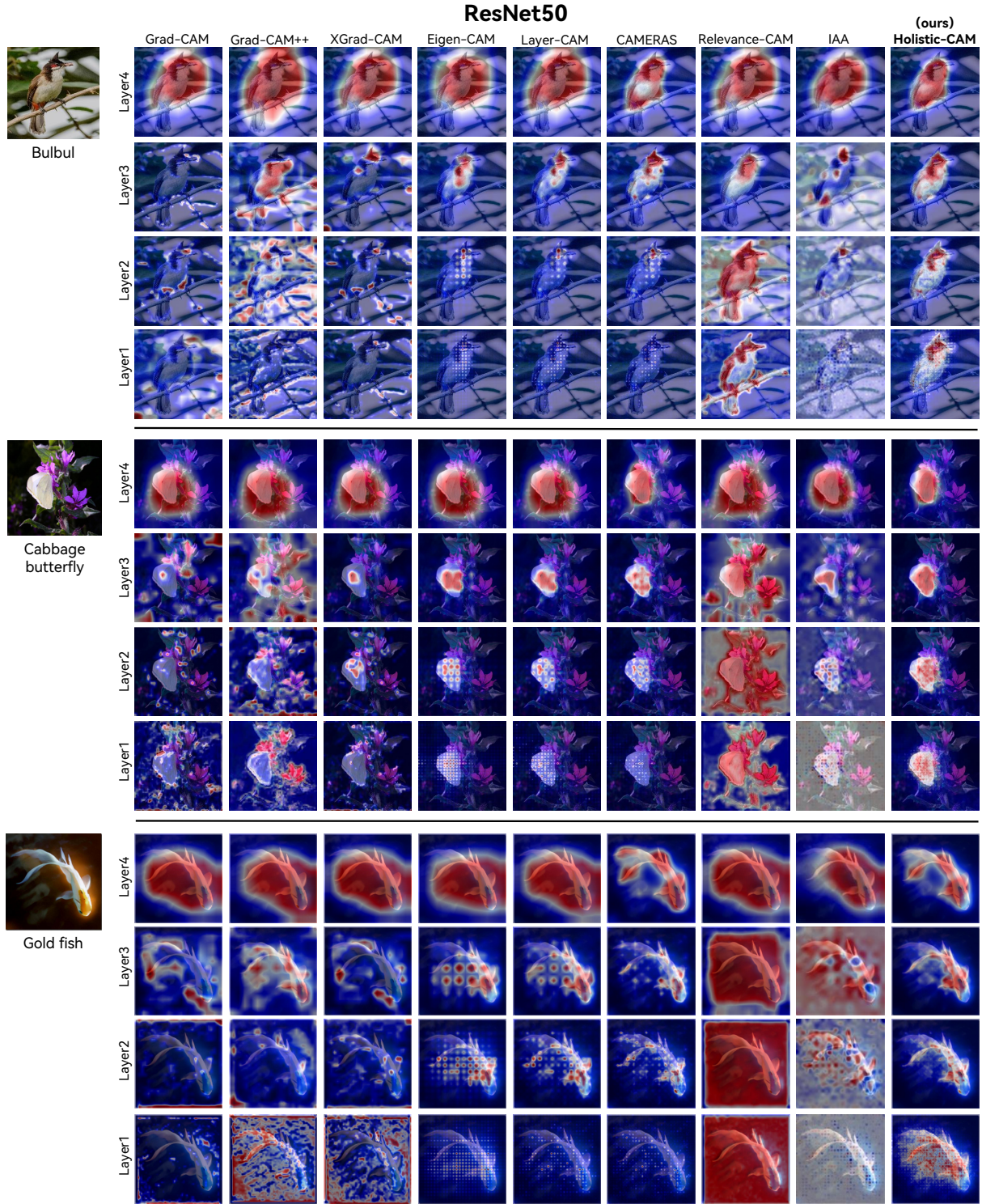


Figure 1: Comparisons of Various Attribution Methods. The columns are divided by the interpretation methods. The rows are divided along layer depth. Layer4 to Layer1 represent the four bottleneck layers of ResNet-50 from deep to shallow, respectively. Holistic-CAM is capable of locating the target object accurately.

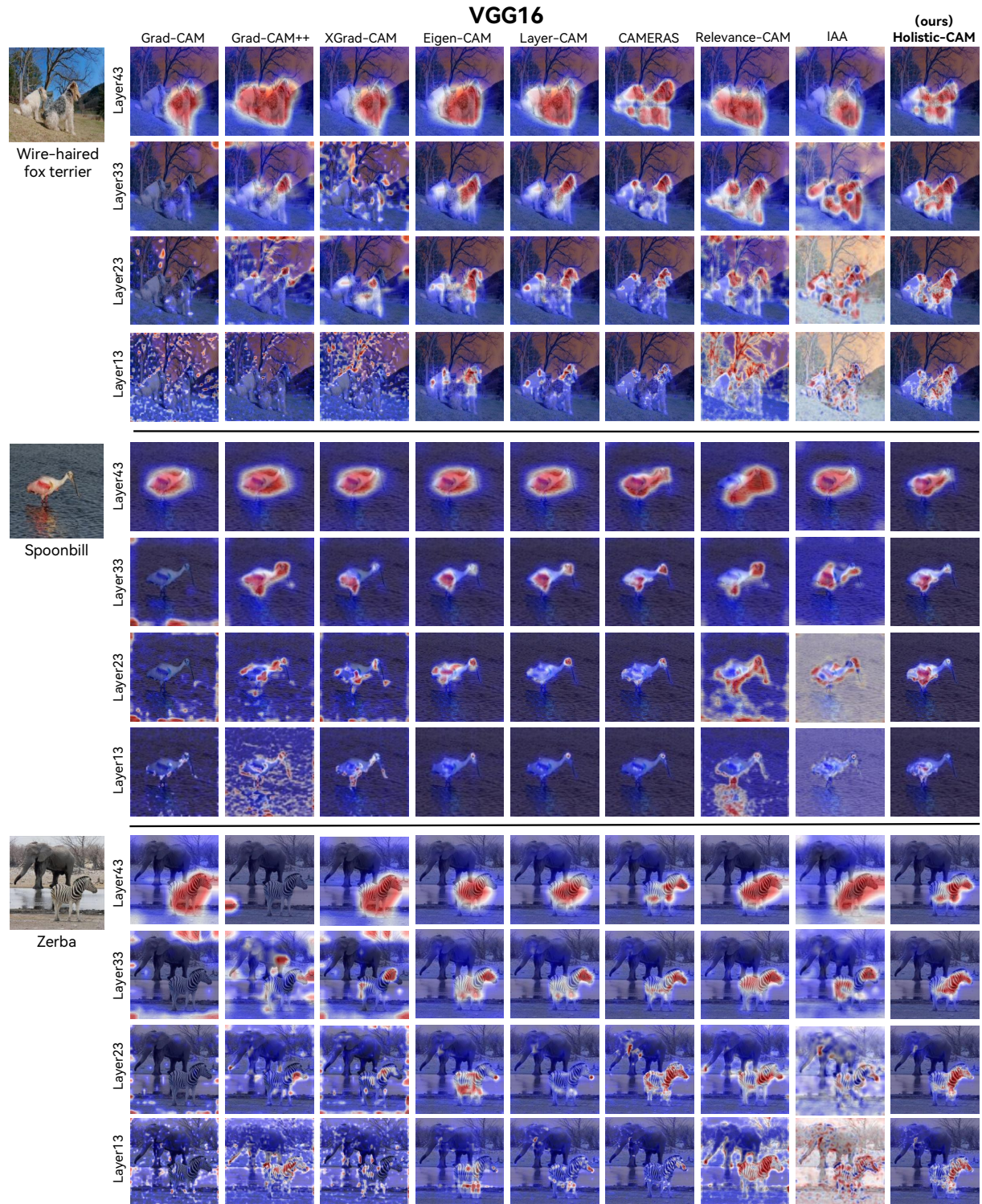


Figure 2: Comparisons of Various Attribution Methods. The columns are divided by the interpretation methods. The rows are divided along layer depth. Layer13 to Layer43 represents the four max-pooling layers of VGG-16 from shallow to deep, respectively.