IDGI: A Framework to Eliminate Explanation Noise from Gradients Integration (Appendix)

1. Theorem

Theorem 1 *Given a function $f_c(x) : R^n \rightarrow R$, points $x_j, x_{j+1}, x_{j_p} \in R^n$, then the gradient of the function with respect to each point in the space R^n forms the conservative vector fields $\overrightarrow{F}$ and further define the hyperplane $h_j = \{x : f_c(x) = f_c(x_j)\}$ in $\overrightarrow{F}$. Assume the Riemann Integration accurately estimates the line integral of the vector field $\overrightarrow{F}$ from points x_j to x_{j+1} and x_{j_p} e.g. $\int_{x_j}^{x_{j_p}} \frac{\partial f_c(x)}{\partial x} dx \approx \frac{\partial f_c(x_j)}{\partial x_j}(x_{j_p} - x_j)$, and $x_j \in h_j$, $x_{j_p}, x_{j+1} \in h_{j+1}$. Then:*

$$\int_{x_j}^{x_{j+1}} \frac{\partial f_c(x)}{\partial x} dx \approx \int_{x_j}^{x_{j_p}} \frac{\partial f_c(x)}{\partial x} dx.$$

Proof.

$$\begin{aligned}\int_{x_j}^{x_{j+1}} \frac{\partial f_c(x)}{\partial x} dx &= f_c(x_{j+1}) - f_c(x_j) \\ &\approx \frac{\partial f_c(x_j)}{\partial x_j}(x_{j+1} - x_j) \\ &= f_c(x_{j_p}) - f_c(x_j) \\ &= \frac{\partial f_c(x_j)}{\partial x_j}(x_{j_p} - x_j) \\ &= \int_{x_j}^{x_{j_p}} \frac{\partial f_c(x)}{\partial x} dx \end{aligned} \tag{1}$$

2. IG with IDGI

The Integrated Gradients algorithm requires a specified reference image to compute the attribution. One often selects a black or white picture as a reference point, resulting in the zero attribution value to pixels with the same value as the reference. This is due to the fact that these pixel values do not change while traveling from the reference image to the original image. However, these pixels may still be crucial for the classifier to make the decision, and merit attribution differs from zero. For example, as shown in Figure 1, the *Xception* model makes the prediction correctly on the given image has a black dog. When utilizing IG for providing an explanation, the body of the dog will be assigned zero attributions since the reference (black) image has the same pixel value as the dog. Intuitively, the explanation method should give non-zero values to these black pixels, since they represent the dog's body and are assumed to be significant characteristics. Alternatively, if the attribution value is zero, it is likely because the feature is insignificant and not because of the explanation method's design. In contrast to the original IG, IG with IDGI might potentially assign non-zero values to pixels with the same value as the reference picture, a desirable trait for a superior explanation technique.

(a) Original Image (b) IG (c) IG+IDGI

Figure 1. Original image is predicted *Tibetan terrier* from Xception classifier. Both 1b and 1c are attributions from IG and IG+IDGI with the black image as reference. Since the pixels are also black for the original image on the dog region, by design, IG is not able to assign important values to those pixels, however, ID+IDGI overcomes the issue.

3. Visual Examples

We present more visual examples in Figs. 2 to 6.

4. Distribution by Normalized Entropy and MS-SSIM

We present more distribution that compares Normalized Entroy and MS-SSIM in Figs. 7 to 17.

5. AIC and SIC with XRAI

Tab. 1 and Tab. 2 show the result of AIC and SIC for all methods and its version with XRAI. Similarly, Tab. 3 and Tab. 4 show the result of AIC and SIC with MS-SSIM for all methods and its version with XRAI.

AUC of AIC							
Models	**IG-based Methods**						**Other**
	IG	+Ours	GIG	+Ours	BlurIG	+Ours	VG
DenseNet121	.161	**.300**	.141	**.252**	.192	**.230**	.087
DenseNet169	.160	**.288**	.154	**.254**	.181	**.216**	.089
DenseNet201	.185	**.307**	.182	**.269**	.213	**.246**	.110
InceptionV3	.203	**.343**	.189	**.338**	.266	**.301**	.127
MobileNetV2	.098	**.233**	.114	**.204**	.145	**.197**	.068
ResNet50V2	.162	**.253**	.162	**.248**	.189	**.210**	.108
ResNet101V2	.177	**.268**	.163	**.253**	.198	**.215**	.116
ResNet151V2	.186	**.281**	.165	**.258**	.205	**.229**	.112
VGG16	.145	**.244**	.141	**.199**	.181	**.222**	.108
VGG19	.153	**.263**	.150	**.219**	.204	**.240**	.117
Xception	.238	**.404**	.239	**.381**	.309	**.355**	.174
	With XRAI						
DenseNet121	.438	**.479**	.460	**.460**	.437	**.452**	.434
DenseNet169	.468	**.508**	.483	**.492**	.466	**.480**	.462
DenseNet201	.439	**.476**	.460	**.468**	.442	**.461**	.449
InceptionV3	.477	**.506**	.472	**.513**	.479	**.503**	.496
MobileNetV2	.407	**.442**	.437	**.435**	.410	**.436**	.424
ResNet50V2	.402	**.433**	.428	**.438**	.409	**.410**	.417
ResNet101V2	.415	**.447**	.433	**.445**	.416	**.422**	.424
ResNet151V2	.410	**.443**	.421	**.435**	.406	**.416**	.412
VGG16	.393	**.423**	.422	**.418**	.402	**.413**	.396
VGG19	.386	**.416**	.417	**.414**	.396	**.408**	.393
Xception	.486	**.521**	.507	**.525**	.492	**.520**	.511

Table 1. AUC of AIC

AUC of SIC							
Models	**IG-based Methods**						**Other**
	IG	+Ours	GIG	+Ours	BlurIG	+Ours	VG
DenseNet121	.054	**.228**	.036	**.157**	.085	**.134**	.015
DenseNet169	.052	**.230**	.045	**.170**	.083	**.130**	.016
DenseNet201	.068	**.241**	.058	**.183**	.109	**.155**	.019
InceptionV3	.087	**.294**	.061	**.286**	.171	**.232**	.029
MobileNetV2	.020	**.145**	.023	**.111**	.043	**.103**	.011
ResNet50V2	.077	**.210**	.067	**.201**	.099	**.158**	.025
ResNet101V2	.095	**.231**	.070	**.201**	.117	**.165**	.026
ResNet151V2	.101	**.249**	.065	**.212**	.122	**.177**	.025
VGG16	.046	**.166**	.039	**.104**	.082	**.141**	.021
VGG19	.046	**.177**	.041	**.115**	.098	**.151**	.023
Xception	.119	**.363**	.107	**.336**	.218	**.296**	.054
	With XRAI						
DenseNet121	.407	**.464**	.435	**.445**	.403	**.428**	.404
DenseNet169	.450	**.496**	.465	**.475**	.439	**.458**	.435
DenseNet201	.427	**.473**	.449	**.462**	.419	**.449**	.432
InceptionV3	.450	**.493**	.449	**.499**	.441	**.481**	.477
MobileNetV2	.351	**.398**	.391	**.394**	.353	**.393**	.374
ResNet50V2	.401	**.439**	.430	**.445**	.404	**.412**	.418
ResNet101V2	.424	**.463**	.445	**.464**	.419	**.428**	.433
ResNet151V2	.413	**.453**	.423	**.445**	.401	**.424**	.414
VGG16	.343	**.382**	.381	**.376**	.352	**.368**	.347
VGG19	.337	**.376**	.374	**.373**	.347	**.362**	.344
Xception	.458	**.502**	.486	**.508**	.465	**.503**	.488

Table 2. AUC of SIC

References

[1] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019. 8, 9, 10, 11

AUC of AIC with MS-SSIM							
Models	**IG-based Methods**						**Other**
	IG	+Ours	GIG	+Ours	BlurIG	+Ours	VG
DenseNet121	.229	**.305**	.231	**.280**	.216	**.277**	.186
DenseNet169	.241	**.314**	.249	**.297**	.218	**.289**	.205
DenseNet201	.254	**.323**	.262	**.303**	.237	**.303**	.216
InceptionV3	.264	**.333**	.268	**.333**	.264	**.323**	.228
MobileNetV2	.179	**.259**	.197	**.238**	.186	**.241**	.150
ResNet50V2	.225	**.277**	.239	**.274**	.209	**.260**	.198
ResNet101V2	.235	**.284**	.243	**.277**	.215	**.265**	.206
ResNet151V2	.247	**.302**	.250	**.292**	.227	**.284**	.212
VGG16	.205	**.271**	.212	**.245**	.204	**.259**	.179
VGG19	.211	**.275**	.220	**.252**	.214	**.266**	.188
Xception	.281	**.362**	.293	**.356**	.284	**.345**	.254
	With XRAI						
DenseNet121	.342	**.376**	.360	**.367**	.336	**.369**	.351
DenseNet169	.375	**.407**	.386	**.397**	.368	**.398**	.382
DenseNet201	.354	**.388**	.370	**.380**	.355	**.387**	.370
InceptionV3	.357	**.384**	.355	**.386**	.348	**.390**	.373
MobileNetV2	.310	**.339**	.333	**.334**	.310	**.339**	.329
ResNet50V2	.302	**.326**	.320	**.330**	.302	**.322**	.317
ResNet101V2	.316	**.342**	.329	**.342**	.312	**.334**	.327
ResNet151V2	.314	**.341**	.321	**.334**	.308	**.335**	.321
VGG16	.314	**.339**	.334	**.334**	.319	**.336**	.319
VGG19	.309	**.333**	.330	**.329**	.315	**.332**	.315
Xception	.370	**.402**	.391	**.406**	.372	**.408**	.396

Table 3. AUC of SIC with MS-SSIM

AUC of SIC with MS-SSIM							
Models	**IG-based Methods**						**Other**
	IG	+Ours	GIG	+Ours	BlurIG	+Ours	VG
DenseNet121	.184	**.263**	.188	**.239**	.172	**.236**	.139
DenseNet169	.205	**.282**	.214	**.263**	.182	**.256**	.166
DenseNet201	.212	**.286**	.221	**.265**	.194	**.266**	.170
InceptionV3	.211	**.287**	.215	**.285**	.214	**.276**	.179
MobileNetV2	.126	**.204**	.144	**.187**	.130	**.188**	.096
ResNet50V2	.196	**.254**	.213	**.250**	.177	**.236**	.167
ResNet101V2	.210	**.265**	.221	**.256**	.188	**.244**	.180
ResNet151V2	.221	**.282**	.227	**.270**	.197	**.261**	.186
VGG16	.163	**.234**	.174	**.210**	.166	**.224**	.137
VGG19	.173	**.240**	.186	**.219**	.177	**.233**	.149
Xception	.223	**.312**	.233	**.304**	.229	**.293**	.194
	With XRAI						
DenseNet121	.290	**.332**	.309	**.324**	.282	**.324**	.306
DenseNet169	.327	**.364**	.338	**.356**	.314	**.353**	.338
DenseNet201	.311	**.349**	.326	**.345**	.301	**.350**	.333
InceptionV3	.300	**.334**	.295	**.342**	.291	**.343**	.323
MobileNetV2	.238	**.270**	.264	**.270**	.239	**.273**	.262
ResNet50V2	.273	**.305**	.294	**.308**	.270	**.299**	.295
ResNet101V2	.291	**.323**	.305	**.322**	.283	**.312**	.306
ResNet151V2	.286	**.322**	.294	**.313**	.277	**.314**	.302
VGG16	.256	**.285**	.280	**.280**	.260	**.284**	.264
VGG19	.252	**.278**	.276	**.277**	.258	**.279**	.262
Xception	.311	**.345**	.331	**.352**	.311	**.353**	.341

Table 4. AUC of SIC with MS-SSIM

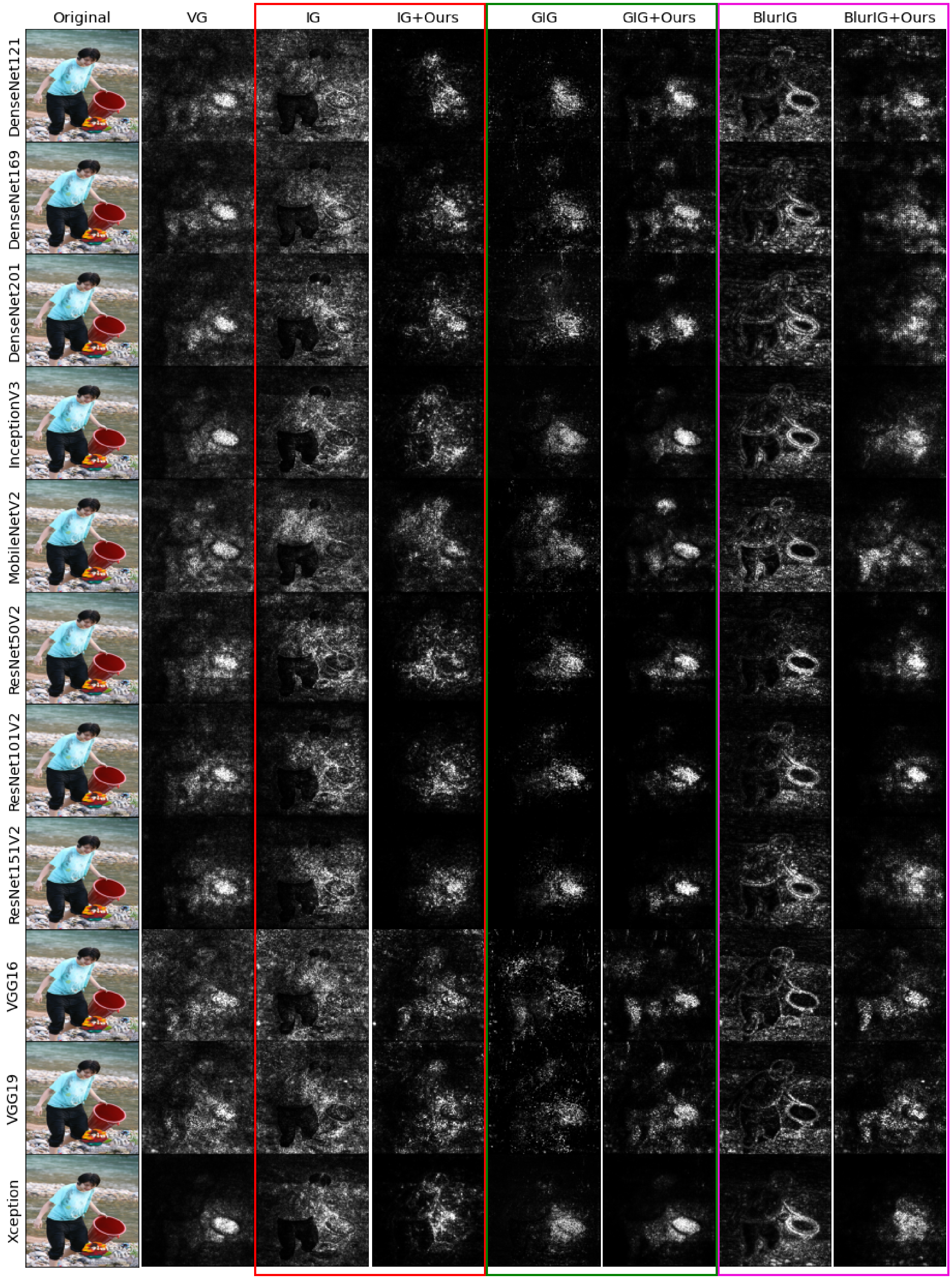

Figure 2. Predicted Label for all models: *bucket*

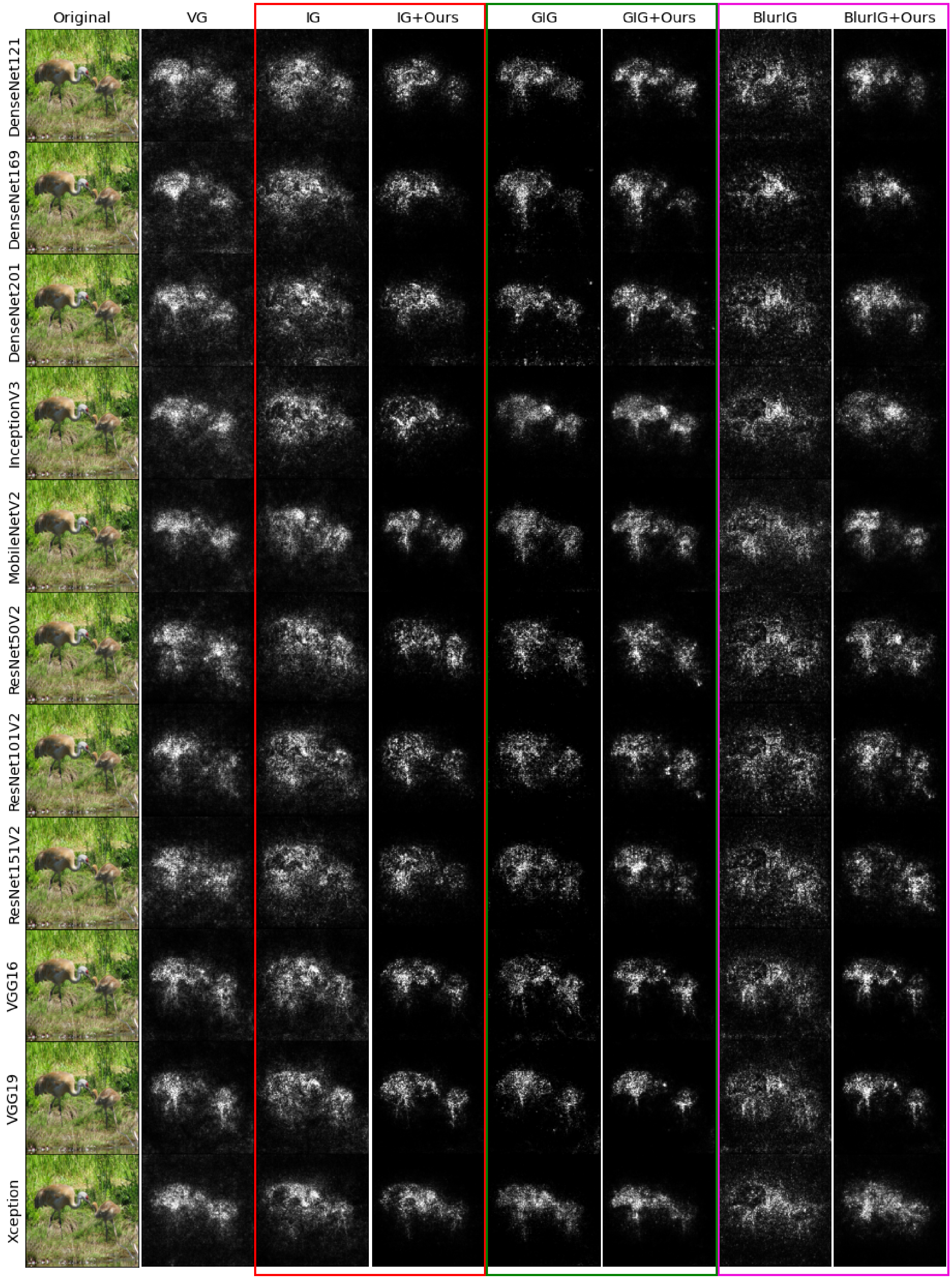

Figure 3. Predicted Label for all models: *crane*

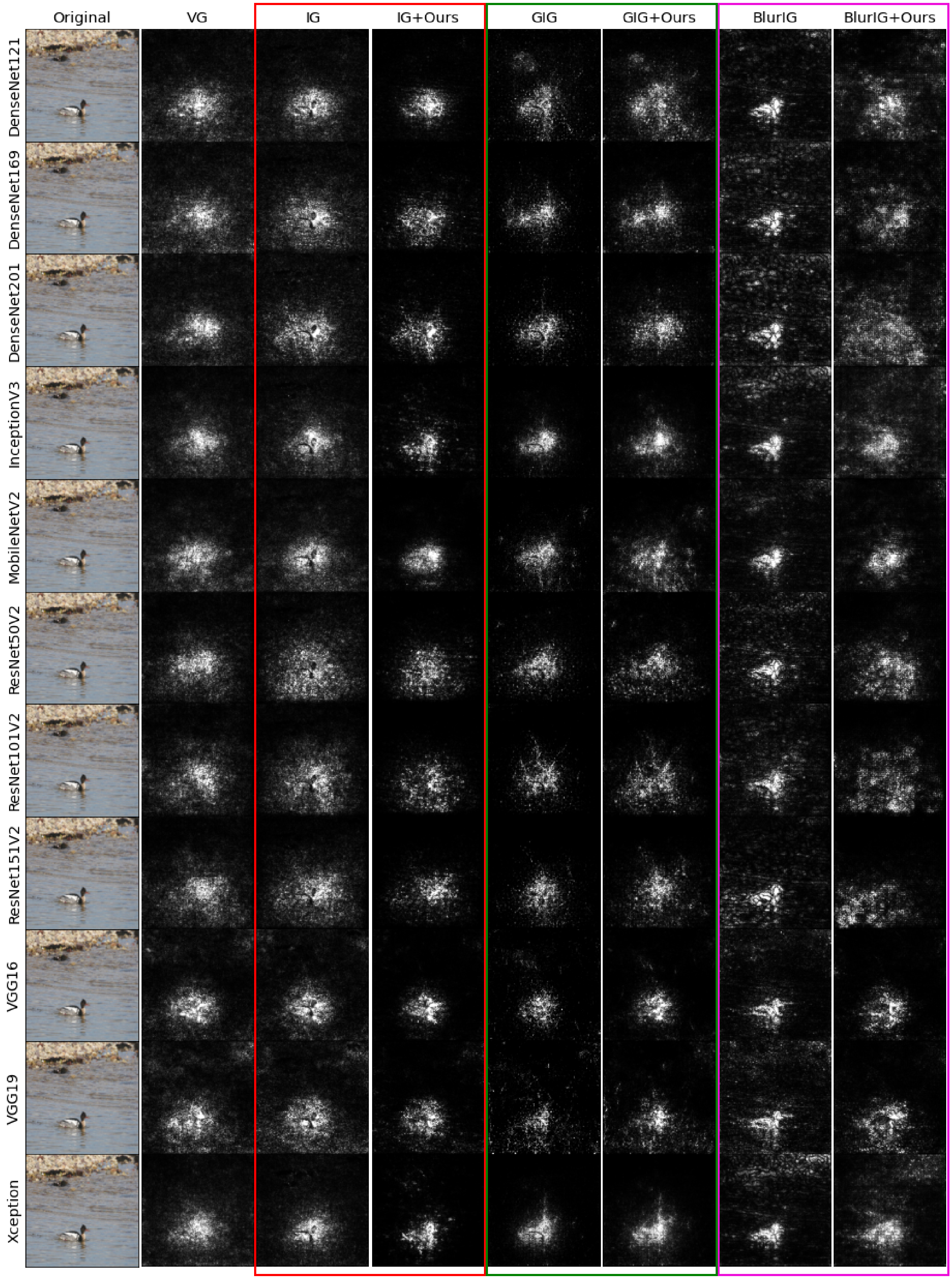

Figure 4. Predicted Label for all models: *mergus serrator*

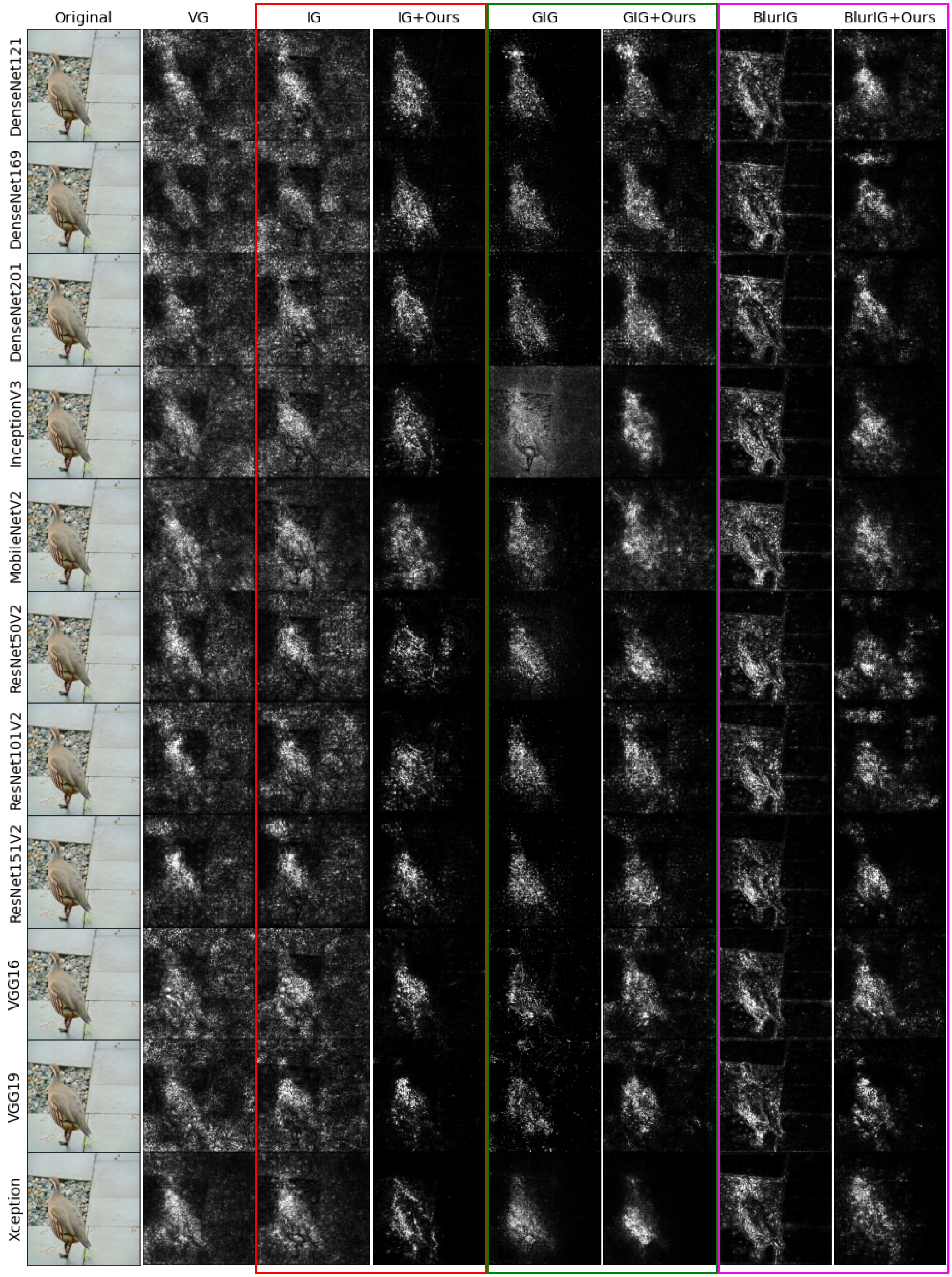

Figure 5. Predicted Label for all models: *partridge*

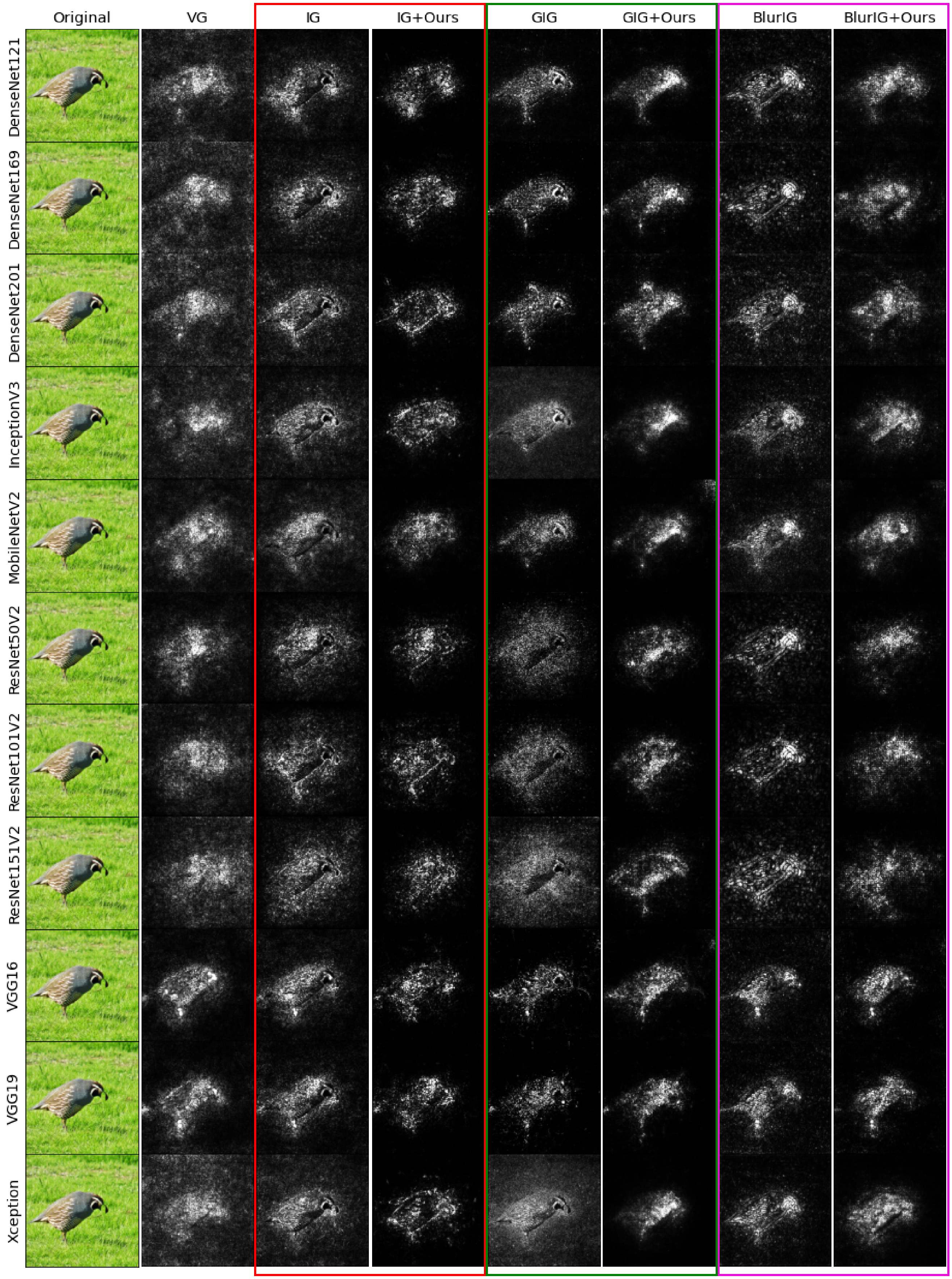

Figure 6. Predicted Label for all models: *quail*

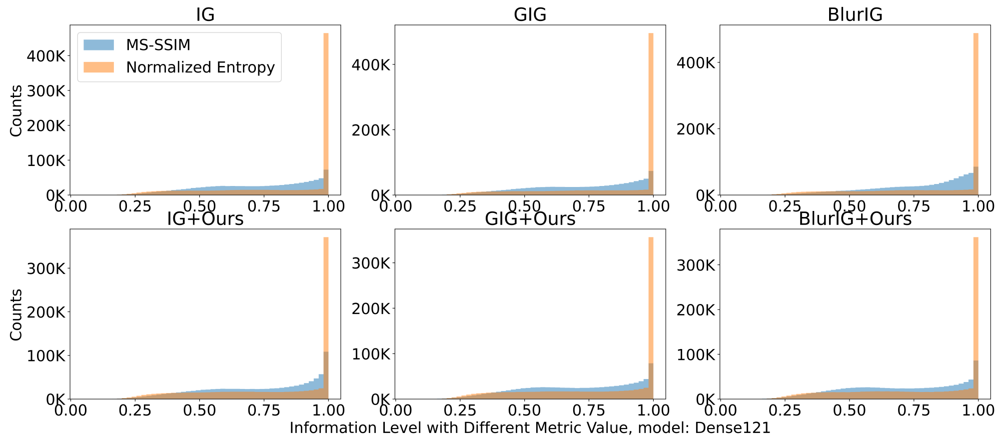

Figure 7. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *DenseNet121*

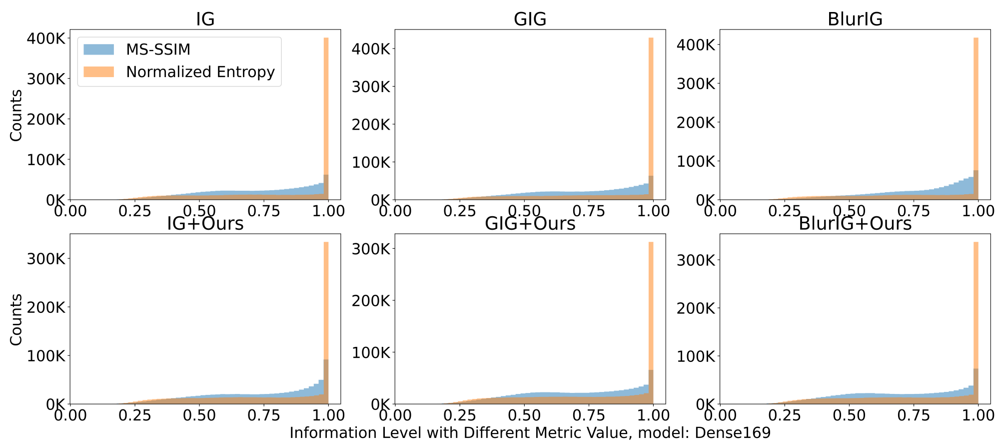

Figure 8. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *DenseNet169*

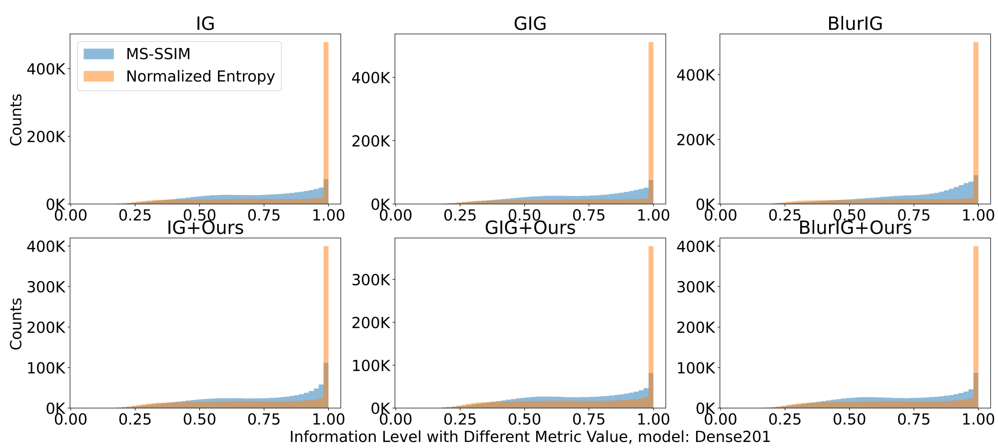

Figure 9. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *DenseNet201*

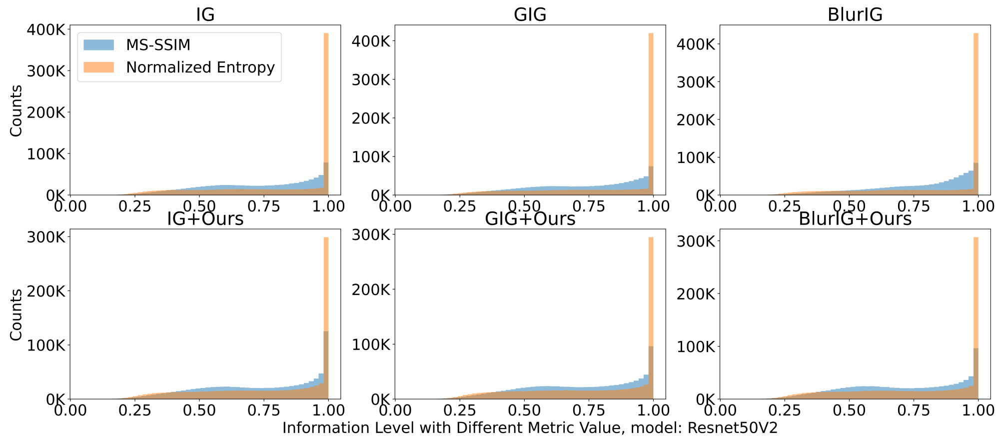

Figure 10. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *Resnet50V2*

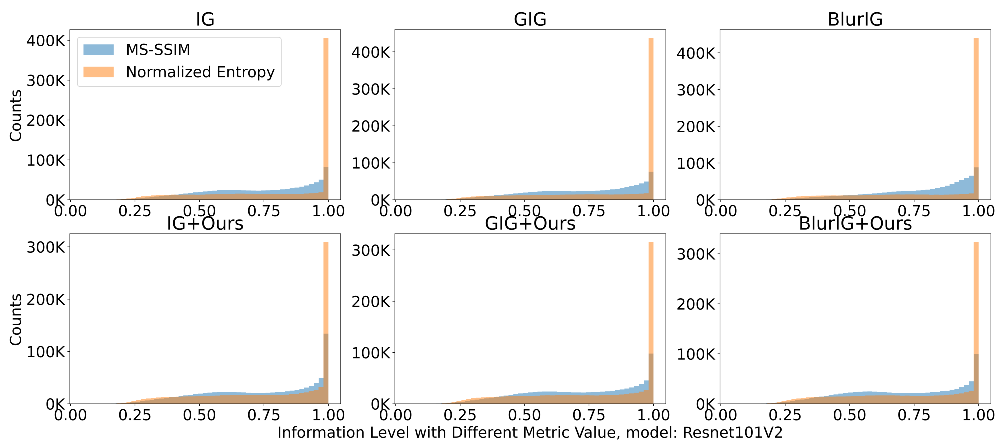

Figure 11. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *Resnet101V2*

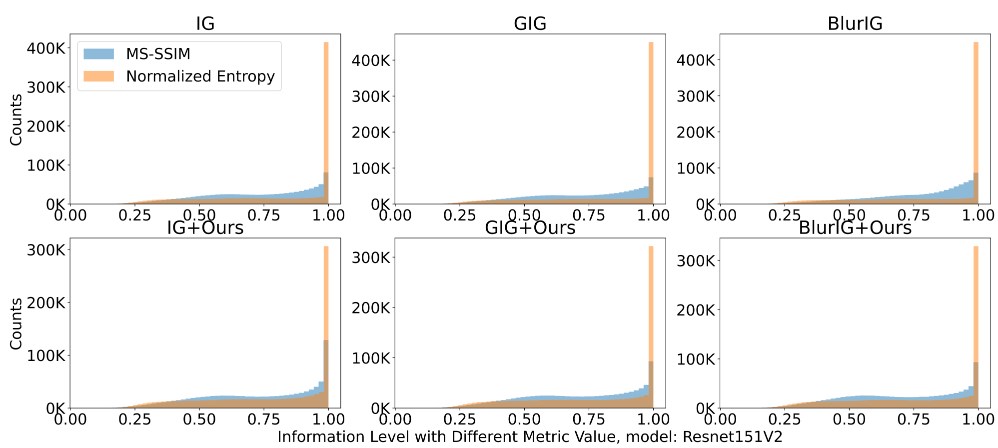

Figure 12. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *Resnet151V2*

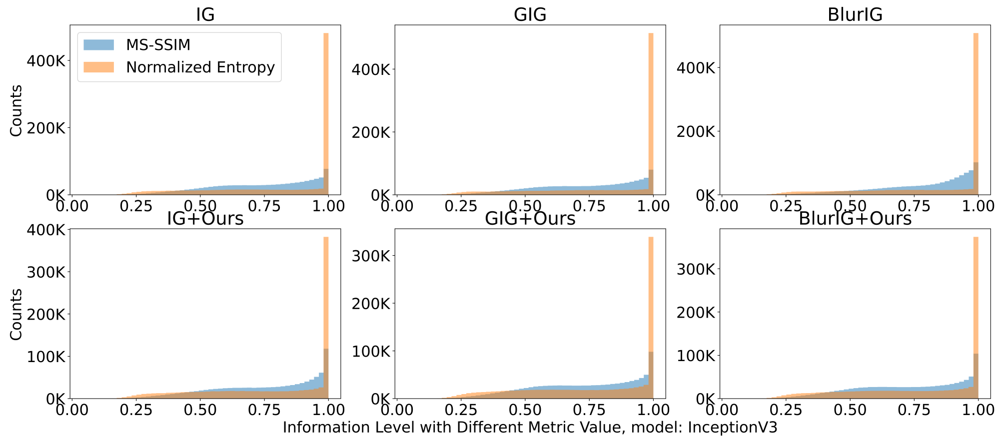

Figure 13. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *InceptionV3*

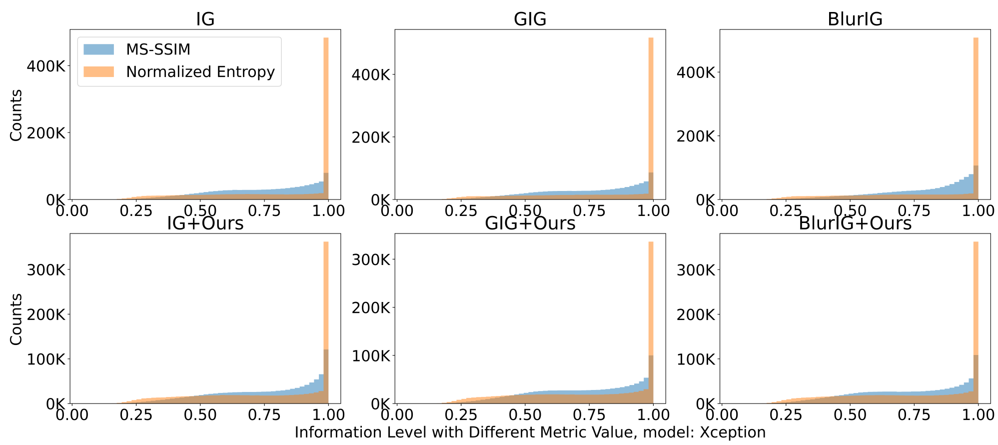

Figure 14. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *Xception*

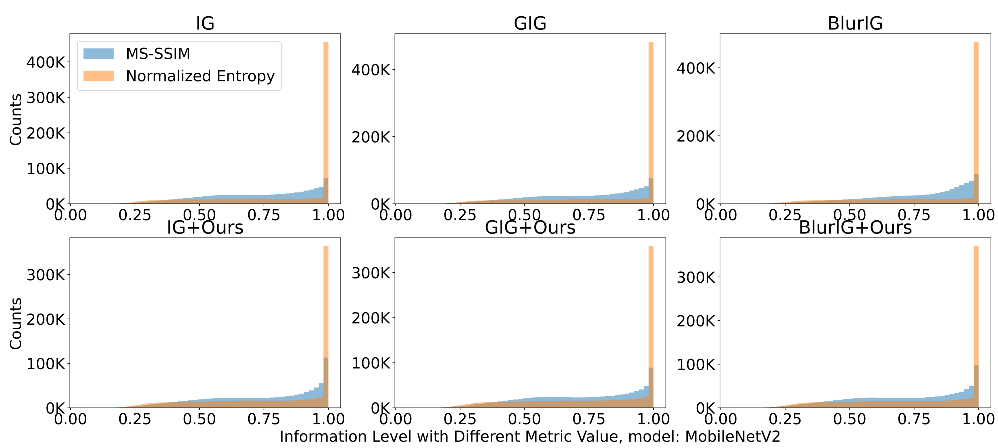

Figure 15. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *MobileNetV2*

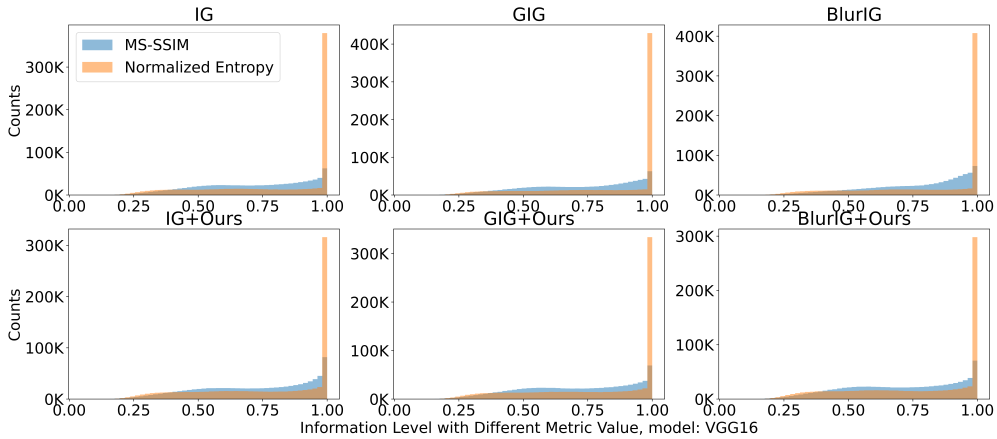

Figure 16. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *VGG16*

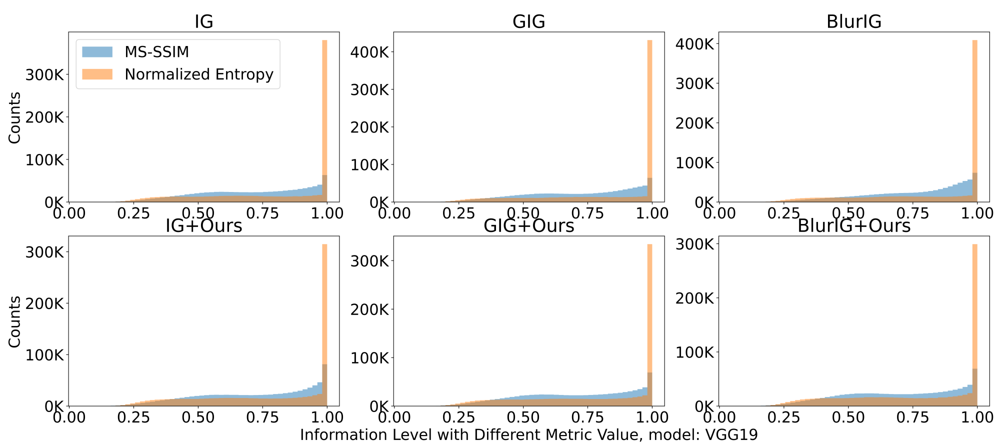

Figure 17. Modified distribution of bokeh images over MS-SSIM and Normalized Entropy [1]. Model: *VGG19*