A Proof of Proposition 1

Proof. The cross entropy loss for binary classification suffices to solve

$$\begin{aligned}
&\min_{\boldsymbol{w}_1,\boldsymbol{w}_2} -\mathbb{E}_{\boldsymbol{x},u}\Big[u\log(p_1)+(1-u)\log(p_2)\Big]\\
\iff &\min_{\boldsymbol{w}_1,\boldsymbol{w}_2} -\mathbb{E}_{\boldsymbol{x},u}\Big[u\log(\frac{e^{\boldsymbol{w}_1^T\boldsymbol{x}}}{e^{\boldsymbol{w}_1^T\boldsymbol{x}}+e^{\boldsymbol{w}_2^T\boldsymbol{x}}})+(1-u)\log(\frac{e^{\boldsymbol{w}_2^T\boldsymbol{x}}}{e^{\boldsymbol{w}_2^T\boldsymbol{x}}+e^{\boldsymbol{w}_1^T\boldsymbol{x}}})\Big]\\
\iff &\min_{\boldsymbol{w}_1,\boldsymbol{w}_2} -\mathbb{E}_{\boldsymbol{x},u}\Big[u\log(\frac{1}{1+e^{-(\boldsymbol{w}_1-\boldsymbol{w}_2)^T\boldsymbol{x}}})+(1-u)\log(\frac{e^{-(\boldsymbol{w}_1-\boldsymbol{w}_2)^T\boldsymbol{x}}}{1+e^{-(\boldsymbol{w}_1-\boldsymbol{w}_2)^T\boldsymbol{x}}})\Big]\\
\overset{\boldsymbol{w}=\boldsymbol{w}_1-\boldsymbol{w}_2}{\iff} &\min_{\boldsymbol{w}} -\mathbb{E}_{\boldsymbol{x},u}\Big[u\log(\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}})+(1-u)\log(\frac{e^{-\boldsymbol{w}^T\boldsymbol{x}}}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}})\Big].
\end{aligned}$$

Denoting the cross entropy loss above by L, the optimization of $\boldsymbol{w}_1,\boldsymbol{w}_2$ are performed by

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i - \eta\nabla_{\boldsymbol{w}_i}L,\quad i=1,2,$$

hence $\boldsymbol{w}=\boldsymbol{w}_1-\boldsymbol{w}_2 \leftarrow \big[(\boldsymbol{w}_1-\boldsymbol{w}_2)-\eta(\nabla_{\boldsymbol{w}_1}L-\nabla_{\boldsymbol{w}_2}L)\big]=\boldsymbol{w}-\eta\nabla_{\boldsymbol{w}}L$, which are optimized equivalently. □

B Proof of Proposition 2

Proof. Suppose that $\boldsymbol{x}\mapsto\boldsymbol{y}$ is a classification model, and $\boldsymbol{y}\mapsto\boldsymbol{p}$ is the softmax activation. Then the gradient attributions of y_t are $\boldsymbol{\phi}^t(\boldsymbol{x})_y := J_{\boldsymbol{x}}y_t$ and the gradient attributions of p_t are

$$\begin{aligned}
\boldsymbol{\phi}^t(\boldsymbol{x})_p :=& J_{\boldsymbol{y}}p_t\cdot J_{\boldsymbol{x}}\boldsymbol{y}=\sum_{s\in[c]}\frac{\partial p_t}{\partial y_s}J_{\boldsymbol{x}}y_s=\sum_{s\in[c]}\frac{\partial p_t}{\partial y_s}\boldsymbol{\phi}^s(\boldsymbol{x})_y\\
=&\frac{\sum_{s\neq t}e^{y_t+y_s}\boldsymbol{\phi}^t(\boldsymbol{x})_y-\sum_{s\neq t}e^{y_t+y_s}\boldsymbol{\phi}^s(\boldsymbol{x})_y}{(\sum_{s\in[c]}e^{y_s})^2}\\
=&\frac{\sum_{s\neq t}e^{y_t+y_s}}{(\sum_{s\in[c]}e^{y_s})^2}(\boldsymbol{\phi}^t(\boldsymbol{x})_y-\sum_{s\neq t}\alpha_s\boldsymbol{\phi}^s(\boldsymbol{x})_y)\\
=&\frac{\sum_{s\neq t}e^{y_t+y_s}}{(\sum_{s\in[c]}e^{y_s})^2}\boldsymbol{\phi}^t(\boldsymbol{x})_{\text{weighted}}\propto\boldsymbol{\phi}^t(\boldsymbol{x})_{\text{weighted}}
\end{aligned}$$

Hence weighted contrastive explanations of class t are essentially direct explanations with respect to the probability p_t value. □

C Proof of Proposition 3

Proof. Let $\boldsymbol{r}^{l,s}\in\mathbb{R}^{d_l}$ be the attribution scores at the l-th step starting from $r^{0,s}$, which is the score for y_s. $\forall l,k\in[L],s\in[c]$, let $\boldsymbol{f}^{l\to k}(\boldsymbol{r}^{l,s})=\boldsymbol{r}^{k,s}$ be the propagation at the from the l-th step to the k-th step, where $l<k$. The weighted contrast of $\boldsymbol{r}^{l,t}$ is $\boldsymbol{r}^{l,t}_{\text{weighted}}=\boldsymbol{r}^{l,t}-\sum_{s\neq t}\alpha_s\boldsymbol{r}^{l,s}$. Then the propagated scores to the k-th step is

$$\begin{aligned}
\boldsymbol{f}^{l\to k}(\boldsymbol{r}^{l,t}_{\text{weighted}})=&\boldsymbol{f}^{l\to k}(\boldsymbol{r}^{l,t}-\sum_{s\neq t}\alpha_s\boldsymbol{r}^{l,s})=\boldsymbol{f}^{l\to k}(\boldsymbol{r}^{l,t})-\sum_{s\neq t}\alpha_s\boldsymbol{f}^{l\to k}(\boldsymbol{r}^{l,s})\\
=&\boldsymbol{r}^{k,t}-\sum_{s\neq t}\alpha_s\boldsymbol{r}^{k,s}=\boldsymbol{r}^{k,t}_{\text{weighted}}
\end{aligned}$$

Let $k=L$, then it equals to the weighted contrast at the input space. □

D Detailed Analysis of fig. 4

Here we present detailed analysis of the weighted contrastive explanations of samples in fig. 4. Different from original GradCAM, which tend to highlight the entire object for both classes, weighted

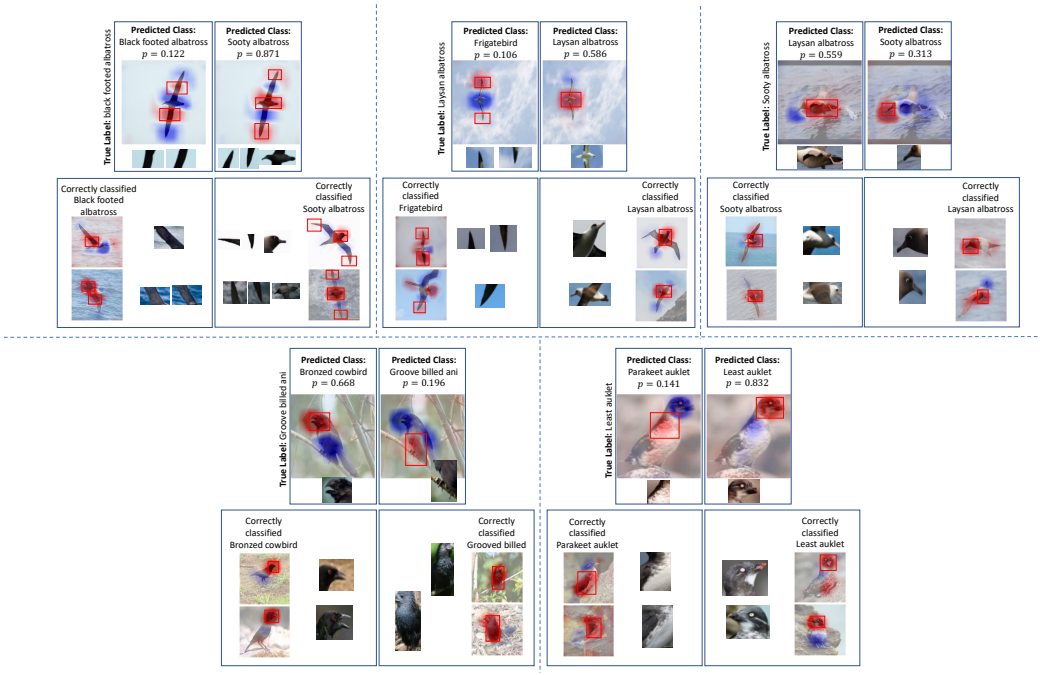

Figure 6: Detailed demonstration of the samples from CUB-200 in fig. 4 of the main paper. Each image is shown in one block separated by dashed lines. In each block, the top two images are weighted contrastive GradCAM explaining the top two classes (with the highest predicted probabilities p). We highlight the top features according to weighted contrastive GradCAM and compare them with 4 correctly classified images shown below, 2 of which from each class, respectively. In contrast to original GradCAM highlighting the entire objects (shown in fig. 4), the weighted contrastive explanations focus more on the contrastive features.

contrastive GradCAM highlights very different areas. We conduct empirical analysis and compare the highlighted areas to check if these areas are semantically meaningful. The results are shown in fig. 6. For each sample, we consider the top 2 classes predicted and check the correctly predicted images of those classes. Then we compare the highlighted areas of those correctly predicted images and the corresponding samples. It can be found in fig. 6 that the highlighted features are visually similar.

E Visualizations of Other Data

Here we demonstrate results of datasets other than CUB-200. The visualizations of FGVC are shown in fig. 7. THe visualizations of Flower-102 are shown in fig. 8. The visualizations of Food-101 are shown in fig. 9. The visualizations of Stanford Cars are shown in fig. 10. The alignment of images follows fig. 4. From these visual results, same conclusion can be drawn. Original explanation methods tend to fail capturing shared features while our weighted contrastive method succeeds.

F Quantitative Results of AlexNet

Here we present the blurring/masking experiments for AlexNet. The only difference between table 1 is the models tested. The results are shown in table 2. It can be found that in general the weighted contrastive methods outperform the original methods.

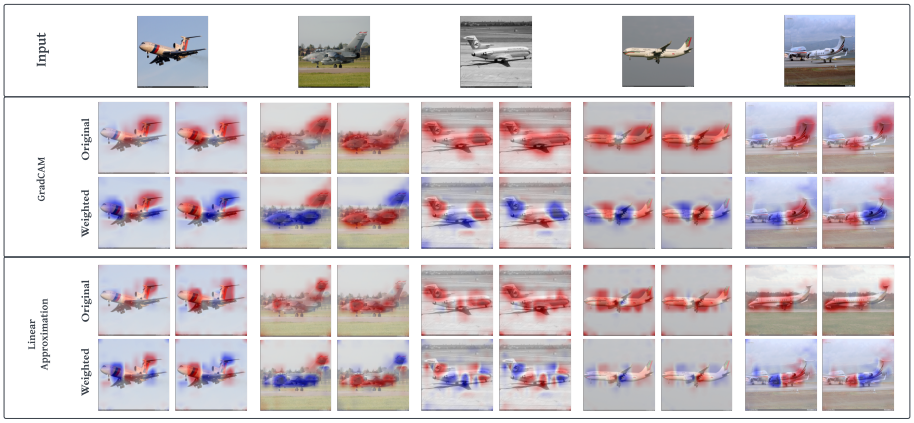

Figure 7: FGVC images of original explanations and weighted contrastive explanations over GradCAM and Linear Approximation. The alignment follows fig. 4.

Figure 8: Flower-102 images of original explanations and weighted contrastive explanations over GradCAM and Linear Approximation. The alignment follows fig. 4.

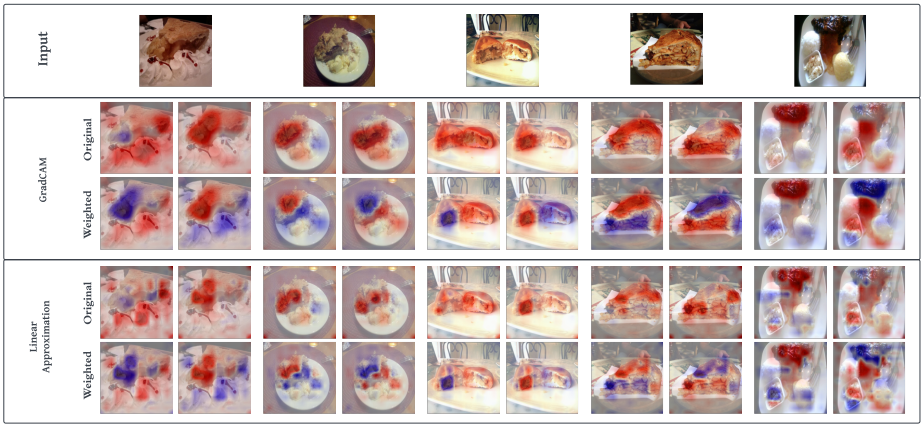

Figure 9: Food-101 images of original explanations and weighted contrastive explanations over GradCAM and Linear Approximation. The alignment follows fig. 4.

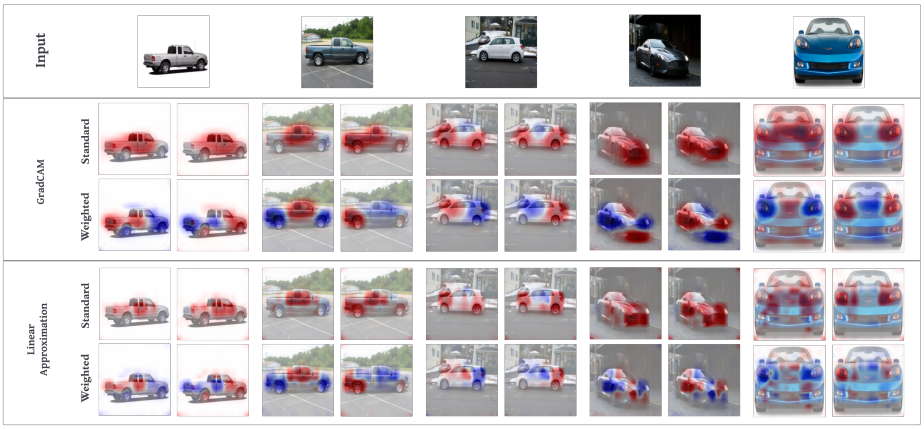

Figure 10: Stanford Cars images of original explanations and weighted contrastive explanations over GradCAM and Linear Approximation. The alignment follows fig. 4.

Table 2: Comparisons between the weighted contrastive method (wtd.) and the original (ord.) method in blurring/masking experiments on 5 datasets. Here the model tested is AlexNet, and other settings follow those of table 1.

				Gaussian Blur				Zeros				Channel-wise Mean			
		p_t		Pos. Features		Neg. Features		Pos. Features		Neg. Features		Pos. Features		Neg. Features	
				ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.
CUB-200	LA	t_1	0.698	0.759	**0.809**	0.396	**0.240**	0.675	**0.699**	0.422	**0.318**	0.727	**0.766**	0.396	**0.275**
		t_2	0.302	0.604	**0.769**	0.422	**0.204**	0.588	**0.695**	0.443	**0.294**	0.605	**0.740**	0.430	**0.233**
	GC	t_1	0.698	0.762	**0.824**	0.415	**0.228**	0.727	**0.764**	0.426	**0.284**	0.758	**0.816**	0.420	**0.236**
		t_2	0.302	0.547	**0.784**	0.393	**0.174**	0.535	**0.736**	0.409	**0.236**	0.550	**0.781**	0.398	**0.192**
FGVC	LA	t_1	0.702	0.769	**0.823**	0.364	**0.218**	0.671	**0.678**	0.419	**0.329**	0.709	**0.736**	0.397	**0.267**
		t_2	0.298	0.697	**0.801**	0.351	**0.169**	0.644	**0.692**	0.423	**0.321**	0.665	**0.754**	0.395	**0.254**
	GC	t_1	0.702	0.769	**0.842**	0.361	**0.195**	0.722	**0.744**	0.398	**0.275**	0.744	**0.800**	0.382	**0.229**
		t_2	0.298	0.673	**0.815**	0.340	**0.157**	0.653	**0.733**	0.398	**0.264**	0.673	**0.785**	0.371	**0.213**
Food-101	LA	t_1	0.707	0.781	**0.791**	0.417	**0.279**	**0.709**	0.702	0.443	**0.333**	0.734	**0.737**	0.430	**0.306**
		t_2	0.293	0.670	**0.738**	0.381	**0.208**	0.651	**0.682**	0.422	**0.296**	0.664	**0.709**	0.409	**0.261**
	GC	t_1	0.707	0.794	**0.813**	0.420	**0.261**	**0.759**	0.757	0.428	**0.293**	0.778	**0.787**	0.423	**0.265**
		t_2	0.292	0.628	**0.752**	0.377	**0.187**	0.637	**0.723**	0.408	**0.242**	0.646	**0.749**	0.394	**0.213**
Flower-102	LA	t_1	0.728	**0.824**	0.819	0.342	**0.234**	0.719	**0.737**	0.462	**0.296**	**0.768**	0.760	0.460	**0.264**
		t_2	0.272	0.686	**0.790**	0.361	**0.182**	0.693	**0.720**	0.378	**0.271**	0.690	**0.754**	0.368	**0.252**
	GC	t_1	0.728	0.814	**0.858**	0.418	**0.189**	0.758	**0.824**	0.404	**0.244**	0.792	**0.863**	0.396	**0.218**
		t_2	0.272	0.516	**0.820**	0.409	**0.136**	0.494	**0.757**	0.423	**0.172**	0.506	**0.775**	0.410	**0.158**
Stanford Cars	LA	t_1	0.698	**0.747**	0.745	0.468	**0.309**	0.694	**0.702**	0.470	**0.356**	0.701	**0.712**	0.465	**0.340**
		t_2	0.302	0.647	**0.693**	0.412	**0.240**	0.656	**0.665**	0.419	**0.303**	0.662	**0.679**	0.421	**0.287**
	GC	t_1	0.698	0.766	**0.782**	0.447	**0.277**	0.748	**0.761**	0.444	**0.291**	0.756	**0.769**	0.439	**0.277**
		t_2	0.302	0.627	**0.740**	0.402	**0.210**	0.640	**0.720**	0.397	**0.240**	0.647	**0.736**	0.399	**0.230**

G Equal Blurring/Masking

It should be noticed that the original methods and corresponding weighted contrastive methods have different highlighted areas, the proportions of areas of positive and negative values are different, too. As a result, the areas of masked/blurred pixels are different. In order to alleviate the bias introduced here, we carry out a *equal blurring/masking* experiment. It can be found that original methods tend to have much larger positive areas than the negative areas, while contrasted weighted methods are more balanced between these two. As a result, when masking/blurring the positive areas, we blur the minimum of the number of pixels that are positive of the two methods. And on the contrary, when masking/blurring the negative areas, we blur the maximum of the number of pixels that are negative of the two methods. In this way, the same number of pixels are masked/blurred. The results are shown in table 3. It can be found that the contrastive weighted methods still outperform original methods by a large margin.

Table 3: Comparisons between the weighted contrastive method (wtd.) and the original (ori.) in blurring/masking experiments on 5 datasets. Here we apply the *equal blurring/masking* technique. Other settings follow thos of table 1.

				Gaussian Blur				Zeros				Channel-wise Mean			
			p_t	Pos.		Neg.		Pos.		Neg.		Pos.		Neg.	
				ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.	ori.	wtd.
CUB-200	LA	t_1	0.710	0.726	**0.809**	0.475	**0.299**	0.691	**0.780**	0.494	**0.325**	0.710	**0.799**	0.486	**0.311**
		t_2	0.290	0.441	**0.728**	0.435	**0.187**	0.445	**0.706**	0.438	**0.221**	0.434	**0.723**	0.429	**0.200**
	GC	t_1	0.710	0.720	**0.832**	0.468	**0.266**	0.702	**0.818**	0.477	**0.278**	0.711	**0.832**	0.474	**0.271**
		t_2	0.290	0.414	**0.761**	0.440	**0.148**	0.411	**0.735**	0.448	**0.167**	0.410	**0.743**	0.445	**0.151**
FGVC	LA	t_1	0.694	0.744	**0.796**	0.407	**0.258**	0.700	**0.733**	0.446	**0.320**	0.716	**0.758**	0.424	**0.290**
		t_2	0.306	0.633	**0.769**	0.373	**0.191**	0.606	**0.695**	0.428	**0.268**	0.620	**0.731**	0.409	**0.239**
	GC	t_1	0.694	0.774	**0.854**	0.392	**0.204**	0.719	**0.801**	0.419	**0.262**	0.738	**0.821**	0.401	**0.238**
		t_2	0.306	0.634	**0.819**	0.351	**0.150**	0.609	**0.747**	0.396	**0.204**	0.614	**0.775**	0.384	**0.182**
Food-101	LA	t_1	0.714	0.766	**0.856**	0.319	**0.179**	0.739	**0.818**	0.348	**0.221**	0.747	**0.827**	0.333	**0.205**
		t_2	0.286	0.699	**0.832**	0.280	**0.138**	0.80	**0.794**	0.317	**0.174**	0.688	**0.805**	0.308	**0.164**
	GC	t_1	0.714	0.794	**0.891**	0.297	**0.130**	0.778	**0.867**	0.309	**0.156**	0.782	**0.875**	0.301	**0.144**
		t_2	0.286	0.709	**0.879**	0.261	**0.099**	0.697	**0.856**	0.287	**0.124**	0.703	**0.867**	0.280	**0.116**
Flower-102	LA	t_1	0.715	0.756	**0.830**	0.367	**0.243**	0.655	**0.746**	0.353	**0.252**	0.647	**0.752**	0.371	**0.269**
		t_2	0.285	0.646	**0.748**	0.391	**0.145**	0.667	**0.751**	0.444	**0.243**	0.651	**0.728**	0.435	**0.232**
	GC	t_1	0.715	0.799	**0.878**	0.361	**0.176**	0.748	**0.844**	0.322	**0.149**	0.751	**0.840**	0.339	**0.166**
		t_2	0.285	0.596	**0.812**	0.358	**0.120**	0.651	**0.825**	0.398	**0.160**	0.619	**0.818**	0.388	**0.153**
Stanford Cars	LA	t_1	0.721	0.749	**0.828**	0.433	**0.262**	0.744	**0.816**	0.439	**0.262**	0.744	**0.819**	0.436	**0.260**
		t_2	0.279	0.574	**0.755**	0.390	**0.166**	0.592	**0.747**	0.398	**0.178**	0.586	**0.754**	0.396	**0.173**
	GC	t_1	0.721	0.73	**0.885**	0.424	**0.208**	0.775	**0.881**	0.423	**0.218**	0.773	**0.886**	0.422	**0.209**
		t_2	0.279	0.574	**0.821**	0.390	**0.112**	0.582	**0.817**	0.394	**0.120**	0.584	**0.821**	0.390	**0.114**