Relational Proxies: Emergent Relationships as Fine-Grained Discriminators

Abhra Chaudhuri[1] **Massimiliano Mancini**[2] **Zeynep Akata**[2,3,4] **Anjan Dutta**[5]*
[1] University of Exeter [2] University of Tübingen [3] MPI for Informatics
[4] MPI for Intelligent Systems [5] University of Surrey

Abstract

Fine-grained categories that largely share the same set of parts cannot be discriminated based on part information alone, as they mostly differ in the way the local parts relate to the overall global structure of the object. We propose *Relational Proxies*, a novel approach that leverages the relational information between the global and local views of an object for encoding its semantic label. Starting with a rigorous formalization of the notion of distinguishability between fine-grained categories, we prove the necessary and sufficient conditions that a model must satisfy in order to learn the underlying decision boundaries in the fine-grained setting. We design Relational Proxies based on our theoretical findings and evaluate it on seven challenging fine-grained benchmark datasets and achieve state-of-the-art results on all of them, surpassing the performance of all existing works with a margin exceeding 4% in some cases. We also experimentally validate our theory on fine-grained distinguishability and obtain consistent results across multiple benchmarks. Implementation is available at `https://github.com/abhrac/relational-proxies`.

1 Introduction

Fine-grained visual categorization (FGVC) primarily requires identifying category-specific, discriminative local attributes [49, 44, 21]. However, the *relationship* of the attributes with the global view of the object is also known to encode semantic information [6, 5]. Such a relationship can be thought of as the way in which local attributes combine to form the overall object. When two categories share a large number of local attributes, this cross-view relational information becomes the only discriminator. To illustrate this in an intuitive example, Figure 1 shows two fine-grained categories of birds, the White-faced Plover (left and top-right) and the Kentish Plover (bottom-right). Along with color and texture information, the two categories share a large number of local features like beak, head, body, tail and wings. Given such constraints of largely overlapping attribute sets, relational information like the distance between the head and the body, or the angular orientation of the legs with respect to the body remain as the only available discriminators. We thus conjecture that the way the global structure (view) of the object arises out of its local parts (views) must be an *emergent* [30] property of the object which is implicitly encoded as the cross-view relationship. However, all existing methods that consider both global and local information, do so in a *relation-agnostic* manner, *i.e.*, without considering cross-view relationships (we formalize relation-agnosticity in Section 3).

We hypothesize that when two categories largely share the same set of local attributes and differ only in the way the attributes combine to generate the global view of the object, relation-agnostic approaches do not capture the full semantic information in an input image. To prove our hypothesis, we develop a rigorous formalization of the notion of distinguishability in the fine-grained setting. Via our theoretical framework, we identify the necessary and sufficient conditions that a learner must satisfy to completely learn a distribution of fine-grained categories. Specifically, we prove

*A. Chaudhuri is with the Department of Computer Science at the University of Exeter. M. Mancini and Z. Akata are with the Cluster of Excellence Machine Learning at the University of Tübingen. A. Dutta is with the Institute for People-Centred AI at the University of Surrey.
36th Conference on Neural Information Processing Systems (NeurIPS 2022).

that a learner must harness both view-specific (relation-agnostic) and cross-view (relation-aware) information in an input image. We also prove that it is not possible to design a single encoder that can achieve both of these objectives simultaneously. Based on our theoretical findings, we design a learner that separately computes metric space embeddings for the relation-agnostic and relation-aware components in an input image, through class representative vectors that we call Relational Proxies.

To summarize, we: (1) provide a theoretically rigorous formulation of the FGVC task and formally prove the necessary and sufficient conditions a learner must satisfy for FGVC, (2) introduce a plug-and-play extension on top of conventional CNNs that helps leverage relationships between global and local views of an object in the representation space for obtaining a complete encoding of the fine-grained semantic information in an input image, (3) achieve state-of-the-art results on all benchmark FGVC datasets with significant accuracy gains.

2 Related Work

Fine-grained visual categorization Prior works have demonstrated the importance of learning localized image features for FGVC [1, 50, 23], with extensions exploiting the relationship between multiple images and between network layers [25]. The high intra-class and low inter-class variations in FGVC datasets can be tackled by designing appropriate inductive biases like normalized object poses [4] or via more data-driven methods like deep metric learning [7]. Analysing part-specific features along with the global context was demonstrated through part detection based on activation regions in CNN feature maps [16, 48] or via context-aware attention pooling [3]. CNNs can also be modified in novel ways for FGVC by incorporating boosting [27], kernel pooling [8], or by randomly masking out a group of correlated channels during training [10]. Vision Transformers [40], with their ability to attend to specific informative image patches, have also shown great promise in FGVC [42, 13, 24]. To the best of our knowledge, we are the first to provide a rigorous theoretical foundation for FGVC and design a cross-view relational metric learning formulation based on the same.

Relation modelling in deep learning Modelling relationships between entities has proven to be a useful approach in many areas of deep learning including deep reinforcement learning [47], object detection [15], question answering [35], graph representation learning [2], few-shot learning [37] and knowledge distillation [31]. The usefulness of modelling relationships between different views of the same image has been demonstrated in the self-supervised context by [33]. All the above works either leverage or aim to learn relationships between entities, the nature of which is assumed to be known *apriori*. Our work breaks free from such assumptions by modelling cross-view relationships as learnable representations that optimize the end-task of FGVC.

Proxy-based deep metric learning Motivated by the fact that pairwise losses for deep metric learning incur a significant computational overhead leading to slow convergence, the idea of using proxies for learning metric spaces was first proposed in [28] and enhanced in [38]. Proxies can also be used to emulate properties of pairwise losses by capturing data-to-data relations (instead of just data-to-proxy) leveraging relative hardness of datapoints [18], by making data representations follow the semantic hierarchy inherent in real-world classes [45], or by regularizing sample distributions around proxies to follow a non-isotropic distribution [34]. However, all the above works perform proxy-based metric learning directly on data representations. In contrast, our approach is designed to learn class proxies that can be used not only to capture isolated, view specific (local/global) information for the underlying class, but also to learn the cross-view *relationships* such that they form embeddings in a metric space.

3 Relational Proxies

Consider an image $\mathbf{x} \in \mathbb{X}$ with a label $\mathbf{y} \in \mathbb{Y}$. Let $\mathbf{g} = c_g(\mathbf{x})$ and $\mathbb{L} = \{\mathbf{l}_1, \mathbf{l}_2, ... \mathbf{l}_k\} = c_l(\mathbf{x})$ be the global and set of local views of an image $\mathbf{x}$ respectively, where c_g and c_l are cropping functions applied on $\mathbf{x}$ to obtain such views. Let f be an encoder that takes as input $\mathbf{v} \in \{\mathbf{g}\} \cup \mathbb{L}$ and maps it to a latent space representation $\mathbf{z} \in \mathbb{R}^d$, where d is the representation dimensionality. Specifically, the representations of the global view $\mathbf{g}$ and local views $\mathbb{L}$ obtained from f are then denoted by $\mathbf{z}_g = f(\mathbf{g})$ and $\mathbb{Z}_\mathbb{L} = \{f(\mathbf{l}) : \mathbf{l} \in \mathbb{L}\} = \{\mathbf{z}_{l_1}, \mathbf{z}_{l_2}, ... \mathbf{z}_{l_k}\}$ respectively. Let $R : (\mathbf{g}, \mathbb{L}) \rightarrow \mathbf{r}$ be a random variable that encodes the relationships $\mathbf{r}$ between the global ($\mathbf{g}$) and the set of local ($\mathbb{L}$) views.

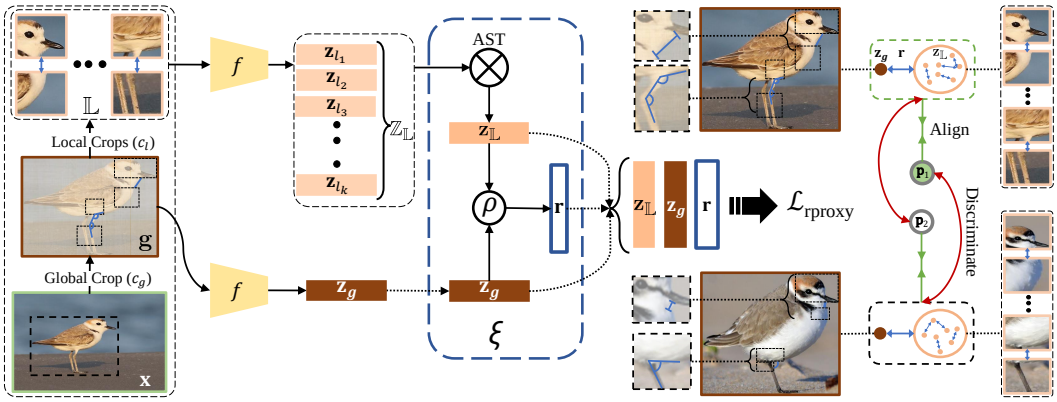


Figure 1: We start by encoding the global and local views using a relation-agnostic encoder f. We then compute the cross-view relational embedding $\mathbf{r}$ between the global $\mathbf{z}_g$ and the summary of local $\mathbf{z}_{\mathbb{L}}$ representations. The AST, in conjunction with ρ, form the cross-view relational function ξ. Finally, the learning of our Relational Proxies is conditioned by both view-specific ($\mathbf{z}_{\mathbb{L}}$ and $\mathbf{z}_g$) and cross-view relational ($\mathbf{r}$) information. Minimizing $\mathcal{L}_{\text{rproxy}}$ helps to align representations from the same category, while discriminating across different categories in a metric space.

3.1 Problem Definition

We leverage the qualitative consistency in the definition of the fine-grained visual categorization (FGVC) problem in the relevant literature [25, 48, 13, 3] to formalize the same in more quantitative terms as follows.

Definition 1 (**k-distinguishability**). *Two categories $\mathcal{C}_1$ and $\mathcal{C}_2$ are said to be k-distinguishable iff along with the global view, a classifier needs at least k local features to tell them apart, i.e., the true hypothesis can only distinguish between $\mathcal{C}_1$ and $\mathcal{C}_2$ if it has access to the complete set $\{\mathbf{z}_g\} \cup \mathbb{Z}_{\mathbb{L}}$, and it fails to distinguish between $\mathcal{C}_1$ and $\mathcal{C}_2$, if it only has access to $\{\mathbf{z}_g\} \cup \mathbb{Z}_{\mathbb{L}} \backslash \mathbf{z}_l, \forall \mathbf{z}_l \in \mathbb{Z}_{\mathbb{L}}$.*

The notion of *k-distinguishability* formalizes what it means for two categories to only be distinguishable in the fine-grained but not in the coarse-grained setting. Given the concept of k-distinguishability, the definition of FGVC problem directly follows from here:

Definition 2 (**Fine-Grained Visual Categorization Problem** - $\mathcal{P}_{\text{FGVC}}$). *A categorization problem is said to belong to the $\mathcal{P}_{FGVC}$ family, iff there exists at least one pair of categories $\mathcal{C}_1$ and $\mathcal{C}_2$ such that they are k-distinguishable.*

Unless otherwise stated, all datapoints $(\mathbf{x}, \mathbf{y})$ are considered to be sampled from k-distinguishable categories of an instance of $\mathcal{P}_{\text{FGVC}}$. In the subsequent sections, we prove that for a learner to completely model the class distribution for an instance of $\mathcal{P}_{\text{FGVC}}$, it must, alongside the view specific representations $\mathbf{z}_g$ and $\mathbb{Z}_{\mathbb{L}}$, also learn a function ξ that models the cross-view relationship between the global and the local views. Thus, a function ξ, to model R, must satisfy the following properties: (1) *View-Unification*: Maps the set of all views $\{\mathbf{g}, \mathbf{l} \in \mathbb{L}\}$ of an image $\mathbf{x}$ to a single output $\mathbf{r}$; (2) *Permutation Invariance*: Produces the same output irrespective of the order of the local attributes, *i.e.*, $\xi(\mathbf{z}_g, \{\mathbf{z}_{l_1}, \mathbf{z}_{l_2}, \dots \mathbf{z}_{l_k}\}) = \xi(\mathbf{z}_g, \{\mathbf{z}_{l_{\pi(1)}}, \mathbf{z}_{l_{\pi(2)}}, \dots \mathbf{z}_{l_{\pi(k)}}\})$, for every permutation π, where $\mathbf{z}_g$ and $\mathbf{z}_{l_i}$ are the representations of the global and the local views respectively, obtained from f. We provide more details on the necessity of these properties in Section 6.1 of the Appendix.

3.2 Relation-Agnostic Representations and Information Gap

In this section, we formally study the nature of the representation spaces learned by models that do not consider the cross-view relational information in the context of $\mathcal{P}_{\text{FGVC}}$. We term such representations as being "relation-agnostic" and prove via Proposition 1 that they suffer from an Information Gap, and thus are unable to capture the complete label information encoded in an input image.

Definition 3 (**Relation-Agnostic Representations - Information Theoretic**). *An encoder is said to produce relation-agnostic representations if it independently encodes the global view $\mathbf{g}$ and local views $\mathbf{l} \in \mathbb{L}$ of $\mathbf{x}$ without considering their relationship information $\mathbf{r}$.*

Lemma 1. *Given a relation-agnostic representation $\mathbf{z}$ of $\mathbf{x}$, the conditional mutual information between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$ can be reduced to $I(\mathbf{x}; \mathbf{r}|\mathbf{z})$.*

Proof. Given a relation-agnostic representation $\mathbf{z}$ of $\mathbf{x}$, the only uncertainty that remains about the label information $\mathbf{y}$ can be quantified as the cross-view relational information $\mathbf{r}$, *i.e.*, $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r})$. The proof of this statement is given in Identity 1 of the Appendix.

Intuitively, the conditional mutual information between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$, *i.e.*, $I(\mathbf{x};\mathbf{y}|\mathbf{z})$ represents the information for predicting $\mathbf{y}$ from $\mathbf{x}$ that $\mathbf{z}$ is unable to capture. Since $\mathbf{z}$ is relation-agnostic, the only uncertainty that remains in $\mathbf{x}$ after $\mathbf{z}$ is the cross-set relationship between the global and the local views, *i.e.*, $\mathbf{r}$. Therefore, we can write $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r})$. Using this equality and further factorizing $I(\mathbf{x};\mathbf{r})$ using the chain rule for mutual information, we get:

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}) + I(\mathbf{r};\mathbf{z}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}),$$

the latter equality following from Definition 3, which implies that $I(\mathbf{r};\mathbf{z}) = 0$, since $\mathbf{z}$ does not explicitly model the local-to-global relationships $\mathbf{r}$. □

Lemma 2. *The mutual information between* $\mathbf{x}$ *and its relation-agnostic representation* $\mathbf{z}$ *does not change with the knowledge of* $\mathbf{r}$.

Proof. Following the chain rule [12], the mutual information between $\mathbf{x}$ and $\mathbf{z}$, *i.e.*, $I(\mathbf{x};\mathbf{z})$ can be expressed as $I(\mathbf{x};\mathbf{z}|\mathbf{r}) + I(\mathbf{z};\mathbf{r})$. However, since $\mathbf{z}$ is *relation-agnostic* (Definition 3), $I(\mathbf{z};\mathbf{r}) = 0$. Thus, $I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{z}|\mathbf{r}) + I(\mathbf{z};\mathbf{r}) = I(\mathbf{x};\mathbf{z}|\mathbf{r})$. □

Proposition 1. *For relation-agnostic representation* $\mathbf{z}$ *of* $\mathbf{x}$*, the label information encoded in* $\mathbf{z}$ *is strictly upper-bounded by the label information in* $\mathbf{x}$*, i.e.,* $I(\mathbf{x};\mathbf{y}) > I(\mathbf{z};\mathbf{y})$ *by an amount* $I(\mathbf{x};\mathbf{r}|\mathbf{z})$.

Proof. The mutual information $I(\mathbf{x};\mathbf{y})$ between a datapoint $\mathbf{x}$ and its ground-truth label $\mathbf{y}$ can be expressed as $I(\mathbf{x};\mathbf{y}|\mathbf{z}) + I(\mathbf{x};\mathbf{z})$ based on the chain rule. Here $I(\mathbf{x};\mathbf{y}|\mathbf{z})$ represents the information for predicting $\mathbf{y}$ from $\mathbf{x}$ that $\mathbf{z}$ is unable to capture, while $I(\mathbf{x};\mathbf{z})$ denotes the predictive information that $\mathbf{z}$ does capture from $\mathbf{x}$. We can thus rewrite $I(\mathbf{x};\mathbf{y})$ using Lemma 1 and Lemma 2 as:

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{x};\mathbf{y}|\mathbf{z}) + I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}) + I(\mathbf{x};\mathbf{z}|\mathbf{r}) \tag{1}$$

Now, using the chain rule of mutual information, $I(\mathbf{z};\mathbf{y}) = I(\mathbf{z};\mathbf{y}|\mathbf{x}) + I(\mathbf{z};\mathbf{x})$. However, as a consequence of the data processing inequality [12], $I(\mathbf{z};\mathbf{y}|\mathbf{x}) = 0$ (since $\mathbf{z}$ cannot encode any more information about $\mathbf{y}$ than $\mathbf{x}$). Applying this and Lemma 2 to Equation (1):

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}) + I(\mathbf{x};\mathbf{z}|\mathbf{r}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}) + I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{r}|\mathbf{z}) + I(\mathbf{z};\mathbf{y})$$

Therefore, $I(\mathbf{x};\mathbf{y}) > I(\mathbf{z};\mathbf{y})$, by an amount $I(\mathbf{x};\mathbf{r}|\mathbf{z})$. □

Intuition: By establishing a strict upper-bound, Proposition 1 shows that relation-agnostic encoders cannot fully capture the label information in an input image. The quantity they are unable to capture is given by $I(\mathbf{x};\mathbf{r}|\mathbf{z})$, which we call the Information Gap.

3.3 Sufficient Learner

Proposition 1 states that the information gap exists *if* the representation space happens to be relation-agnostic. We now explore if there is really the need to learn relation-agnostic representations in the first place. From there, we identify the necessary and sufficient conditions for a complete learning of $I(\mathbf{x};\mathbf{y})$, and derive the requirements for a learner to do the same.

Definition 4 (**Relation-Agnostic Representations - Geometric**). *Let* $n_\epsilon(\cdot)$ *represent the* ϵ*-neighbourhood around a point in the limit* $\epsilon \to 0$ [2]*. A representation space is relation-agnostic if and only if* $\forall \mathbf{z}_l \in \mathbb{Z}_L : n_\epsilon(\mathbf{z}_l) \cap n_\epsilon(\mathbf{z}_g) = \phi$.

An intuitive explanation of Definition 4 can be found in Section 6.3 of the Appendix.

Axiom 1. f *learns representations* $\mathbf{z}$ *such that a classifier operating on the domain of* $\mathbf{z}$ *learns a distribution* $\hat{\mathbf{y}}$*, minimizing its cross-entropy with the true distribution* $-\sum_i \mathbf{y}_i \log(\hat{\mathbf{y}}_i)$*, where* i *denotes the* i*-th class.*

Lemma 3. *For an instance of* $\mathcal{P}_{FGVC}$*, the representation space learned by* f *is relation-agnostic, i.e., the global view* $\mathbf{g}$ *and the set of local views* $\mathbf{l} \in \mathbb{L}$ *are mapped to disjoint locations in the representation space.*

[2] The choice of ϵ determines the degree of relation-agnosticity of the representation space.

Proof. From Definition 4, a representation space is not relation-agnostic *iff* $\exists \mathbf{z}_l \in \mathbb{Z}_L : n_\epsilon(\mathbf{z}_l) \cap n_\epsilon(\mathbf{z}_g) \neq \phi$. Under this condition, the classifier only has the information from $\{\mathbf{z}_g\} \cup \mathbb{Z}_L \backslash \mathbf{z}_l$ instead of the required $\{\mathbf{z}_g\} \cup \mathbb{Z}_L$. Thus, for instances of $\mathcal{P}_{\text{FGVC}}$, according to Definition 1, removing the relation-agnostic nature from the representation space of f would cause a downstream classifier to produce misclassifications across the instances of k-distinguishable categories, leading to a violation of Axiom 1. Hence, f can only learn relation-agnostic representations. □

We can thus conclude from Lemma 3 and Proposition 1 that the necessary and sufficient conditions for a learner to capture the complete label information $I(\mathbf{x}; \mathbf{y})$, are to consider both (1) the relation-agnostic information $\mathbf{z}$ and (2) the cross-view relational information $\mathbf{r}$.

Proposition 2. *An encoder f trained to learn relation-agnostic representations $\mathbf{z}$ of datapoints $\mathbf{x}$ cannot be used to model the relationship $\mathbf{r}$ between the global and local views of $\mathbf{x}$.*

Proof. $f : \mathbf{x}_v \rightarrow \mathbf{z}_v$ is a unary function that takes as input a (global or local) view $\mathbf{x}_v$ of an image $\mathbf{x}$ and produces view-specific (Lemma 3) representations $\mathbf{z}_v$ for a downstream function $g : \mathbf{z}_v \rightarrow \mathbf{y}$.

For f to model the cross-view relationships, it must output the same vector $\mathbf{r}$ irrespective of whether $\mathbf{x}_v = \mathbf{g}$ or $\mathbf{x}_v = \mathbf{l} \in \mathbb{L}$, *i.e.* whether $\mathbf{x}_v$ is a global or a local view of the input image $\mathbf{x}$ (*view-unification* property of ξ). However, Lemma 3 prevents this from happening by requiring the output space of f to be relation-agnostic. Hence, f cannot be used to model $\mathbf{r}$. □

Thus, to bridge the information gap, a learner must have distinct sub-models that individually satisfy the properties of being relation-agnostic and relation-aware. Only such a learner could qualify as being sufficient for an instance of $\mathcal{P}_{\text{FGVC}}$.

Intuition: In this section, we have effectively proven that the properties of relation-agnosticity and relation-awareness are dual to each other. We show that while relation-agnosticity is not sufficient, it is a necessary condition for encoding the complete label information $I(\mathbf{x}; \mathbf{y})$. We also show that a disjoint encoder cannot be used to model the two properties alone without violating one of the necessary criteria. The requirement of a separate, relation-aware sub-model follows from here.

3.4 Learning Relation-Agnostic and Relation-Aware Representations

Figure 1 depicts the end-to-end design of our framework. Derived from our theoretical findings, it comprises of both the relation-agnostic agnostic encoder f, and the cross-view relational function, ξ, expressed as a composition of the Attribute Summarization Transformer, AST, and a network for view-unification, ρ. Below, we elaborate on each of these components.

Relation-Agnostic Representations: We follow recent literature [43, 48] for localizing the object of interest in the input image $\mathbf{x}$ and obtaining the global view $\mathbf{g}$ by thresholding the final layer activations of a CNN encoder f and detecting the largest connected component in the thresholded feature map. We obtain the set of local views $\{\mathbf{l}_1, \mathbf{l}_2 \ldots, \mathbf{l}_k\}$ as sub-crops of $\mathbf{g}$ (more details in Section 4.1). Following the primary requirement of Proposition 1, we produce relation-agnostic representations by propagating $\mathbf{g}$ and $\mathbf{l}_i$ through a CNN encoder f that independently encodes the two view families as $\mathbf{z}_g = f(\mathbf{g})$ and $\mathbf{z}_{l_i} = f(\mathbf{l}_i)$.

Relational Embeddings: The second requirement, according to Proposition 1, for completely learning $I(\mathbf{x}; \mathbf{y})$ is to minimize $I(\mathbf{x}; \mathbf{r}|\mathbf{z})$, *i.e.*, the uncertainty about the relational information $\mathbf{r}$ encoded in $\mathbf{x}$, given a relation-agnostic representation $\mathbf{z}$. However, according to Proposition 2, we cannot perform the same using the relation-agnostic encoder f. Contrary to existing relational learning literature [31, 33] that assumes the nature of relationships to be known beforehand, we take a novel approach that models cross-view relationships as learnable representations of the input $\mathbf{x}$. We follow the definition of the relationship modelling function $\xi : (\mathbf{g}, \mathbb{L}) \rightarrow \mathbf{r}$, that takes as input relation-agnostic representations of the global view $\mathbf{z}_g$ and the set of local views $\mathbb{Z}_\mathbb{L} = \{\mathbf{z}_{\mathbf{l}_1}, \mathbf{z}_{\mathbf{l}_1}, \ldots \mathbf{z}_{\mathbf{l}_\mathbf{k}}\}$, and outputs a relationship vector $\mathbf{r}$, satisfying the *View-Unification* and *Permutation Invariance* properties.

We satisfy the Permutation Invariance property by aggregating the local representations via a novel Attribute Summarization Transformer (AST). We form a matrix whose columns constitute a learnable summary embedding $\mathbf{z}_\mathbb{L}$ followed by the local representations $\mathbf{z}_{l_i}$ as $\mathbf{Z}'_\mathbb{L} = [\mathbf{z}_\mathbb{L}, \mathbf{z}_{\mathbf{l}_1}, \mathbf{z}_{\mathbf{l}_1}, \ldots \mathbf{z}_{\mathbf{l}_\mathbf{k}}]$. We compute the self-attention output $\mathbf{z}'_*$ for each column $\mathbf{z}_*$ in $\mathbf{Z}'_\mathbb{L}$ as $\mathbf{z}'_* = \mathbf{a} \cdot \mathbf{Z}_\mathbb{L}\mathbf{W}$, where $\mathbf{a} = \sigma\left((\mathbf{z}_*\mathbf{W}_q) \cdot (\mathbf{Z}_\mathbb{L}\mathbf{W})^T / \sqrt{D}\right)$, and D is the embedding dimension. By iteratively performing

self-attention operations among the columns of $\mathbf{Z}'_{\mathbb{L}}$, AST aggregates information across all the local attributes into the final learnable output of $\mathbf{z}_{\mathbb{L}}$. Unlike the usual vision transformer [40], we omit the usage of positional embeddings, as doing so provides better permutation invariance [29].

For satisfying the View-Unification property, we introduce a simple feed-forward multilayer perceptron that learns the mapping $\rho : (\mathbf{z_g}, \mathbf{z}_{\mathbb{L}}) \rightarrow \mathbf{r}$. It takes as input the representation of the global view $\mathbf{z}_g$ and the summary of the set of local views $\mathbf{z}_{\mathbb{L}}$, and outputs the relationship as a learned vector $\mathbf{r}$. Thus, in our construction, the AST along with ρ, constitute the relation modelling function ξ.

Learning Relational Proxies: The representations $\mathbf{z}_g, \mathbf{z}_{\mathbb{L}}$ and r in unison encode the full semantic information $\mathbf{y}$ in $\mathbf{x}$ (Proposition 1). To alleviate the low inter-class variance in $\mathcal{P}_{\text{FGVC}}$, metric learning has been shown to be an effective [7] approach. Furthermore, approaches like [28] and [18] for metric learning have shown that substituting pairwise comparisons with assignment to a fixed set of learnable class proxies reduces the training-time complexity from a large polynomial like $\mathcal{O}(n^2)$ or $\mathcal{O}(n^3)$ to near linear $\mathcal{O}(c.n)$, where c is the number of classes in a dataset, n is the number of train-set datapoints, and $c \ll n$. For this purpose, we contrast instance representations across classes through class proxy vectors that are informed by both the view specific and relational representations via learning the conditional distribution $p(\mathbf{y}|\mathbf{z}_g, \mathbf{z}_{\mathbb{L}}, \mathbf{r})$. We term such class proxies, Relational Proxies, as they leverage cross-view relationship information for encoding class semantics.

Consider a set of c learnable class proxy vectors $\mathbb{P} = \{\mathbf{p}_1, \mathbf{p}_2, ... \mathbf{p}_c\}$, where c is the number of fine-grained classes. Here, we present a novel formulation of the proxy-anchor loss [18] in cross-entropic terms that allows us to conform to the requirement of Axiom 1 in the fine-grained setting. Specifically, for each of the representations $\omega \in \{\mathbf{z}_g, \mathbf{z}_l, \mathbf{r}\}$ for all $\mathbf{x} \in \mathbb{X}$, we minimize the following:

$$\mathcal{L}_{\text{rproxy}} = -\frac{1}{c}\sum_{\mathbf{p}\in\mathbb{P}} \log\left(\frac{1}{\psi^+ \cdot \psi^-}\right), \quad \begin{cases} \psi^+ = 1 + \sum\limits_{\omega\in\Omega} e^{-\alpha(s(\omega,\mathbf{p})-\delta)} \\ \psi^- = 1 + \sum\limits_{\omega\in\bar{\Omega}} e^{\alpha(s(\omega,\mathbf{p})+\delta)} \end{cases} \tag{2}$$

where Ω is the set of representations in a mini-batch for which $\mathbf{p}$ is the true class proxy, $\bar{\Omega}$ is one for which $\mathbf{p}$ is not the true class proxy, and $s(\cdot,\cdot)$ computes the cosine distance. ψ^+ helps align matching $(\omega, \mathbf{p})$ pairs close together in the representation space (since $\mathcal{L}_{\text{rproxy}}$ follows a cross-entropic form, it does not violate the relation-agnosticity of f, as proven in Lemma 3, with a more detailed note in Section 6.4 of the appendix), while ψ^- helps embedding non-matching $(\omega, \mathbf{p})$ pairs farther apart. The scaling parameter α along with the margin parameter δ control the intensity with which the alignment and discrimination are performed. $1/(\psi^+ \cdot \psi^-)$ gives a probability indicating how closely the learned representation space reflects the semantic structure in $\mathbb{X}$. $\mathcal{L}_{\text{rproxy}}$ thus computes the cross-entropy loss between the ground-truth and the predicted class distributions over the set of proxies.

Inference: Given an input image $\mathbf{x}$, we compute its global ($\mathbf{z}_g$), summary of local ($\mathbf{z}_{\mathbb{L}}$), and relational ($\mathbf{r}$) representations using f, AST and ρ as explained above. We then predict the class probability distribution $\hat{\mathbf{y}}$ of these representations by computing their soft-assignment scores across the relational proxies. The assignment score for each proxy $\mathbf{p} \in \mathbb{P}$ is computed as follows:

$$\hat{\mathbf{y}}_{\mathbf{p}} = \sum_{\omega\in\{\mathbf{z}_L,\mathbf{z}_g,\mathbf{r}\}} \frac{e^{s(\omega,\mathbf{p})}}{\sum\limits_{\mathbf{p}'\in\mathbb{P}} e^{s(\omega,\mathbf{p}')}}$$

The class corresponding to the relational proxy with the highest assigned score is returned as the prediction.

4 Experiments

We now present the implementation details of Relational Proxy, and the results obtained upon evaluating it on benchmark FGVC datasets. We also discuss observations from ablation studies that we performed to validate our theoretical foundations, as well as the implementation specific choices that we made, along with qualitative visualizations of the learned cross-view local relationships.

4.1 Experimental Settings and Datasets

Implementation details – We implement our Relational Proxy model using the PyTorch [32] deep learning framework, on an Ubuntu 20.04 workstation with a single NVIDIA GeForce RTX 3090 GPU,

Method	Benchmark					Cultivar	
	FGVC Aircraft	Stanford Cars	CUB	NA Birds	iNaturalist	Cotton	Soy
MaxEnt [11] NeurIPS'18	89.76	93.85	86.54	-	-	-	-
DBTNet [51] NeurIPS'19	91.60	94.50	88.10	-	-	-	-
StochNorm [19] NeurIPS'20	81.79	87.57	79.71	74.94	60.75	45.41	38.50
MMAL [48] MMM'21	94.70	95.00	89.60	87.10	69.85	65.00	47.00
FFVT [42] BMVC'21	79.80	91.25	91.65	89.42	70.30	57.92	44.17
CAP [3] AAAI'21	94.90	95.70	91.80	91.00	-	-	-
TransFG [13] AAAI'22	80.59	94.80	91.70	90.80	71.70	45.84	38.67
Ours (Relational Proxy)	**95.25** ± 0.02	**96.30** ± 0.04	**92.00** ± 0.01	**91.20** ± 0.02	**72.15** ± 0.03	**69.81** ± 0.04	**51.20** ± 0.02

Table 1: Comparison of classification accuracies obtained by our method (averaged over 5 independent runs) on standard FGVC datasets with current state-of-the-art approaches.

an 8-core Intel Xeon processor and 32 GBs of RAM. Since we proposed a proxy-based approach for learning the relational metric space, we do not have a dependency on batch-size for the purpose of negative sampling as part of our metric learning phase, which enables us to train our entire model end-to-end on a single GPU. Also by the virtue of using class-proxies, the convergence time is reduced by a significant amount compared to pairwise losses.

Hyperparameter settings – For initial training stability, we consider five disjoint locations (four corners and the centre) of $\mathbf{g}$ to be the set of local views. As training progresses, we also allow the model to learn from an increased number views obtained via random cropping. In the same way at inference time, the local views constitute a combination of the five disjoint crops along with some random crops. We found that the optimal number of local views k to be equal to 7 for FGVC Aircraft, Stanford Cars and both the cultivar datasets. For CUB and NABirds, $k = 8$ gave the best performance. We use ResNet50 [14] pretrained on ImageNet [9] as the backbone of our relation-agnostic encoder f. In Sec. 1.3 of the supplementary, we also provide evaluations using VGG-16 [36] to show that the performance gains achieved by our model do not depend on the specific backbone. We train our full Relational Proxy model end-to-end for 200 epochs using the stochastic gradient descent optimizer with an initial learning rate of 0.001 (decayed by a factor of 0.1 every 50 epochs), a momentum of 0.9, and a weight decay of 10^{-4}.

Datasets and Evaluation – We evaluate our model on the four most common fine-grained visual categorization benchmarks (number of classes and train/test splits respectively in brackets): FGVC Aircraft [26] (100 | 6667/3333), Stanford Cars [20] (196 | 8144/8041), CUB [41] (200 | 5994/5794), and NA Birds [39] (555 | 23,929/24,633). For large scale benchmark evaluation, we choose the iNaturalist 2017 dataset which consists of 13 super-categories that have been split into a total of 5089 fine-grained categories with 675,170 training and 182,707 test images. We also perform experiments on two challenging datasets of the cultivar domain that offer very low inter-class variations, namely Cotton Cultivar [46] (80 | 240/240) and Soy Cultivar [46] (200 | 600/600). We use classification accuracy as our metric for evaluating the performance of a model.

4.2 Comparison with State of the Art

Benchmark Datasets – In Table 1, we report the performance of our method on benchmark datasets along with existing SotA approaches. StochNorm [19] presents a novel way to refactor batch normalization that helps prevent overfitting for the task of FGVC. The novel training routine proposed in MaxEnt[11] improves FGVC performance by maximizing the entropy of the output probability distribution of a CNN. By designing a computationally inexpensive bilinear feature transformation mechanism for CNNs, DBT [51] achieves competitive performance on benchmark FGVC datasets. MMAL [48] is one of the most competitive models for FGVC Aircraft and Stanford Cars, which extracts the most informative global and local views by analyzing the activation maps of the final layer of a CNN, and embeds them in a relation-agnostic representation space. TransFG [13] proposes a vision transformer based technique for extracting informative local patches, achieving SotA performance on iNaturalist, and promising results on CUB and NA Birds. By learning a context aware attention pooling mechanism, CAP [3] reports SotA performance on all benchmark datasets other than iNaturalist. From Table 1, we see that our method surpasses the SotA on all four benchmarks by significant margins. Specifically, we beat the SotA on Stanford Cars by 0.60%, on iNaturalist by 0.45%, on FGVC Aircraft by 0.35%, and on both CUB and NA Birds by 0.20%.

Cars and Aircrafts can largely vary in color, texture and custom, part-specific styles within a category. However, the geometry of the overall object (represented by cross-view relationships) within a class remains fairly constant. This leaves room for a large amount of relational information to be captured. This also holds true for the iNaturalist dataset, as the local-to-global emergent relationships can be used to discriminate between both coarse-grained (super) and fine-grained (sub) categories. For the

ID	Relation-Agnostic Encoder	AST	RelationNet	Learnable Relation	Proxies	Aircraft	CUB	Stanford Cars
1.	✓					94.60	91.25	95.21
2.	✓	✓				94.91	91.50	95.62
3.	✓	✓	✓	✓		95.13	91.90	96.15
4.	✓			✓	✓	94.92	91.55	95.70
5.	✓	✓		✓	✓	95.10	91.81	96.05
6.	✓		✓	✓	✓	95.05	91.73	95.93
7.	✓	✓	✓	✓	✓	**95.25**	**92.00**	**96.30**

Table 2: Results of ablating the key components of our Relational Proxy model (sufficient learner). Grouped so as to better illustrate the effect of learning cross-view relationships. Note that the meaning of non-existence of a component may vary according to context / other row elements. Refer to the corresponding paragraph in Section 4.3 of the main text for further details.

ID	Attribute Summary ($\mathbf{z}_{\mathbb{L}}$)	Global Representation ($\mathbf{z}_g$)	Relational Representation ($\mathbf{r}$)	FGVC Aircraft	CUB	Stanford Cars
1.	✓	✓		94.91	91.50	95.62
2.	✓		✓	94.85	91.58	95.75
3.		✓	✓	94.60	91.47	95.51
4.	✓	✓	✓	**95.25**	**92.00**	**96.30**

Table 3: Results of ablating the Proxy Conditioning Representations. Note that only the output representation vectors were ablated here, and not the entire model component producing them. The latter has been studied independently with findings reported in Table 2.

bird datasets (CUB, NABirds), although this relational information is still there, most categories can be told apart by color, texture and local-attribute specific information, if they are clearly visible. For this reason, the accuracy gains obtained in the Cars and Aircraft datasets surpass those obtained for the birds.

Cultivar Datasets – For the highly challenging datasets of the cultivar domain, *i.e.*, Cotton and Soy Cultivar, FFVT [42] provides state-of-the-art results by using a specialized feature fusion technique for vision transformers. As can be seen in Table 1, our model, by leveraging cross-view relational embeddings, manages to provide a performance boost exceeding 4% over the current SOTA on the cultivar datasets. Cultivar datasets have very low inter-class differences. Cross-view relational information like edge curvature, relative angles between leaf sub-parts, width to height ratio, convergence patterns of leaf ends, etc., largely determine the uniqueness of a category. For this reason, our method is extremely effective when applied to such domains.

4.3 Ablation Studies

We perform the following three classes of ablation studies:

Key components of the sufficient learner – Table 2 shows the results of ablating the key components of our model. The relation agnostic encoder being the most fundamental component, cannot be removed, and therefore appears in all the rows. Row 1 thus represents training a simple classification head on top of the representations obtained from the relation-agnostic encoder. Row 2 denotes the result of aggregating the local views, computing a predefined relationship function, specifically the distance between the local and global representations, and minimizing a Huber loss between the relational distance value between instances of the same class. Row 3 introduces the idea of learnable relational vectors (instead of predefined functions like distances). Since cross-view relationships are unique to a class, we aim to embed the relational vectors in a metric space by minimizing a pairwise contrastive loss across classes. However, as noted in recent metric learning literature [28, 18], computing pairwise losses can be expensive and lead to slower convergence. Motivated by this, we introduce the idea of relational proxies in Rows 4 - 7. Row 4 replaces the AST and RelationNet by simple concatenation of the inputs and propagation through a linear layer. Row 5 and 6 individually show the effects of replacing the AST and RelationNet with linear layers. Finally, Row 7 denotes the performance of our model with all components included.

Rows 2 and 3 demonstrate the importance of modelling cross-view relationships specifically as a learnable metric space embedding. Rows 5, 6 and 7 show the contribution of our AST in summarizing the local attributes, as well as the fact that the cross-view relationship is non-linear in nature.

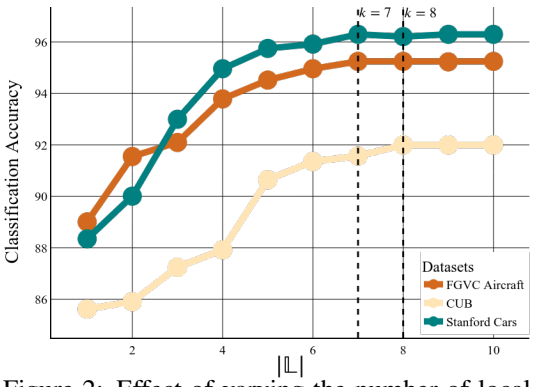


Figure 2: Effect of varying the number of local views $|\mathbb{L}|$

Method	T-ImageNet	D-ImageNet
$\mathbf{z}_g, \mathbf{z}_{\mathbb{L}}$	88.75	91.30
$\mathbf{z}_g, \mathbf{z}_{\mathbb{L}}, \mathbf{r}$	88.91	92.75
Δ	0.16	**1.45**

Table 4: Comparison of accuracy gains obtained by using the relational information $\mathbf{r}$ along with the view-specific representations $\mathbf{z}_g$ and $\mathbf{z}_{\mathbb{L}}$ (second row) over using only relation-agnostic representations (first row) on the coarse-grained Tiny (T) and fine-grained Dogs (D) subsets of ImageNet. As can be observed, the performance gain Δ by accounting for the relational information is significantly more in the fine-grained setting than in the coarse-grained one.

Conditioning of the relational proxies – The relational proxies in our model are conditioned by three representations of the input $\mathbf{x}$ – the summary of the local attributes $\mathbf{z}_{\mathbb{L}}$, the representation of the global view $\mathbf{z}_g$, and the relational vector $\mathbf{r}$. We study the contribution of each of these representations and summarize our findings in Table 3. These results demonstrate that the information encoded in all three representations are necessary for learning the complete set of class attributes.

Results on ImageNet subsets – In order to validate whether our findings are in fact particularly applicable to the fine-grained setting, we perform experiments to compare the performance boost provided by our method over a vanilla relation-agnostic encoder, between coarse-grained (Tiny ImageNet[22]) and fine-grained (Dogs ImageNet / Stanford Dogs [17]) subsets of ImageNet. Our findings are summarized in Table 4, which shows that our method does in fact provide a more significant improvement over a relation-agnostic encoder in the fine-grained setting.

The optimal value of k for the k-distinguishability criterion – Figure 2 empirically illustrates the idea of k-distinguishability for a given local-crop size on the FGVC Aircraft, Stanford Cars and CUB datasets. For an instance of $\mathcal{P}_{\text{FGVC}}$, the performance of a model is strongly dependent on the number of local views it has access to. When the number of local views $|\mathbb{L}|$ is less than the minimum required number k, the classification performance is poor as the model does not have access to the minimum set of required fine-grained information. As $|\mathbb{L}|$ approaches k, the performance increases, reaching its maximum at k (= 7 for FGVC Aircraft and Stanford Cars, and 8 for CUB). However, if $|\mathbb{L}|$ is increased beyond k, there is no further gain in performance, as the extra information is either redundant or semantically irrelevant.

Correlation between $|\mathbb{L}|$ and local patch size – To determine the right computational trade-offs for our method, we perform a study to identify possible correlations between the number of local views $\mathbb{L}$ and size of local patches. We trained our model on FGVC Aircraft [26] by varying the number of local views $|\mathbb{L}|$ and the size of each local patch to identify their correlations. We present our results in Table 5, where rows represent the number of local views $|\mathbb{L}|$ and the columns represent the side-length of each local patch. So, if the global view has spatial dimensions $N \times N$, each local patch would be of $N/t \times N/t$, where t is the scaling factor that is varied across the columns. In summary, the rows represent increasing the number of local views top-down, and the columns represent increasing the patch-size left-to-right. The numbers are expressed as relative deviations from a reference of 95.25%, i.e., the setting corresponding to our reported accuracy for FGVC Aircraft in Table 1.

From Table 5, we can see that increasing the patch size beyond a certain point has a detrimental effect as the local views tend to lose their granularity and degenerate into global views. Increasing the number of crops has a stronger improvement effect on performance if the patch size is small, thus influencing the value of k accordingly. However, decreasing the patch size at the cost of an increased number of local views also has its downsides - the number of attention computations in the attribute summarization step increases quadratically. Thus $|\mathbb{L}|$ and the local patch size needs to be determined based on application specific accuracy requirements and the available computational resources.

$\lvert\mathbb{L}\rvert$	$\mathbf{N/5}$	$\mathbf{N/4}$	$\mathbf{N/3}$	$\mathbf{N/2}$
7	-0.03	-0.02	0.00	-0.14
12	+0.01	+0.02	0.00	-0.11
15	+0.05	+0.03	+0.01	-0.11
18	+0.05	+0.03	+0.00	-0.10

Table 5: Correlation between number of local views and size of local patches

4.4 Visual Representations of Cross-View Local Relationships

Our AST-based aggregation scheme allows us to visualize the local relationships that lead to the emergence of the global-view. We aim to construct a graph of local views for depicting the cross-view local relationships. The graph represents the manner in which the local views combine to form the overall object. The nodes of the graph represent the local views. Two nodes are connected via an edge if there exists a relationship between them. The thickness of the edges in the illustration is proportional to the degree of relatedness. We compute the topology of this graph by analyzing the final layer mutual attention values of the Attribute Summarization Transformer (AST). We add an edge between two local views if their mutual attention score is higher than a threshold (which we choose to be the average of all pairwise attention scores). The weight of the edge is proportional to the magnitude of attention. For the purpose of simplicity, we depict fewer local views in the visualization, than are actually used for computation. Figure 3 shows example graphs on images from the Stanford Cars dataset. In Appendix 6.5, we provide more such qualitative results and based on these graphs, we provide an analysis of scenarios under which even relational information cannot distinguish between certain fine-grained categories.

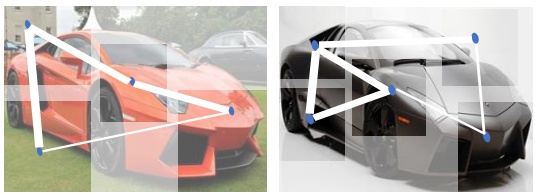


Figure 3: Visualization of the learned relationships across local-views. Despite close similarities between the two car categories, it can be seen that our model leverages discriminative cross-view relationships to tell the instances apart.

5 Conclusion and Discussion

Starting with the idea of k-distinguishability, we derived the necessary and sufficient conditions that a model must satisfy in order to completely capture the fine-grained information in an image. We proved that a model needs to simultaneously encode both view-specific and cross-view relational properties of an object in order to bridge the information gap that its representations have with the semantic content in the input image. Based on our theoretical findings, we designed Relational Proxies, a method that achieves state-of-the-art results on benchmark FGVC datasets by learning class representations conditioned with cross-view relationships. By introducing a theoretically rigorous framework, we believe that our work opens up new avenues for studying the problem of FGVC in a more systematic manner. One immediate potential outcome of our work that we foresee is the development of explainable fine-grained features. Such features can be used for computing a minimal set of fine-grained attributes to limit compute time/resources, or to perform tasks like cross-modal retrieval in domains with large modality gap.

Limitations – The process of obtaining local views in our method is somewhat of an uninformed, generic cropping methodology on the global view of the object, which may not necessarily always yield the best set of local object parts. More informed ways of detecting novel object parts from which the global view emerges can lead to obtaining at par performance but with fewer local views.

Societal Impacts – The rigorous theoretical basis of our work has a positive societal impact, which not only makes our methodology transparent and easy to analyze, but also provides a framework to study the foundations of FGVC in general. So far, we are not aware of any negative societal impact that is specific to our methodology. However, as with all data-driven approaches, underlying biases in the datasets on which our model is trained would influence the patterns learned by it.

Acknowledgements

This work has been partially supported by the ERC 853489–DEXIM, by the DFG–EXC number 2064/1–Project number 390727645, and as part of the Excellence Strategy of the German Federal and State Governments.

References

[1] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013.

[2] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv*, 6 2018.

[3] Ardhendu Behera, Zachary Wharton, and Asish Bera. Context-aware Attentional Pooling (CAP) for Fine-grained Visual Classification. In *AAAI*, 2021.

[4] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *NeurIPS*, 2021.

[6] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised Part Discovery from Contrastive Reconstruction. In *NeurIPS*, 2021.

[7] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *CVPR*, 2016.

[8] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

[10] Yifeng Ding, Shuwei Dong, Yujun Tong, Zhanyu Ma, Bo Xiao, and Haibin Ling. Channel DropBlock: An Improved Regularization Method for Fine-Grained Visual Classification. In *BMVC*, 2021.

[11] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeurIPS*, 2018.

[12] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*, 2020.

[13] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. TransFG: A Transformer Architecture for Fine-grained Recognition. In *AAAI*, 2022.

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[15] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation Networks for Object Detection. In *CVPR*, 2018.

[16] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, 2016.

[17] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPRW*, 2011.

[18] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, 2020.

[19] Zhi Kou, Kaichao You, Mingsheng Long, and Jianmin Wang. Stochastic normalization. In *NeurIPS*, 2020.

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *3DRR*, 2013.

[21] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2017.

[22] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Tiny image net. `https://www.kaggle.com/c/tiny-imagenet`, 2017.

[23] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 2015.

[24] Di Lu, Jinpeng Wang, Ziyun Zeng, Bin Chen, Shudeng Wu, and Shu-Tao Xia. SwinFGHash: Fine-grained Image Retrieval via Transformer-based Hashing Network. In *BMVC*, 2021.

[25] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry Davis, Jun Li, Jian Yang, and Ser Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019.

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv*, 2013.

[27] Mohammad Moghimi, Mohammad Saberian, Jian Yang, Li Jia Li, Nuno Vasconcelos, and Serge Belongie. Boosted convolutional neural networks. In *BMVC*, 2016.

[28] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No Fuss Distance Metric Learning Using Proxies. In *ICCV*, 2017.

[29] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. In *NeurIPS*, 2021.

[30] Timothy O'Connor. Emergent Properties. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[31] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPSW*, 2017.

[33] Massimiliano Patacchiola and Amos Storkey. Self-supervised relational reasoning for representation learning. In *NeurIPS*, 2020.

[34] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy Regularization for Proxy-based Deep Metric Learning. In *CVPR*, 2022.

[35] Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. In *CVPR*, 2018.

[38] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. ProxyNCA++: Revisiting and Revitalizing Proxy Neighborhood Component Analysis. In *ECCV*, 2020.

[39] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. An image is worth 16x16 words: transformers for image recognition at scale. In *NeurIPS*, 2017.

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. In *California Institute of Technology, CNS-TR-2010-001*, 2011.

[42] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. In *BMVC*, 2021.

[43] Xiu Shen Wei, Jian Hao Luo, Jianxin Wu, and Zhi Hua Zhou. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. *IEEE TIP*, 2017.

[44] Xiu Shen Wei, Chen Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *PR*, 2018.

[45] Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Doug Gray, and Dimitris Samaras. Hierarchical Proxy-based Loss for Deep Metric Learning. In *WACV*, 2022.

[46] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Patchy image structure classification using multi-orientation region transform. In *AAAI*, 2020.

[47] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, and Peter Battaglia. Deep reinforcement learning with relational inductive biases. In *ICLR*, 2019.

[48] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization. In *MMM*, 2021.

[49] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, 2016.

[50] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

[51] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In *NeurIPS*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] Code and pre-trained models will be made public upon paper acceptance. Details of all experimental settings required to reproduce our results are provided in Section 4.1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] But the numbers we report are the means of 5 runs with different random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

6 Appendix

6.1 Properties of the Relationship Modelling Function

Intuitive Analogy: The problem of local-to-global relation computation can be viewed as a bit-string-to-integer matching problem. Consider 3 bits, say b_1, b_2 and b_3, corresponding to 3 local views. Let the global view be represented by an integer that can be encoded with 3 bits, say with a value of g = 6, for this example. The problem then is to find the association of the integer 6 with its corresponding binary representation of 110. This association represents the cross-view relationship.

The first step towards solving this problem is to *enumerate* all the possible ways in which the local views can combine (to produce any global view, not specifically g). The set of all such combinations will be given by $S = \{000, 001, 010, ..., 110, 111\}$. The bit values encode the presence or absence of a particular view in the cross-view relationship. So, no matter what order we observe b_1, b_2 and b_3 in, we must output the same set S, as it is required to be an exhaustive enumeration. This is exactly what the property of permutation invariance achieves. Once we have S, the next step is to *find* the mapping $S, g \mapsto 110$, i.e, the correct binary encoding for the integer $g = 6$, which is accomplished by the property of view-unification.

Purpose: As illustrated through the above analogy, one can view the local-to-global relationship modelling function as an enumerative search algorithm - given a set of local views, it first *enumerates* all possible ways in which they can combine to form a meaningful global view. Given that enumeration, it then *finds* the target solution by learning to identify the correct combination that matches with the global-view representation. Thus, the enumerate operation needs to be permutation invariant, as it has to consider all possible combinations of the inputs, and the find operation needs to be a view-unifier by construction.

Motivation: Behind our specific design choice was the motivation to keep the enumerate and find steps separate. This allows the model to have dedicated representation spaces for the two distinct sub-tasks, which in turn facilitates better convergence.

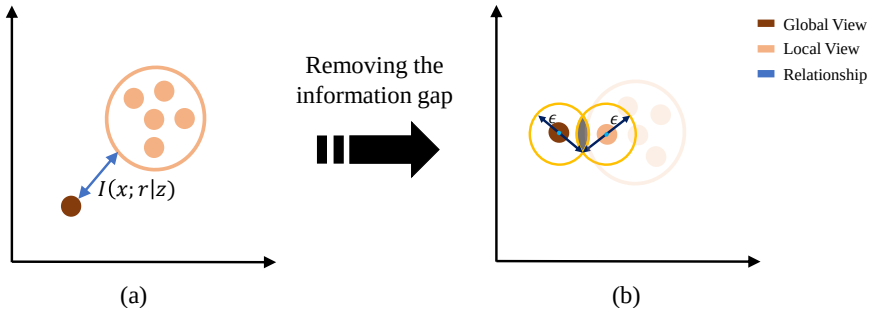


Figure 4: (a) A relation-agnostic representation space. (b) The ϵ-neighborhoods of the global and local views begin colliding as the information gap is reduced.

6.2 Proofs of Additional Identities

Identity 1. *Given a relation-agnostic representation* $\mathbf{z}$ *of* $\mathbf{x}$*, the only uncertainty that remains about the label information* $\mathbf{y}$ *can be quantified as the cross-view relational information* $\mathbf{r}$*, i.e.,* $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r})$.

Proof. Using the chain rule for mutual information [12], we can factorize the label information $\mathbf{y}$ contained in $\mathbf{x}$, *i.e.*, $I(\mathbf{x};\mathbf{y})$ as:

$$I(\mathbf{x};\mathbf{y}) = I(\mathbf{x};\mathbf{y}|\mathbf{z}) + I(\mathbf{x};\mathbf{z}) \tag{3}$$

As evidenced by recent literature [25, 48, 3, 6], the label information in $\mathbf{x}$ can be expressed exclusively as a function of its global ($\mathbf{g}$) and local ($\mathbf{l}_i$) views. Thus, in quantitative terms, the label information $\mathbf{y}$ in $\mathbf{x}$ can also be factorized into *relation-agnostic* and *relation-aware* components as follows:

$$I(\mathbf{x};\mathbf{y}) = \underbrace{I(\mathbf{x};\mathbf{g}) + \sum_{\mathbf{l}\in\mathbb{L}} I(\mathbf{x};\mathbf{l})}_{\text{relation-agnostic}} + \underbrace{I(\mathbf{x};\mathbf{r})}_{\text{relation-aware}} \tag{4}$$

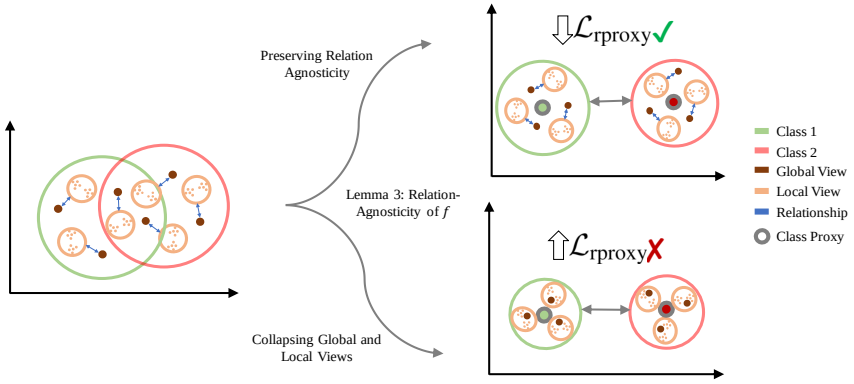


Figure 5: Embeddings of datapoints from two classes obtained from f before training (left). As f is trained with an end objective of minimizing $\mathcal{L}_{\text{rporxy}}$, it has two potential choices (right). However, as proven in Lemma 3, the relation-agnostic nature of f prevents the collapse of the global and local embeddings even when they share the same set of class proxies.

The relation-aware representation $\mathbf{r}$ is, unlike relation-agnostic representations, obtained explicitly based on the cross-view relationship. However, since f computes $\mathbf{z}$ without considering any relational information, it only models the relation-agnostic component of Equation (4). Thus,

$$I(\mathbf{x};\mathbf{z}) = I(\mathbf{x};\mathbf{g}) + \sum_{\mathbf{l}\in\mathbb{L}} I(\mathbf{x};\mathbf{l}) \tag{5}$$

Substituting the *relation-agnostic* component of Equation (4) with the L.H.S. of Equation (5), and comparing it with Equation (3), we get:

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = I(\mathbf{x};\mathbf{r}) \tag{6}$$

□

6.3 Geometric Relation Agnosticity

Definition 4 is based on the fact that the information gap (derived in Proposition 1) between the global and the local views has the effect that the two view families would be mapped to distinct locations in the representation space, and the separation between them would be proportional to the information gap, *i.e.*, $I(\mathbf{x};\mathbf{r}|\mathbf{z})$. Definition 4 also mentions that relation-agnostic embeddings of the local and the global views must thus be well separated, *i.e.*, the ϵ-neighborhood of the global embedding $n_\epsilon(\mathbf{z}_g)$ must not intersect with those of the local embeddings $n_\epsilon(\mathbf{z}_l)$. In other words, the global embedding must be sufficiently far apart from each of the local embeddings.

Figure 4 depicts the geometric effect of removing the information gap from a relation-agnostic representation space. As proven in Lemma 3, if the information gap is reduced using the same encoder f that was used to obtain $\mathbf{z_g}$ and $\mathbf{z}_l$, the model starts mapping the global and the local views to identical regions in the representation space. This could potentially lead to the requirement of k-distinguishability to not be satisfied, as the unique information pertaining to at least one of the local views is lost upon merger with the global view (and vice-versa). It is thus a requirement for a sufficient learner to preserve the relation-agnosticity in the representation space of f.

6.4 Relation-Agnosticity of Relational Proxies

The representations $\mathbf{z}_g$ and $\mathbf{z}_{l_i}$ are computed in a relation-agnostic manner and no explicit operation is performed to reduce the domain gap between the global and the set of local views. This natural domain gap thus manifests in the representation space of $\mathbf{z}$ as its relation-agnostic nature.

Figure 5 diagrammatically illustrates this idea. Given an entangled representation space where the classes are not entirely separable (left), the encoder f has two choices to map the local and global views of the corresponding datapoints to completely separable proxy neighborhoods. It could either:

1. Preserve the relation-agnosticity by maintaining the information gap (equal to the cross-view relational information) even within the proxy neighborhood (top right), or

2. Collapse the local and global representations in the process of alignment (bottom right) by mapping them to ϵ-neighborhoods of each other.

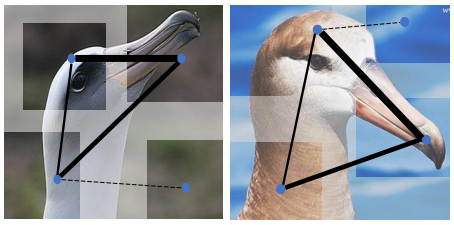
However, since the end objective of our model is to minimize $\mathcal{L}_{\text{rproxy}}$, which is cross-entropic in nature, we prove via Lemma 3 that f cannot collapse the local and global representations, as that would lead to an increase in the downstream cross-entropy loss. f would thus choose to preserve the relational gap in the representation space while mapping them to the neighborhood of their corresponding proxy.

6.5 Visual Representations of Cross-View Local Relationships

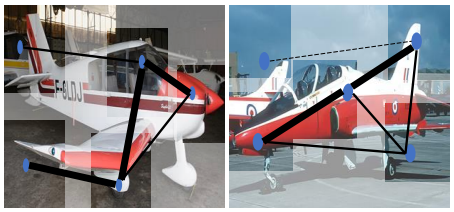
Figure 6 depicts examples of graphs depicting cross-view local relationships. It can be seen that images that provide a diverse set of local views, and thus, a larger space of possible cross-view relationships are the ones that get classified correctly with full certainty. However, as the number of unique local views get limited (possibly due to occlusion or an incomplete photographing of the object), it reduces the amount of relational information that can be mined. Under situations when even the individual local-views are largely shared between classes, there remains no discriminative premise (neither local/global, nor relational) for telling their instances (with limited depiction of local views) apart. It is under such circumstances that the classifier gets confused.

Example: For instance, in the example from the CUB dataset (the top row in Figure 6), the images of the Acadian Flycatcher and Bank Swallow depict sufficient numbers of local views like the head, tail, belly and wings, which provide a large space of potential cross-view relationships that favor classification outcome. On the other hand, the images of the Black-footed Albatross and Laysan Albatross only depict the head and the neck, thus limiting the number of computable relationships that can act as discriminators. Moreover, the head and the neck look largely similar between the two categories, thereby leading to cross-category confusion causing a subsequent misclassification. However, we believe that such a situation can be addressed by learning different distributional priors over the set of local views, which we plan to take up as future work.

Correct Classifications: Acadian Flycatcher (left) and Bank Swallow (right).

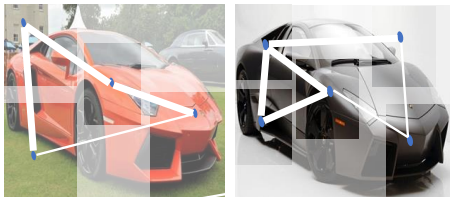


Misclassifications: Black-footed Albatross (left) and Laysan Albatross (right) confused with each other.

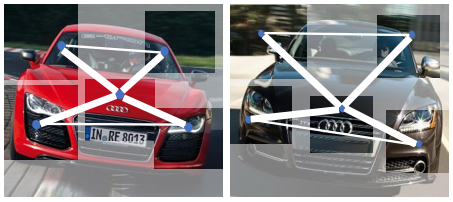


Correct Classifications: DR-400 (left) and Hawk T1 (right).

Misclassifications: Boeing-727 (left) and Falcon-900 (right) confused with each other.



Correct Classifications: Lamborghini Aventador (left) and Lamborghini Reventón (right).



Misclassifications: Audi R8 (left) and Audi TT (right) confused with each other.

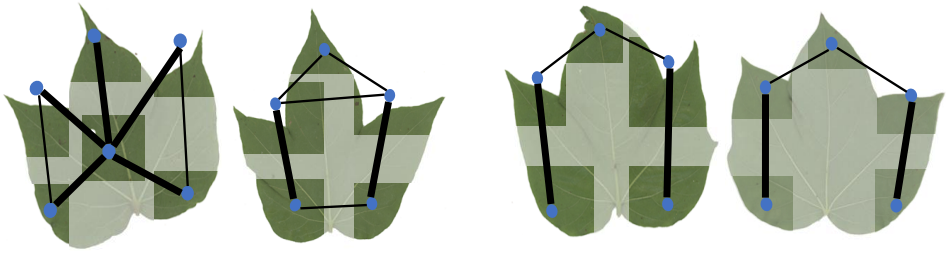


Figure 6: Visual representations of cross-view relationships along with qualitative classification results on (in order from top) CUB, FGVC Aircraft, Stanford Cars and Cotton Cultivar datasets. The pairs on the left correspond to correct classifications made by our model, while the ones on the right are misclassifications occurring out of cross-category confusions.