Transitivity Recovering Decompositions: Interpretable and Robust Fine-Grained Relationships

Abhra Chaudhuri[1,5,6] **Massimiliano Mancini**[2] **Zeynep Akata**[3,4] **Anjan Dutta**[5,6] *
[1] University of Exeter [2] University of Trento [3] University of Tübingen
[4] MPI for Informatics [5] The Alan Turing Institute [6] University of Surrey

Abstract

Recent advances in fine-grained representation learning leverage local-to-global (emergent) relationships for achieving state-of-the-art results. The relational representations relied upon by such methods, however, are abstract. We aim to deconstruct this abstraction by expressing them as interpretable graphs over image views. We begin by theoretically showing that abstract relational representations are nothing but a way of recovering transitive relationships among local views. Based on this, we design Transitivity Recovering Decompositions (TRD), a graph-space search algorithm that identifies interpretable equivalents of abstract emergent relationships at both instance and class levels, and with no post-hoc computations. We additionally show that TRD is *provably robust* to noisy views, with empirical evidence also supporting this finding. The latter allows TRD to perform at par or even better than the state-of-the-art, while being fully interpretable. Implementation is available at `https://github.com/abhrac/trd`.

1 Introduction

Identifying discriminative object parts (local views) has traditionally served as a powerful approach for learning fine-grained representations [95, 85, 43]. Isolated local views, however, miss out on the larger, general structure of the object, and hence, need to be considered in conjunction with the global view for tasks like fine-grained visual categorization (FGVC) [94]. Additionally, the way in which local views combine to form the global view (local-to-global / emergent relationships [59]) has been identified as being crucial for distinguishing between classes that share the same set of parts, but differ only in the way the parts relate to each other [32, 97, 10, 9, 61, 12]. However, representations produced by state-of-the-art approaches that leverage the emergent relationships are encoded in an abstract fashion – for instance, through the summary embeddings of transformers [9, 10], or via aggregations on outputs from a GNN [97, 27]. This makes room for the following question that we aim to answer through this work: *how can we make such abstract relational representations more human-understandable?* Although there are several existing works that generate explanations for localized, fine-grained visual features [11, 78, 36, 63], providing interpretations for abstract representations obtained for discriminative, emergent relationships still remains unaddressed.

Illustrated in Figure 1, we propose to make existing relational representations interpretable through bypassing their abstractions, and expressing all computations in terms of a graph over image views, both at the level of the image as well as that of the class (termed *class proxy*, a generalized representation of a class). At the class level, this takes the form of what we call a Concept Graph: a generalized relational representation of the salient concepts across all instances of that class. We use graphs, as they are a naturally interpretable model of relational interactions for a variety of tasks [86, 18, 1, 46], allowing us to visualize entities (e.g., object parts, salient features), and their relationships.

*Abhra Chaudhuri (ac1151@exeter.ac.uk) is the corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Figure 1: Instead of learning representations of emergent relationships that are abstract aggregations of views, our method deconstructs the input, latent, and the class representation (proxy) spaces into graphs, thereby ensuring that all stages along the inference path are interpretable.

We theoretically answer the posed question by introducing the notion of transitivity (De nition 3) for identifying subsets of local views that strongly in uence the global view. We then show that the relational abstractions are nothing but a way of encoding transitivity, and design Transitivity Recovering Decompositions (TRD), an algorithm that decomposes both input images and output classes into graphs over views by recovering transitive cross-view relationships. Through the process of transitivity recovery (Section 3.2), TRD identi es maximally informative subgraphs that co-occur in the instance and the concept (class) graphs, providing end-to-end relational transparency. In practice, we implement TRD by optimizing a GNN to match a graph-based representation of the input image to the concept graph of its corresponding class (through minimizing the Hausdorff Edit Distance between the two). The input image graph is obtained by decomposing the image into its constituent views [10, 84, 94], and connecting complementary sets of views. The concept graph is obtained by performing an online clustering on the node and edge embeddings of the training instances of the corresponding class. The process is detailed in Section 3.3.

We also show that TRD is provably robust to noisy views in the input image (local views that do not / negatively contribute to the downstream task [18]). We perform careful empirical validations under various noise injection models to con rm that this is indeed the case in practice. The robustness allows TRD to always retain, and occasionally, even boost performance across small, medium, and large scale FGVC benchmarks, circumventing a known trade-off in the explainability literature [31, 73, 23, 37]. Finally, the TRD interpretations are ante-hoc – graphs encoding the learned relationships are produced as part of the inference process, requiring no post-hoc GNN explainers [88, 92, 34, 68, 90].

In summary, with the purpose of bringing interpretability to abstract relational representations, we make the following contributions – (1) show the existence of graphs that are equivalent to abstract local-to-global relational representations, and derive their information theoretic and topological properties; (2) design TRD, a provably robust algorithm that generates such graphs in an ante-hoc manner, incorporating transparency into the relationship computation pipeline at both instance and class levels; (3) extensive experiments demonstrating not only the achieved interpretability and robustness, but also state-of-the-art performance on benchmark FGVC datasets.

2 Related Work

Fine-grained visual categorization: Learning localized image features [2, 96, 47], with extensions exploiting the relationship between multiple images and between network layers [52, 53] were shown to be foundational for FGVC. Strong inductive biases like normalized object poses [6, 83] and data-driven methods like deep metric learning [14] were used to tackle the high intra-class and low inter-class variations. Unsupervised part-based models leveraged CNN feature map activations [35, 94] or identi ed discriminative sequences of parts [4]. Novel ways of training CNNs for FGVC included boosting [55], kernel pooling [15], and channel masking [19]. Vision transformers, with their ability to learn localized features, had also shown great promise in FGVC [31, 29, 50]. FGVC interpretability has so far focused on the contribution of individual parts [11, 20]. However, the usefulness of emergent relationships for learning ne-grained visual features was demonstrated by

[83, 97, 9, 10]. Our method gets beyond the abstractions inherent in such approaches, and presents a fully transparent pipeline by expressing all computations in terms of graphs representing relationships, at both the instance and the class-level.

Relation modeling in deep learning: Relationships between entities serve as an important source of semantic information as has been demonstrated in graph representation learning [72, 18], deep reinforcement learning [93], question answering [67], object detection [33], knowledge distillation [60], and few-shot learning [70]. While the importance of learning relationships between different views of an image was demonstrated in the self-supervised context, [61, 8] and [10] showed how relational representations themselves can be leveraged for tasks like FGVC. However, existing works that leverage relational information for representation learning typically (1) model all possible ways the local views can combine to form the global view through a transformer, and distill out the information about the most optimal combination in the summary embedding [9, 10], or (2) model the underlying relations through a GNN, but performing an aggregation on its output [27, 97]. Both (1) and (2) produce vector-valued outputs, and such, cannot be decoded in a straightforward way to get an understanding of what the underlying emergent relationships between views are. This lack of transparency is what we refer to as "abstract". The above set of methods exhibit this abstraction not only at the instance, but at the class-level as well, which also appear as vector-valued embeddings. On the contrary, an interpretable relationship encoder should be able to produce graphs encoding relationships in the input, intermediate, and output spaces, while also representing a class as relationships among concepts, instead of single vectors that summarize information about emergence. This interpretability is precisely what we aim to achieve through this work.

Explainability for Graph Neural Networks: GNNExplainer [88] was the first framework that proposed explaining GNN predictions by identifying maximally informative subgraphs and subset of features that influence its predictions. This idea was extended in [92] through exploring subgraphs via Monte-Carlo tree search and identifying the most informative ones based on their Shapley values [68]. However, all such methods inherently provided either local [88, 34, 68] or global explanations [90], but not both. To address this issue, PGExplainer [54] presented a parameterized approach to provide generalized, class-level explanations for GNNs, while [82] obtained multi-grained explanations based on the pre-training (global) and fine-tuning (local) paradigm. [48] proposed a GNN explanation approach from a causal perspective, which led to better generalization and faster inference. On the other hand, [3] improved explanation robustness by modeling decision regions induced by similar graphs, while ensuring counterfactuality. Although ViT [27] presents an algorithm for representing images as a graph of its views, it exhibits an abstract computation pipeline lacking explainability. A further discussion on this is presented in Appendix A.10. Also, generating explanations for graph metric-spaces [99, 45] remains unexplored, which we aim to address through this work.

3 Transitivity Recovering Decompositions

Our methodology is designed around three central theorems - (i) Theorem 1 shows the existence of semantically equivalent graphs for every abstract representation of emergent relationships, (ii) Theorem 2 establishes an information theoretic criterion for identifying such a graph, and (iii) Theorem 3 shows how transitivity recovering functions can guide the search for candidate solutions that satisfy the above criterion. Additionally, Theorem 4 formalizes the robustness of transitivity recovering functions to noisy views. Due to space constraints, we defer all proofs to the Appendix A.9.

Preliminaries: Consider an image $x \in X$ with a categorical label $y \in Y$ from an FGVC task. Let $g = c_g(x)$ and $L = \{l_1, l_2, \dots l_k\} = c_l(x)$ be the global and set of local views of an image x respectively, jointly denoted as $V = \{g\} \cup L$, where c_g and c_l are cropping functions applied on x to obtain such views. Let f be a semantically consistent, relation-agnostic encoder (Appendices A.4 and A.7) that takes as input $v \in V$ and maps it to a latent space representation $z \in \mathbb{R}^n$, where n is the representation dimensionality. Specifically, the representations of the global view g and local views L obtained from f are then denoted by $z_g = f(g)$ and $Z_L = \{f(l) : l \in L\} = \{z_{l_1}, z_{l_2}, \dots z_{l_k}\}$ respectively, together denoted as $Z_V = \{z_g, z_{l_1}, z_{l_2}, \dots z_{l_k}\}$. Let Γ be a function that encodes the relationships $\gamma \in \mathbb{R}^n$ between the global g and the set of local L views. DiNo [9] and Relational Proxies [10] can be considered as candidate implementations for Γ. Let $\mathcal{G}$ be a set of graphs $\{(V, E_1, F_{V_1}, F_{E_1}), (V, E_2, F_{V_2}, F_{E_2}), \dots\}$, where the nodes in each graph constitute of the view set V, and the edge set $E_i \in P(V \times V)$, where P denotes the power set, $|\mathcal{G}| = |P(V \times V)|$, meaning

that $\mathcal{G}$ is the set of all possible graph topologies with V as the set of nodes. The node features F_{v_i}, and the edge features $E_{e_i} \in \mathbb{R}^n$.

3.1 Decomposing Relational Representations

Definition 1 (Semantic Relevance Graph)**.** A graph $G_s \in \mathcal{G}$ of image views where each view-pair is connected by an edge with strength proportional to their joint relevance in forming the semantic label y of the image. Formally, the weight of an edge $E_{S_{ij}}$ connecting v_i and v_j is proportional to $I(v_i v_j; y)$.

Intuitively, $E_{S_{ij}}$ is a measure of how well the two views pair together as compatible puzzle pieces in depicting the central concept in the image.

Theorem 1. For a relational representation r that minimizes the information gap $\mathcal{U}(x; y | Z_V)$, there exists a metric space $(M; d)$ that defines a Semantic Relevance Graph such that:

$$d(z_i; y) > \quad \wedge\ d(z_j; y) > \quad)$$
$$\exists\ : \mathbb{R}^n \to^{\text{sem}} \mathbb{R}^n \mid r = (\ (z_i);\ (z_j)); d(r; y)$$

where $M \in \mathbb{R}^n$, $\to^{\text{sem}}$ and d denote semantically consistent transformations and distance metrics respectively, $: Z_V \to M$, and is some finite, empirical bound on semantic distance.

Intuition: The theorem says that, even if z_i and z_j individually cannot encode anymore semantic information, the metric space $(M; d)$ encodes their relational semantics via its distance function d. Assuming the theorem to be false would mean that the outputs of the aggregation do not allow for a distance computation that is semantically meaningful, and hence, there is no source from which r (the relational embedding) can be learned. This would imply that either (i) the information gap (Appendix A.2) does not exist, which is false, because is a relation-agnostic encoder, or (ii) all available metric spaces are relation-agnostic and hence, the information gap would always persist. The latter is again in contrary to what is shown in Appendix A.6, which derives the necessary and sufficient conditions for a model to bridge the information gap. Thus, $(M, ; d)$ defines G_s with weights $1/d(\ (v_i);\ (v_j)) \propto I(v_i v_j; y)$, where $: V \to M$.

Corollary 1.1 (Proxy / Concept Graph)**.** Proxies can also be represented by graphs.

Intuition: Since $(M; d)$ is equipped with a semantically relevant distance function, the hypersphere enclosing a locality/cluster of local/global view node embeddings can be identified as a semantic relevance neighborhood. The centers of such hyperspheres can be considered as a proxy representation for each such view, generalizing some salient concept of a class. The centers of each such view cluster can then act as the nodes of a proxy/concept graph, connected by edges, that are also obtained in a similar manner by clustering the instance edge embeddings.

Theorem 2. The graph G underlying r is a member of $\mathcal{G}$ with maximal label information. Formally,

$$G = \arg\max_{G \in \mathcal{G}} I(G; y)$$

Intuition: The label information $I(x; y)$ can be factorized into relation-agnostic and relation-aware components (Appendix A.2). The node embeddings in already capture the relation-agnostic component (being derived from). From the implication in Theorem 1, we know that given the node embeddings, is able to reduce further uncertainty about the label by joint observation of view pairs (or in other words, edges in). Hence must be relation-aware, as the relation-agnostic component is already captured by . Thus, , which is obtained as a composition of and , must be sufficient representation of (Appendices A.2 and A.3). Since sufficiency is the upper-bound on the amount of label information that a representation can encode (Appendix A.3), the graph that satisfies the condition in the theorem. In the section below, we show that the way to obtain from is by recovering transitive relationships among local views.

3.2 Transitivity Recovery

Definition 2 (Emergence)**.** The degree to which a set of local views $(v_1, v_2, ..., v_k)$ contributes towards forming the global view. It can be quantified as $I(v_1 v_2 ... v_k; g)$.

Figure 2: The TRD pipeline: We begin by computing a Complementarity Graph (G_c) from the view embeddings obtained from f. The Graph Encoder derives the Semantic Relevance Graph (G_s) through transitivity recovery of G_c. We perform a class-level clustering of the instance node and edge embeddings, producing a proxy graph, representing the salient concepts of a class and their interrelationships. The sufficiency of G_s is guaranteed by minimizing its Hausdorff distance with its proxy graph.

Definition 3 (Transitivity). Local views v_1, v_2, and v_3 are transitively related iff, when any two view pairs have a high contribution towards emergence, the third pair also has a high contribution. Formally,

$$I(v_i v_j; g) > \tau \wedge I(v_j v_k; g) > \tau \implies I(v_i v_k; g) > \tau,$$

where $(i, j, k) \in \{1, 2, 3\}$ and τ is some threshold for emergence. A function $\psi(v)$ is said to be transitivity recovering if the transitivity between v_1, v_2, and v_3 can, in some way, be inferred from the output space of ψ. This helps us quantify the transitivity of emergent relationships across partially overlapping sets of views.

Theorem 3. A function $\psi(z)$ that reduces the uncertainty $H(v_i v_j; g \mid \psi(z_i), \psi(z_j))$ about the emergent relationships among the set of views can define G_s by being transitivity recovering.

Intuition: Reducing the uncertainty about transitivity among the set of views is equivalent to reducing the uncertainty about the emergent relationships, and hence, bridging the information gap in I_Δ. If v_1 and v_3 are not transitively related, then it would imply that, at most, only one of the three views (v_j) is semantically relevant, and the other two (v_i, v_k) are not. This leads to a degenerate form of emergence. Thus, transitivity is the key property in the graph space that filters out such degenerate subsets in I_Δ, helping identify view triplets where all the elements contribute towards emergence. A further discussion can be found in Appendix A.11. Now, we know that relational information must be leveraged to learn a sufficient representation. So, for sufficiency, a classifier operating in the graph space must leverage transitivity to match instances and proxies. In the proof, we hence show that transitivity in G is equivalent to the relational information in $z \in \mathbb{R}^n$. Thus, the explanation graph G_s is a maximal MI member of G obtained by applying a transitivity recovering transformation on the Complementarity Graph.

3.3 Generating Interpretable Relational Representations

Depicted in Figure 2, the core idea behind TRD is to ensure relational transparency by expressing all computations leading to classification on a graph of image views (Theorem 1), all the way up to the class-proxy, by decomposing the latter into a concept graph (Corollary 1.1). We ensure the sufficiency (Theorem 2) of the instance and proxy graphs through transitivity recovery (Theorem 3) by Hausdorff Edit Distance minimization.

Complementarity Graph: We start by obtaining relational-agnostic representations Z_V from the input image x by encoding each of its views $v \in V$ as $f(v)$. We then obtain a Complementarity Graph $G_c \in G$ with node features $F_V = Z_V$, and edge strengths inversely proportional to the value of the mutual information between the local view embeddings $h(z_i, z_j)$ that the edge connects. Specifically, we instantiate the edge features F_E as learnable n-dimensional vectors, all with the same value of

$1/|z_i - z_j|$. This particular choice of calculating the edge weights is valid under the assumption that f is a semantically consistent function. Intuitively, local views with low mutual information (MI) are complementary to each other, while ones with high MI have a lot of redundancy. The inductive bias behind the construction of such a graph is that, strengthening the connections among the complementary pairs suppresses the flow of redundant information during message passing, thereby reducing the search space size for finding G_t. The global view, however, is connected to all the local-views with a uniform edge weight of 1. This is because the local-to-global emergent information, i.e., $I(l_i, l_j, \ldots; g)$, is what we want the model to discover through the learning process, and hence, should not incorporate it as an inductive bias.

Semantic Relevance Graph: We compute the Semantic Relevance Graph G_s by propagating G_t through a Graph Attention Network (GAT) [77]. The node embeddings obtained from the GAT correspond to the $z(z)$ in our theorems. We ensure that its transitivity recovering by minimizing the Learnable Hausdorff Edge Distance between the instance and the proxy graphs (Corollary 1.1), which is known to take into account the degree of local-to-global transitivity for dissimilarity computation [64]. Below, we discuss this process in further detail.

Proxy / Concept Graph: We obtain the proxy / concept graph for each class via an online clustering of the Semantic Relevance node and edge embeddings ($z_s(z)$, $e_e(z)$ respectively) of the train-set instances of that class, using the Sinkhorn-Knopp algorithm [16, 8]. We set the number of node clusters to be equal to $|V|$. The number of edge clusters (connecting the node clusters) is equal to $|V|(|V| - 1)/2$. Note that this is the case because all the graphs throughout our pipeline are complete. Based on our theory of transitivity recovery, sets of semantically relevant nodes should form cliques in the semantic relevance and proxy graphs, and remain disconnected from the irrelevant ones by learning an edge weight of ∞. We denote the set of class proxy graphs by $\mathcal{P} = \{G_{P_1}, G_{P_2}, \ldots, G_{P_k}\}$, where k is the number of classes.

Inference and Learning objective: To recover end-to-end explainability, we need to avoid converting the instance and the proxy graphs to abstract vector-valued embeddings at all times. For this purpose, we perform matching between G_s and $G_p \in \mathcal{P}$ via a graph kernel [22]. This kernel trick helps us bypass the computation of the abstract relations $h(G_s) \to \mathbb{R}^{n^2}$, and perform the distance computation directly in G, using only the graph-based decompositions, and thereby keeping the full pipeline end-to-end explainable.

We choose to use the Graph Edit Distance (GED) as our kernel [56]. Although computing GED has been proven to be NP-Complete [79], quadratic time Hausdorff Edit Distance (HED) [25] and learning-based approximations [65] have been shown to be of practical use. However, such approximations are rather coarse-grained, as they are either based on linearity assumptions [65], or consider only the substitution operations [25]. Learnable (L)-HED [64] introduced two additional learnable costs for node insertions and deletions using a GNN, thereby making it a more accurate approximation. We observe that our α already performs the job of the encoder pre-α in L-HED (Appendix A.12). Thus, we simply apply a fully connected MLP head $\mathbb{M}: \mathbb{R} \to \mathbb{R}$ with shared weights on the $z(z)$ embeddings and the vertices of $G_p \in \mathcal{P}$ to compute the node insertion and deletion costs. Thus, in our case, L-HED takes the following form:

$$h(G_s, G_p) = \sum_{u \in G_s} \min_{v \in G_p} c(u, v) + \sum_{v \in G_p} \min_{u \in G_s} c(u, v) \tag{1}$$

$$c(u, v) = \begin{cases} \alpha \|\mathbb{M}(u)\|, & \text{for deletions } (u \to \varepsilon), \\ \alpha \|\mathbb{M}(v)\|, & \text{for insertions } (\varepsilon \to v), \\ \|u - v\|/2, & \text{for substitutions,} \end{cases}$$

where $\alpha = 1/(2|V|)$, and ε is the empty node [64]. We use $h(\cdot, \cdot)$ as a distance metric for the Proxy Anchor Loss [39], which forms our final learning objective, satisfying Theorem 2. Below we discuss how transitivity recovery additionally guarantees robustness to noisy views.

3.4 Robustness

Let $\mathcal{D}$ be the data distribution over $\mathcal{X} \times \mathcal{Y}$. Consider a sample $x \in \mathcal{X}$ composed of views $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$ with label $y \in \mathcal{Y}$. A view $v \in \mathcal{V}$ is said to be noisy if label$(v) \neq y$ [13]. The

[2] Graph kernels, in general, help bypass mapping $G \to \mathcal{H}$, where $\mathcal{H}$ is the Hilbert space [22].

Figure 3: Sample TRD explanations on CUB. Top: 8 nearest neighbors (columns) of the top-2 proxy nodes / concepts (rows) with highest emergence (Definition 2). Bottom: Instance (global view with green border) and proxy (global view with blue border) explanation graphs constructed using the top-6 nodes (nearest train set neighbor for the proxy nodes) with highest emergence. The edge-weights are L_2-norms of the edge embeddings (best visible on screen with zooming).

fraction of noisy views for a given sample is denoted by η. Let f be a classifier that minimizes the empirical risk of the mapping $g(x) \mapsto y$ in the ideal noise-free setting. Let $\mathcal{V}$ be the set of noisy views in V with topology $\mathcal{T}$. Let $V = V \setminus \mathcal{V}$ be the set of noise free (and hence, transitively related - Theorem 3) views with topology $\mathcal{T}$. Recall that G is the optimal graph of V satisfying the sufficiency criterion in Theorem 2, and $G_p \in \mathbb{P}$ is its proxy graph.

Theorem 4. In the representation space Φ of f, the uncertainty in estimating the topology $\mathcal{T}$ of $\mathcal{V}$ is exponentially greater in the error rate η, than the uncertainty in estimating the topology $\mathcal{T}$ of V.

Intuition: This theorem is based on the fact the number of topologies that a set of noisy views can assume, is exponentially more (in the desired error rate) to what can be assumed by a set of transitive views. Since f is explicitly designed to be transitivity recovering, making predictions using h based on noisy views would go directly against its optimization objective of reducing the output entropy (due to the exponentially larger family of topologies to choose from). So, to minimize the uncertainty in Equation (1), f would always make predictions based on transitive views, disregarding all noisy views in the process, satisfying the formal requirement of robustness (Definition 4). In other words, the following property of TRD allowed us to arrive at this result – the structural priors that we have on the transitive subgraphs make them the most optimal candidates for reducing the prediction uncertainty, relative to the isolated subgraphs of noisy views which do not exhibit such regularities.

4 Experiments

4.1 Experimental Settings & Datasets

Implementation Details: For obtaining the global view, we follow [84, 94] by selecting the smallest bounding box containing the largest connected component of the thresholded final layer feature map obtained from an ImageNet-1K [17] pre-trained ResNet50 [30], which we also use as the relation-agnostic encoder h. The global view is resized to 224×224. We then obtain 64 local views by randomly cropping 28×28 regions within the global crop and resizing them to 224×224. We use a 8-layer Graph Attention Network (GAT) [77], with 4 attention heads in each hidden layer, and normalized via GraphNorm [7] to obtain the Semantic Relevance Graph. We train TRD for 1000 epochs using the Adam optimizer, at an initial learning rate of 0.005 (decayed by a factor of 0.1 every 100 epochs), with a weight decay of 10^{-4}. We generally followed [7, 10] for choosing the above settings. We implement TRD with a single NVIDIA GeForce RTX 3090 GPU, an 8-core Intel Xeon processor, and 32 GBs of RAM.

Datasets: To verify the generalizability and scalability of our method, we evaluate it on small, medium and large-scale FGVC benchmarks. We perform small-scale evaluation on the Soy and Cotton Cultivar datasets [9], while we choose FGVC Aircraft [53], Stanford Cars [42], CUB [80],

and NA Birds [75] for medium scale evaluation. For large-scale evaluation, we choose the iNaturalist dataset [76], which has over 675K train set and 182K test set images.

4.2 Interpretability & Robustness

Qualitative Results: Figure 3 shows sample explanations obtained using TRD on the CUB dataset. The top rows represent proxy nodes (concepts), and the columns represent the 8 nearest neighbors of the corresponding proxy nodes, which is highly consistent across concepts and classes. In the bottom are instance and proxy explanation graphs (top-6 nodes with highest emergence) from very similar-looking but different classes. For the proxies, the nearest train set neighbors to the node embeddings are visualized. Even with similar appearance, the graphs have very different structures. This shows that TRD is able to recover the discriminative relational properties of an image, and encode them as graphs. We provide additional visualizations in the supplementary.

Figure 4: Fidelity vs Sparsity curves of TRD and SOTA GNN interpretability methods.

Quantitative Evaluation: We quantitatively evaluate the explanations obtained using TRD based on the Fidelity (relevance of the explanations to the downstream task) vs Sparsity (precision of the explanations) plots for the generated explanations, which is a standard way to measure the efficacy of GNN explainability algorithms [12, 62, 91]. We compare against SOTA, generic GNN explainers, namely SubgraphX [92], PGExplainer [51], and GNNExplainer [88], on candidate small (Cotton), medium (FGVC Aircraft), and large (iNaturalist) scale datasets, and report our findings in Figure 4. TRD consitently provides the highest levels of Fidelity not only in the dense, but also in the high sparsity regimes, outperforming generic GNN explainers by significant margins. Unlike the generic explainers, since TRD takes into account the edit distance with the proxy graph for computing the explanations, it is able to leverage the class-level topological information, thereby achieving higher precision. The supplementary contains results on the remaining datasets, details on the Fidelity and Sparsity metrics, and experiments on the functional equivalence of TRD with post-hoc explanations.

Figure 5: Classification accuracies of relational learning based methods with different noise models on FGVC Aircraft. Left: Increasing number of local views sampled randomly across the entire image (instead of just the global view). Right: Increasing the proportion of noisy views in a controlled manner by sampling from the region outside of the global view.

Robustness to Noisy-Views: Figure 5 shows the performance of TRD on FGVC Aircraft under the following two noise injection models – (1) The local views are sampled uniformly at random across

the entire image, instead of just the global view. This model thus randomizes the amount of label information present in L_j. (2) A fraction α of the local views are sampled from outside the global view, and the remaining ones $(1-\alpha)$ come from within the global view, where α is the variable across experiments. This puts a fixed upper bound to the amount of label information in L_j. It can be seen that under model (1), TRD is significantly more stable to changing degrees of uncertainty in label information compared to other relational learning-based methods. Under model (2), TRD outperforms existing methods as the amount of label information is decreased by deterministically increasing the noise rate. As discussed in Section 3.4 and formally proved in Appendix A.9.1, transitivity acts as a semantic invariant between the instance and the proxy graph spaces, which is a condition that subgraphs induced by the noisy views do not satisfy, leading to the observed robustness.

Effect of Causal Interventions: Here, we aim to understand the behaviour of our model under corruptions that affect the underlying causal factors of the data generating process. To this end, we train TRD by replacing a subset of the local views for each instance with local views from other classes, both during training and inference. As the proxies are obtained via a clustering of the instance graphs, these local views consequently influence the proxy graphs. We report our quantitative and qualitative findings in Table 1 and Figure 6 (Appendix A.17) respectively. TRD significantly outperforms Relational Proxies at all noise rates (percentage of local views replaced), and the gap between their performances widens as the percentage of corruption increases (Table 1). Qualitatively (Figure 6), our model successfully disregards the views introduced from the negative class at both the instance and proxy level. Such views can be seen as being very weakly connected to the global view, as well as the correct set of local views that actually belong to that class. Under this causal intervention, the TRD objective is thus equivalent to performing classification while having access to only the subgraph of clean views from the correct class.

	10	20	30	40	50
Relational Proxies	93.22	87.12	79.35	70.99	63.60
TRD (Ours)	94.90	91.54	82.80	76.35	70.55

Table 1: Quantitative performance comparison between Relational Proxies and TRD with increasing rates of corruption (: percentage local views from different classes).

4.3 FGVC Performance

Comparison with SOTA: Table 2 compares the performance of TRD with SOTA on benchmark FGVC datasets. Some of the most commonly used approaches involve some form of regularizer [21, 41], bilinear features [38], or transformers [29, 81]. However, methods that use multi-view information [94, 49], especially their relationships [97, 10, 5, 71] for prediction have been the most promising, although their final image and class representations are abstract. TRD can be seen to provide the best performance across all benchmarks, while also being fully interpretable. We attribute this to the robustness of our method to noisy views, which is known to have a significant impact on metric learning algorithms [13].

Ablation Studies: Table 3 shows the contributions of the components of TRD, namely Complementarity Graph: CG (Section 3.3), Proxy Decomposition: PD (Corollary 1.1), and Transitivity Recovery: TR (Theorem 3) to its downstream classification accuracy on the FGVC Aircraft, Cotton, and Soy Cultivar datasets. Row 0 is the baseline with none of our novel components. Row 1 corresponds to the setting where the proxy representation is still abstract, i.e., in $\mathbb{R}^n$. We concatenate the nodes in G_s and propagate it through a 1-hidden layer MLP to summarize it in $\mathbb{R}^n$. As we decompose the proxy into a graph (PD, Row 2) based on the instance node and edge clustering-based approach as discussed in Section 3.3 for class-level interpretability, we can see that the performance

ID	CG	PD	TR	Aircraft	Cotton	Soy
0.				94.60	67.70	46.00
1.	✓			95.10	69.60	47.70
2.		✓		94.94	68.31	46.70
3.	✓	✓		95.15	69.70	50.32
4.		✓	✓	95.40	70.10	50.99
5.	✓	✓	✓	95.60	70.90	52.15

Table 3: Ablations on the key components of TRD namely Complementarity Graph (CG), Proxy Decomposition (PD), and TR (Transitivity Recovery).

Method	Small		Medium				Large
	Cotton	Soy	FGVC Aircraft	Stanford Cars	CUB	NA Birds	iNaturalist
MaxEnt, NeurIPS'18	-	-	89.76	93.85	86.54	-	-
DBTNet, NeurIPS'19	-	-	91.60	94.50	88.10	-	-
StochNorm, NeurIPS'20	45.41	38.50	81.79	87.57	79.71	74.94	60.75
MMAL, MMM'21	65.00	47.00	94.70	95.00	89.60	87.10	69.85
FFVT, BMVC'21	57.92	44.17	79.80	91.25	91.65	89.42	70.30
CAP, AAAI'21	-	-	94.90	95.70	91.80	91.00	-
GaRD, CVPR'21	64.80	47.35	94.30	95.10	89.60	88.00	69.90
WTFocus, ECCV'22	-	-	94.70	95.30	90.80	-	-
SR-GNN, TIP'22	-	-	95.40	96.10	91.90	-	-
TransFG, AAAI'22	45.84	38.67	80.59	94.80	91.70	90.80	71.70
Relational Proxies, NeurIPS'22	69.81	51.20	95.25	96.30	92.00	91.20	72.15
PMRC, CVPR'23	-	-	94.80	95.40	91.80	-	-
TRD (Ours)	70.90± 0.22	52.15± 0.12	95.60± 0.08	96.35± 0.03	92.10± 0.04	91.45± 0.12	72.27± 0.05

Table 2: Comparison with SOTA on standard small, medium, and large scale FGVC benchmarks.

drops. This happens because of a lack of prior knowledge about the proxy-graph structure, which increases room for noisy views to be more in uential, harming classi cation robustness. To alleviate this, we introduce TR based on our ndings from Theorem 3, which effectively wards off the in uence of noisy views. With PD, Rows 3 and 4 respectively show the individual contributions of CG, and enforcing the Transitivity invariant between instance and proxy graphs. Row 5 shows that TRD is at its best with all components included.

5 Conclusion & Discussion

We presented Transitivity Recovering Decompositions (TRD), an approach for learning interpretable relationships for FGVC. We theoretically showed the existence of semantically equivalent graphs for abstract relationships, and derived their key information theoretic and topological properties. TRD is a search algorithm in the space of graphs that looks for solutions that satisfy the above properties, providing a concrete, human understandable representation of the learned relationships. Our experiments revealed that TRD not only provides end-to-end transparency, all the way up to the class-proxy representation, but also achieves state-of-the-art results on small, medium, and large scale benchmark FGVC datasets, a rare phenomenon in the interpretability literature. We also showed that our method is robust to noisy input views, both theoretically and empirically, which we conjecture to be a crucial factor behind its effectiveness.

Limitations: Although robust to noisy views, our method is not fully immune to spurious correlations (Figure 3, rightmost graph – the wing and the sky are spuriously correlated; additional examples in the supplementary). Combined with recent advances in learning decorrelated representations [63], we believe that our method can overcome this limitation while remaining fully interpretable.

Societal Impacts: Our method makes a positive societal impact by adding interpretability to existing ne-grained relational representations, in a provably robust manner. Although we are not aware of any speci c negative societal impacts that our method might have, like most deep learning algorithms, our method is susceptible to the intrinsic biases of the training set.

Acknowledgements

This work was supported by the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

References

[1] Aniket Agarwal, Ayush Mangal, and Vipul. Visual relationship detection using scene graphs: A survey. arXiv, 2020.

[2] Anelia Angelova and Shenghuo Zhu. Ef cient object detection and segmentation for ne-grained recognition. In CVPR, 2013.

[3] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In NeurIPS, 2021.

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

[5] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. Sr-gnn: Spatial relation-aware graph neural network for ne-grained image categorization. TIP, 2022.

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

[7] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-Yan Liu, and Liwei Wang. Graphnorm: A principled approach to accelerating graph neural network training. ICML, 2021.

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS, 2020.

[9] 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.

[10] Abhra Chaudhuri, Massimiliano Mancini, Zeynep Akata, and Anjan Dutta. Relational proxies: Emergent relationships as ne-grained discriminators. In NeurIPS, 2022.

[11] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In NeurIPS, 2019.

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

[13] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In CVPR, 2022.

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

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

[16] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In NeurIPS, 2013.

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

[18] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Single-stage visual relationship learning using conditional queries. In NeurIPS, 2022.

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

[20] Jonathan Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classi er using deformable prototypes. CVPR, 2021.

[21] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy ne grained classi cation. In NeurIPS, 2018.

[22] Anjan Dutta and Hichem Sahbi. Stochastic graphlet embedding. IEEE TNNLS, 2019.

[23] Gintare Karolina Dziugaite, Shai Ben-David, and Daniel M. Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. arXiv, 2020.

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

[25] Andreas Fischer, Ching Y. Suen, Volkmar Frinken, Kaspar Riesen, and Horst Bunke. Approximation of graph edit distance based on hausdorff matching. PR, 2015.

[26] Aritra Ghosh, Naresh Manwani, and P.S. Sastry. Making risk minimization tolerant to label noise. Neurocomputing, 2015.

[27] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision GNN: An image is worth graph of nodes. In NeurIPS, 2022.

[28] David Haussler. Quantifying inductive bias: AI learning algorithms and valiant's learning framework. AI, 1988.

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

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

[31] Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, and Christian Janiesch. Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. IJIM, 2023.

[32] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In ICANN, 2011.

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

[34] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, and Yi Chang. Graphlime: Local interpretable model explanations for graph neural networks. IEEE TKDE, 2022.

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

[36] Zixuan Huang and Yin Li. Interpretable and accurate ne-grained recognition via region grouping. In CVPR, 2020.

[37] Ulf Johansson, Cecilia Sönströd, Ulf Norinder, and Henrik Boström. Trade-off between accuracy and interpretability for predictive in silico modeling. FMC, 2011.

[38] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for ne-grained image categorization. CVPRW, 2011.

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

[40] Daniel Kleitman and B. Rothschild. The number of nite topologies. Proceedings of The American Mathematical Society - PROC AMER MATH SOC, 1970.

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

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

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

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

[45] Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. Let invariant rationale discovery inspire graph contrastive learning. ICML, 2022.

[46] Yanan Li, Jun Yu, Yibing Zhan, and Zhi Chen. Relationship graph learning network for visual relationship detection. In CMA, 2021.

[47] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classi cation for ne-grained recognition. In CVPR, 2015.

[48] Wanyu Lin, Hao Lan, and Baochun Li. Generative causal explanations for graph neural networks. In ICML, 2021.

[49] Yang Liu, Lei Zhou, Pengcheng Zhang, Xiao Bai, Lin Gu, Xiaohan Yu, Jun Zhou, and Edwin R. Hancock. Where to focus: Investigating hierarchical attention relationship for ne-grained visual classi cation. In ECCV, 2022.

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

[51] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In NeurIPS, 2020.

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

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

[54] Antonio Martinón and Józef Baśa. Some properties of the hausdorff distance in metric spaces. Extracta mathematicae, 1990.

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

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

[57] Elaine D. Nelson. Separation axioms in topology. Graduate Student Theses, Dissertations, & Professional Papers. 6117, 1966.

[58] Michel Neuhaus and Horst Bunke. Bridging the gap between graph edit distance and kernel machines. In MPAI, 2007.

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

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

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

[62] Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In CVPR, June 2019.

[63] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In ICCV, 2021.

[64] Pau Riba, Andreas Fischer, Josep Lladós, and Alicia Fornés. Learning graph edit distance by graph neural networks. PR, 2021.

[65] Elena Rica, Susana Álvarez, and Francesc Serratosa. On-line learning the graph edit distance costs. PRL, 2021.

[66] Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. In ICLR, 2023.

[67] 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.

[68] Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller, and Grégoire Montavon. Higher-order explanations of graph neural networks via relevant walks. IEEE TPAMI, 2022.

[69] L. S Shapley. Notes on the n-person game – ii: The value of an n-person game. The RAND Corporation, 1951.

[70] 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.

[71] Zhenchao Tang, Hualin Yang, and Calvin Yu-Chian Chen. Weakly supervised posture mining for fine-grained classification. In CVPR, 2023.

[72] Komal Teru, Etienne Denis, and Will Hamilton. Inductive relation prediction by subgraph reasoning. In ICML, 2020.

[73] Matt Turek. Explainable artificial intelligence (xai). https://www.darpa.mil/program/explainable-artificial-intelligence .

[74] Pál Turán. On an extremal problem in graph theory. Matematikai és Fizikai Lapok (in Hungarian), 1941.

[75] 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.

[76] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In CVPR, 2018.

[77] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In ICLR, 2018.

[78] Jörg Wagner, Jan Mathias Köhler, Tobias Gindele, Leon Hetzel, Jakob Thaddäus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In CVPR, 2019.

[79] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. JACM, 1974.

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

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

[82] Xiang Wang, Yingxin Wu, An Zhang, Xiangnan He, and Tat-Seng Chua. Towards multi-grained explainability for graph neural networks. In *NeurIPS*, 2021.

[83] Zhuhui Wang, Shijie Wang, Haojie Li, Zhi Dou, and Jianjun Li. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *AAAI*, 2020.

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

[85] 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.

[86] Guang Yang, Juan Cao, Zhineng Chen, Junbo Guo, and Jintao Li. Graph-based neural networks for explainable image privacy inference. *PR*, 2020.

[87] Xuhui Yang, Yaowei Wang, Ke Chen, Yong Xu, and Yonghong Tian. Fine-grained object classification via self-supervised pose alignment. In *CVPR*, 2022.

[88] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, 2019.

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

[90] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *KDD*, 2020.

[91] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *arXiv*, 2020.

[92] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *ICML*, 2021.

[93] 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.

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

[95] 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.

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

[97] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *CVPR*, 2021.

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

[99] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. Fewer is more: A deep graph metric learning perspective using fewer proxies. In *NeurIPS*, 2020.

A Appendix

Below, we provide some theoretical results and quanti cation of empirical observations from the existing literature [24, 9, 10] that are used in the main text.

A.1 Identities of Mutual Information

Let three random variables x, y, and z form the Markov Chain $x \to z \to y$. Then,

- Positivity:
$$I(x;y) \geq 0; I(x;y|z) \geq 0$$
- Chain Rule:
$$I(x;y) = I(x;y|z) + I(x;z)$$
- Data Processing Inequality:
$$I(x;z) \leq I(x;y)$$

A.2 Information Gap

The Information Gap [10] is the uncertainty that remains about the label information y given a relation-agnostic representation z. Quantitatively,

$$I(x;r|z) = I(x;y) - I(z;y)$$

An implication of the above is that the label information x in can be factorized as follows:

$$I(x;y) = I(x;r|z) + I(z;y)$$

A.3 Suf ciency

A representation z of x is suf cient [24] for predicting its label y if and only if, given z, there remains no uncertainty about the label y of x. Formally,

$$I(x;y|z) = 0 \iff I(x;y) = I(z;y)$$

A.4 Relation-Agnostic Representations - Information Theoretic

An encoder f is said to produce relation-agnostic representations [10] if it independently encodes the global view g and local views $v \in L$ of x without considering their relationship information. Quantitatively,

$$I(x;y|z) = I(x;r) = I(v_1 v_2 ::: v_{|L|}; g)$$

A.5 Distributive property of f

Since $I(x;r)$ refers to local-to-global relationships $I(v_1 v_2 ::: v_{|L|}; g)$, the de nition of relation agnosticity in Appendix A.4 implies that $z = f(x)$ does not reduce any uncertainty about emergence, i.e.,

$$I(x;r) = I(v_1 v_2 ::: v_{|L|}; g) = I(f(v_1) f(v_2) ::: f(v_{|L|}); f(g)) = I(z_1; z_2; ::: z_{|L|}; z_g)$$

A.6 Suf cient Learner

Following are the necessary and suf cient conditions for a suf cient learner [10]:

- It should have a relation-agnostic encoder producing mappings $x \mapsto z$.
- There must be a separate encoder that takes relation-agnostic representations as input, and bridges the information gap $I(x;r|z)$.

A.7 Semantic Consistency

A metric space $(M; d)$ along with its embedding function $\psi : X \to M$ are called semantically consistent, if the following holds:

$$I(x; y|\psi(x)) = d(\psi(x); y)$$

In other words, M encodes the semantics of x via its distance function d.

A.8 Semantic Emergence

Without loss of generality, let us assume that g only contains semantic information. Semantic Emergence refers to the fact that if, for a particular view, its embedding, z_v, encodes a large amount of semantic information, it must also have a large amount of information about the global-view (local-to-global correspondence in DiNO [10]). Intuitively, this is the case because all views $v \in V$, in some way, arise from g, and local-to-global relationships encode semantic information. Quantitatively,

$$d(z_v; y) = \frac{1}{I(v; g)}$$

A.9 Proofs of Theorems

Theorem 1: For a relational representation r that minimizes the information gap $I(x; y|Z_V)$, there exists a metric space $(M; d)$ that defines a Semantic Relevance Graph such that:

$$d(z_i; y) > \alpha \wedge d(z_j; y) > \alpha)$$
$$\exists\, \phi : R^n \xrightarrow{sem} R^n \mid r = \psi(\phi(z_i); \psi(z_j)); d(r; y)$$

where $M \subseteq R^n$, $\xrightarrow{sem}$ and d denote semantically consistent transformations and distance metrics respectively, $\psi : Z_V \to M$, and α is some finite, empirical bound on semantic distance.

Proof. By contradiction.

Let us assume that there exists no metric space (M, d) that induces a Semantic Relevance Graph on Z_V satisfying the given constraint. In other words, given Z_V, we get an upper-bound of α on the amount of label information y that can be captured from x, and there is no function ϕ that can reduce the uncertainty about the label any further.

Since f and d are semantically consistent (Appendix A.7), in information theoretic terms, Z_V is a sufficient representation (Appendix A.3), i.e., both sides of the following are true:

$$I(x; y) = I(Z_V; y) \tag{1}$$
$$I(x; y|Z_V) = \sum_{z \in Z_V} d(z; y) = \alpha; \tag{2}$$

where $\alpha > 0$ (Appendix A.1). However, since f is only a relation-agnostic encoder (as stated in the Preliminaries), it exhibits an information gap (Appendix A.2),

$$I(x; y) = I(Z_V; y) + I(x; r|Z_V); \tag{3}$$

where $I(x; r|Z_V) > 0$. This leads to a contradiction between Equation (2) and Equation (3), establishing the falsity of the initial assumption and proving Theorem 1. □

Theorem 2: The graph G underlying the output relational representation is a member of $\mathcal{G}$ for which the mutual information with the label is maximal. Formally,

$$G = \arg\max_{G \in \mathcal{G}} I(G; y)$$

Proof. By contradiction.

Let us assume the existence of a graph $G' \in \mathcal{G}$ such that $I(G'; y) > I(G; y)$. As shown in Theorem 1, since G encodes both relation-agnostic (nodes) and relation-aware (edges) information, it is a sufficient representation (Appendix A.3) of x. Hence, for $z \in Z_V$,

$$I(G; y) = I(x; r|z) + I(z; y) = I(x; y)$$

Thus, for a G^0 to exist,

$$I(G^0; y) > I(x; y) \quad (4)$$

However, computing G^0 involves the following Markov Chain:

$$x \rightarrow Z_V \rightarrow G^0$$

Thus, from the Data Processing Inequality (Appendix A.1):

$$I(x; Z_V) \geq I(Z_V; G^0)$$

Using the chain rule of mutual information (Appendix A.1):

$$\Rightarrow I(x; y) - I(x; y|Z_V) \geq I(Z_V; G^0)$$
$$\Rightarrow I(x; y) - I(x; y|Z_V) \geq I(G^0; y) - I(G^0; y|Z_V)$$

Now, $I(G^0; y|Z_V) = 0$, since $Z_V \rightarrow G^0$ (G^0 is derived from Z_V),

$$\Rightarrow I(x; y) - I(x; y|Z_V) \geq I(G^0; y) \Rightarrow I(x; y) \geq I(G^0; y) + I(x; y|Z_V)$$

By the positivity property of mutual information [24], $I(x; y|Z_V) \geq 0$. Hence,

$$I(x; y) \geq I(G^0; y);$$

which is in contradiction to Equation (4). Hence the initial assumption must be false. □

Theorem 3: A function $\psi(z)$ that reduces the uncertainty $I(v_i v_j; g|\psi(z_i)\psi(z_j))$ about the emergent relationships among the set of views can define a Semantic Relevance Graph by being transitivity recovering.

Proof. Without loss of generality, consider three transitively related views v_1, v_2, and v_3 such that:

$$d(z_2; y) < \epsilon \wedge d(z_1; y) > \epsilon \wedge d(z_3; y) > \epsilon$$

In qualitative terms, v_2 plays a strong role in determining y given Z_V, while the other views do not. From the property of Semantic Emergence (Appendix A.8),

$$d(z_2; y) < \epsilon \Rightarrow I(v_1 v_2; g) > \epsilon \wedge I(v_2 v_3; g) > \epsilon$$

However, since v_1, v_2, and v_3 are transitively related,

$$I(v_1 v_3; g) > \epsilon$$

Now, f being a relation agnostic encoder, it has no way to minimize the uncertainty in the emergence $I(v_1 v_3; g)$. Hence, using the distributive property for I (Appendix A.5),

$$I(v_1 v_3; g) > \epsilon \Rightarrow I(z_1 z_3; z_g) > \epsilon$$

We know that $I(x; r) = I(v_1 v_3; g) = I(z_1 z_3; g)$ (Appendix A.4).

Consider a function $\psi(z)$ that maps elements in Z_V to (M, d) such that $d(\psi(z_i); \psi(z_j)) \propto 1/I(z_i z_j; g)$. ψ is thus transitivity recovering, as it allows for the identification of transitivity among v_1, v_2 and v_3 via the distance metric d. By allowing for such a distance computation, ψ minimizes $I(z_1 z_3; g|\psi(z_1)\psi(z_3))$, meaning that the knowledge obtained from ψ reduces the uncertainty about the emergent relationships. This in turn also minimizes $I(v_1 v_3; g|\psi(z_1)\psi(z_3))$, and consequently $I(x; r|\psi(z_1)\psi(z_3))$.

However, $I(x; r) = I(x; y|z_1 z_3)$ (Appendix A.4). Hence, ψ also minimizes the information gap $I(x; y|z_1 z_3\psi(z_1)\psi(z_3))$, which implies that:

$$\exists\, r = \psi(\psi(z_1), \psi(z_3)); d(r; y) < \epsilon;$$

thus defining a Semantic Relevance Graph (Theorem 1). □

A.9.1 Robustness

Definition 4 (Robustness). A classifier f is said to be robust (noise-tolerant) if it has the same classification accuracy as that of f under the distribution D [26]. Formally,

$$P_D[f(x)] = P_D[f(x')];$$

where P_D defines the probability distribution over D, and x' is the noisy version of the input x. In other words, f would base its prediction on the $(1-\epsilon)|V|$ noise-free views only.

Lemma 1. Let V' be the set of noisy views in V. Let $V = V \setminus V'$ be the set of noise free (and hence, transitively related - Theorem 3) views. Then, the cut-set $C = (V', V) = \emptyset$ when mapped to the output space of $\psi : V \to G$.

Proof. ψ restricts G to the Hausdorff space (Theorem 3, Inference and Learning objective). In the general case, since the Hausdorff distance defines a pseudometric [54], it satisfies the triangle inequality. In our specific case, it also defines a complete metric. This is because ψ also satisfies the requirement of geometric relation agnositicity in the fine-grained scenario under the minimization of the cross-entropy loss with a categorical ground-truth [10]:

$$\forall z_l \in Z_L : n(z_l) \setminus n(z_g) = \emptyset$$

The above implies that the representation space ψ satisfies the separation axioms [57], and hence:

$$h(x, y) > 0, \forall x \neq y;$$

where $h(\cdot, \cdot)$ is the Hausdorff edit distance [25, 64]. Therefore, for $\psi(z)$, the Hausdorff distance defines a complete metric while satisfying the triangle inequality. Now, consider a set of views $\{v_i, v_j, v_k\}$, where $v_i, v_k \in V'$, and $v_j \in V$. Suppose these nodes define a 3-clique (clique of size 3) in the instance graph. The completeness of the Hausdorff distance $h(\psi(z))$, as well as the satisfaction of the triangle inequality thus gives us:

$$d(\psi(v_i), \psi(v_k)) < d(\psi(v_i), \psi(v_j)) + d(\psi(v_j), \psi(v_k))$$

Hence, for the aforementioned clique formed by v_i, v_j, and v_k, the only edge that survives is (v_i, v_k). In other words, upon convergence, ψ would never produce a path between two transitively related views that involves a noisy view, rendering V' and V disconnected. This completes the proof of the lemma. □

Corollary 4.1. There is no path in G connecting two views in V' that involves a noisy view from V. Thus, the elements of V' and V respectively form homogeneous subgraphs of transitive and noisy views in G and G_p.

Definition 5 (Continents and Islands). A continent is a complete subgraph (clique) of G exclusively composed of transitively related views. An island in G is a subgraph that exclusively contains noisy views.

Lemma 2. Let E_{12} be the set of edges connecting continents C_1 and C_2. The elements of E_{12} are mutually redundant, i.e.,

$$H(e_j | e_i) = 0; \quad \forall (e_i, e_j) \in E_{12}$$

Proof. Consider a pair of vertices $v_i, v_j \in C_1$, such that both are connected to C_2 via edges e_i, and e_j respectively. However, since C_1 is a clique, v_i and v_j (and all other nodes in C_1) are also connected. Thus, the complete set of information that v_i has access to (and can potentially encode in e_i) is the transitivity of (local-to-global relationship emerging from) all the local views in C_1 given by:

$$H(e_i) = I(v_1, v_2, \ldots, v_{|C_1|}, g)$$

Let γ be a message passing function (like a GNN) that carries information across connected nodes. Thus, for γ, the following holds:

$$e_j \xrightarrow{\gamma} v_j \xrightarrow{\gamma} v_i \xrightarrow{\gamma} e_i \implies I(e_j, C_2) = I(e_i, C_2) \implies H(e_j | e_i) = 0;$$

meaning that the knowledge of a single edge between C_1 and C_2 is sufficient to determine the connectivity of the two continents. No other edge in E_{12} can thus provide any additional information, thereby concluding the proof. □

Theorem 4: In the representation space of, the uncertainty in estimating the topology of V is exponentially greater than the uncertainty in estimating the topology of V, in terms of the error rate ϵ.

Proof. It is known that the total number of possible topologies for a graph with n nodes is given by [40]:

$$|\mathcal{T}| = 2^{(n^2/4)} \tag{5}$$

In what follows, we show that it is possible to dramatically reduce this number $O(n)$ for continents, but not for islands.

According to Lemma 1 and Corollary 4.1, the groups of transitive and noisy view nodes can be separated into disjoint sets with no edges between them. Let $V_1 = \{v_1, v_2, \ldots, v_k\}$ be a set of transitively related views (a continent) with topology $\mathcal{T}_1$. Let $V_2 = \{v_{k+1}, v_{k+2}, \ldots, v_n\}$ be a set of noisy views (an island) with topology $\mathcal{T}_2$. Let $\mathcal{D}$ denote the probability mass of a topology in X (fraction of samples in X with topology $\mathcal{T}$).

Let $\mathcal{T}$ be the ground truth topology of a class, determined exclusively by the continents (since noisy views should not determine the label). Let $\hat{\mathcal{T}}$ be the best approximation of $\mathcal{T}$ that can be achieved by observing M unique samples, such that:

$$h[G(\mathcal{T}) \triangle G(\hat{\mathcal{T}})] \leq \epsilon;$$

where $G(\mathcal{T})$ is the graph with topology $\mathcal{T}$. In other words, ϵ is the upper-bound on the statistical error in the approximation of the ground truth topology that can be achieved with M unique samples.

Now, uniquely determining $G(\mathcal{T})$ involves determining all the k-cliques involving the global-view (Transitivity Recovery from Theorem 3). However, the Túran's theorem [74] puts an upper-bound on the number of edges $|E|$ of a graph with clique number k and n nodes at:

$$|E| = (1 - \frac{1}{k})\frac{n^2}{2};$$

which gives a direct upper-bound on the sample complexity M for learning $\hat{\mathcal{T}}$.

If there are $|E|$ edges representing transitive relationships among n views, and the maximum clique size for a class is k, the number of k-cliques to encode all possible clique memberships with this configuration would be given by:

$$\log_k e = \frac{\log_2 e}{\log_2 k} = \frac{\log_2[(1 - \frac{1}{k})\frac{n^2}{2}]}{\log_2 k} = \frac{\log_2[(1 - \frac{1}{k})n^2]}{\log_2 k} = \log_2(1 - \frac{1}{k}) - \log_2 k + 2\log_2 n = O(\log_2 n) \tag{6}$$

Since the maximum clique size k is a constant for a given class. The above number of samples would be required to gain the knowledge of all the cliques in the graph.

To analyze the connectivity across different continents in G, we can apply Lemma 2 to reduce all the continents in G into single nodes, connected by an edge if there is at least one node connecting the two continents in G. Using this observation, and Lemma 2, the number of topologies $|R|$ of such a graph can be given by:

$$|R| = 2^{(O(\log_2 n)/4)} = O(n) \tag{7}$$

Thus with a clique number of k, G could have a maximum of $O(\log_2 n)$ continents (Equation (6)), connected with each other in $O(n)$ possible ways (Equation (7)). Since there is only 1 possibility for the topologies of each clique, the only variation would come from the cross-continent connectivity, which has $O(n)$ potential options. Using the Haussler's theorem [28], the number of samples M required to achieve a maximum error rate of ϵ with probability $(1 - \delta)$ for a continent is given by:

$$M = \frac{\log_2 O(n) + \log_2 \frac{1}{\delta}}{\epsilon} = \frac{\log_2 O(n) + \log_2 \frac{1}{\delta}}{(1/c)} \propto \log_2 O(n) + \log_2 \frac{1}{\delta}; \tag{8}$$

where c is the number of classes, and in practice $1/c \gg \epsilon > 0$. Note that the above bound is tight since we restrict our hypothesis class to only having transitivity recovering decompositions (Theorem 3).

Thus, one would need a number of samples that is linear in the number of transitive views ( ) as well as the the error rate ( ), and the certainty ( ), to determine for a class. It can also be seen that is completely independent of the noise rate .

However, for , since its constituent views are noisy, no prior assumptions can be made about its structure. Thus, the number of possible topologies for can be obtained by substituting with in Equation (5) as follows:

$$|\mathcal{T}| = 2^{(n^2/4)} = O(2^{n^2})$$

Thus, based on the Haussler's theorem, the sample complexity for learning the topology of an island would be given by:

$$M \geq \frac{\log_2 O(2^{n^2}) + \log_2 \frac{1}{\delta}}{\epsilon} \geq \frac{\log_2 O(2^{n^2}) + \log_2 \frac{1}{\delta}}{2\epsilon}; \qquad (9)$$

since the probability of finding an instance with a specific noisy topology is very small, and hence can be approximated with the error-rate ( ). Note that the small probability mass for a specific comes from the worst-case assumption that the dataset would have all possible forms of noise, and hence, the probability for a specific kind would be very low. Also note that unlike Equation (8), the bound cannot be tightened as the premise of transitivity does not hold.

Thus, while the sample complexity for learning a transitive topology is linear in the number of views and the error rate Equation (8), that for a noisy topology is exponential in the number of views and quadratic in the error rate Equation (9). Hence, in the worst case, when [26], . This implies that it is significantly more difficult for to learn the topology of an island than it is to learn the topology of a continent.

From Equation (9), the probability of making a correct prediction based on noisy views would be given by:

$$\begin{aligned}
M \geq \frac{\log_2 O(2^{n^2}) + \log_2 \frac{1}{\delta}}{2\epsilon} &\Rightarrow \log_2 \frac{1}{\delta} \leq M 2\epsilon - \log_2 O(2^{n^2}) \\
&\Rightarrow \log_2 \frac{1}{\delta} \leq M 2\epsilon \\
&\Rightarrow \frac{1}{\delta} \log_2 \frac{1}{\delta} \leq \frac{M}{\epsilon} 2\epsilon^3 \leq 1 \\
&\Rightarrow \log_2 \frac{1}{\delta} \leq 3 \\
&\Rightarrow \frac{1}{\delta} \leq 2^{(3-1)} \\
&\Rightarrow \delta \geq O(2^{-3})
\end{aligned}$$

where the noise rate is a constant for a particular dataset, and , since , as . From Equation (8), the probability of making a correct prediction based on transitive views would be given by:

$$\begin{aligned}
M = \frac{\log_2 O(n) + \log_2 \frac{1}{\delta}}{\epsilon} &\Rightarrow \log_2 \frac{1}{\delta} = M\epsilon - \log_2 O(n) \geq M\epsilon - \log_2 n \\
&\Rightarrow \log_2 \frac{1}{\delta} = (\epsilon - \log_2 n)\epsilon \geq \log_2 n \\
&\Rightarrow \log_2 \frac{1}{\delta} = \log_2 n^{\epsilon} \geq \log_2 n \\
&\Rightarrow \log_2 \frac{1}{\delta} = \log_2 \frac{n^{\epsilon}}{n} = \log_2 n^{\epsilon-1} \\
&\Rightarrow \frac{1}{\delta} = n^{\epsilon-1} \geq 2^{\epsilon-1} \\
&\Rightarrow \delta \leq O(2^{-\epsilon});
\end{aligned}$$

where , which is a constant for a particular class in a dataset, and so is the number of transitively related views , for a class, and . Thus the relative uncertainty

between and $_*$ would be given by:

$$\frac{}{_*} \quad \frac{O(2^{-\ 3})}{O(2^{-})} \quad \frac{2^{-\ 3}}{2^{-}} = \frac{(2^{-})^{\ 2}}{2^{-}} = (2^{-})^{\ 2-1} = 2^{-\ 3+} = O(2^{-\ 3})$$

This completes the proof of the theorem. □

A.10 Discussion on Graph-Based Image Representations and Proxy-Based Graph Metric Learning

Graph-based image representation using patches was recently proposed in ViG [27]. Although the initial decomposition of an image into a graph allowed for flexible cross-view relationship learning, the image graph is eventually abstracted into a single representation in R^n. This prevents one from identifying the most informative subgraph responsible for prediction, at both the instance level, and at the class-level. Although interpreting ViGs may be possible via existing GNN explainablity algorithms [88, 51, 82], they are typically post-hoc in nature, meaning that only the output predictions can be explained and not the entire classification pipeline. Incurred computational and time overheads that come with post-hoc approaches in general, are additional downsides. Our approach preserves the transparency of the entire classification pipeline by maintaining a graph-based representation all the way from the input to the class-proxy, thereby allowing us to explain predictions at both instance and class levels in an ante-hoc manner.

Although graph metric learning has been shown to be effective for learning fine-grained discriminative features from graphs [45], instance-based contrastive learning happens to be (1) computationally expensive and (2) unable to capture the large intra-class variations in FGVC problems [56, 39]. Proxy-based graph metric learning [99] has been shown to address both issues, while learning generalized representations for each class. However, the metric space embeddings learned by the above methods are not fully explainable in terms of identifying the most influential subgraph towards the final prediction, which is an issue we aim to address in our work.

A.11 Role of Transitivity

The idea of transitivity allows us to narrow down sets of views whose relation-agnostic information have been exhaustively encoded in Z, and the only available sources of information comes only when all the views are observed *collectively* as the emergent relationship. On the other hand, dealing with a general local-to-global relationship recovery would additionally include view-pairs, *only one* of which might be contributing to the relational information ($I(\mathsf{v}_1\mathsf{v}_2;\mathsf{g})$ and $I(\mathsf{v}_2\mathsf{v}_3;\mathsf{g})$), thus not exhibiting emergence. Using only the former in our proof helps us to identify the set of nodes responsible for bridging the information gap (via emergence) as the ones satisfying transitivity.

A.12 Learning the Graph Edit Costs

The GNN proposed in [64] would directly learn the mapping $: \mathsf{Z}_\mathsf{V} \ ! \ \mathsf{R}$ for computing the edit costs. In TRD, we instead split the GNN for edit cost computation into a **prefix network** $' : \mathsf{Z}_\mathsf{V} \ ! \ M$ and a **cost head** $: M \ ! \ \mathsf{R}$. Note that this factorization is just a different view of that allows us to draw parallels of our implementation with the theory of TRD.

A.13 Further comparisons with Relational Proxies

Same number of local views: We evaluate both Relational Proxies and TRD with the same number of local views in the normal (no explicit addition of noise) setting on the FGVC Aircraft dataset, and report our findings in Table 4. TRD marginally outperforms Relational Proxies for all values of the number of local views, and exhibits a trend of scaling in accuracy with increasing number of local views. Relational Proxies, on the other hand, does not seem to benefit from increasing the number of local views, possibly due to its lack of robustness to noisy views.

Compute cost: In Table 5, we provide the computational costs of Relational Proxies and TRD in terms of wall clock time evaluated on FGVC Aircraft (same experimental settings including the number of local views). We can see that TRD is significantly more efficient than Relational Proxies in terms of both single sample inference as well as training time until convergence. This is because of the following reasons:

# Local Views	8	16	32	64
Relational Proxies	95.25	95.30	95.29	95.31
TRD (Ours)	**95.27**	**95.45**	**95.52**	**95.60**

Table 4: Comparison of TRD with Relational Proxies under the same number of local views.

- The Complementarity Graph in TRD is constructed exactly once before training, and the semantic relevance graph, as well as the proxy graph are learned as part of the training process.
- TRD does not involve updating the relation-agnostic encoder, which is a ResNet50, as part of the training process. Relational Proxies requires it to be updated, thereby exhibiting computationally heavier forward (as local view embeddings cannot be pre-computed) and backward passes.

	Average Inference Time (ms)	Training Time (hrs)
Relational Proxies	130	22
TRD (Ours)	110	15

Table 5: Compute cost of TRD relative to Relational Proxies.

A.14 Over-smoothing

To understand whether the process of modelling emergent relationships is vulnerable to the over-smoothing phenomenon in GNNs, we evaluated TRD using GATs of up to 64 layers on FGVC Aircraft, presenting our findings in Table 6. We see that the performance does drop as the number of layers are increased beyond 8. To validate whether this is due to the oversmoothing phenomenon, we measure the degree of distinguishability among the nodes by taking the average of their pairwise L^2 distances. The table shows that the distinguishability also decreases as we increase the number of layers, suggesting that the over-smoothing phenomenon does occur. Under the light of the above experiments, the 8-layer GAT is an optimal choice for our problem.

GAT-Depth	4	8	16	32	64
Accuracy	95.05	**95.60**	95.32	94.78	94.20
Distinguishability	0.87	0.63	0.49	0.21	0.09

Table 6: Effect of over-smoothing in GAT on learning emergent relationships.

A.15 Results on ImageNet subsets

Following existing FGVC literature [10, 21], we evaluate the contribution of our novel Transitivity Recovery objective in the coarse-grained (multiple fine-grained subcategories in a single class) and fine-grained subsets of ImageNet, namely Tiny ImageNet [44] and Dogs ImageNet (Stanford Dogs [38]) respectively, and report our findings in Table 7. Although our method can surpass existing SOTA in both the settings, larger gains ( ) are achieved in the fine-grained setting, suggesting that TRD is particularly well suited for that purpose.

A.16 Dependence on input image size

We follow recent SOTA FGVC approaches that use relation-agnostic encoders to extract global and local views [10, 94, 84]. In particular, our view extraction process is exactly the same as Relational Proxies [10], with the same input image resolution and backbone (relation-agnostic) encoder. The

	Tiny ImageNet	Dogs ImageNet
MaxEnt [21]	82.29	75.66
Relational Proxies [10]	88.91	92.75
a : **w/o Transitivity Recovery**	88.10	91.03
b : **with Transitivity Recovery**	89.02	93.10
= (b a)	0.92	**2.07**

Table 7: Coarse-grained vs fine-grained classification results.

above approaches first extract the global view from the input image, and crop out local views from the global view. Since there are two scales at which the image is cropped, all the crops are resized to 224x224. In practice, this provides a similar resolution to resizing the full input image to a higher resolution at the start. To evaluate this hypothesis, we re-trained and re-evaluated our model with the input images resized to 448x448, keeping the remainder of the process of view extraction the same. We provide our results on multiple datasets in Table 8, which shows that the performance gains at the higher input resolution are minor, thus supporting our hypothesis.

		Soy	FGVC Aircraft	Stanford Cars
TRD: 224	224	52.15	95.60	96.35
TRD: 448	448	52.23	95.62	96.39

Table 8: Dependence of TRD on input image size.