'Eyes of a Hawk and Ears of a Fox': Part Prototype Network for Generalized Zero-Shot Learning

Joshua Feinglass[1] Jayaraman J. Thiagarajan[2] Rushil Anirudh[2] T.S. Jayram[2] Yezhou Yang[1]
[1] Arizona State University
[2] Lawrence Livermore National Lab
{joshua.feinglass,yz.yang}@asu.edu, {jjayaram,anirudh1,thathachar1}@llnl.gov

Abstract

Many approaches in Generalized Zero-Shot Learning (GZSL) are built upon base models which consider only a single class attribute vector representation over the entire image. This is an oversimplification of the process of novel category recognition, where different regions of the image may have properties from different seen classes and thus have different predominant attributes. With this in mind, we take a fundamentally different approach: a pre-trained Vision-Language detector (VINVL) sensitive to attribute information is employed to efficiently obtain region features. A learned function maps the region features to region-specific attribute attention used to construct class part prototypes. We conduct experiments on a popular GZSL benchmark consisting of the CUB, SUN, and AWA2 datasets where our proposed Part Prototype Network (PPN) achieves promising results when compared with other popular base models. Corresponding ablation studies and analysis show that our approach is highly practical and has a distinct advantage over global attribute attention when localized proposals are available.

1. Introduction

Generalized Zero-Shot Learning (GZSL) has become a popular research topic with a wide variety of different approaches [28]. As benchmarks have become more competitive, many researchers have begun to develop increasingly sophisticated models requiring a large amount hyperparameter tuning and multiple stages of computationally expensive training. This hinders progress and introduces a need for potentially prohibitively expensive model reconfiguration and computational overhead in a field originally motivated by the expense of human annotation. As such, we explore the potential of using a general pre-trained Vision-Language (VL) detector model combined with a specialized base model that can perform well out-of-the-box and be trained in a single stage to potentially serve as an improved foundation for more sophisticated enhancements like generative models [25] and graph networks [33].

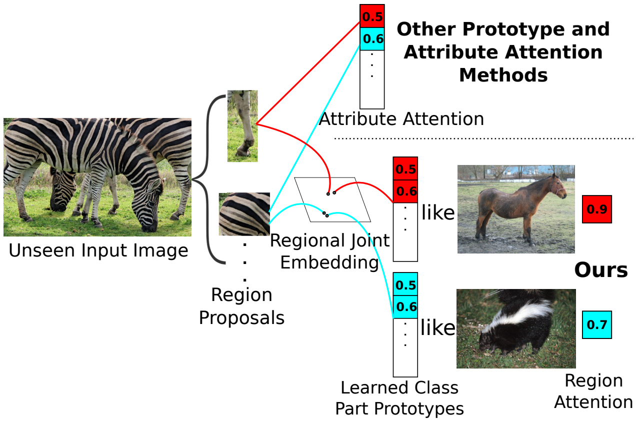

Figure 1. A comparison between the proposed Prototype Proposal Network (PPN) approach and existing approaches which utilize global attribute attention like the base model DAZLE [13].

Localization has been shown to be a key step in many Vision-Language (VL) tasks, especially detail-oriented tasks like fine-grained Zero-Shot Learning [11, 13, 14, 19, 40, 43]. In current approaches, all attribute-specific localization is performed by the GZSL model after either extracting learned features or utilizing popular visual features pre-trained using image classification on ImageNet [38]. To perform this localization, GZL models must correlate global representations of attributes and regions. However, we argue that localized and attribute-specific features can be obtained from VL pretrained detectors like VINVL and GZSL performance can be improved by instead constructing region-specific attribute representations using part prototypes. This is a natural extension of two commonly used base models: SJE [2] which utilizes global joint embeddings of visual features and attributes and DAZLE [13] which employs visual and attribute attention to generate global representations compared with similarity scoring. These two base models are selected for comparison and enhancement due to their simplicity in terms of both computational complexity and hyperparameter tuning. Further motivation for part-based representations can be found in recent GZSL works which seek to learn more discriminative attribute localizations [36] based on part structures [44]. Figure 1 shows an example where a model must generalize to the unseen class, zebra. While prior approaches are limited to estimating the likelihood that an attribute (e.g. stripes) is present in the entire image, our approach can deduce that the shape of a zebra's legs resemble those of a horse, while the color and patterns of its hair is similar to that of a skunk.

We evaluate our approach using the popular GZSL benchmark provided by [38] which utilizes the AWA2 [38], CUB [37], and SUN [27] datasets. Ablation studies of different region localization sources, two proposed regularizers, and a proposed post-processing based calibration technique are performed in support of our framework. We observe that the performance of PPN greatly improves when utilizing pre-trained visual feature extractors with more localized information, resulting in very promising performance against comparable base models.

Contributions. We demonstrate the potential using a pre-trained VL detector for GZSL and propose a novel base model, PPN, designed to better leverage the localized region proposals to achieve promising results on popular GZSL benchmarks.

2. Related Work

Pre-Trained Localization with object detectors for GZSL has been previously been attempted with dataset-specific detectors [42], yielding sub-optimal results when compared against dataset-specific attention mechanisms [13, 40]. This performance gap when using an object detector may be due to a reliance on dataset-specific part annotations, which is both impractical and biased compared to general vision-language detector pre-training. Furthermore, pre-trained Vision-Language models trained to provide global representations [7] have shown promising results in certain GZSL settings [23, 24] when combined with sophisticated transformer-based architectures and retrieved class information. While there is evidence to suggest that the performance of pre-trained localization methods may suffer when applied to unaligned tasks [5, 9], pre-trained models have shown surprising resilience in generalization tasks like zero-shot classification [30] and out-of-distribution detection [4, 10]. We postulate that the object and attribute pre-training utilized by VINVL [43] can potentially serve as a strong foundation for localization in GZSL and allow for methods to explore improvements in other components of the inference process, like attribute representation.

Specialized GZSL Localization is usually a computationally expensive process for competitive methods [29], with popular approaches often utilizing Generative Adversarial Networks [25, 39, 45] and Variational Auto-Encoders [20, 21, 31]. These methods require dataset-specific effort like precise hyperparameter tuning and multiple-stages of training.

Structural Misalignment between seen and unseen data continues to be a major bottleneck for GZSL performance [35]. Prior works have attempted to address these issues with image domain transformations [6, 35], part-object relations [18], and attribute attention [13, 14]. Our work further extends attribute attention by associating attributes with part prototypes which aim to describe the typical [8, 26] characteristics of a class and can potentially serve as more robust primitives than less localized image information.

3. Preliminaries

3.1. (Generalized) Zero-Shot Learning

The (Generalized) Zero-Shot Learning (GZSL) task evaluates model performance under label shift. The training set is defined as $\{(x_n, y_n)|x_n \in X^s, y_n \in Y^s\}_{n=1}^{N_s}$ where images x_n and their corresponding labels y_n are sampled from seen classes Y^s exclusively. In the Zero-Shot Learning (ZSL) setting, models predict on new examples such that $\mathcal{X} \rightarrow \mathcal{Y}^u$ where $\mathcal{Y}^u$ refers to unseen classes not used for training. In the generalized setting, models predict on examples from both the seen and unseen classes $\mathcal{X} \rightarrow \mathcal{Y}^u \cup \mathcal{Y}^s$. ZSL differs from other label shift generalization benchmarks in that for each class label $y \in \mathcal{Y}^u \cup \mathcal{Y}^s$, attribute information $\phi(y)$ is available for use at both train and test time.

3.2. Existing Approaches in GZSL

For classification, prior Zero-Shot Learning methods rely on a compatibility function ψ_c [1, 2] between image features $\theta(x)$ and class attribute vectors $\phi_c(y)$ to estimate the likelihood of y being class c for $c \in \mathcal{C}$

$$P(y = c) = \psi_c(x, y) = \theta(x) \otimes \phi(y_c). \tag{1}$$

Regularization of the image feature representation has been the primary focus of GZSL research [25, 39], while more recent methods have begun to apply grid cell [13] attention and pixel-level [40] attention to the image representation $\Theta(x) \rightarrow \Theta_{\mathcal{R}}(x)$.

3.3. Vision-Language Detectors

Vision-Language (VL) Detectors [3, 43] differ from general object detectors in that they are designed to provide detailed labels and features for both objects and their corresponding attributes. This makes them uniquely suited for common Vision-Language tasks like image captioning, VQA, and text-to-image/image-to-text retrieval. The VINVL [43] detector is extensively pre-trained across Open Images [16],

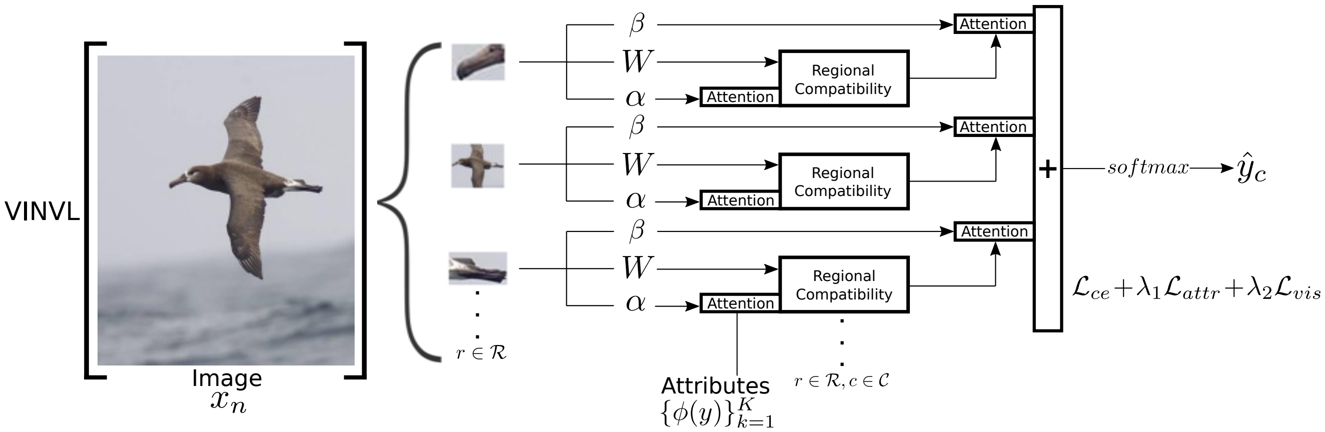

Figure 2. A visualization of the proposed Part Proposal Network (PPN) methodology. α, W, and β represent learned parameters and correspond to the part prototypes, regional embedding, and the mapping function for regional attention, respectively.

Objects375 [32], COCO [17], Visual Genome [15] with standard object annotations before being fine-tuned on Visual Genome with both an object detection and attribute classification loss has demonstrated State-of-the-Art results across a wide variety of Vision-Language tasks. The region-specific features provided by VINVL set serve as a flexible image-region feature extraction method, providing regional features θ_r for all proposed regions of interest, $r = \{1, ..., R\}$, as shown for an image x

$$\{\theta_r\}_{r=1}^{R} = \text{VINVL}(x). \tag{2}$$

4. VL Detector based ZSL Architectures

Encoder-decoder architectures, where a general encoder is trained across a large and diverse set of data and a specialized decoder is fine-tuned on a small dataset, dominate the current Machine Learning landscape. We theorize that a similar paradigm should be employed to ZSL, where data scarcity makes fine-grained classification tasks exceptionally challenging. To this end, our work establishes that VL detectors can serve as an encoder for the ZSL tasks, effectively replacing the visual attention and feature regularization approaches of prior works with object and attribute localization and classification for pre-training. Thus, our work places greater emphasis on compatible decoders to create modular ZSL architectures.

With its use of grid-cells as input, DAZLE [13] serves as the only VINVL compatible decoder in existing ZSL architectures. In an ablation study in Section 6.2, we show that simply substituting the ImageNet classify grid-cell features with the regional features provided by VINVL provides a significant boost to the achievable performance of DAZLE. However, the immediate aggregation of VINVL region proposals performed by DAZLE in this architecture removes the regional information provided by VINVL before measuring the compatibility between visual and attribute features. Thus, we explore the use of regional information in compatibility measures with our proposed network.

5. Part Prototype Network

5.1. Proposed Architecture

Priors of each attribute for a given class ϕ^a in space $\mathcal{R}^{C \times A}$ are provided by human annotation by ZSL datasets where $a \in A$ represent the human selected attributes. Using the names provides for each attribute, a semantic representation of embedding length K can be constructed for each attribute ϕ_k in $\mathcal{R}^{A \times K}$ using word2vec [13, 22]. We proceed to combine these two information sources by expanding the attribute priors along the semantic embedding dimension and the semantic embeddings of the attributes along the class dimension and performing a hadamard product which yields a semantic class attribute tensor ϕ_k^a in the space $\mathcal{R}^{C \times A \times K}$.

Visual features are provided by VINVL [43] for each proposed region of interest. These are used for both region-specific class semantic representation $\mathbf{f}_c^r$ extracted from the semantic class attribute tensor using attribute attention and a visual semantic embedding which is compared against the corresponding class semantic embedding for that region as shown in Figure 2.

$$\mathbf{f}_c^r = g(x, y_c) = \sum_{a=1}^{A} [\alpha(\theta_r(x))]_a \times \phi(y_c)_k^a, \tag{3}$$

where a linear combination in the word2vec semantic space of all A attribute embeddings is taken for each class $c \in \mathcal{C}$ using learned class part prototype α. α serves as a part-specific extension from the global attribute attention utilized in DAZLE [13] and is intended to extract attribute information relevant to the parts present in the input region. Note that α

		Zero-Shot Learning			Generalized Zero-Shot Learning								
Visual		AWA2	CUB	SUN	AWA2			CUB			SUN		
Features	ZSL Model	T1	T1	T1	u	s	H	u	s	H	u	s	H
ResNet101 [12]	SJE [2]	61.9	53.9	52.7	8.0	73.9	14.4	23.5	59.2	33.6	14.4	29.7	19.4
	DAZLE [13]	66.2	66.0	60.3	57.5	76.2	**65.5**	56.8	59.7	58.2	48.4	26.4	34.1
	PPN (Ours)	58.6	55.8	54.1	51.4	70.4	59.4	46.2	50.5	48.3	21.7	24.8	23.2
APN(feats) [40]	SJE [2] (APN-base)	**68.4**	72.0	61.6	56.5	78.0	**65.5**	65.3	69.3	**67.2**	41.9	34.0	37.6
	DAZLE [13]	-	52.9	-	-	-	-	42.7	57.8	49.1	-	-	-
	PPN (Ours)	-	64.8	-	-	-	-	50.2	66.8	57.3	-	-	-
VINVL [43]	DAZLE [13]	63.2	**74.8**	**66.0**	54.7	70.4	61.6	66.9	63.1	64.9	53.8	33.1	**41.0**
	PPN (Ours) *base*	54.7	64.8	62.9	30.4	85.8	44.9	50.2	66.8	57.3	45.2	34.7	39.3
	↳ $\mathcal{L}_{vis}$	63.9	65.9	65.1	40.3	82.6	54.2	54.9	65.5	59.7	50.6	33.5	40.3
	↳ $\mathcal{L}_{attr}$	70.4	72.1	63.5	60.1	62.8	61.4	61.9	65.0	63.4	48.6	31.2	38.0
	↳ $\mathcal{L}_{attr}+\mathcal{L}_{vis}$	**70.4**	**76.0**	**65.0**	59.2	75.9	**66.6**	65.8	67.8	**66.8**	48.6	32.5	**39.0**

Table 1. A comparison of the performance of PPN and popular base models when utlizing different visual feature extractors in the human-annotated attribute ZSL and GZSL setting with the proposed split from [38]. For GZL, accuracy per unseen class (**u**), accuracy per seen class (**s**), and their harmonic mean (**H**) are all reported. Visual feature extractors lower in the table provide more localized feature information. Parameters are set at $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $z = 10^8$ (multiplicative calibrated stacking) for all PPN variants. Word2vec embeddings [22] are used for the attribute representations of all reported results to ensure a fair comparison. The best results are highlighted in **blue** (best) and **red** (second best) for each evaluation with the combination of VINVL and PPN utilizing both regularizers consistently achieving either the best or second best results when compared with other approaches.

shares the same parameters for all attention mappings. Extending from the global compatibility functions proposed in prior ZSL works [1, 2], we aggregate the compatibility computed for each region and class into an overall compatibility function for each class as shown

$$\psi_c(x, y_c) = \sum_{r=1}^{R} [\theta_r(x)]^T \mathbf{W}\mathbf{f}_c^r \beta(\theta_r(x)), \tag{4}$$

where $\mathbf{W}$ is the trained regional embedding which maps the region proposals from VINVL into the word2vec semantic space and β is a learned function which maps a region proposal to the attention for its compatibility function. Like α, parameters of function W and β are also shared across all regions such that a universal mapping between the image and semantic space is learned. Applying a softmax across each compatibility function of each class

$$\hat{y}_c = \frac{\exp\{\psi_c(x, y_c)\}}{\sum_i \exp\{\psi_c(x, y_i)\}}, \tag{5}$$

results in probability mass function $\hat{y}$ which estimates the likelihood that the image example belongs to any given class.

5.2. Loss and Regularization

The cross-entropy between the probability mass functions of our model's prediction and the one-hot ground-truth class label

$$\mathcal{L}_{ce}(\{\hat{y}_c\}_{c \in \mathcal{C}}) = -\sum_c y_c \log(\hat{y}_c), \tag{6}$$

serves as the primary task for our optimization.

For our regularization terms, we construct a penalty function to ensure our attribute and visual representations of the image are relevant to unseen class attributes based on the average of the unseen attribute priors $\phi^a(y \in \mathcal{Y}^u)$ provided by human annotation

$$H(\phi^a) = 1 - \frac{1}{|\mathcal{Y}^u|} \sum_{y \in \mathcal{Y}^u} \phi^a(y). \tag{7}$$

Penalties for each attribute vary in the range of 0 to 1, with a lower value indicating that the attribute in question occurs more frequently in unseen classes. For attributes, this penalty is multiplied by the square of the attribute attention. Squaring the attention weight ensures outliers incur a greater penalty. This scaled penalty is then summed over all attributes and averaged over all regions as shown

$$\mathcal{L}_{attr}(\phi^a, x) = \frac{1}{R} \sum_{r=1}^{R} \sum_{a=1}^{A} H(\phi^a)[\alpha(\theta_r(x))]_a^2. \tag{8}$$

For visual semantic features, the penalty is projected into the w2v space using the w2v embeddings of the attribute classes ϕ_k and a cosine embedding loss is used to contrast the attention-weighted aggregation of the VINVL region proposals projected into the word2vec space $\sum_{r=1}^{R} [\theta_r(x)]^T \mathbf{W} \beta(\theta_r(x))$ with the penalty with the projected penalty

$$\mathcal{L}_{vis} = max(0, cos(\sum_{r=1}^{R}[\theta_r(x)]^T \mathbf{W}\beta(\theta_r(x)), \sum_{a=1}^{A} H(\phi^a)\phi_k)). \tag{9}$$

The cross-entropy, prior confidence, and self-calibration loss functions are combined to form the training objective

$$\min_{\mathbf{W},\alpha,\beta} \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{attr} + \lambda_2 \mathcal{L}_{vis}, \tag{10}$$

which optimizes the joint embedding and attention parameters.

5.3. Pre/Post-Processing

L_2 normalization is performed across the attribute dimension A for our attribute input tensor $\mathcal{R}^{C \times A \times K}$ as described in [34] and the feature dimension for region proposals. Because ZSL models are only exposed to seen classes at training time, their confidence is typically biased towards predicting seen classes. Calibrated stacking is a standard post-processing approach for adjusting confidence bias of seen classes [40]. Since part prototypes in PPN are constructed based on seen classes, the prediction confidence of PPN is significantly higher for seen classes, even when compared against prior methods. For example, APN [40] report using additive calibrated stacking value of 0.8 for CUB while our method would require a value of 0.9995 to achieve its best validation set performance. To address this phenomenon, we propose multiplicative calibrated stacking as a means of adjusting confidence bias where prediction confidences corresponding to a seen class are adjusted by dividing by a constant z as shown

$$\hat{y} = \begin{cases} \frac{\hat{y}_c}{z} & if\ y \in \mathcal{Y}^s, \\ \hat{y}_c & otherwise. \end{cases} \tag{11}$$

With addition, all predictions experience the same adjustment while multiplication applies less and potentially no adjustment to predictions with little or no confidence, respectively. This allows us to apply more significant confidence adjustments without impacting false positive rate of unseen selections as significantly.

6. Experiments

6.1. Experimental Setup

We perform our experiments using the three widely adopted ZSL and GZSL benchmark datasets provided by [38]. CUB [37] is a fine-grained bird species classification dataset with 150 seen and 50 unseen classes. With 312 human-annotated attributes, it is the most heavily annotated of the 3 benchmarked datasets and provides 7,057 training examples and 4,731 testing examples. SUN [27] is a scene classification dataset with 645 seen and 72 unseen classes. Only 10,320 training samples and 4,020 testing samples are provided along with 102 human-annotated attributes, meaning SUN has the least images per class of the 3 benchmarked datasets. AWA2 [38] is an animal species classification dataset and is relatively coarse when compared to CUB and SUN. 23,527 training images, 13,795 test images, and 85 human-annotated attributes are provided. The traditional ZSL setting benchmark uses the top-1 accuracy (T1) performance on unseen class test set. The GZSL setting challenges models to classify both unseen and seen images in a single test set, such that unseen images may be misclassified as being from a seen category and vise versa. The harmonic mean (H) measures the trade-off between unseen and seen test set performance and serves as the primary metric for GZSL, with the unseen (u) and seen (s) accuracy included for additional transparency.

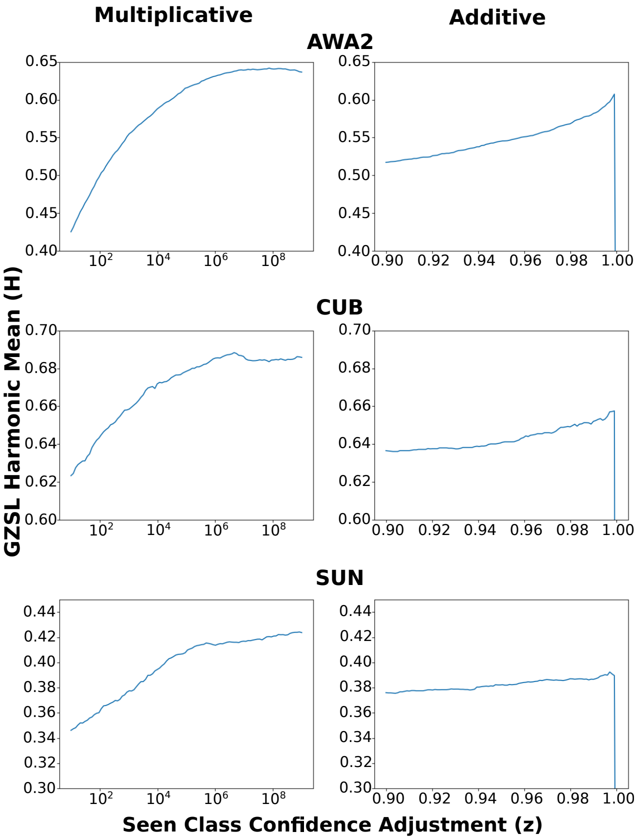

Figure 3. An ablation study of the GZSL harmonic mean performance of DAZLE (with VINVL features) and RAJE when using addition and multiplication for calibrated stacking. The same vertical axes are used when plotting the multiplicative and additive performance in each dataset. Our proposed multiplicative approach for calibration exhibits greater performance over a larger portion of the graph while the previous additive approach has reduced performance and a sharp dip after its peak. Furthermore, additive calibration has the potential to sharply dip as it approaches 1 since it will begin classifying all examples as unseen.

In our implementations, visual features from 3 sources are tested: ResNet101 pre-trained on ImageNet [12, 38] providing the most least localized information, 2048x7x7 grid features (R=49) of ResNet101 fine-tuned using APN (model weights are only available for CUB) providing more localized information, and the top 30 (R=30) most confident proposals provided by VINVL [43] of dimension 2048 providing the most localized information. Combinations of the aforementioned visual features with 3 different base models are tested: SJE [2], DAZLE [13], and our proposed PPN. Label attributes are sourced from human annotations provided by [38]. As done in DAZLE [13], we semantically embed the human annotated attributes using word2vec [22]. The optimization procedure for PPN utilizes the Adam optimizer with a learning rate of 0.001 following the procedure from [41]. The proposed hyperparameters obtain their highest validation set performance at 0.1 across all tested benchmarks. Thus, all tested base models are out-of-the-box, meaning no dataset-specific hyperparameter settings are utilized.

6.2. (Generalized) Zero-Shot Learning

Table 1 explores the relationship between different visual feature extractors and specialized base models. With the use of VINVL features, DAZLE becomes a more competitive method for both the CUB and SUN dataset, while a slight decrease in performance occurs in AWA2. This may be caused in part by similarities between the categories in AWA2 and the ImageNet dataset, on which the ResNet101 feature extraction is pretrained. Consistent improvements in the performance of PPN can be observed as more features with more localized information are utilized. The use of both regularizers also consistently improves the performance of PPN with the exception of the SUN dataset, where only the $\mathcal{L}_{vis}$ regularizer contributes to improved performance The multiplicative corrections used to compensate for the large discrepancy between seen and unseen confidence for GZSL can be seen in Figure 3. The plots show significant improvements in performance when using the multiplicative correction compared to the previously proposed additive correction.

7. Conclusion and Future Work

We propose a novel approach for localization in GZSL using region proposals from a pre-trained VL detector (VINVL) and utilize the provided proposals in to create part prototype which extract relevant information from attributes for each of these regions. Our ablation and analysis show that VINVL is a highly effective visual information source for GZSL and that our proposed Part Prototype Network can potentially serve as an improved foundation for future GZSL works. One potential avenue not explored in this work is enhancements of the visual features provided by VINVL, either through regularization like in generative approaches or additional training tailored to the zero-shot tasks.

Acknowledgments: The authors acknowledge support from the NSF RI Project VR-K #1750082, CPS Project #2038666. Any opinions, findings, and conclusions in this publication are those of the authors and do not necessarily reflect the view of the funding agencies.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1453–1463, 2023.

[5] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kuehne, and Michael Vössing. What a MESS: Multi-domain evaluation of zero-shot semantic segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[6] Shiming Chen, GuoSen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Joshua Feinglass and Yezhou Yang. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online, 2021. Association for Computational Linguistics.

[9] Joshua Feinglass and Yezhou Yang. Towards addressing the misalignment of object proposal evaluation for vision-language tasks via semantic grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4397–4407, 2024.

[10] Tejas Gokhale, Joshua Feinglass, and Yezhou Yang. Covariate shift detection via domain interpolation sensitivity. In

First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022, 2022.

[11] Harald Hanselmann and Hermann Ney. Elope: Fine-grained visual classification with efficient localization, pooling and embedding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1247–1256, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. *Conference on Neural Information Processing Systems*, 2020.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[18] Man Liu, Chunjie Zhang, Huihui Bai, and Yao Zhao. Part-object progressive refinement network for zero-shot learning. *IEEE Transactions on Image Processing*, 33:2032–2043, 2024.

[19] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803, 2021.

[20] Yang Liu, Xinbo Gao, Jungong Han, and Ling Shao. A discriminative cross-aligned variational autoencoder for zero-shot learning. *IEEE Transactions on Cybernetics*, 2022.

[21] Peirong Ma and Xiao Hu. A variational autoencoder with deep embedding model for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11733–11740, 2020.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013.

[23] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. *arXiv preprint arXiv:2212.02291*, 2022.

[24] Muhammad Ferjad Naeem, Yongqin Xian, Luc V Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. In *Advances in Neural Information Processing Systems*, pages 12283–12294. Curran Associates, Inc., 2022.

[25] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020.

[26] Daniel Osherson and Edward E Smith. On typicality and vagueness. *Cognition*, 64(2):189–206, 1997.

[27] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.

[28] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

[29] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[33] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Invertible zero-shot recognition flows. In *Computer Vision – ECCV 2020*, pages 614–631, Cham, 2020. Springer International Publishing.

[34] Ivan Skorokhodov and Mohamed Elhoseiny. Class normalization for (continual)? generalized zero-shot learning. In *International Conference on Learning Representations*, 2021.

[35] Chenwei Tang, Zhenan He, Yunxia Li, and Jiancheng Lv. Zero-shot learning via structure-aligned generative adversarial network. *IEEE transactions on neural networks and learning systems*, 2021.

[36] Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. In *Advances in Neural Information Processing Systems*, 2021.

[37] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech, 2010.

[38] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[39] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *Advances in Neural Information Processing Systems*, pages 21969–21980. Curran Associates, Inc., 2020.

[41] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9316–9325, 2022.

[42] Shiqi Yang, Kai Wang, Luis Herranz, and Joost van de Weijer. Simple and effective localized attribute representations for zero-shot learning. *arXiv preprint arXiv:2006.05938*, 2020.

[43] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021.

[44] Yang Zhang and Songhe Feng. Enhancing domain-invariant parts for generalized zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6283–6291, 2023.

[45] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018.