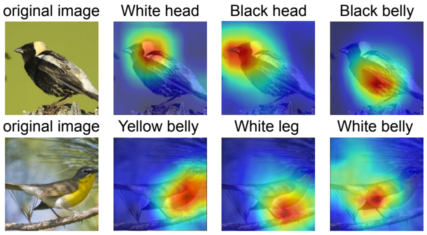

**Figure 1.** Localizing partially different attributes, e.g. “white head” v.s. “black head”.

| Method | **CUB** | **AWA2** | **SUN** |
|---|---|---|---|
| `[A]` | 49.5 | 43.5 | 53.4 |
| `[B]` | 68.9 | 65.2 | 43.1 |
| `[C]` | 47.5 | 66.7 | - |
| `APN+[A] (Ours)` | 67.3 | 46.5 | **56.7** |
| `APN+[B] (Ours)` | **70.0** | 65.6 | 43.7 |
| `APN+[C] (Ours)` | 67.7 | **71.0** | - |

**Table 1.** Results of applying our learned APN features to [A], [B] and [C] in GZSL (harmonic mean reported). We did not reproduce results of [C] on SUN due to time limit (will include in final paper).

**R1, R2, R3: novelty.** We thank the reviewers for mentioning a few pioneering works [8, 58, D] that learn to localize object parts. Here we compare our work with these papers and clarify our novelty. [8] learns multiple prototypes for each object class with image labels to improve the model interpretability but is not tailed for ZSL. However, we aim to improve the image representation for ZSL by learning attribute prototypes with class-level attributes. [58] proposes to improve the image features by learning channel-wise part attention. Similarly, [D] uses the channel grouping model [52] to learn part-based representations and part prototypes. In contrast, we treats each channel equally, and we use spatial features associated with input image patches to learn prototypes for attributes (see Fig.1 in the main paper). We argue that these methods are limited when localizing object parts and visual attributes. Specifically, they learn latent attention/prototypes during training, whose meaning is posteriorly inducted by observation, which is not deterministic. Besides, they can only localize a small number of object parts, i.e., [58] for 2 parts, [D] for 4 parts, and are not able to localize attributes, which play an important role for ZSL. Very different from these publications, our work is innovative in two aspects which overcome the above mentioned limitations. 1) We improve the attribute localization and image features by learning prototypes to regress attributes, where each prototype corresponds to a specific attribute. Therefore, our model can localize all the attributes and parts, which is essential in building trustworthy ZSL model to associate attributes with correct image regions. Our claim can be supported by the results in Fig. 1 and other results (Fig. 2, 3, Tab. 3) in main paper. 2) We decorrelate prototype learning and enforce the localization compactness. In R2’s words, “the employment of these ideas together for attribute localization and ZSL is quite interesting and seems to lead to consistent good performance” (see Tab. 1 for improvement over SOTA). The discussion will be added in the final paper.

**R1: differences with CAM.** CAM is a post-hoc method to investigate the model attention by computing the channel-wise weighted sum of last layer CNN feature maps, which didn’t improve the image feature. In contrast, our APN aims to improve the image representation for zero-shot learning by learning prototypes that predict attributes from intermediate features. Besides, while the original CAM generates one global attention map for predicting an object class in a given image, our APN can generate multiple attention maps that localize different attributes in an image. Our APN can obtain attention maps that better localize visual attributes compared to CAM (See Fig. 3 in the main paper).

**R2: impact of binary and continuous attributes.** We follow [21] to generate class binary attributes by thresholding the continuous attributes. Our method works equally well for both kinds of attributes in terms of ZSL accuracy (continuous 73.3% vs binary 73.1% on CUB) and part localization accuracy (continuous 52.8% vs binary 52.1% on CUB).

**R2: evaluate attribute prediction** As suggested by R2, we build an attribute prediction baseline by training a standard CNN to predict binary attributes without prototypes. Our APN (also trained with binary attributes) can classify binary attributes with a threshold of 0.5. The evaluation is conducted on the image-level attributes of test images on CUB dataset. Our APN achieves attribute prediction accuracies of 87% on unseen classes and 90% on seen classes, which significantly outperforms the baseline model without prototype that obtains 82% (unseen) and 84% (seen).

**R3: ProtoMod needs more evaluation.** For each attribute, we learn a prototype on intermediate CNN features to regress the desired attribute. So it is expected to learn specific attributes rather than just the color or parts, which is empirically confirmed by the attribute prediction accuracy which is higher than the attribute prediction model (See previous response to R2), and the qualitative results in Figure. 1, where our APN is able to precisely localize “white head”, “black head” and “black belly” in the first image and “yellow belly”, “white leg” and “white belly” in the second.

**R2: comparing with [A], [B] and [C].** While [A], [B] and [C] propose strong ZSL classifiers with fixed ImageNet-pretrained (or finetuned) features, our APN aims to improve the image features for ZSL. In Table 1, we show new SOTA GZSL results when applying the image features extracted from our learned APN to [A], [B] and [C]. For example, on SUN, our `APN+[A]` achieves 56.7% vs 53.4% of `[A]`. On CUB, our `APN+[B]` obtains 70.0% vs 68.9% of `[B]`. On AWA, our `APN+[C]` reaches 71.0% vs 66.7% of `[C]`. We will include these papers and results in the final paper.

**[A]** Huang et al., Generative Dual Adversarial Network for Generalized Zero-shot Learning, In CVPR 2019
**[B]** Xian et al., f-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning, In CVPR 2019
**[C]** Li et al., Rethinking Zero-Shot Learning: A Conditional Visual Classification Perspective, In ICCV 2019
**[D]** Zhu et al., Learning classifiers for target domain with limited or no labels, In ICML 2019