We thank the reviewers for their comments. This work presents a novel HGTL task along with a baseline BigSPN framework. As agreed by all reviewers, it is an interesting and novel task, which will attract more researches to follow.

To Review #1:
Q1:Unclear formulation of HGTL. Given the basic-category annotated data $\{\boldsymbol{x}, y_b \in \mathcal{Y}_b\}$ and sub-category set $\mathcal{Y}_s$, HGTL aims to train a model which minimizes testing recognition error on both $\{\boldsymbol{x}, y_b\}$ and $\{\boldsymbol{x}, y_s \in \mathcal{Y}_s\}$. Some available side information includes affiliation relationship between $\mathcal{Y}_b$ and $\mathcal{Y}_s$, and category descriptions $\boldsymbol{a}(\cdot)$.
Q2:A reasonable DA baseline. Yes, we actually have adjusted DA to HGTL with a two-stage manner. A base DA/ZSL model first trains a visual encoder $f_v(\cdot)$ and a semantic interpreter $g(\cdot)$ by minimizing $d[f_v(\boldsymbol{x}), g(\boldsymbol{a}(y_b))]$ on $\{\boldsymbol{x}, y_b\}$, and then directly transfer $f_v(\cdot)$ and $g(\cdot)$ to target domain by argmax $d[f_v(\boldsymbol{x}), g(\boldsymbol{a}(y_s))]$, where $y_s \in \Omega_{b\to s}$. Notably, $\Omega_{b\to s}$ (line 171) is the subordinate categories of predicted y_b of $\boldsymbol{x}$ by argmax $d[f_v(\boldsymbol{x}), g(\boldsymbol{a}(y_b))]$.
Q3:Unclear motivation of BigSPN. Given two images $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with the same basic category y_b but different sub-categories y_{s1} and y_{s2}. For y_b, we expect $f_v(\boldsymbol{x}_1)$ and $f_v(\boldsymbol{x}_2)$ to be invariant, but for y_{s1} and y_{s2}, $f_v(\boldsymbol{x}_1)$ and $f_v(\boldsymbol{x}_2)$ should be discriminative. This conflict motivates us to learn a new visual encoder $f_{pv}(\cdot)$ to specifically recognize y_{s1} and y_{s2}. To capture the subtle visual clues between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, multi-head attention and bilinear pooling are used to build $f_{pv}(\cdot)$ in Fig. 2. Due to unavailable annotated $\{(\boldsymbol{x}_1, y_{s1}), (\boldsymbol{x}_2, y_{s2})\}$, we design an entropy minimization loss $\mathcal{L}_{se}$ in Eq. (7) to train the weights of $f_{pv}(\cdot)$ using only $\{(\boldsymbol{x}_1, \boldsymbol{x}_2, y_b)\}$. The experimental results have proved the effectiveness of BigSPN in HGTL, which surpasses all related DA&ZSL methods.
Q4: Two-stage solution for HGTL. All the compared methods and BigSPN are the two-stage models, as stated in Q2.

To Review #2:
Q1:Illustrating why a new visual encoder is needed. Please refer to Q3 of R#1.
Q2:Explaining how ZSL is applied to HGTL. A baseline ZSL model can refer to Q2 of R#1, which first trains $\{f_v(\cdot), g(\cdot)\}$ on basic-domain data $\{\boldsymbol{x}, y_b\}$, and then merely transfers $\{f_v(\cdot), g(\cdot)\}$ to $\{\boldsymbol{x}, y_s\}$ by argmax $d[f_v(\boldsymbol{x}), g(\boldsymbol{a}(y_s))]$. Other related ZSL methods in Table 2 usually design stronger $f_v(\cdot)$ or $g(\cdot)$.
Q3:Why BigSPN performs not well on Basic_R1? Since BigSPN focuses on designing a new visual extractor $f_{pv}(\cdot)$ for the sub-category recognition, we just adopt a simple 1-layer convolution as the visual extractor $f_v(\cdot)$ for basic-category. Compared to well-designed $f_v(\cdot)$ in ZSL models, e.g., SPAEN [5] uses the auto-encoder architecture, BigSPN obtains comparable results in Basic_R1, but much better results in Sub_R1.
Q4:How to determine training strategy. Grid searching is used to determine the training strategies from the general training strategies in ZSL and FGVC. Exponential Decay is used with step 30 and decay=0.1.
Q5:Other issues. The motivation and design for attention map part please refers to Q3 of R#1. More recognition examples will be added in revision.

To Review #3:
Q1:applicability in real world about categorical hierarchy and text description. HGTL focuses on object recognition applications, such as animals, goods, and plants, whose categorical hierarchy naturally exists and is the basic knowledge for scientists in specific fields. Compared with massive image annotations, the hierarchy for limited categories is easier for scientists to build. For the text description of categories, it is cheap to collect from Internet, e.g., querying in wiki, though it maybe noisy. Thus, besides using "word2vec" to encode wiki text for Flower-HGTL, stronger NLP models, e.g., Bert, will be explored to generate more compact semantic embeddings. With scientist's categorical hierarchy and category description from wiki-Bert, HGTL can be applied to various real-world applications. For example, with only basic-category annotations, HGTL can automatically pre-annotate sub-categories to reduce human labeling burden.
Q2:Limited technique novelty in BigSPN. Different from ZSL&DA, each image in HGTL has two categories, as our answer to Q3 of R#1. Thus, one main contribution of BigSPN is to learn a two-branch architecture for respective basic- and sub-category recognition without sub-category annotations. Specifically, for sub-category branch, the multi-head attention f_{pv} is used to capture detailed visual clues, and entropy minimization is leveraged to train the weights of f_{pv} in an unsupervised manner. Although the separate concept of each component is not new, **BigSPN is the first work that designs specific visual representations for basic- and sub-category recognitions with only basic-category annotations.** Thus, BigSPN is a novel and effective baseline for the new HGTL task, compared to related ZSL&DA methods.
Q3: Use of multi-attention and bilinear pooling. The motivation of multi-attention and bilinear pooling refers to Q2. On CUB, they bring obvious 2.7% grain from Table 3, and the tuned hyper-params are suitable for most of datasets.
Q4: Other issues. Hierarchical classification requires hierarchical category annotations, which is unavailable in HGTL. Detailed discussion about hierarchical classification and broader impact will be added in revised version.

To Review #4:
Q1:Extension to larger iNat dataset. We are building the category descriptions and hierarchy for iNat dataset, and detailed experiments will be published.
Q2:Typos. We have carefully revised the typos and incorrect sentences.