In this document, we include supplementary materials for "Improving Visual Recognition with Hyperbolical Visual Hierarchy Mapping". We first provide more concrete implementation details (Sec. A), a theoretical baseline (Sec. B), and additional experimental results (Sec. C). Finally, we visualize more visual hierarchy trees from the selected images to provide solid evidence of the proposed method (Sec. D).

A. Network Architecture

A.1. Classification

To demonstrate the scalability of our proposed method, we deploy our method on ResNet50 [1], EfficientNet [2], DeiT [3], and Swin [4]. All encoders are pre-trained on ImageNet-1K [5] dataset. Then, the hierarchy decomposition module $\mathcal{D}$ is composed of two transformer layer, and then the Hierarchy encoding module $\mathcal{G}$ is composed of four transformer layers, in which each transformer layer has four attention heads with 128 embedding dimensions. For hyperbolic embedding, we initialize curvature parameter $c = 1$. For dense prediction tasks, including semantic segmentation and object detection, we feed $\mathbf{v}_{\text{map}}$ into the hierarchy encoding stage instead of $\mathbf{v}_{\text{cls}}$. We obtain $\mathbf{v}_{\text{cls}}$ by applying global average pooling on $\mathbf{v}_{\text{map}}$ for ResNet [1] and Swin [4], and utilize the `[CLS]` embedding for DeiT [3].

A.2. Dense prediction.

In the main paper, we report utilizing UperNet [6] and Mask-RCNN [7] for semantic segmentation and object detection & instance segmentation tasks. For each task's decoder, we incorporated the penultimate feature map $\hat{\mathbf{v}}_{\text{map}}$, computed by our Hi-Mapper, along with the vanilla output from the intermediate layers of encoder $\mathcal{F}$, as illustrated in Fig. 1.

B. Theoretical Baseline

B.1. Mixture of Gaussians

For hierarchy tree $\mathbf{T}$, each node distribution at level $l+1$ is represented as a Mixture of Gaussians (MoG) of its child node distributions, which are independent. Hence, we can derive the semantic seed distribution at level $l+1$ through a simple computation.

$$f_k^{l+1}(z) = \frac{1}{2}\sum_{i=0}^{1} f_{2k-i}^{l}(z),$$

where f_k^{l+1} is the PDF for $\mathbf{c}_k^{l+1}$. Then, the mean of the MoG is formulated follow as:

Table 1. Performance comparisons between *full-trining* and *fine-tuning* across various DNNs on the ImageNet-1K dataset [5].

Backbone	*full-training*	*fine-tuning*	Δ
DeiT-T [3]	74.5%	**74.8%**	+0.3
DeiT-S [3]	**82.8%**	82.6%	-0.2
DeiT-B [3]	83.3%	**83.4%**	+0.1
Swin-T [4]	83.2%	**83.5%**	+0.3
Swin-S [4]	83.6%	**84.1%**	+0.5

$$\begin{aligned}\mu_k^{l+1} &= \int z f_k^{l+1}(z)dz \\ &= \frac{1}{2}\sum_{i=0}^{1}\int z f_{2k-i}^{l}(z)dz \\ &= \frac{1}{2}\sum_{i=0}^{1}\mu_{2k-i}^{l}.\end{aligned}$$

The standard deviation $(\sigma_k^{l+1})^2$ is derived as follow:

$$\begin{aligned}(\sigma_k^{l+1})^2 &= \int z^2 f_k^{l+1}(z)dz - (\mu_k^{l+1})^2 \\ &= \frac{1}{2}\sum_{i=0}^{1}\int z^2 f_{2k-i}^{l}(z)dz - (\mu_k^{l+1})^2 \\ &= \frac{1}{2}\sum_{i=0}^{1}((\mu_{2k-i}^{l})^2 + (\sigma_{2k-i}^{l})^2) - (\mu_k^{l+1})^2.\end{aligned}$$

C. Additional Results

C.1. Fine-tuning vs. full-training.

We also investigate the effectiveness of our proposed method when it is applied to training the model from scratch. For fair comparisons, we evaluate the classification performance of Hi-Mapper trained with the *full-training* scheme (350 epochs) and *fine-tuning* scheme (baseline + 50 epochs) of the same learning objectives on ImageNet-1K [5]. As shown in Tab. 1, the experimental results demonstrate that the *fine-tuning* scheme is better-suitable than *full-training* in terms of understanding the structural organization of visual scenes.

D. Additional visualization

For a more comprehensive understanding, we will provide additional visualization results that are included in the main paper and also examine the visual hierarchy in CNNs [2], as shown in Figure 2, 3. This will offer insights into the feature representation aspects in transformer structures and CNNs, as well as the benefits of applying our method.

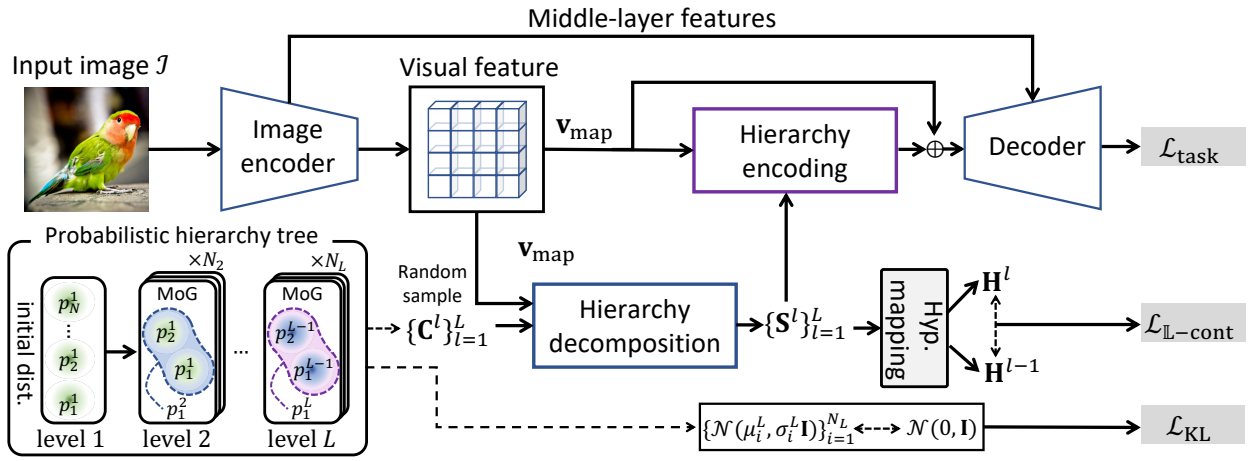

Figure 1. Illustration for overall procedure of Hi-Mapper for dense prediction tasks.

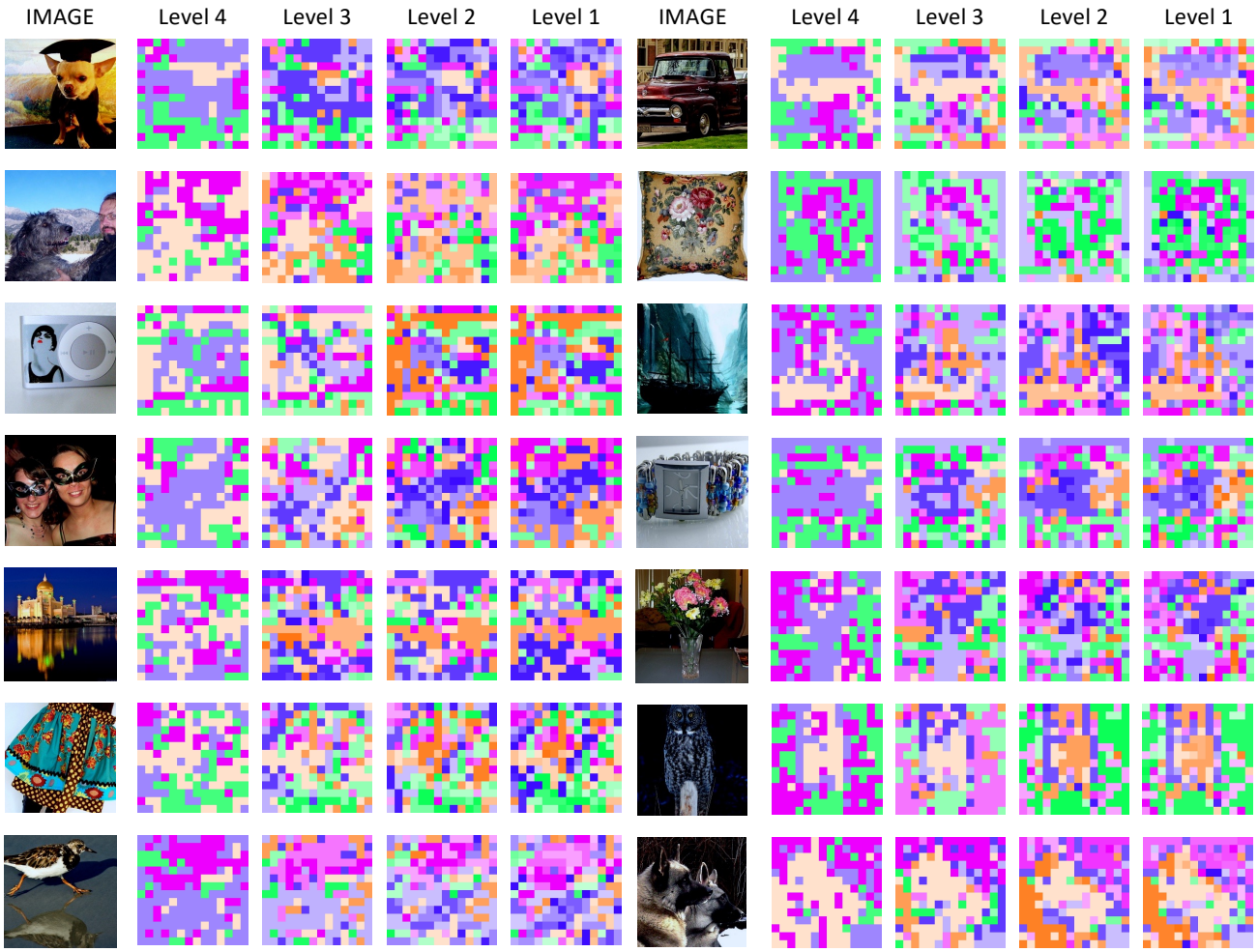

Figure 2. Visualization of visual hierarchy trees decomposed by Hi-Mapper(DeiT-S) trained on ImageNet-1K with classification objective. The same color family represents the same subtree.

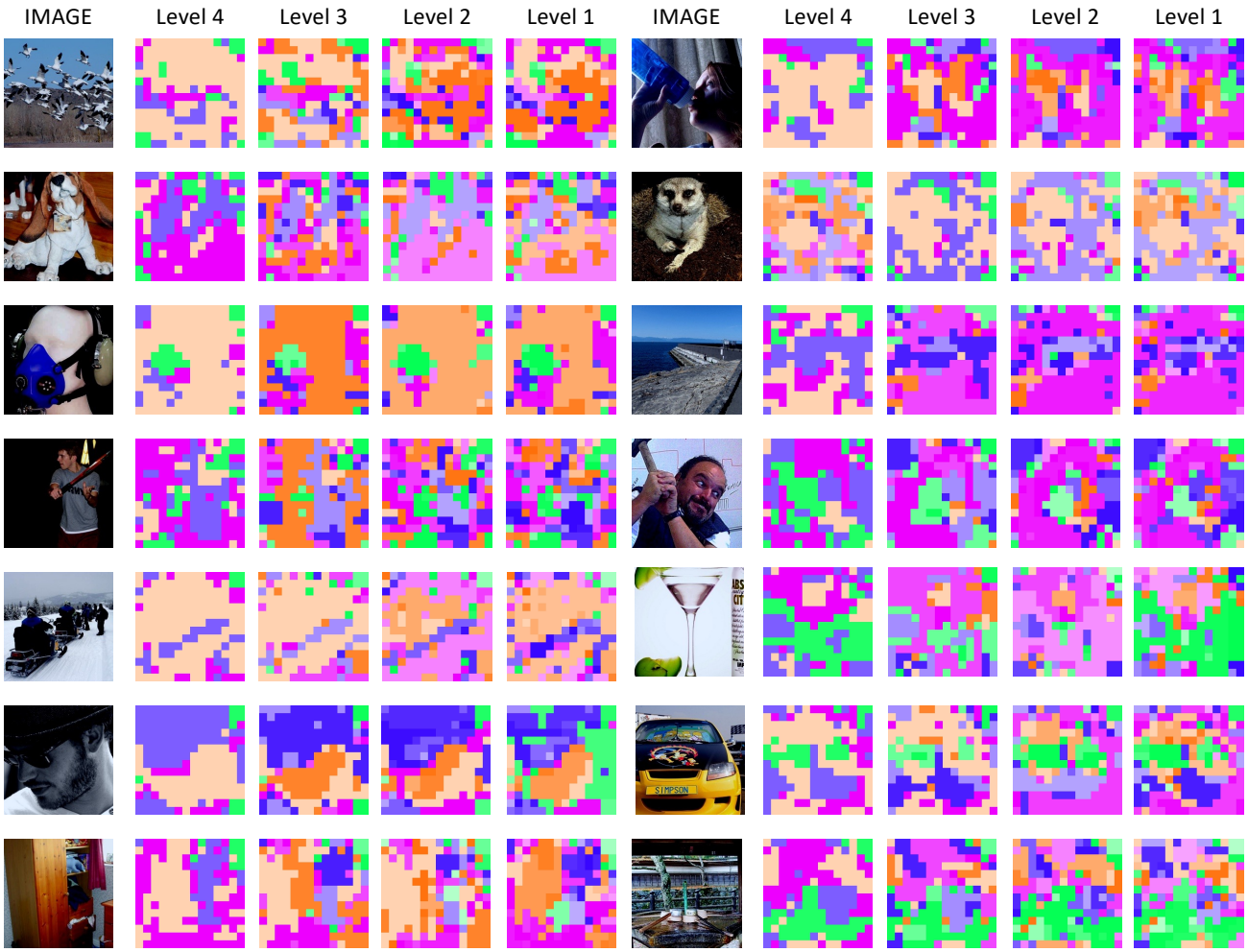

Figure 3. Visualization of visual hierarchy trees decomposed by Hi-Mapper(ENB4) trained on ImageNet-1K [5] with classification objective. The same color family represents the same subtree.

References

[1] 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*, pages 770–778, 2016. 1

[2] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[3] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[6] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1