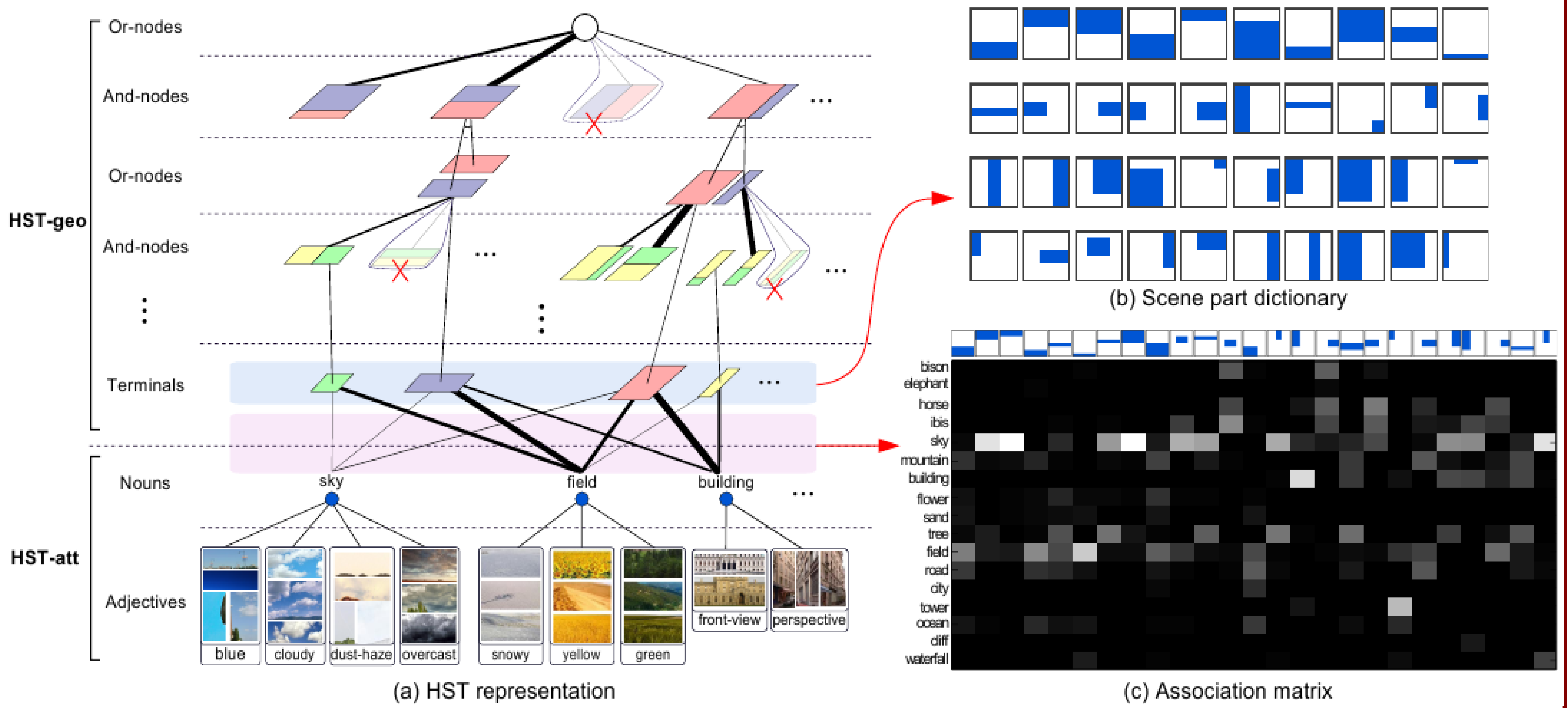
Weakly Supervised Learning for Attribute Localization in Outdoor Scenes

Shuo Wang[1,2], Jungseock Joo[2], Yizhou Wang[1] and Song Chun Zhu[2]

[1]Nat'l Engineering Lab for Video Technology, Peking University; [2]Center for Vision, Cognition, Learning and Arts, UCLA

Introduction



(a) HST representation

(b) Scene part dictionary

(c) Association matrix

Scene configuration: *spatial layouts* of a scene which are composed by the objects and regions of varying shapes

Scene attributes: described by text, contain the *nouns and adjectives*, corresponding to semantic meanings of the objects/regions and their characteristics

Hierarchical Space Tiling (HST):

- *HST-geo:* quantizes the huge scene configuration space
- *HST-att:* model the noun attribute as an appearance-Or node having a mixture of adjectives

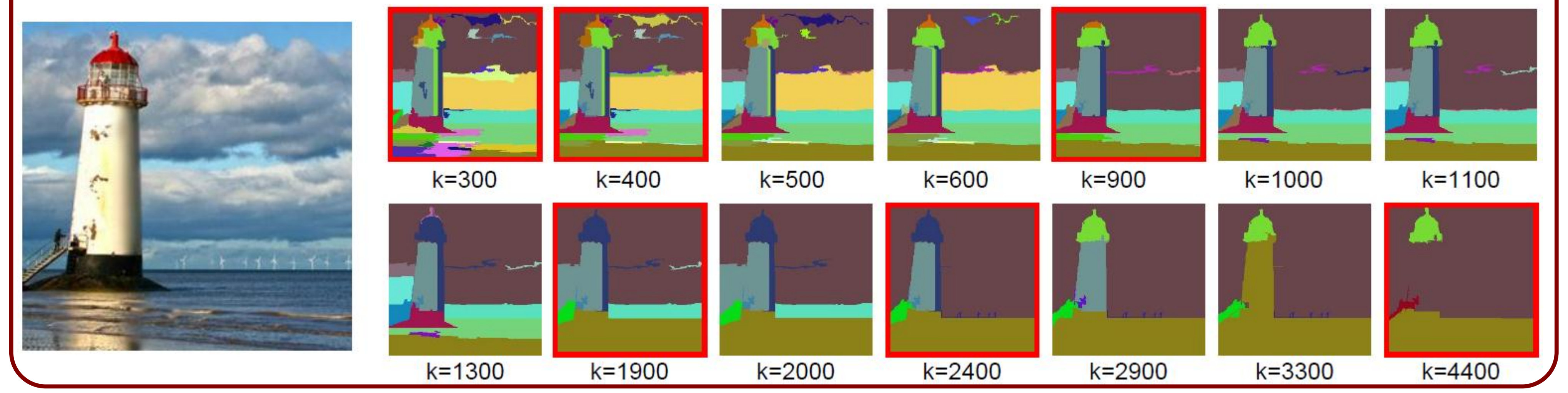
Association matrix: measures the co-occurrence of a local region and an object, e.g., the grassland always appears at the bottom area of an image.

Weakly supervised method: given a collection of natural images associated with attributes in text, where the precise localization of each attributes left unknown, we simultaneously learn the scene configurations and attributes

Learning & Inference

Input: images + text descriptions

1. Learn HST-geo:
   (a) For each image, do multi-scale segmentation
   (b) Infer the optimal configuration for each image based-on the multi-scale segmentation
   (c) Update HST-geo, then repeat (b) and (c) until convergence
2. Pursue association matrix by non-maximum suppression
3. Jointly inference: for an image and its text description, infer the optimal configuration and attribute localization
4. Repeat 2 & 3 until convergence



Dataset

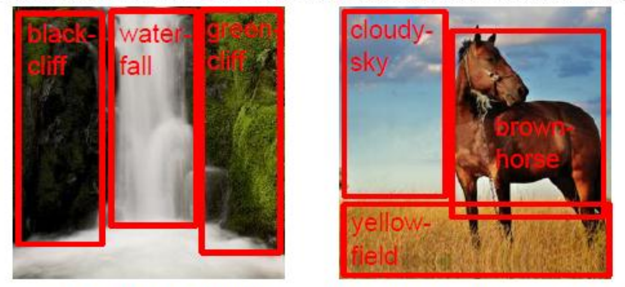
- 1226 images (256×256) from 12 categories
- 17 noun attributes and 30 noun+adjective attribute pairs
- Ground truth bounding box for evaluation
- *http://www.stat.ucla.edu/shuo.wang/SceneAtt.rar*

cloudy sky
yellow field, mountain

green field, blue sky
white tower

dust-haze sky, road
green trees
prespective buildings

white horse
green grassland



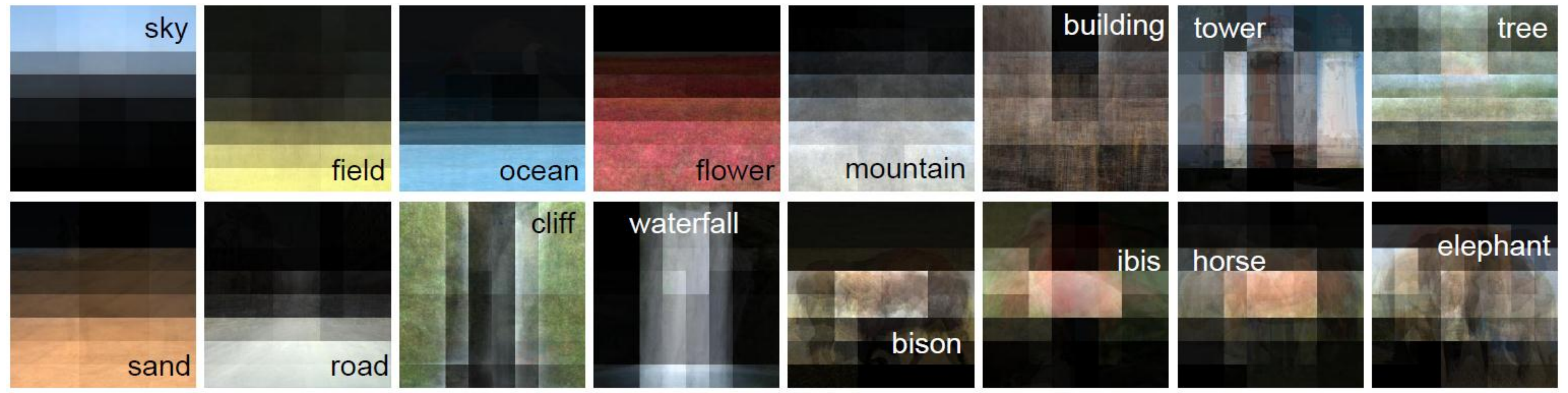
waterfall, black cliff
green cliff

cloudy sky, brown horse
yellow field

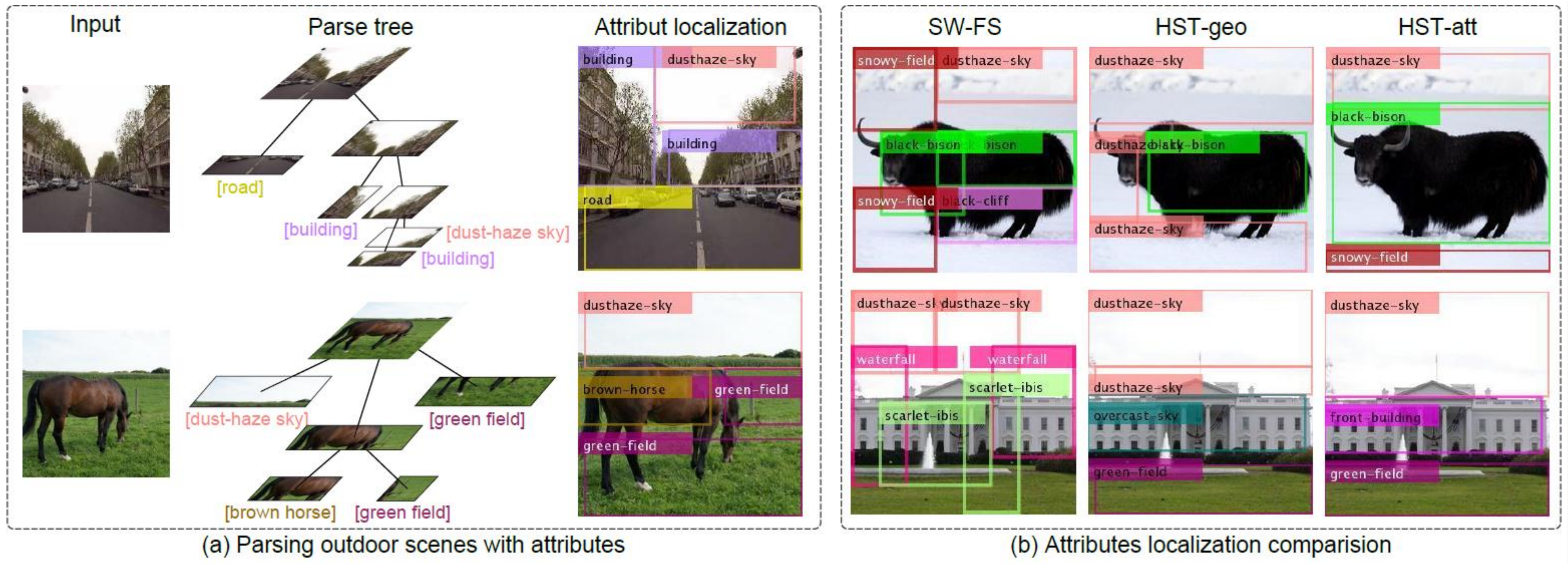
Ground-truth for evaluation

Experiments

- The association of noun attributes and scene parts

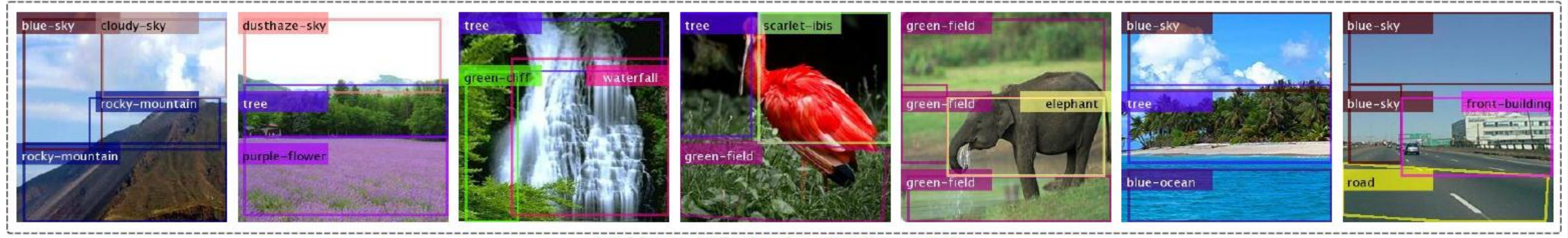


- Attribute localization results and comparison



(a) Parsing outdoor scenes with attributes

(b) Attributes localization comparision



(c) More attributes localization results

- Scene attribute recognition

	cKernel+SVM	BoW+SPM	HST-geo	HST-att
MAP(%)	64.48	53.11	51.67	**67.58**

- Scene attribute localization

	SW-FS	HST-geo	HST-att
MAP(%)	33.88	32.55	**50.22**