Show or Tell? A Benchmark To Evaluate Visual and Textual Prompts in Semantic Segmentation

Supplementary Material

6. Semantic support set generation

In Algorithm 1 we describe in detail how the semantic support set $\mathcal{S}^{sem}$ has been generated.

Algorithm 1 Generation of Semantic Support Set

Input: Set of class IDs $\mathcal{C}$, Set of training images names for each class $\mathcal{D}_c$, Number of visual prompts k
Output: Semantic support set $\mathcal{S}^{sem}$

$\mathcal{S}^{sem} \leftarrow \emptyset$
for each $c \in \mathcal{C}$ **do** ▷ Iterate over class IDs
  $\mathcal{I}_s \leftarrow \emptyset$ ▷ Support images
  $\mathcal{M}_s \leftarrow \emptyset$ ▷ Support masks
  $\mathcal{N}_s \leftarrow \emptyset$ ▷ Support names
  while $|\mathcal{I}_s| < k$ **do**
    Sample $n_s \sim \mathit{Uniform}(\mathcal{D}_c \setminus \mathcal{N}_s)$
    $i_s \leftarrow \texttt{LoadImage}(n_s)$
    $m_s \leftarrow \texttt{LoadMask}(n_s)$
    $m_s \leftarrow \mathbb{1}(m_s = c)$ ▷ Select only the class c
    $\mathcal{I}_s \leftarrow \mathcal{I}_s \cup \{i_s\}$
    $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \{m_s\}$
    $\mathcal{N}_s \leftarrow \mathcal{N}_s \cup \{n_s\}$
  end while
  $\mathcal{S}^{sem} \leftarrow \mathcal{S} \cup \{(\mathcal{I}_s, \mathcal{M}_s, \mathcal{N}_s)\}$
end for

7. Additional implementation details

Visual reference prompt methods. Following the original implementations, visual reference prompt methods are evaluated using DINOv2 [43] with ViT-L/14 [14] (when applicable), and SAM [25] with ViT-H [14] (when applicable).

Open-vocabulary methods. To ensure a fair comparison, we report the results for open-vocabulary methods without applying any mask refinement step (*e.g.* PAMR [1]). Furthermore, we use ViT-L/14 [14] as the visual backbone for NACLIP [21] and ProxyCLIP [27], while all other methods are evaluated using ViT-B/16 [14].

8. Dataset classes

For each dataset comprised in our SoT benchmark, we report the list of classes.

ADE20K The ADE20K [75] dataset is made up of 150 classes. The classes are: `wall`, `building`, `sky`, `floor`, `tree`, `ceiling`, `road`, `bed`, `windowpane`, `grass`, `cabinet`, `sidewalk`, `person`, `earth`, `door`, `table`, `mountain`, `plant`, `curtain`, `chair`, `car`, `water`, `painting`, `sofa`, `shelf`, `house`, `sea`, `mirror`, `rug`, `field`, `armchair`, `seat`, `fence`, `desk`, `rock`, `wardrobe`, `lamp`, `bathtub`, `railing`, `cushion`, `base`, `box`, `column`, `signboard`, `chestofdrawers`, `counter`, `sand`, `sink`, `skyscraper`, `fireplace`, `refrigerator`, `grandstand`, `path`, `stairs`, `runway`, `case`, `pooltable`, `pillow`, `screendoor`, `stairway`, `river`, `bridge`, `bookcase`, `blind`, `coffeetable`, `toilet`, `flower`, `book`, `hill`, `bench`, `countertop`, `stove`, `palm`, `kitchenisland`, `computer`, `swivelchair`, `boat`, `bar`, `arcademachine`, `hovel`, `bus`, `towel`, `light`, `truck`, `tower`, `chandelier`, `awning`, `streetlight`, `booth`, `televisionreceiver`, `airplane`, `dirttrack`, `apparel`, `pole`, `land`, `bannister`, `escalator`, `ottoman`, `bottle`, `buffet`, `poster`, `stage`, `van`, `ship`, `fountain`, `conveyerbelt`, `canopy`, `washer`, `plaything`, `swimmingpool`, `stool`, `barrel`, `basket`, `waterfall`, `tent`, `bag`, `minibike`, `cradle`, `oven`, `ball`, `food`, `step`, `tank`, `tradename`, `microwave`, `pot`, `animal`, `bicycle`, `lake`, `dishwasher`, `screen`, `blanket`, `sculpture`, `hood`, `sconce`, `vase`, `trafficlight`, `tray`, `ashcan`, `fan`, `pier`, `crtscreen`, `plate`, `monitor`, `bulletinboard`, `shower`, `radiator`, `glass`, `clock`, and `flag`.

PASCAL VOC 2012 The PASCAL VOC 2012 [16] dataset is made up of 21 classes. The classes are: `background`, `aeroplane`, `bicycle`, `bird`, `boat`, `bottle`, `bus`, `car`, `cat`, `chair`, `cow`, `diningtable`, `dog`, `horse`, `motorbike`, `person`, `pottedplant`, `sheep`, `sofa`, `train`, and `tvmonitor`.

Cityscapes The Cityscapes [13] dataset is composed of 19 classes. The classes are: `road`, `sidewalk`, `building`, `wall`, `fence`, `pole`, `trafficlight`, `trafficsign`, `vegetation`, `terrain`, `sky`, `person`, `rider`, `car`, `truck`, `bus`, `train`, `motorcycle` and `bicycle`.

UAVid The UAVid [37] dataset is made up of 7 classes. The classes are: `Building`, `Road`, `Static Car`, `Tree`, `Vegetation`, `Human` and `Moving Car`. A background class is also present in the dataset.

Trash The Trash [42] dataset is made up of 12 classes. The classes are: `Aluminium foil`, `Cigarette`, `Clear plastic bottle`, `Corrugated carton`,

Disposable plastic cup, Drink Can, Egg Carton, Foam cup, Food Can, Garbage bag, Glass bottle, Glass cup, Metal bottle cap, Other carton, Other plastic bottle, Paper cup, Plastic bag - wrapper, Plastic bottle cap, Plastic lid, Plastic straw, Pop tab and Styrofoam piece. A background class is also present in the dataset.

ZeroWaste The ZeroWaste [3] dataset is made up of 4 classes. The classes are: rigid plastic, cardboard, metal and soft plastic. A background class is also present in the dataset.

Pizza The Pizza [68] dataset is made up of 5 classes. The classes are: Mushroom, Pepper, Pepperoni, Tomato and pizza. A background class is also present in the dataset.

UECFOOD The UECFOOD [15] dataset is made up of 102 classes. The classes are: rice, eels on rice, pilaf, chicken-'n'-egg on rice, pork cutlet on rice, beef curry, sushi, chicken rice, fried rice, tempura bowl, bibimbap, toast, croissant, roll bread, raisin bread, chip butty, hamburger, pizza, sandwiches, udon noodle, tempura udon, soba noodle, ramen noodle, beef noodle, tensin noodle, fried noodle, spaghetti, Japanese-style pancake, takoyaki, gratin, sauteed vegetables, croquette, grilled eggplant, sauteed spinach, vegetable tempura, miso soup, potage, sausage, oden, omelet, ganmodoki, jiaozi, stew, teriyaki grilled fish, fried fish, grilled salmon, salmon meuniere, sashimi, grilled pacific saury, sukiyaki, sweet and sour pork, lightly roasted fish, steamed egg hotchpotch, tempura, fried chicken, sirloin cutlet, nanbanzuke, boiled fish, seasoned beef with potatoes, hambarg steak, beef steak, dried fish, ginger pork saute, spicy chili-flavored tofu, yakitori, cabbage roll, rolled omelet, egg sunny-side up, fermented soybeans, cold tofu, egg roll, chilled noodle, stir-fried beef and peppers, simmered pork, boiled chicken and vegetables, sashimi bowl, sushi bowl, fish-shaped pancake with bean jam, shrimp with chill source, roast chicken, steamed meat dumpling, omelet with fried rice, cutlet curry, spaghetti meat sauce, fried shrimp, potato salad, green salad, macaroni salad, Japanese tofu and vegetable chowder, pork miso soup, chinese soup, beef bowl, kinpira-style sauteed burdock, rice ball, pizza toast, dipping noodles, hot dog, french fries, mixed rice, goya chanpuru, others and beverage. A background class is also present in the dataset.

Toolkits The Toolkits [39] dataset is made up of 8 classes. The classes are: Allen-key, block, gasket, plier, prism, screw, screwdriver and wrench. A background class is also present in the dataset.

PIDray The PIDray [73] dataset is made up of 12 classes. The classes are: Baton, Pliers, Hammer, Powerbank, Scissors, Wrench, Gun, Bullet, Sprayer, HandCuffs, Knife and Lighter. A background class is also present in the dataset.

House-Parts The House-Parts [55] dataset is made up of 22 classes. The classes are: aluminium door, aluminium window, cellar window, mint cond roof, plaster, plastic door, plastic window, plate fascade, wooden door, wooden fascade, wooden window and worn cond roof. A background class is also present in the dataset.

MHPv1 The MHPv1 [29] dataset is made up of 17 classes. The classes are: hat, hair, sunglasses, upper clothes, skirt, pants, dress, belt, left shoe, right shoe, face, left leg, right leg, left arm, right arm, bag and scarf. A background class is also present in the dataset.

LoveDA The LoveDA [60] dataset, which is composed by LoveDA-Rural and LoveDA-Urban, is made up of 6 classes. The classes are: building, road, water, barren, forest and agriculture. A background class is also present in the dataset.