Report tecnico: Pre-labeling dataset addestramento (SegFormer)

11 Agosto 2025

1 Obiettivo

Preparare un set di immagini di cabine primarie per la fase di labeling, con due vincoli fondamentali:

1. Distribuzione geografica bilanciata (quote per regione).
2. Diversità morfologica garantita (selezione via clustering sulle feature visive).

Sono state escluse tutte le cabine già etichettate e quelle senza coordinate valide.

2 Origine dei dati

- Database: `PostgreSQL/PostGIS`, tabella `public.cabine`.
- Condizioni di selezione:

$$tipo\_cabina\ IS\ NULL \quad \wedge \quad geom\_centered \neq NULL$$

- Colonne utilizzate: `chk`, `lat`, `lng`, `regione`, `provincia`, `area_regionale`.

3 Filtraggio iniziale

Esclusione già etichettati

Si è effettuato il parsing dei file in `LABELED_DIR` estraendo il campo `chk` dal nome file tramite espressione regolare:

$$\text{regex} : (?P < chk > [A - Za - z0 - 9]+)_{latlonz}oom19$$

Tutti i `chk` presenti sono stati rimossi dal set del DB.

4 Estrazione feature (fase "prepare_features_dataset")

Per ciascun candidato:

1. Download thumbnail satellitare (Esri World Imagery) a **zoom 15**, dimensione 256 × 256 px, centrata sulle coordinate.

2. Calcolo feature RGB:

$$\begin{aligned}
\mu_R, \mu_G, \mu_B &\quad \text{(media per canale)} \\
\sigma_R, \sigma_G, \sigma_B &\quad \text{(deviazione standard per canale)} \\
\text{green_frac} &= \frac{\#\{g > r \wedge g > b\}}{N_{pixel}} \\
\text{dry_frac} &= \frac{\#\{v < 0.45 \wedge \neg(g > r \wedge g > b)\}}{N_{pixel}} \\
v &= \max(R, G, B) \\
\text{edge_density} &= \frac{\overline{|\nabla_x v|} + \overline{|\nabla_y v|}}{2} \\
\text{histogrammi RGB :} &\quad \text{16 bin normalizzati per canale}
\end{aligned}$$

   Totale feature per immagine: $3 + 3 + 1 + 1 + 1 + 48 = 57$.

3. Salvataggio in `features.csv` con colonna `features` in formato JSON.

5 Allocazione quote regionali

Dato:

$$priority = 0.6 \cdot \frac{1}{1 + labeled\_reg} + 0.4 \cdot \frac{1}{1 + labeled\_prov}$$

dove `labeled_reg` e `labeled_prov` sono conteggi etichettati per regione/provincia.

Correzione macro-aree:

$$w_{macro} = \begin{cases} 1.0 & \text{Nord} \\ 1.1 & \text{Centro} \\ 1.3 & \text{Sud} \\ 1.4 & \text{Isole} \end{cases}$$

Quota per regione:

$$quota\_regione = \frac{\overline{priority}_{reg} \cdot n_{reg}}{\sum_{tutte} \overline{priority} \cdot n} \cdot N_{target}$$

Arrotondamento e cap alla disponibilità reale.

6 Diversificazione morfologica con KMeans

Per ogni regione:

1. Selezione di q candidati secondo la quota.

2. Standardizzazione feature:
$$x' = \frac{x - \mu}{\sigma}$$

3. Clustering KMeans con $k = q$:
$$\text{minimizza} \quad \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$
dove μ_i è il centroide del cluster C_i:
$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

4. Per ogni cluster C_i, scelta del punto più vicino al centroide:
$$x^* = \arg\min_{x_j \in C_i} \|x_j - \mu_i\|$$

7 Download immagini finali

- Zoom 19, dimensione 768×768 px.
- Salvataggio in `new_images_segV2`.
- Nome file: `CHK_lat_lon_zoom19.png`.

8 Risultati

- Record iniziali DB: 1997
- Dopo esclusione etichettati: 1910
- Con feature valide: 1910
- Candidati post-join: 1394
- Selezione finale: 100 immagini

9 Distribuzione per regione

Regione	Disponibili	Selezionati	Quota %
LOMBARDIA	299	14	4.7
PIEMONTE	178	11	6.2
TOSCANA	167	8	4.8
VENETO	157	8	5.1
SICILIA	149	8	5.4
PUGLIA	136	8	5.9
CAMPANIA	139	6	4.3
CALABRIA	94	6	6.4
LAZIO	105	5	4.8
SARDEGNA	84	4	4.8
...	...	...	...

10 File prodotti

- `prepare_features_dataset/features.csv` — feature per tutti i candidati.
- `new_images_segV2/to_label_selection.csv` — metadati immagini finali.
- Cartella `new_images_segV2` — immagini z19 per labeling.