Data Leakage Detection and De-duplication in Large Scale Geospatial Image Datasets

Supplementary Material

6. Data Leakage and Overfitting

To quantify the effect of train–test leakage, we partition the test set into seen images (exact duplicates of training images) and unseen images (no overlap with training). Using the AICrowd pretrained checkpoint of HiSup [25], we observe a pronounced discrepancy in polygonal segmentation performance between these two subsets, shown in Table 5. Performance on seen test images is much higher than on unseen ones, indicating that the model primarily memorizes seen samples rather than generalizing to novel images. As a result, the originally reported test performance of HiSup [25] is inflated, demonstrating overfitting induced by data leakage.

Table 5. Polygonal segmentation results of HiSup [25] on the AICrowd test set, split into seen & unseen images.

Subset	AP	AP_{50}	AP_{75}	AR
All	79.4	92.7	85.3	81.5
Seen	79.7	92.7	85.4	81.9
Unseen	65.4	87.1	74.5	69.3

7. PHash vs AHash

To evaluate the effect of the hashing algorithm on the duplicate detection pipeline, we constructed a benchmark dataset of 10,000 images with known ground truth duplicates. The source images were drawn from a unique subset of the AICrowd validation set, ensuring no pre-existing duplicates. The dataset comprises 2,501 duplicate groups (7,501 images) & 2,499 purely unique images, yielding a 50% duplicate ratio. Each duplicate group contains one original image and two augmented variants, for a total of 5,000 intentionally created duplicates. The augmentations were sampled uniformly from six transformation types: exact copies (781), rotations of 90° (841), 180° (861), 270° (856), horizontal flips (802), & vertical flips (859). As shown in Table 6, perceptual hashing (pHash) achieved near-perfect performance while average hashing (aHash) showed lower performance. This shows that pHash is more robust to geometric transformations and less prone to false positives on this dataset.

8. Additional Qualitative Examples

In this supplementary material, we show qualitative examples of data leakage and duplication discovered in the AICrowd Mapping Challenge dataset [18] in Figures 5 and 6. Additionally, in the case of the INRIA Aerial Image Labelling Dataset [17] and the SpaceNet 2 Building Detection v2 dataset [4], we also depict some examples of the false positive duplicates identified by the deduplication pipeline in Figures 7 and 8 respectively.

Table 6. Duplicate detection on the benchmark dataset.

Method	Precision	Recall	F1 Score	FP	FN
pHash	1.0000	0.9997	0.9999	0	2
aHash	0.9164	0.9709	0.9429	664	218

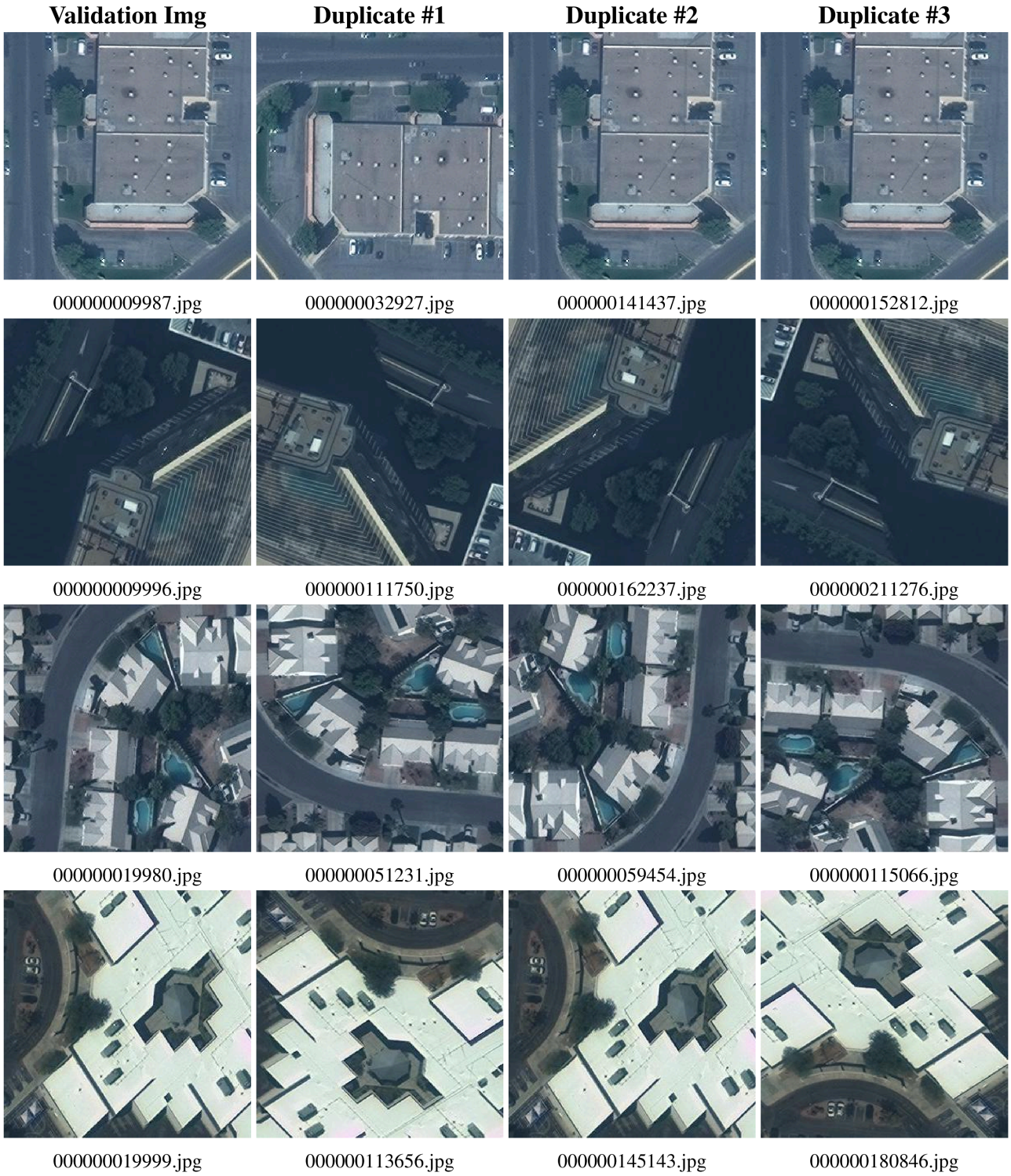

Figure 5. **Additional examples of data leakage.** Here we show additional examples of data leakage in the AICrowd Mapping Challenge dataset [18] (CC BY-NC-SA 4.0). We sample four images from the **validation split** in column 1 and show duplicates occurring in the **training split** in columns 2, 3, and 4.

Figure 6. **Additional examples of data leakage.** Here we show additional examples of data leakage in the AICrowd Mapping Challenge dataset [18] (CC BY-NC-SA 4.0). We sample four images from the **test split** in column 1 and show duplicates occurring in the **training split** in columns 2, 3, and 4.

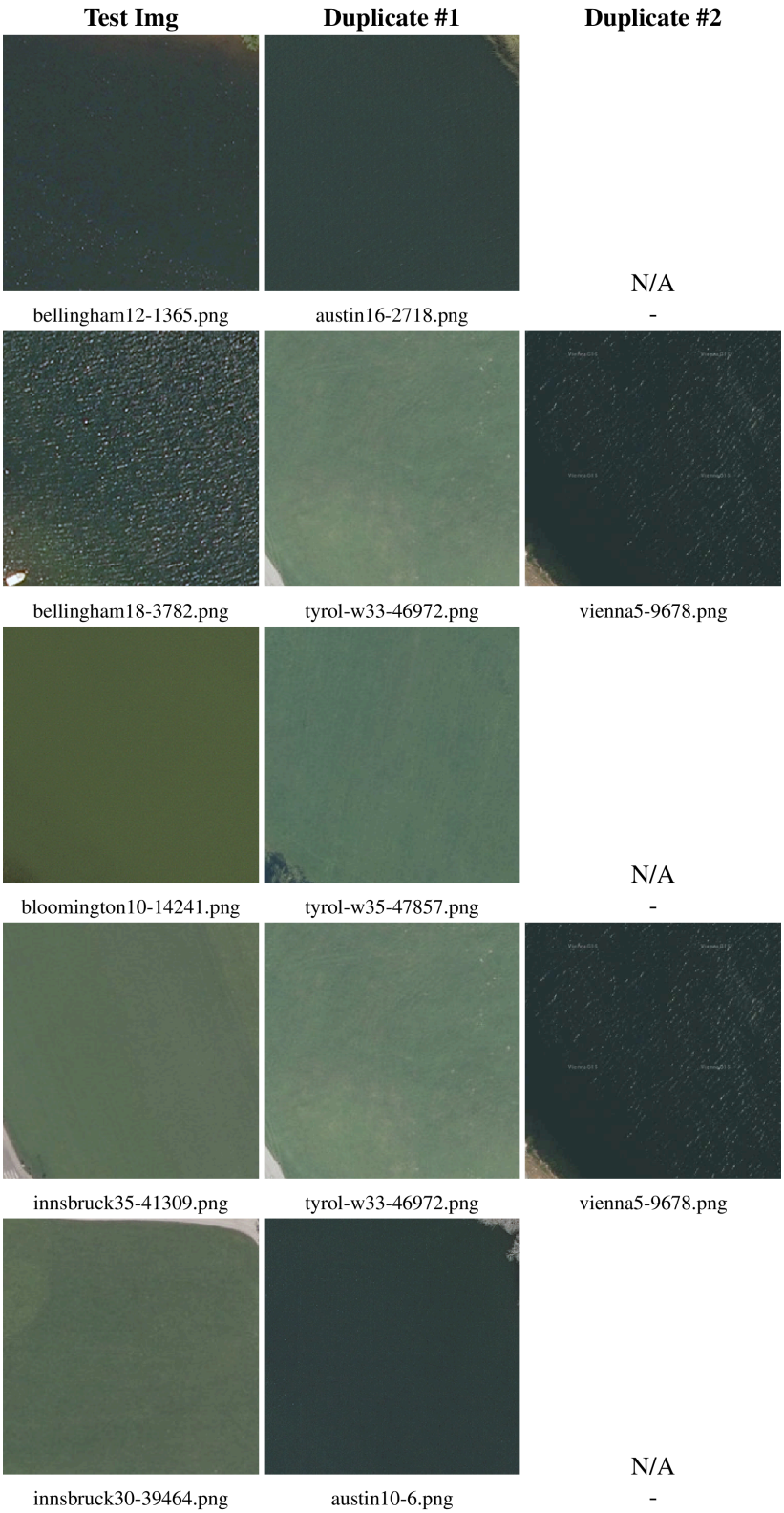

Figure 7. **False positive examples of data leakage.** Here we show falsely detected examples of data leakage in the INRIA Aerial Image Labelling dataset [17]. We sample images from the **test split** in column 1 and show duplicates occurring in the **training split** in columns 2 and 3.

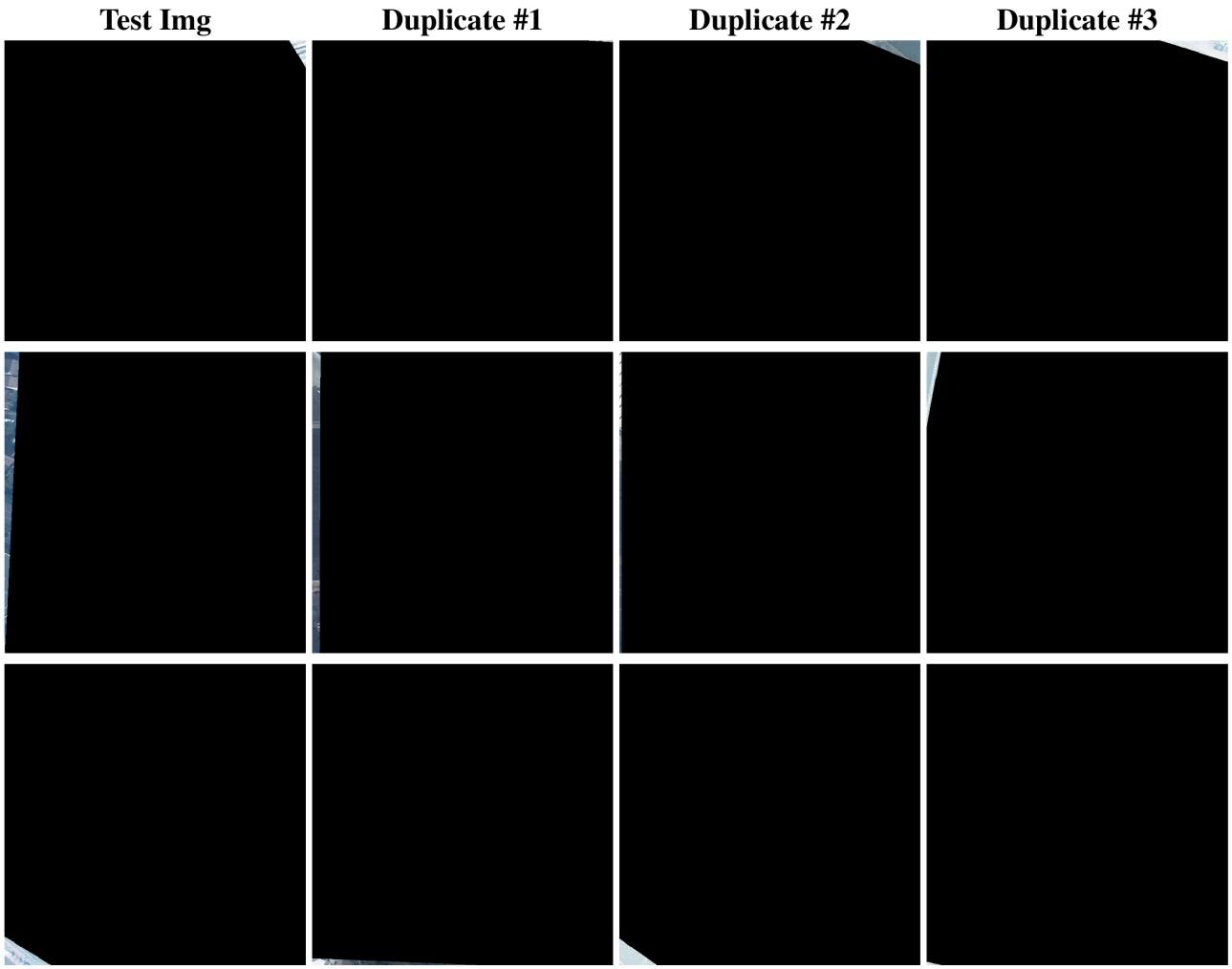

Figure 8. **False positive examples of data leakage.** Here we show examples of falsely detected examples of data leakage in the SpaceNet 2: Building Detection v2 dataset [4] (CC BY-SA 4.0). We sample four images from the **test split** in column 1 and show duplicates occurring in the **training split** in columns 2, 3, and 4.