The *Sesame Street* Archive: a labeled image repository of educational children's television, 1969-2018

Karol Sadkowski[1], Siyuan Guo[1], Chen Yu[2], Sophia Vinci-Booher[1]*

[1]Department of Psychology and Human Development, Vanderbilt University — Nashville, TN, USA

[2]Department of Psychology, University of Texas at Austin — Austin, TX, USA

{karol.sadkowski, siyuan.guo, sophia.vinci-booher}@vanderbilt.edu, chen.yu@austin.utexas.edu

Abstract

The **Sesame Street** ***Archive*** **(SSA)** *is the first image repository based on educational children's television, comprising over 35,000 primary-source film frames extracted from 4,397* Sesame Street *episodes broadcast from 1969 to 2018. Using* Sesame Street*, perhaps the most enduring and internationally adapted educational television series in the world, the SSA provides the computer vision community with curated scenes of sociocultural and educational importance televised to children throughout six historical decades. SSA film frames contain highly variable object instances, including diverse characters, languages, numeric patterns, and built environments, situated in both real-world and stylized settings. This first-look report details the creation of the SSA with a training dataset of images to showcase its breadth and potential. The SSA will be made available for research only. The training dataset discussed in this report is temporarily accessible via* `https://github.com/muppetAnon/ssa-cvpr2025`*.*

1. Introduction

Children's educational media remain underrepresented in computer vision research, despite offering remarkable intra-category variation and scene compositions capable of challenging current state-of-the-art methods and models. For example, objects categorized as faces may belong to humans, animals, or even anthropomorphized puppets of letters and digits. Scenes may depict the real world, its creative stylizations, or a mix of both. While computer vision datasets may feature categories like letters or digits in isolation, educational scenes present them together to support early literacy and numeracy. As such, children's educational media constitute an untapped resource to train models on the same real and imaginary environments through which children learn.

Leveraging *Sesame Street*'s prolific content volume spanning decades of public programming, the *Sesame Street* Archive (SSA) (Fig. 1) is the first image repository derived from children's educational media capable of supporting computer vision research. The SSA is characterized by high object variability, uncommon symbolic representations, and portrayals of a widely successful early learning curriculum. It also captures lower-level object variations, such as frequent occlusion, truncation, and label rotation and nesting. Furthermore, the SSA reflects historical changes in recording technology, including differences in frame rate, resolution, colorization, and aspect ratio (Fig. 2).

Figure 1. Logo of the *Sesame Street* Archive (SSA).

Through its careful annotations of curated film frames, the SSA emulates *Sesame Street*'s commitment to inclusive representation, highlighting various social, cultural, economic, racial and ethnic, and other human identities. *Sesame Street* began as a novel attempt to expand early childhood education through public television, focusing on disadvantaged households without preschool access [4]. Since its first episode aired in 1969, it has sought to fairly portray the demographics, homes, and experiences of its child audience, now also reflected in the SSA and available to the computer vision community.

This first-look report details the methods used to create the SSA, including the labeling of its target object categories—faces, places, words, and numbers. It then presents a benchmark for the `ssa_face` subset, and concludes with a discussion of ongoing and future efforts to scale and disseminate the SSA for computer vision research.

2. Related Works

Other datasets explicitly based on children's educational media either do not exist or are extremely difficult to locate. Near-alternatives show children as research partici-

*Corresponding author.

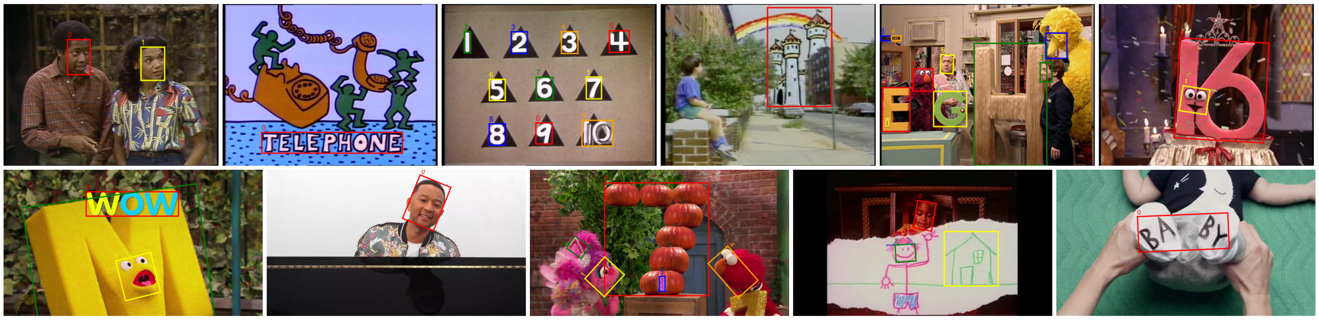

Figure 2. Example *Sesame Street* film frames in the SSA dataset, each with bounding boxes around faces, words, numbers, and buildings (places). Bottom-row frames have slightly wider aspect ratios, reflecting a historical shift in film production and broadcasting technology.

pants, or students with teachers in learning spaces, but do not focus on any educational content being learned. For example, ChildPlay [7, 22] captures child and adult gaze patterns and hand-object manipulations in children's playrooms, where interactions are centered around learning materials, but the materials themselves are not labeled and are rarely featured. Datasets of children's egocentric viewpoints, such as BabyView [13] and Toybox [23], incidentally include learning materials but without explicit labeling or consistent prevalence. Perhaps most relatable to the SSA are datasets that showcase media created by children or implicitly suitable for them, including Pencils to Pixels [17], which evaluates creativity in children's drawings, and the Hand Shadow Puppet Image Repository (HaSPeR) [20], which supports the detection of hand shadow animal puppets. Like the SSA, both Pencils to Pixels and HaSPeR feature unique object stylizations familiar to children, but still fall short in explaining their direct educational value.

Beyond the absence of datasets explicitly based on children's educational media, an absence also exists in datasets that simultaneously support applications in face detection, optical character recognition, and scene understanding. Instead, thematically comparable datasets tend to specialize in a single application: Young Labeled Faces in the Wild (YLFW) [15] and FairFace [12] focus exclusively on faces; DrawEduMath [1] and Street View House Numbers (SVHN) [18] are limited to word and number characters; and Scene UNderstanding (SUN) [25] is dedicated to built environments. Embracing these frequently isolated applications, the SSA is a multipurpose resource with strong potential to advance computer vision.

3. The *Sesame Street* Archive Dataset

The SSA is projected to contain approximately 113,751 labeled object instances across at least 35,000 *Sesame Street* film frames. As a first step toward this scale, human coders created a training dataset of 15,705 labeled instances across 4,832 frames, targeting four object categories: faces, places, words, and numbers (Tab. 1). These categories were selected for their relevance across research domains, including computer vision [27] and brain sciences [10, 11]. As current out-of-the-box models fail to identify many ground truth instances in the SSA (Fig. 3), models aligned with the characteristics of each training subset (e.g., `ssa_face`, `ssa_word`) will be strategically fine-tuned to support scalable labeling across the full SSA.

To calculate the projected distributions of the four training categories, a scaling factor of 7.243 was applied based on the assumption of similar object distributions and co-occurrence rates across instances and frames during scaling (Tab. 1). This relationship can be expressed as:

$$\text{Scaling Factor} = \frac{\text{Proj. Frames}}{\text{Curr. Frames}} = \frac{35{,}000}{4{,}832} \approx 7.243 \quad (1)$$

	Instances		Frames	
Subset	Current	Projected	Current	Projected
`face`	7,214	52,251	3,433	24,865
`place`	1,486	10,763	1,000	7,243
`word`	4,672	33,839	1,359	9,843
`number`	2,333	16,898	1,182	8,561
Total	15,705	113,751	4,832*	35,000*

Table 1. Current and projected compositions of `ssa_` subsets based on instance and frame counts. Asterisks indicate total frame counts are lower than the sums of their individual subsets, as object instances of different categories co-occur within frames.

3.1. Object Definitions and Annotation Schemas

Creating the SSA training dataset started with custom definitions and annotation schemas for the four target object categories, guiding the sampling of *Sesame Street* film frames and the labeling of highly variable object instances.

Within each object category label, a fixed schema of annotation attributes captures the qualitative characteristics of a given object instance. See Supplementary Material Sec. 6 (Tabs. 3 to 6) for complete object definitions and annotation schemas across the four categories.

3.2. Data Collection

GBH Media Library and Archives [9] provided 4,397 MP4 copies of original *Sesame Street* episodes, of which 343 were convenience sampled for data preprocessing. Provided metadata of the sampled episodes were cross-checked against corresponding documentation on Muppet Wiki for accuracy [16]. Dynamic string construction (e.g., concatenation of columnar values) and text normalization (e.g., removal of punctuation from episode names) were applied to generate structured episode-level identifiers for consistent file naming. For example, a row for Season 48's Episode 4835 ("The Count's Counting Error") yielded the directory identifier `S48-E4835_The-Counts-Counting-Error`. Using FFmpeg [8], film frames were then extracted as PNG image files at 1 frame per second and were named sequentially, starting from `S48-E4835_00001.png` to support both historical and temporal traceability.

Preprocessing of the 343 sampled episodes yielded 1,091,370 total *Sesame Street* film frames considered for inclusion in the SSA training dataset. Complete frame sets for the 343 sampled episodes were then individually uploaded to the Computer Vision Annotation Tool (CVAT) [5], where a data reduction process was carried out to minimize visual redundancy in frames selected for further annotation. Four trained annotators iteratively scrubbed through sequential frames, identifying and retaining those containing the clearest and most distinct instances of target objects. On average, approximately 14 frames were retained from each episode's set of 3,182 frames.

As PNGs, frames display in two resolutions: 480×360 pixels in a 4:3 aspect ratio for episodes from Seasons 1–38 (1969–2007), and 640×360 pixels in a 16:9 aspect ratio for episodes from Seasons 39–48 (2008–2018) (Fig. 2).

3.3. Image Annotation

The same four annotators applied bounding box labels to all target object instances identified in qualifying frames, following their annotation schemas preconfigured in CVAT. To support each schema with interpretive context, object-specific annotation criteria were continuously refined as rare and ambiguous edge cases (e.g., category instances with subjective attribute specifications) were flagged. These case-by-case criteria were often articulated in brief 'if–then' logic statements to make granular handling decisions increasingly exhaustive and transparent. For example:

> "If a `face` has a `frontal orientation`—in that it is not turned leftward or rightward—but it is still noticeably turned upward or downward, then its facial geometry remains non-canonical, and its `orientation` attribute value should be set to `other` instead of `frontal`."

All annotation criteria used for the training dataset will be provided together with the fully scaled SSA dataset for potential downstream use with language models.

3.4. Quality Assurance

Once preliminary annotations totaling at least 1,000 images per category subset were completed, four new annotators were assigned to independently conduct consensus annotation for each subset. By specializing exclusively in one subset, each consensus annotator remained impartial to prior annotation decisions. To further reduce bias, frames were presented to consensus annotators out of sequence by prepending file names with random numerical prefixes.

Errors from the preliminary annotation process were corrected during consensus annotation to align with established annotation criteria. However, consensus annotators also encountered previously overlooked edge cases, prompting the further refinement of existing criteria; to accommodate such refinements without altering original annotation configurations, four custom principles were formed to maintain procedural consistency across consensus annotators. These principles state that any new criterion should be:

1. Aligned with the conceptual framework of its respective object category, including its definition and schema;
2. Supplementary to and textually streamlined with its respective object category's existing criteria;
3. Compatible with previously consensus-coded annotations and generalizable to future, unseen instances; and
4. Syntactically consistent with other object categories' criteria to facilitate cross-category knowledge transfer.

Following consensus annotation, the final dataset was exported from CVAT as LabelMe 3.0 [21] XML metadata files paired with their corresponding PNGs.

3.5. Ethics and Copyright

Sesame Workshop and the Joan Ganz Cooney Center granted explicit permission to process *Sesame Street* episodes for this work. Permission was also granted to share curated *Sesame Street* film frames and their associated metadata contained in the SSA through a controlled platform, in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles [24], to support future not-for-profit research.

4. Analysis and Results

Modern computer vision systems suffer from a representation gap in child-centered understanding, stemming from

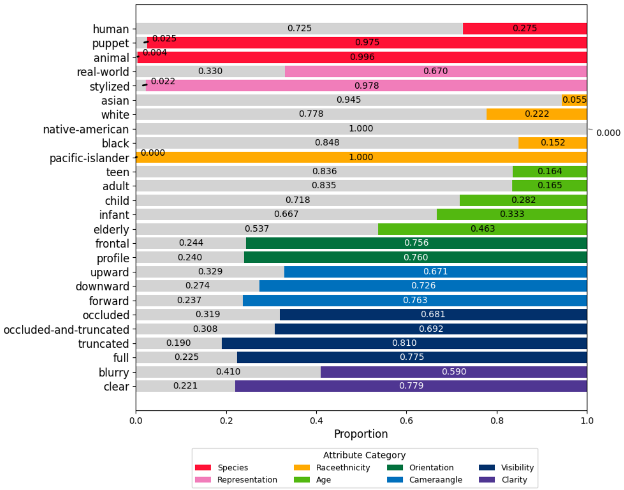

Figure 3. Normalized proportions of `ssa_face` instances missed by RetinaFace, evaluated at IoU threshold 0.5. Color-coded segments show proportions missed, while gray segments show proportions detected. Left-side attribute value names that share a common attribute are grouped by color.

an overreliance on adult-centered data. To quantify this gap, the `ssa_face` subset of 3,433 frames featuring 7,214 `face` instances was analyzed with a focus on core detection tasks. The analysis reveals two key dimensions of diversity in child-centered datasets: (1) facial diversity, with 31.31% of faces being nonhuman (e.g., puppets, animals) and 49% representing children, sharply contrasting with adult-centered datasets; and (2) diversity in task semantics, as SSA scenes are designed for educational engagement, departing from the cluttered scenes that dominate adult benchmarks. These differences introduce a domain shift in both visual appearance and task structure, posing challenges for models trained solely on adult data.

4.1. Detection Model Benchmarking

Two widely used face detection models, RetinaFace [6] and Single-Shot Scale-Invariant Face Detector (S^{3}FD) [26], were employed to evaluate the `ssa_face` subset. As both models were trained exclusively on real-world human faces, their performance declined significantly when applied to `ssa_face`, where 50% of ground truth face instances are nonhuman (e.g., puppets, stylized characters) and 26.78% of human faces are also stylized representations.

When applied to `ssa_face`, both RetinaFace and S^{3}FD exhibit a marked discrepancy between average precision (AP) and recall. While AP remains high—indicating accurate prediction ranking—recall is lower, reflecting a failure to detect a substantial portion of `face` instances. This pattern suggests the models produce sparse yet precise predictions, functioning effectively only within narrow, adult-centered domains. The resulting performance gap underscores that current face detection models are poorly equipped for the child-centered domain, where representations of human faces are often non-photorealistic.

Fig. 3 displays distributions of `ssa_face` ground truth instances missed by RetinaFace, revealing that undetected instances are typically those reflecting attribute values that are more common in child-centered than adult-centered media (e.g., puppets, stylized characters). Tab. 2 reports the precision, recall, and AP of RetinaFace and S^{3}FD on the `ssa_face` subset, evaluated using an intersection over union (IoU) threshold of 0.5, as well as the mean average precision (mAP) across IoU thresholds ranging from 0.5 to 0.95 in 0.05 increments. See Supplementary Material Sec. 7 (Figs. 4 and 5) for detailed AP trends across IoU thresholds.

	IoU@0.5			IoU@0.5:0.95
Model	Precision	Recall	AP	mAP
RetinaFace	0.80	0.24	0.94	0.57
S^{3}FD	0.79	0.22	0.95	0.51

Table 2. Performance of RetinaFace and S^{3}FD on the `ssa_face` subset, evaluated at IoU thresholds 0.5 and 0.5:0.95.

5. Ongoing and Future Work

As of this first-look report, the next step in advancing the *Sesame Street* Archive (SSA) is to scale it to at least 35,000 labeled images. In parallel, a dedicated research platform is being developed to support multidisciplinary SSA use, as well as the formation of an independent consortium to govern data use and content moderation. Future efforts may also include enhancing the SSA's multimodal data capabilities, incorporating scenes from international *Sesame Street* adaptations, establishing a benchmark to assess the educational elements of other child-centered resources, and exploring the theoretical potential for improving vision model generalization by first introducing models to inputs that are developmentally appropriate for children.

Downstream use cases of the SSA that support socially impactful research could involve pose and gesture analysis to help preserve disappearing puppetry traditions; facial expression and affective state analysis to inform cognitive-behavioral play therapy interventions; and intercultural scene analysis to develop culturally responsive AI. Popular children's television series that followed *Sesame Street*, including *Reading Rainbow* (1983–2006) [2], *Bill Nye the Science Guy* (1993–1998) [19], and *The Magic School Bus* (1994–1997, 2017-2021) [3, 14], could become foundations for complementary datasets, further positioning educational children's media as a dynamic resource for the computer vision community.

References

[1] Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T. Heffernan, and Kyle Lo. DrawEduMath: Evaluating Vision Language Models with Expert-Annotated Students' Hand-Drawn Math Images, 2025. arXiv:2501.14877.

[2] LeVar Burton, Stephen Horelick, and Jennifer Betit Yen. Reading Rainbow. https://www.imdb.com/title/tt0085075/, 1983. Accessed: 2025-04-07.

[3] Joanna Cole, Bruce Degen, Kristin Laskas Martin, Lily Tomlin, Daniel DeSanto, and Lisa Jai. The Magic School Bus. https://www.imdb.com/title/tt0108847/, 1994. Accessed: 2025-04-07.

[4] Joan Ganz Cooney. A report to Carnegie Corporation of New York: The Potential Uses of Television for Preschool Education. https://joanganzcooneycenter.org/publication/potential-uses-of-television-in-preschool-education/, 2019. Accessed: 2025-03-10.

[5] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). https://www.cvat.ai/, 2023. Accessed: 2025-04-03.

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild, 2019. arXiv:1905.00641.

[7] Arya Farkhondeh, Samy Tafasca, and Jean-Marc Odobez. ChildPlay-Hand: A Dataset of Hand Manipulations in the Wild, 2024. arXiv:2409.09319.

[8] FFmpeg Developers. FFmpeg 7.0 "Dijkstra". https://ffmpeg.org/, 2024. Accessed: 2025-04-03.

[9] GBH Media Library and Archives. Archives. https://www.wgbh.org/foundation/archives. Accessed: 2025-04-06.

[10] Golijeh Golarai, Alina Liberman, Jennifer M. D. Yoon, and Kalanit Grill-Spector. Differential development of the ventral visual cortex extends through adolescence. *Frontiers in Human Neuroscience*, 3, 2010.

[11] Kalanit Grill-Spector, Golijeh Golarai, and John Gabrieli. Developmental neuroimaging of the human ventral visual cortex. *Trends in Cognitive Sciences*, 12(4):152–162, 2008.

[12] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age, 2019. arXiv:1908.04913.

[13] Bria Long, Violet Xiang, Stefan Stojanov, Robert Z. Sparks, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, Daniel L. K. Yamins, and Michael C. Frank. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences, 2024. arXiv:2406.10447.

[14] Kate McKinnon, Gabby Clarke, and Lynsey Pham. The Magic School Bus Rides Again. https://www.imdb.com/title/tt3869122/, 2017. Accessed: 2025-04-07.

[15] Iurii Medvedev, Farhad Shadmand, and Nuno Gonçalves. Young Labeled Faces in the Wild (YLFW): A Dataset for Children Faces Recognition, 2023. arXiv:2301.05776.

[16] Muppet Wiki contributors. Sesame Street Episodes. https://muppet.fandom.com/wiki/Category:Sesame_Street_Episodes. Muppet Wiki (Fandom). Accessed: 2025-01-05.

[17] Surabhi S. Nath, Guiomar del Cuvillo y Schröder, and Claire E. Stevenson. Pencils to Pixels: A Systematic Study of Creative Drawings across Children, Adults and AI, 2025. arXiv:2502.05999.

[18] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[19] Bill Nye, James McKenna, Erren Gottlieb, Bill Nye, Pat Cashman, and Michaela Leslie-Rule. Bill Nye, the Science Guy. https://www.imdb.com/title/tt0173528/, 1993. Accessed: 2025-04-07.

[20] Syed Rifat Raiyan, Zibran Zarif Amio, and Sabbir Ahmed. HaSPeR: An Image Repository for Hand Shadow Puppet Recognition, 2025. arXiv:2408.10360.

[21] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008.

[22] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour, 2023. arXiv:2307.01630.

[23] Xiaohan Wang, Tengyu Ma, James Ainooson, Seunghwan Cha, Xiaotian Wang, Azhar Molla, and Maithilee Kunda. The Toybox Dataset of Egocentric Visual Object Transformations, 2018. arXiv:1806.06034.

[24] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, and et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016. Publisher: Nature Publishing Group.

[25] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[26] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^3FD: Single Shot Scale-invariant Face Detector, 2017. arXiv:1708.05237.

[27] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. Conference Name: Proceedings of the IEEE.

The *Sesame Street* Archive: a labeled image repository of educational children's television, 1969-2018

Supplementary Material

See Supplementary Material Pages 2–4 for additional tables and figures.

6. Supplement to Sec. 3.1: Object Definitions and Annotation Schemas

A **face** is the expressive focal area of any real-world, stylized, or fantastical being's head, minimally characterized by an eye and a mouth, bounded by a perceptible or estimated hairline, chin, and ears, and potentially adorned with accessories such as a hat, glasses, or face paint, or embedded within other labeled or unlabeled objects as a static fixture or anthropomorphic element. Labeled instances of human faces aim to be inclusive of all human identities but may not fully capture a real person's self-assigned identity or a stylized figure's artistic portrayal.

(a) Object Definition

Attribute	Values
representation:	real-world; stylized; other
species:	human; puppet; animal; other
race-ethnicity:	white; black; asian; native-american; pacific-islander; other
age:	infant; child; teen; adult; elderly; other
orientation:	frontal; profile; other
camera-angle:	forward; downward; upward; other
visibility:	full; occluded; truncated; occluded-and-truncated; other
clarity:	clear; blurry; other

(b) Annotation Schema

Table 3. Object definition (a) and annotation schema (b) for identifying and labeling face instances. Annotation schema attributes and values reflect object label configurations made in CVAT.

A **place** is an externally perceptible, distinctly bordered geometric structure with architectural or anthropomorphized building elements—such as an entrance, window, roof, patterned design, or material texture—interpretable as a finished or unfinished, permanent or semi-permanent, fully or predominantly static source of shelter, storage, or cultural value for human, nonhuman, or fantastical beings, depicted in any combination of photorealistic real-world settings, stylized mediums, or imaginative environments. Building entrances, whole buildings, and skylines comprise the three types of places and receive mutually exclusive object labels, even if one label appears within another.

(a) Object Definition

Attribute	Values
representation:	real-world; stylized; other
function:	domicile; business; attraction; institution; other
domicile-type:	house; row-house; apartment; castle; other
scope:	entrance; building; skyline; other
orientation:	cardinal; oblique; other
camera-angle:	forward; downward; upward; other
visibility:	full; occluded; truncated; occluded-and-truncated; other
clarity:	clear; blurry; other
text-box:	[write-in]

(b) Annotation Schema

Table 4. Object definition (a) and annotation schema (b) for identifying and labeling place instances. Annotation schema attributes and values reflect object label configurations made in CVAT.

A **`word`** is a sequence of one or more Latin-alphabet letters, optionally forming meaningful language units in English or other languages, and permissive of accent marks and internal punctuation. Words may appear in any case, typeface, or stylized form—including distorted representations such as overlapping characters or irregular spacing—and may be two- or three-dimensional, tangible or intangible, and real or invented. Words may also serve as bodies of anthropomorphized beings, outlines of stylized places, or components within larger structures that offer semantic cues, such as spoof names referencing cultural artifacts.

(a) Object Definition

Attribute	Values
`single-letter:`	`a; b; c; d; e; f; g; h; i; j; k; l; m; n; o; p; q; r; s; t; u; v; w; x; y; z; other`
`multi-letter:`	`word; nonword-pronounceable; nonword-unpronounceable; other`
`case:`	`lowercase; uppercase; mixed; other`
`proper-noun:`	`true; false`
`language:`	`english; spanish; french; other`
`visibility:`	`full; occluded; truncated; occluded-and-truncated; other`
`clarity:`	`clear; blurry; other`
`text-box:`	`[write-in]`

(b) Annotation Schema

Table 5. Object definition (a) and annotation schema (b) for identifying and labeling `word` instances. Annotation schema attributes and values reflect object label configurations made in CVAT.

A **`number`** is a sequence of one or more Arabic-alphabet digits symbolizing a numerical magnitude, regardless of mathematical correctness or formatting. Numbers may appear in any typeface or stylized form—including distorted representations such as overlapping characters or irregular spacing—and may be two- or three-dimensional, tangible or intangible. They may also serve as the bodies of anthropomorphized beings, outlines of stylized places, or components within larger structures. Symbolic numerals may cooccur with nonsymbolic representations of magnitude, such as countable arrays or clusters of objects.

(a) Object Definition

Attribute	Values
`single-digit:`	`0; 1; 2; 3; 4; 5; 6; 7; 8; 9; other`
`multi-digit:`	`true; false`
`non-symbolic-pair:`	`true; false`
`visibility:`	`full; occluded; truncated; occluded-and-truncated; other`
`clarity:`	`clear; blurry; other`
`text-box:`	`[write-in]`

(b) Annotation Schema

Table 6. Object definition (a) and annotation schema (b) for identifying and labeling `number` instances. Annotation schema attributes and values reflect object label configurations made in CVAT.

7. Supplement to Sec. 4.1: Detection Model Benchmarking

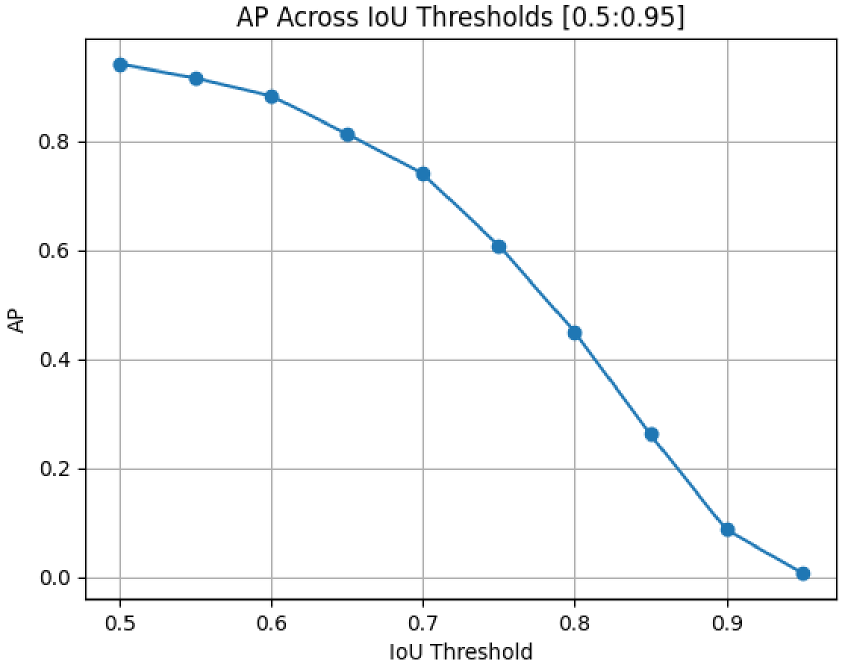

Figure 4. Average precision of RetinaFace on `ssa_face`, evaluated across IoU thresholds [0.5:0.95].

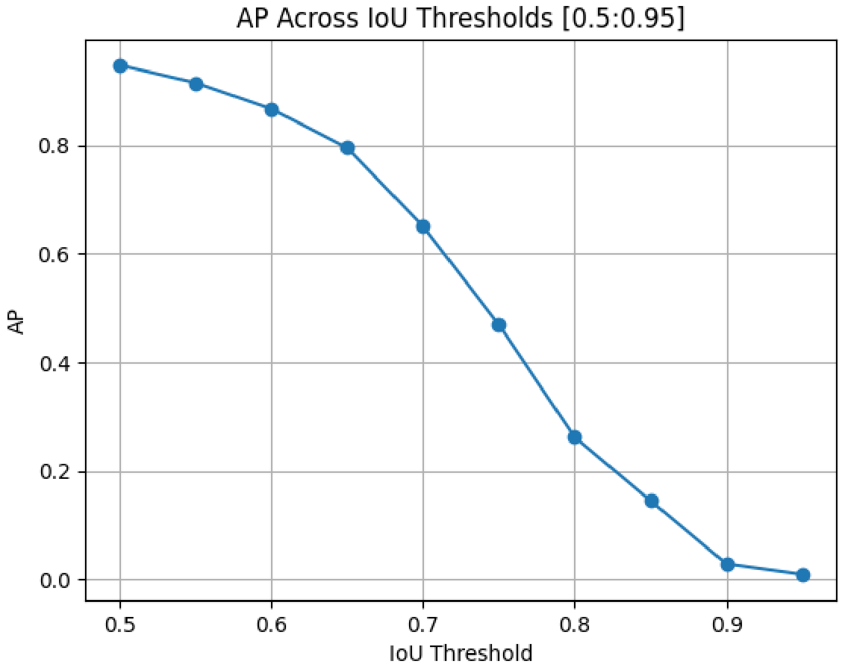

Figure 5. Average precision of S^{3}FD on `ssa_face`, evaluated across IoU thresholds [0.5:0.95].