Kippas and *Keffiyehs*: Measuring Political Polarization in Israel Through Interactions Between Jewish and Arab Citizens

Cantay Caliskan, Yuanzhu Li, and Timothy Loftus

University of Rochester

Abstract. Political fissures in Israel are often captured in sweeping metaphors—"tribes," "two camps," "parallel publics." We convert those metaphors into physical distances. Drawing on over 30 million seconds (approximately one year) of video footage across all 73 Israeli electoral districts, we apply a YOLOv9–DepthAnything pipeline to detect 449,992 pedestrians, infer their visible ethnoreligious identities, and compute 76,604 inter-pedestrian distances. The numbers are stark: Arab–Arab dyads stand 23–33 % closer to one another than to nearby Jews, while Jewish–Jewish dyads reduce the gap by 7–11 %. Insert a conspicuous religious marker—a kippah, keffiyeh, or niqab—and mixed-group distances widen by up to 60 %. These patterns hold on Tel Aviv's promenades, in ultra-Orthodox Bnei Brak, and along the narrow streets of Nazareth alike, revealing a nationwide geometry of polarization that echoes recent survey evidence of record mutual distrust. By treating sidewalk proximity as a behavioral barometer, our study offers two contributions: a scalable metric for tracking polarization *in situ* and a public, 3.047 million-frame "Israel Street-View Reservoir 2025" that can seed comparative research on divided societies. The same metric, we argue, could alert peace-building agencies to early signs of social thaw—or warn of new ideological chill—long before ballots are cast or clashes erupt.

Keywords: Computer Vision · Political Polarization · Object Classification · Computational Social Science · Religious Conflict · Israel

"Space communicates."—E. T. Hall

1 Introduction

On a busy afternoon in Jerusalem, a young Jewish woman in secular attire strides past an ultra-Orthodox man and a Palestinian Muslim mother in a hijab. They share the same crowded sidewalk, yet subtle distances separate their paths. Such quotidian spatial patterns—who walks alongside whom, and who veers away—offer a tangible microcosm of the societal cleavages that define Israel. They raise a compelling question: *to what extent do Israel's Jewish, Muslim,*

and other visible communities inhabit the same physical space, or live in parallel worlds partitioned by invisible barriers?

As (26) argues, post-Cold-War conflicts are often rooted less in ideology or economics than in deep-seated cultural identities; Israel's streets thus become a stage on which a broader civilizational rift plays out.

Culture alone, however, does not determine whether pedestrians mingle or keep their distance. Institutional contexts powerfully mediate how cultural divides translate into everyday segregation. Drawing on the insights of (24), we contend that Israel's legal regimes, housing policies, and civic institutions either reproduce or bridge ethnoreligious boundaries. Whether Jerusalem's sidewalks embody coexistence or silent segregation is therefore contingent on political choices and institutional incentives rather than fate.

Israel's urban experience parallels other divided societies. In Belfast, for example, decades of Protestant–Catholic conflict have produced segregated neighborhoods and "peace walls," with one estimate placing the annual economic cost of division at £1.5 billion (17). Studies of contested cities further demonstrate that boundaries are maintained not only by formal barriers but by the *daily throng* of ordinary people navigating urban space (44). Comparative evidence thus suggests that polarization is an evolving process shaped by both top-down policies and bottom-up interactions, admitting the possibility of change.

Despite abundant qualitative testimony, quantitatively capturing these micro-level dynamics has proved difficult. We address this gap with an innovative methodology that combines computer vision and statistical modeling. Adapting the approach of (11), hundreds of hours of street-camera footage from Jerusalem, Tel Aviv, and mixed cities are analyzed to detect pedestrians, infer visible ethnoreligious affiliation, and convert two-dimensional video into real-world inter-pedestrian distances. The resulting *polarization metric*, expressed in physical distances, reconceptualizes segregation as a dynamic, moment-to-moment phenomenon and enables comparison across conflict zones.

Beyond its academic contribution, a rigorous behavioral indicator of polarization offers practical value for peace-building. If Jews and Arabs begin to walk closer, sit nearer, and interact more freely in public, a crucial social boundary is softening—an early signal that opinion polls or election results may overlook. Conversely, widening distances can warn of escalating tensions. As (50) demonstrates, dense intergroup civic networks restrain conflict entrepreneurs; monitoring physical proximity therefore provides policymakers and NGOs with a timely barometer of communal relations and a basis for evaluating interventions.

1.1 Political Polarization in Israel

Israel's contemporary politics are acutely polarized, most visibly along the Jewish–Arab cleavage. Arab citizens comprise roughly 20 % of the population, yet decades of scholarship document persistent gaps in resources, political power, and civic trust between Jews and Arabs (47; 53). These disparities—magnified by the protracted Israeli–Palestinian conflict—have bred mutual suspicion: segments of the Jewish right portray Arab citizens as a security threat, whereas

many Arabs perceive Israel's Jewish character as constraining their full membership in the polity (47).

Institutional design intensifies these identity tensions. Israel's nationwide proportional-representation system routinely returns a dozen or more parties to the *Knesset*, embodying President Rivlin's metaphor of competing "tribes." Between April 2019 and November2022, Israelis voted five times because rival blocs could not assemble stable majorities (39). Paradoxically, this profusion of parties coalesced into two rigid blocs—pro-Netanyahu and anti-Netanyahu—"freezing" societal fractures into an increasingly zero-sum contest (9; 40).

Prime Minister BenjaminNetanyahu stands at the center of this polarization. Lauded by supporters as a security stalwart and condemned by opponents as an indicted populist, Netanyahu has leaned on ultranationalist and ultra-Orthodox partners. His 2023 judicial-overhaul initiative—which would curb Supreme Court powers and give the governing coalition decisive sway over judicial appointments—triggered the largest street protests in Israel's history (45; 41). Critics labeled the plan "executive aggrandizement" characteristic of democratic backsliding (6; 33), whereas proponents framed it as a necessary majoritarian reform, deepening partisan mistrust and raising existential questions about Israel's liberal-democratic future (43).

Jewish–Arab relations now occupy the epicenter of this polarization. Arab parties were long excluded from governing coalitions; only in 2021 did an independent Arab list (Ra'am) enter government—an unprecedented step that briefly suggested greater integration but provoked backlash from Jewish right-wing parties (47). Subsequent elections saw Arab lists fragment and lose seats, underscoring the fragility of such inclusion. Policies like the 2018 Nation-State Law, disputes over land use, and policing practices continue to reinforce feelings of second-class citizenship among Arabs and exacerbate intergroup distrust (53).

Public-opinion data confirm a sharp rise in affective polarization. Right- and left-leaning Jews increasingly report dislike and distrust of one another, while Arab citizens oscillate between engagement and alienation (9). Institutional trust has grown partisan—liberals defend the Supreme Court as a democratic bulwark, conservatives condemn it as an elitist obstacle to majority rule—mirroring trends in other democracies facing erosion (6). The October 2023 Hamas attacks, which struck amid intense domestic turmoil, further exposed the security risks of internal division and reignited debate over unity versus pluralism.

In sum, Israel illustrates how ethnic cleavages, populist leadership, and electoral rules can interact to create what comparative theorists term *pernicious polarization*. Whether cross-group trust can be rebuilt without sacrificing either Israel's Jewish identity or its liberal-democratic institutions remains an open question for scholars and practitioners alike.

These dynamics also make Israel a compelling setting for studying the spatial effects of polarization. As Dr. Jesse Ferris of the Israel Democracy Institute warns, Israel is "dangerously polarized," with opinions about Netanyahu generating "potentially unbridgeable polarization" (2023). Indeed, five elections were held between April92019 and November 1 2022 (25). Although Netanyahu's

Likud party has been the most electorally successful, his tenure has been clouded by bribery charges and conflict-of-interest allegations, and such frequent cabinet turnovers point to democratic dysfunction.

Israel's political complexity, high party fragmentation, and religious diversity therefore provide a fertile case for measuring affective polarization. The region's long, multifaceted, and contentious history—replete with competing claims and interpretations—carries profound political implications (3; 38). The conflict shapes how Palestinians and Israelis think, feel, and act, as well as their attitudes toward various religious and ethnic groups (3). Evidence that political and religious cues strongly influence perception (10; 31), alongside advances in computer-vision analysis of body language (42; 28; 51), suggests our approach can successfully detect and quantify polarization in this context. Modeling affective polarization through distance estimations thus promises to illuminate Israel's complex political atmosphere and contribute to broader comparative debates on democratic resilience.

1.2 Spatial Approaches to Understanding Polarization

While this study's methodologies are immediately pertinent to the Israeli–Palestinian context, their utility extends to other regions experiencing entrenched social and political conflicts. By harnessing advanced machine-vision techniques to quantify social interactions and degrees of polarization, our findings enrich the broader scholarly discourse on measuring and mitigating societal divisions. The empirical insights derived herein may inform the design of universally applicable interventions aimed at fostering social cohesion and preempting the intensification of conflict within diverse, multicultural settings.

The analysis of spatial comportment—specifically, the interindividual distances maintained during social engagements—yields nuanced perspectives on the underpinnings of conflict and social fragmentation. This line of inquiry, termed *proxemics*, was first conceptualized by Edward T. Hall as "the study of how human beings communicate through their use of space" (2). Proxemic research systematically examines the spatial buffers individuals adopt in interactional contexts, offering quantifiable proxies for relational intimacy and social perception (35). Empirical investigations demonstrate that conversational partners in casual standing encounters adjust their proximity based on relational closeness, suggesting that physical distance serves as a valid indicator of underlying social bonds (16). Moreover, McCall et al. (2009) report that variability in interpersonal distance and head orientation during face-to-face meetings with Black participants significantly predicted subsequent aggressiveness in virtual shooting tasks, underscoring the broader behavioral ramifications of proxemic patterns (36).

Research employing computer-vision techniques within the social sciences remains scarce, especially regarding the use of street-level imagery to quantify political polarization. A notable exception is Caliskan et al. (2024), who apply object detection and inter-pedestrian distance estimation to assess social

polarization empirically (11). Analyzing over 1,400 YouTube videos from Turkish streets—comprising roughly 170,000 frames—the authors trained Hijab-Net (based on YOLOv5) to classify individuals by religiosity and compute relative distances between detected pedestrians. Their results demonstrate pronounced in-group clustering: religious females maintain distances that are on average 23.2% shorter, religious males 18.4% shorter, and non-religious females 11.0% shorter, compared to those observed between ideologically opposed groups. These findings underscore the potential of computer vision to capture subtle manifestations of social boundary maintenance and provide a methodological template for studying polarization in other highly divided contexts, such as urban Israel.

Dietrich and Sands (2023) extend this line of inquiry by employing "bird's eye view" traffic-camera footage to measure racial avoidance behaviors on New York City sidewalks (20). Their analysis of 3,419 individuals shows that non-Black pedestrians maintain distances that are between 3.30% and 3.70% of the sidewalk width greater from Black pedestrians than from White non-Hispanic pedestrians. While this approach achieves precise distance measurement, it depends on fixed camera viewpoints and manual classification of race. Joseph et al. (2022) propose a complementary technique—transforming oblique perspective images into virtual bird's eye views via homography and Python scripts leveraging YOLOv5 and OpenCV—to automate distance computation and reduce reliance on consistent camera placement and manual annotation (29).

1.3 Spatial Approaches to Understanding Various Political Phenomena

Existing research also illuminates the effects of political and religious symbolism on political perceptions, and how computer vision has been applied to assess these computationally. Bucy (2023) analyzes partisan symbolism and affective polarization of political protests, using computer vision to show partisan divides (10). Peng (2018) used computer vision on the facial attributes of people from news footage of presidential candidates, proving that human detection models can accurately reveal political attitudes (42). Bucy et al. (2019) showed that computational methods can replicate the effectiveness of manual human perception in assessing the body language and physical gestures of people (28). Wang et al. (2017) used computer vision techniques to demonstrate how gender can inform elections and political beliefs (51). Krogstad (2017) affirms the importance of political and religious presentation in shaping perceptions (31).

Dietrich (2021) analyzed 1,413 hours of C-SPAN coverage using motion detection and found that fewer instances of members of Congress crossing the physical aisle between Democrats and Republicans were associated with more partisan voting behavior (19). Bassan-Nygate and Weiss (2022) have focused on affective polarization—"the gap between in-party affect and out-party dislike"—using Israeli National Election Studies data (5). Affective polarization intensifies as elections approach, driven by increased in-party affect rather than out-party hostility.

2 Research Questions

Building on the theoretical framework and empirical gaps outlined in the *Introduction* and *Literature Review*, this study moves beyond macro-level accounts of Jewish–Arab division by directly observing how members of each ethnonational group navigate shared urban space. We address three interrelated research questions:

RQ1: Ethnonational Proxemics. To what extent does visibly signaled ethnonational identity (Jewish vs. Arab) predict inter-pedestrian distance on mixed sidewalks in Israel's major cities?

RQ2: Gendered Dynamics. How does perceived gender interact with ethnonational identity in structuring proxemic behavior? Specifically, do women display different in-group and out-group spacing patterns than men?

RQ3: Contextual Moderators. Do situational factors—e.g., neighborhood demography, time of day, or time of the year—amplify or attenuate ethnonational and gender effects?

To guide the empirical analysis, we advance the following hypotheses, each stated in terms of expected average inter-pedestrian distance measured by our computer-vision pipeline shown in Figure 1.

H1: Visible ethnonational identity will shape everyday social interactions between Jews and Arabs.

- **H1a:** *Jewish pedestrians will, on average, walk closer to other Jews (in-group) than to Arabs (out-group).*
- **H1b:** *Arab pedestrians will likewise walk closer to other Arabs than to Jews, but the in-group/out-group gap will be smaller than that observed for Jews.*

H2: Visible gender identity will condition social interactions and interact with ethnonational cues.

- **H2a:** *Individuals perceived as sharing the same gender will maintain shorter interpersonal distances than those perceived as different genders.*
- **H2b:** *The gender effect will be stronger for women than for men.*

Together, these research questions and hypotheses allow us to test whether Israel's roads and sidewalks simply mirror entrenched cleavages documented by surveys and electoral returns, or whether micro-level spatial interactions reveal pockets of latent integration that macro indicators overlook.

Our second objective is to assemble a comprehensive *visual* resource for computational social-science researchers studying Jewish–Arab interaction and developing novel behavioral indicators of political polarization in Israel. The resulting corpus—provisionally titled **Israel Street-View Reservoir 2025** — includes 2,497 human-centered YouTube clips recorded in Jerusalem, Tel

Aviv–Yafo, Haifa, and Nazareth. To assemble ISR, we first queried YouTube in Hebrew and English using the pattern "[electoral district name] street video," yielding an initial pool of 4,020 candidate clips. Because keyword searches alone can retrieve videos from similarly named locations, we then employed three generative-AI models (ChatGPT 3.5, Gemini 1.5 Pro, and Cohere R) to infer each clip's city of origin based solely on its title and description. We retained only those videos for which all three models concurred, producing our final set of 2,497 clips. In total, these videos comprise 30,470,670 seconds of footage—equivalent to 352 days, 16 hours, 4 minutes, and 30 seconds of continuous recording. The videos were then sampled at a rate of one frame per second, yielding *3,047,067* frames. To our knowledge, **Israel Street-View Reservoir 2025** represents both the largest thematic street-level image collection in computational social science and the most extensive behavioral dataset assembled to date on Jewish–Arab polarization.

Dataset curation and model development proceeded in four stages. First, video links were scraped, filtered for camera stability and resolution, and segmented into still frames. Second, after randomly selecting ten videos from each of the four target cities (Haifa, Nazareth, Tel Aviv, and Jerusalem), a stratified subset of 12,261 frames was hand-labelled to create the training set **IS-12261**. While the class distribution in **IS-12261** is markedly imbalanced, this asymmetry reflects the composition and visibility of groups in Israeli public space; to mitigate its effects during learning we applied targeted data augmentation (horizontal flips, random crops and rescaling, and brightness/contrast jitter), with higher augmentation rates for sparse classes. Third, among several candidate object detectors we selected YOLOv9n for its favorable speed–accuracy trade-off on crowded scenes; fine-tuned on **IS-12261**, the resulting classifier **Coexist-Net** achieved a mean average precision (mAP@0.5) of 60.1 across *sixteen* target classes.[1] Finally, bounding-box centroids were projected into a homography-based bird's-eye view to convert pixel distances into real-world distances, enabling precise measurement of intergroup proxemics.

Training and inference were executed on a high-performance desktop workstation.[2] This computationally intensive workflow required approximately three months to complete. The distribution of six demographic categories within the **Israel Street-View Reservoir 2025** is presented in Figure 2.

[1] The training set comprises the following classes: *Neutral male* (6835), *Non-religious female* (4484), *Neutral female* (3083), *Non-religious male* (2654), *Religious Jewish male* (2357), *Very religious Jewish male* (975), *Religious Arab female* (373), *Religious Jewish female* (338), *Soldier male* (137), *Very religious Arab female* (130), *Very religious Jewish female* (54), *Religious Arab male* (51), *Officer male* (51), *Soldier female* (21), *Very religious Arab male* (16), and *Officer female* (6).

[2] Model training was carried out on NVIDIA RTX 3060, NVIDIA RTX 4090, and NVIDIA TITAN V GPUs; inference was performed on a desktop equipped with an Intel Core i9-13900K CPU and an NVIDIA RTX 4070 Ti GPU.

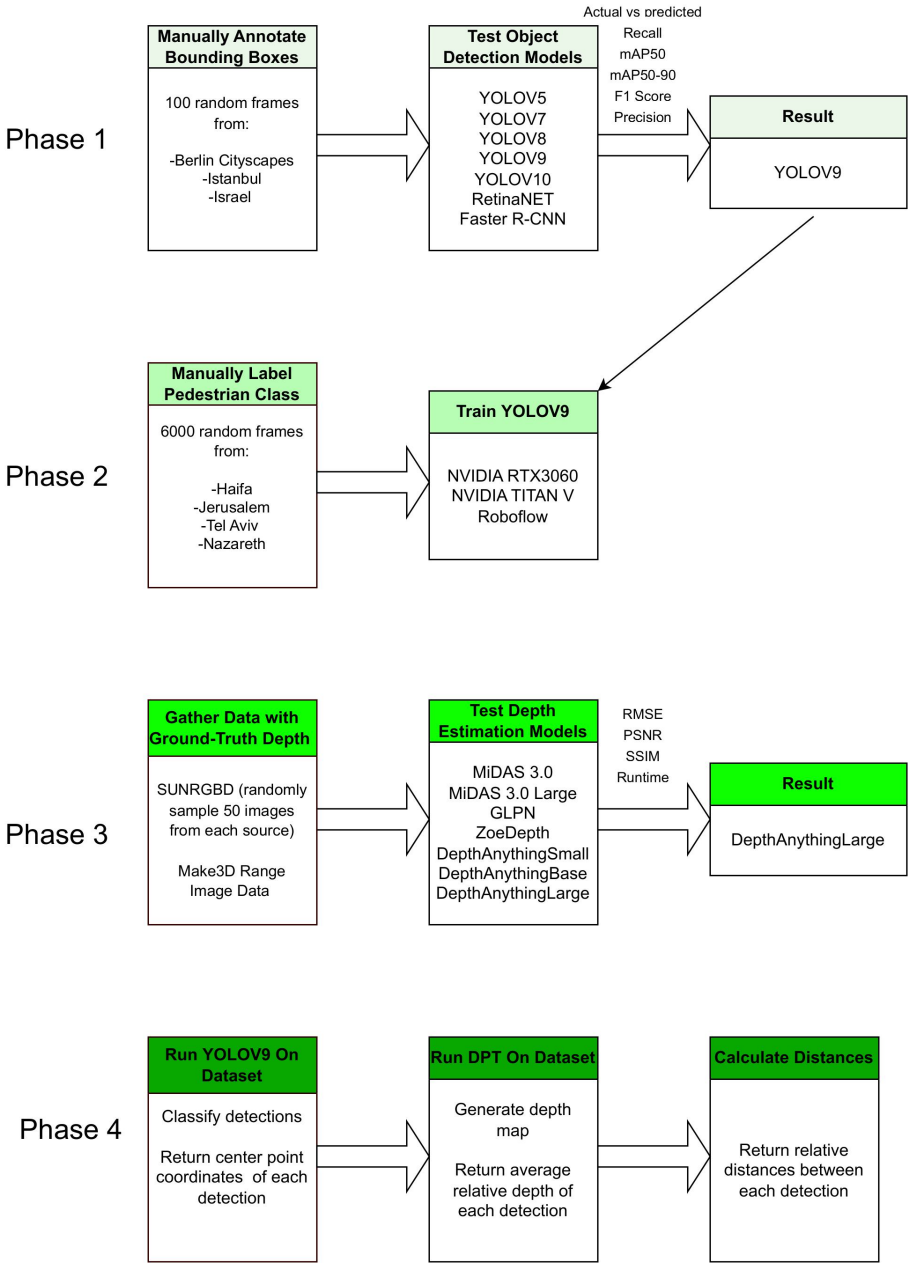
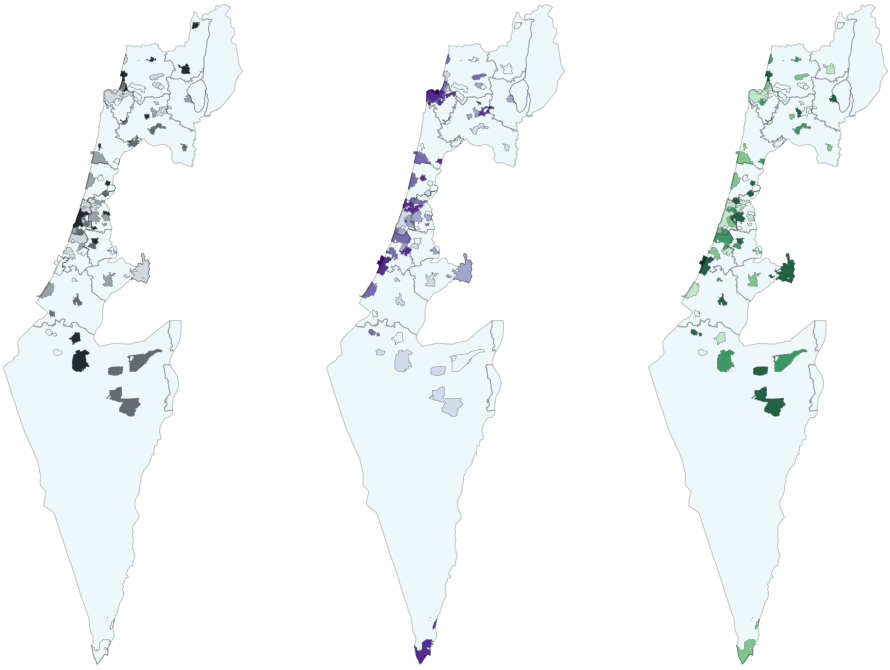
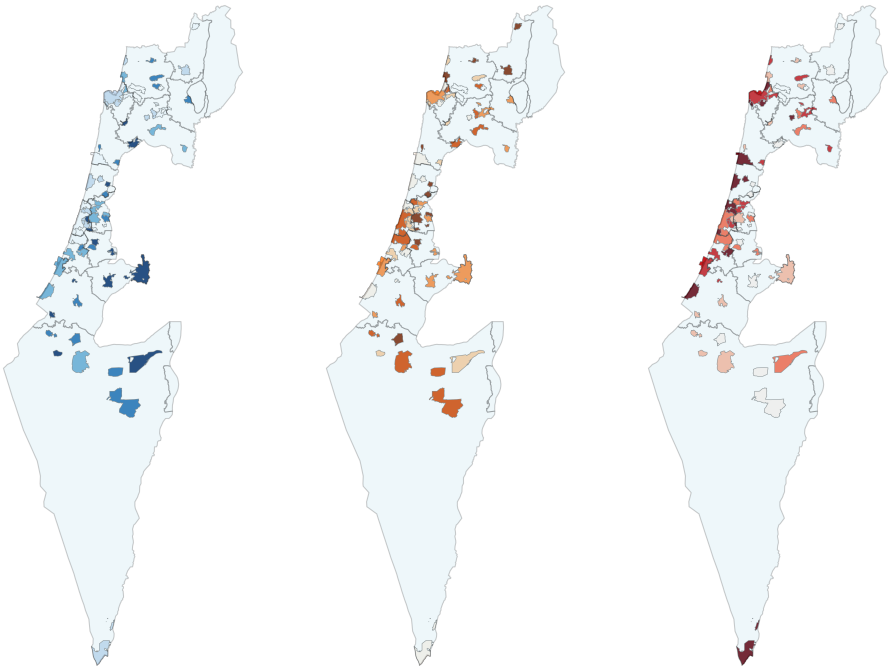


Fig. 1: Methods Pipeline



(a) Male (b) Female (c) Religious Arab



(d) Religious Jew (e) Neutral (f) Non-Religious

Fig. 2: Municipality-level Percentage Choropleths for Six Demographic Categories

3 Pedestrian Detection Methodology and Results

Object detection models define object classes based on annotated training examples. In a comparative evaluation of several pre-trained object detection architectures, prior researchers observed that no single model consistently outperforms the others across all datasets. Caliskan et al. (2024) reported that YOLOv5 and YOLOv8 achieved the highest detection accuracy on their benchmark tasks (11). Sanchez et al. (2020) further argued that broad, direct comparisons of diverse object detection models are inherently challenging due to differences in their underlying architectures and intended applications (48). Together, these findings underscore the necessity of empirically evaluating multiple object detection models on our own data to identify the most appropriate solution for our research objectives.

Faster R-CNN, a two-stage region-based convolutional neural network, first generates candidate object regions via a region proposal network and then classifies and refines each proposal (48). By contrast, the YOLO family of models divides each image into a grid and predicts bounding boxes and class probabilities simultaneously in a single forward pass (34). RetinaNet addresses the challenge of detecting objects at different scales by integrating a feature pyramid network with a novel focal loss function that down-weights easy negatives and focuses training on hard examples (34).

Given the diversity of existing people-detection models, we evaluated six pretrained architectures on a random sample of 100 frames from our dataset. To establish ground truth, we used Roboflow to manually annotate all instances of people in the selected frames. We then compared each model's outputs against these annotations using precision, recall, mean average precision at 50% IoU (mAP_{50}), mean average precision across IoU thresholds from 50% to 95% ($mAP_{50–95}$), F_1 score, and the total number of predicted versus actual people.

Precision quantifies the proportion of correct positive detections, thereby reflecting each model's ability to minimize false positives. Recall measures the proportion of true positives identified, indicating how effectively the model detects all relevant instances. Intersection over union (IoU) evaluates the overlap between predicted and ground-truth bounding boxes; mAP_{50} summarizes precision and recall at a fixed IoU threshold of 50%, while $mAP_{50–95}$ aggregates performance across multiple overlap thresholds to assess localization accuracy (see Appendix A for further details). The F_1 score harmonizes precision and recall into a single balanced metric, which is especially useful when class distributions are uneven. Finally, comparing the total predicted count of people against the actual count reveals any systematic tendencies toward over- or under-detection. By employing this comprehensive set of metrics, we aim to rigorously assess each model's strengths and limitations in detecting people within our imagery.

Prior research has demonstrated that both YOLO variants and R-CNN architectures achieve strong performance in pedestrian detection, although their relative effectiveness varies across datasets (11; 48). Caliskan et al. (2024) further observed that earlier YOLO iterations—specifically YOLOv5 and YOLOv7—can outperform more recent models on human detection tasks. To identify the opti-

mal object detection model for street-level pedestrian recognition, we sampled 100 frames from three datasets: the Israel Street-View Reservoir 2025 (hereafter "Israel Dataset"), the Berlin Cityscapes Dataset (13) (hereafter "Berlin Dataset"), and the Istanbul Street-View Reservoir 2024 (11) (hereafter "Istanbul Dataset"). The Istanbul Street-View Reservoir closely parallels the imagery characteristics of our current work. The Cityscapes Berlin Dataset is a large-scale benchmark for semantic segmentation and object detection in urban environments, comprising high-resolution street scenes from diverse areas of Berlin, Germany, which makes it particularly well suited for evaluating pedestrian detection in complex metropolitan settings.

Our evaluation demonstrated that the YOLO family of models consistently outperformed Faster R-CNN and RetinaNet in pedestrian detection. The substantially lower performance of Faster R-CNN and RetinaNet, relative to YOLO variants, may be attributable to architectural differences. YOLO architectures are optimized for real-time, high-throughput inference by processing the entire image in a single forward pass, whereas Faster R-CNN and RetinaNet employ more complex, region-based mechanisms (34). Although performance varied across datasets, integrating quantitative detection metrics with runtime efficiency considerations indicates that YOLOv9 constitutes the optimal choice for this study.

Model	YOLOv5	YOLOv8	YOLOv9	YOLOv10	Faster R-CNN	RetinaNET
Precision	0.675	0.672	0.762	0.643	0.029	0.031
Recall	0.555	0.516	0.513	0.5	0.069	0.069
mAP50	0.601	0.581	0.601	0.559	0.029	0.031
mAP50-95	0.322	0.327	0.343	0.325	0.007	0.01
F1 Score	0.609	0.584	0.613	0.563	0.041	0.043
Actual Instance	318	318	318	318	318	318
Predicted Instance	261	244	214	247	314	302

Table 1: Comparison of different models on Israel Dataset (n=100)

Model	YOLOv5	YOLOv8	YOLOv9	YOLOv9	Faster R-CNN	RetinaNET
Precision	0.72	0.787	0.816	0.749	0.025	0.029
Recall	0.661	0.682	0.72	0.682	0.055	0.06
mAP50	0.724	0.765	0.781	0.755	0.025	0.029
mAP50-95	0.312	0.327	0.331	0.332	0.008	0.01
F1 Score	0.689	0.731	0.765	0.714	0.034	0.039
Actual Instance	236	236	236	236	236	236
Predicted Instance	204	228	223	187	213	204

Table 2: Comparison of different models on Berlin Cityscapes Dataset (n=100)

Model	YOLOv5	YOLOv8	YOLOv9	YOLOv10	Faster R-CNN	RetinaNET
Precision	0.798	0.831	0.802	0.753	0.028	0.03
Recall	0.728	0.748	0.735	0.737	0.06	0.071
mAP50	0.816	0.839	0.831	0.802	0.028	0.03
mAP50-95	0.5	0.521	0.539	0.518	0.01	0.012
F1 Score	0.761	0.787	0.767	0.745	0.038	0.042
Actual Instance	547	547	547	547	547	547
Predicted Instance	575	643	585	545	620	514

Table 3: Comparison of different models on Istanbul Dataset (n=100)

4 Annotating Training Data and Classifying by Religious Categories

Building on the dataset curation and sampling procedures introduced and the model selection results demonstrated above, we constructed a rigorously annotated dataset (IS-12261) to train and evaluate our custom detector, **CoexistNet**. A stratified sample of 12,261 frames was drawn equally from Haifa, Jerusalem, Nazareth, and Tel Aviv in the Israel Street-View Reservoir 2025, capturing varied lighting conditions, camera angles, pedestrian poses, occlusions, and scales.

All annotation was performed in Roboflow. Each of the 21,572 pedestrian instances was enclosed in a tight bounding box—drawn exclusively around visible human features from head to toe, excluding background elements and avoiding excessive padding—and labeled along two orthogonal dimensions:

- **Perceived Ethno-Religious Status:** Religious Jew/Muslim, Very Religious Jew/Muslim, or Non-Religious Person.
- **Perceived Gender and Role:** Male, Female, Soldier, or Officer.

To ensure consistency and accuracy, we applied the following labeling rules:

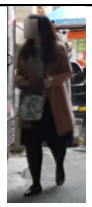
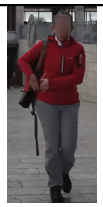
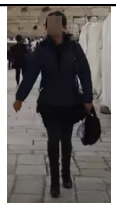
1. **Age Restriction.** Exclude all minors to focus on adults whose attire more reliably signals religious status.
2. **Visibility Threshold.** Annotate only pedestrians with unobstructed, clearly visible features; omit partially occluded individuals or those wearing full-face coverings.
3. **Reliance on Religious Attire.** Base labels solely on observable garments and symbols (e.g., kippot, kufis, hijabs, tzitzit, modest robes); disregard contextual cues such as companions or facial features to minimize bias.

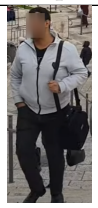
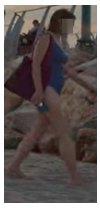
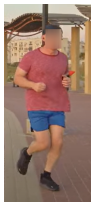
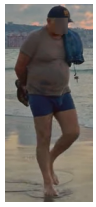
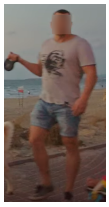
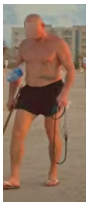
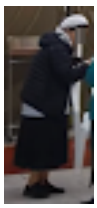
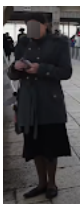
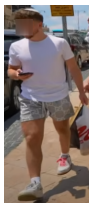
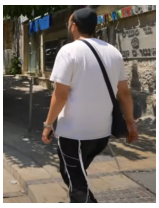
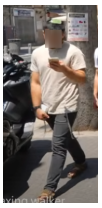
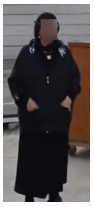
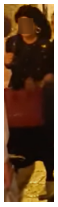
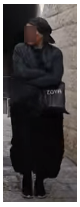
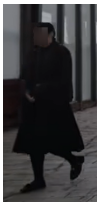
The combination of these attributes yields twelve mutually exclusive classes for model training:

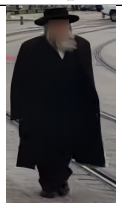
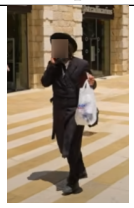
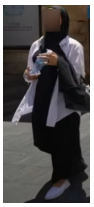
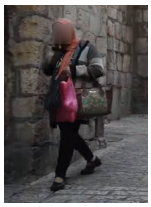
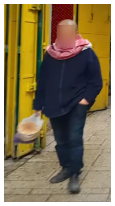
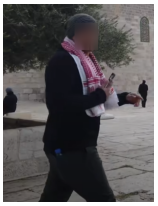
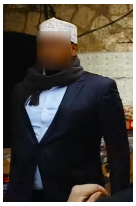
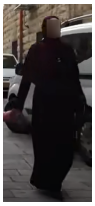
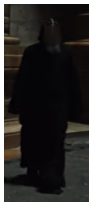
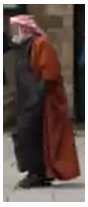
- **Neutral Individual:** Wears everyday, non-religious clothing that is neither overtly modest nor revealing (e.g., men in T-shirts and long pants; women in knee-length skirts or casual dresses without head coverings).
- **Non-Religious Individual:** Displays secular attire with more liberal exposure (e.g., men in shorts or tank tops; women in short sleeves, open-neck tops, or skirts above the knee) and no visible religious symbols.

- **Religious Jewish Male:** Sports a kippah and may wear tzitzit; clothing is modest but not strictly traditional (e.g., collared shirt and slacks).
- **Very Religious Jewish Male:** Adheres to ultra-Orthodox dress norms, including a black suit or long coat (rekel), white shirt, wide-brimmed hat or shtreimel, and visible tzitzit; may have a beard and sidelocks.
- **Religious Jewish Female:** Dresses modestly with a head covering (e.g., tichel or scarf), long skirt (below knee), and sleeves to the elbow; colors may vary.
- **Very Religious Jewish Female:** Wears ultra-Orthodox women's attire such as a full tichel or sheitel, calf-length skirt, high-neck blouse, and opaque stockings; garments are generally dark or muted.
- **Religious Muslim Male:** Wears a kufi or taqiyah and modest clothing (e.g., long-sleeve shirt and pants); may don a simple thobe in lighter fabrics.
- **Very Religious Muslim Male:** Follows conservative Islamic dress, including a long robe (thobe or dishdasha), keffiyeh or ghutra secured with an agal, and often a full beard.
- **Religious Muslim Female:** Covers hair with a hijab and wears loose, modest attire such as an abaya or long tunic with trousers; garments may feature modest patterns or colors.
- **Very Religious Muslim Female:** Observes stricter dress codes, such as wearing a niqab or jilbab, covering all but the eyes, with long, dark outer garments and gloves.
- **Soldier:** Wears Israeli Defense Forces (IDF) combat uniform—camouflage fatigues, helmet or cap, and visible insignia or equipment (e.g., rifle, utility belt).
- **Officer:** Identified by IDF dress uniform or officer's attire with rank insignia on shoulders or collar, polished boots, and formal headgear.

By adhering to these criteria, the IS-12261 dataset attains a high level of annotation consistency, thereby enabling rigorous training and evaluation of the emph{Coexist-Net detector. A visual summary of the annotated categories is presented in Figure 4. The bounding boxes applied to a variety of ethnoreligious categories can be found in Figure 3.

	Example 1	**Example 2**	**Example 3**	**Example 4**	**Example 5**
Neutral Female					

	Example 1	Example 2	Example 3	Example 4	Example 5
Neutral Male					
Non-Religious Female					
Non-Religious Male					
Religious Jewish Female					
Religious Jewish Male					
Very Religious Jewish Female					

	Example 1	Example 2	Example 3	Example 4	Example 5
Very Religious Jewish Male					
Religious Muslim Female					
Religious Muslim Male					
Very Religious Muslim Female					
Very Religious Muslim Male					

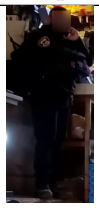
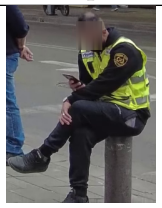
	Example 1	Example 2	Example 3	Example 4	Example 5
Officer Male					

Table 4: Objects Categorized by Religious Classification and Gender



Fig. 3: Roboflow annotation example showing bounding boxes for diverse ethnoreligious and gender/role classes.

4.1 Training Dataset Overview and Analytics

This annotation framework provides a consistent, systematic approach for assessing and classifying visible religiosity. Applying it, we assembled a comprehensive training dataset of 12,071 frames and 21,572 pedestrian instances to fine-tune our custom **Coexist-Net** classifier. We contend that this corpus broadly represents Israel's resident population. One limitation is that it may also capture non-resident individuals. However, although transient tourists could introduce noise, daily visitor counts in Israel have ranged from approximately 2,000 during conflict-affected periods (e.g., January 2024) to around 12,500 at peak levels (e.g., 2019) (49; 22). Even at maximum influx, tourists accounted for less than 0.2% of the national population on any given day, rendering their impact on observed proxemic patterns negligible.

To capture Israel's religious, ethnic, and socioeconomic heterogeneity, our dataset draws equally from four urban centers:

– **Jerusalem:** The nation's capital and pilgrimage center for Jewish, Muslim, and Christian communities, characterized by highly diverse neighborhoods.

– **Nazareth:** Israel's principal Arab-majority city, providing insight into proxemic behavior within predominantly Arab contexts.
– **Haifa:** A northern port city with substantial Jewish and Arab populations, exemplifying everyday coexistence.
– **Tel Aviv:** A secular, densely populated metropolis, where urban congestion and cosmopolitan dynamics prevail.



Fig. 4: Map of Israel highlighting the four sampled cities.

Of the 12,071 frames, 6,029 contained no detectable pedestrians and were marked as null. In the remaining 6,042 frames, most individuals exhibited no visible religious markers and were classified as *neutral*, while a substantial minority displayed identifiable religious attire. Table 5 reports the frequency of each annotated class. This geographically and culturally diverse corpus provides a robust foundation for modeling interpersonal distances and assessing social polarization across Israel's major urban environments.

4.2 CoExist-Net: Model Performance

We evaluate four aggregation schemes for pedestrian detection using two YOLOv9 backbones: the compact *YOLOv9c* and the extended *YOLOv9e*. All models are

Class	**Count**
Neutral Male	6,835
Non-Religious Female	4,484
Neutral Female	3,083
Non-Religious Male	2,654
Religious Jewish Male	2,357
Very Religious Jewish Male	975
Religious Muslim Female	373
Religious Jewish Female	338
Soldier Male	137
Very Religious Muslim Female	130
Very Religious Jewish Female	54
Religious Muslim Male	51
Officer Male	51
Soldier Female	21
Very Religious Muslim Male	16
Officer Female	6

Table 5: Number of instances per annotated class in the training dataset.

trained on images resized to 1024×1024 with a batch size of 4. We report precision, recall, mAP at IoU 0.5 (mAP_{50}), and mAP averaged over IoU thresholds 0.5:0.95 ($\text{mAP}_{50\text{–}95}$). Results on the validation set are given in Table 6.

We compare four supervision regimes that differ only in how the underlying subclasses are grouped. *All Classes* preserves the full fine-grained taxonomy (16 labels). *Merging Gender* collapses female and male variants within each semantic category (e.g., *Religious Jewish Male* and *Religious Jewish Female*), yielding 8 gender-agnostic labels. *Merging Religiousness* merges religious affiliation/ethnicity within each gender (e.g., *Religious Female Jewish* and *Religious Female Muslim*), producing 12 labels—six categories per gender—distinguishing *Neutral*, *Non-Religious*, *Religious*, *Very Religious*, *Soldier*, and *Officer*. *Merging Gender & Religiousness* collapses both gender and affiliation, further merging *Religious* with *Very Religious* into a single *Religious* class and *Neutral* with *Non-Religious* into a single *Non-Religious* class, and drops the remaining categories, yielding a 2-label space. These progressively coarser spaces reduce within-class ambiguity and attenuate long-tail sparsity at the cost of semantic specificity.[3]

Two patterns stand out. First, coarser label spaces deliver clear gains. Collapsing both gender and affiliation and unifying *Religious* with *Very Religious* produces the strongest recall and the highest mAP values within each backbone. For YOLOv9c, mAP_{50} increases from 0.192 to 0.576 and $\text{mAP}_{50\text{–}95}$ from 0.122 to 0.348 relative to the full label set; YOLOv9e follows the same trend, reaching 0.560 and 0.345, respectively. These improvements indicate that many errors un-

[3] In some regressions, not all categories appear. Extremely sparse classes can be effectively undetectable for a given detector/backbone pair, yielding zero or near-zero detections; such categories are dropped from the corresponding specification and comparisons are made on the remaining estimable set.

Table 6: Validation performance for four aggregation schemes. Inputs are 1024 × 1024; batch size is 4. Best result within each backbone is in **bold**.

Aggregation scheme	YOLOv9c				YOLOv9e			
	Precision	Recall	mAP_{50}	$mAP_{50–95}$	Precision	Recall	mAP_{50}	$mAP_{50–95}$
All Classes	0.460	0.232	0.192	0.122	**0.546**	0.245	0.208	0.133
Merging Gender	0.496	0.488	0.455	0.284	0.476	0.528	0.463	0.291
Merging Religiousness	0.375	0.508	0.407	0.253	0.428	0.478	0.416	0.261
Merging Gender & Religiousness	**0.531**	**0.626**	**0.576**	**0.348**	0.516	**0.617**	**0.560**	**0.345**

der the fine-grained regime arise from confusion among closely related religious subclasses rather than missed localizations.

Second, backbone capacity shifts the precision–recall balance. YOLOv9e attains the highest precision in the most demanding setting (0.546 on *All Classes*), consistent with better calibration when separating similar categories. After aggregation, YOLOv9e emphasizes sensitivity, with recall rising from 0.245 to 0.617 and corresponding gains in mAP. YOLOv9c benefits even more from aggregation, ultimately achieving its best scores under the five-label scheme. In practice, the five-label configuration is well suited for downstream tasks that prioritize reliable counts and coverage, whereas the full sixteen-class model with YOLOv9e is preferable when subclass specificity is essential and a modest loss in recall is acceptable.

5 Depth Estimation

5.1 Introduction to Monocular Depth Estimation

Monocular depth estimation predicts per-pixel depth from a single two-dimensional image (30). Unlike stereo or multi-view approaches, monocular methods rely on learned cues—such as texture gradients, shading, and object recognition—to infer relative distances. This capability is critical for our study, as our street-view dataset lacks ground-truth depth maps, multiple viewpoints, or reference objects.

To select an appropriate model, we evaluated performance and runtime on publicly available datasets with accurate depth annotations. First, we used the SUN RGB-D[4] dataset—a large collection of over 10,000 indoor scenes drawn from NYU Depth v2, Berkeley B3DO, and SUN3D—with each image paired with a high-resolution depth map (30). We randomly sampled 50 images from each source (totaling 150) to assemble an indoor testing set. Second, we incorporated the UrbanSyn dataset, which comprises simulated outdoor urban scenes with wide depth ranges and variable lighting conditions. The combination of indoor and outdoor test sets allows us to assess model generalization across both small- and large-scale depth variations.

[4] The SUN RGB-D dataset, widely used in computer vision research for evaluating depth estimation methods, offers a diverse set of scenes with accurate ground-truth depth data, making it well-suited for benchmarking and comparing model performance across varied real-world environments.

We benchmarked four monocular depth estimation architectures:

- **DepthAnything** emphasizes scale-up of training data and heavy augmentation to obtain robustness without novel architectural changes (52). We evaluate both the standard and the outdoor-fine-tuned variants, each in small, base, and large configurations.
- **Global–Local Path Network (GLPN)** integrates global scene features with local texture information via a hybrid convolutional architecture (18).
- **MiDaS 3.0 DPT** extends the original MiDaS models by adopting a transformer-based encoder–decoder design that captures both fine details and global context (21).
- **ZoeDepth** adapts the DPT backbone and is pre-trained on twelve diverse datasets, yielding strong zero-shot generalization (8).

Most models offer multiple size variants—"small," "base," and "large"—to trade off parameter count, representational capacity, and computational cost. Larger versions typically achieve lower root-mean-square error (RMSE), higher peak signal-to-noise ratio (PSNR), and greater structural similarity (SSIM) at the expense of increased inference time and potential overfitting. Given our dataset's size and the importance of runtime efficiency, we include both large and smaller variants in our evaluation.

Model performance was quantified using standard depth-estimation metrics, including RMSE, absolute relative error, PSNR, and SSIM (definitions and formulas in Appendix B). These metrics collectively assess accuracy, scale-invariant error, and perceptual quality, enabling a rigorous comparison of candidate models for outdoor street-view depth estimation.

- **Root Mean Square Error (RMSE):** Defined as the square root of the mean squared difference between predicted and ground-truth depth values, RMSE quantifies overall estimation error. Lower RMSE indicates higher prediction accuracy.
- **Peak Signal-to-Noise Ratio (PSNR):** The ratio between the maximum possible value of the depth map and the power of the distortion (noise) in the prediction. Expressed in decibels, higher PSNR reflects closer agreement with the ground-truth depth.
- **Structural Similarity Index (SSIM):** A perceptual metric evaluating structural, luminance, and contrast fidelity between predicted and ground-truth depth maps. SSIM values range from –1 to 1, with values nearer to 1 indicating stronger preservation of structural information.
- **Threshold Accuracy (δ):** The fraction of predicted depths d satisfying $\max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < \delta$, where d^* is the ground-truth depth. We report accuracy at $\delta = 1.25$, 1.25^2, and 1.25^3 to assess prediction robustness across varying depth scales.

5.2 Analysis of Depth Estimation Results

On the SUN RGB-D dataset, the MiDaS 3.0 base and large variants, as well as all three DepthAnything configurations, demonstrated superior performance.

Large models achieved the lowest RMSE and the highest PSNR and SSIM, indicating their capacity to capture fine structural details and minimize overall error. In contrast, threshold accuracy metrics (= 1.25, 1.25^2, 1.25^3) revealed that smaller or base models often maintain more consistent relative accuracy across diverse depth ranges. Both GLPN and ZoeDepth underperformed on SUN RGB-D, suggesting that their architectures may be less effective at resolving small-scale depth variations in confined indoor environments.

To evaluate generalization to outdoor contexts, we conducted parallel experiments on UrbanSyn, which presents large-scale depth variations typical of urban scenes. By benchmarking models on both SUN RGB-D and UrbanSyn, we identify those that handle the full spectrum of depth scales—ranging from close-quarters indoor settings to expansive outdoor vistas—an essential requirement given the heterogeneous depth characteristics of our Israel street-view frames.

The results for the SUN RGB-D and UrbanSyn datasets dataset are listed in 7 and 8 respectively.

Model	RMSE	PSNR	SSIM	Accuracy @ 1.25	Accuracy @ 1.25^2	Accuracy @ 1.25^3
MiDAS 3.0	0.210	14.432	0.823	0.421	0.653	0.791
MiDAS 3.0 Large	0.208	14.497	0.823	0.405	0.654	0.800
GLPN	0.341	10.032	0.641	0.083	0.188	0.343
ZoeDepth	0.217	13.854	0.793	0.286	0.544	0.765
DepthAnything Small	0.213	14.526	0.830	0.424	0.648	0.799
DepthAnything Base	0.208	14.694	0.834	0.440	0.655	0.803
DepthAnything Large	0.210	14.677	0.834	0.433	0.650	0.799
DepthAnything Outdoor Large	0.188	15.183	0.831	0.351	0.610	0.798
DepthAnything Outdoor Base	0.176	15.848	0.847	0.449	0.696	0.853
DepthAnything Outdoor Small	0.203	14.571	0.824	0.324	0.629	0.820

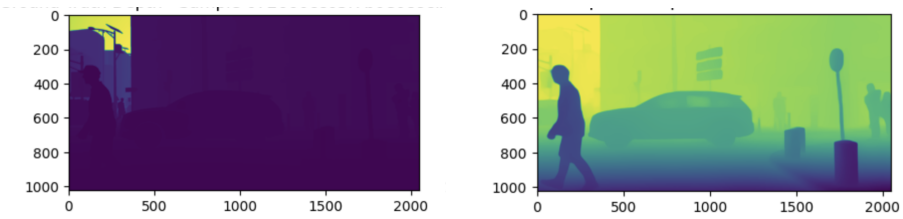
Table 7: Normalized comparison of depth estimation models on SUN RGB-D dataset (n=150)

Model	RMSE	PSNR	SSIM	Accuracy @ 1.25	Accuracy @ 1.25^2	Accuracy @ 1.25^3
MiDAS 3.0	0.600	4.547	0.162	0.035	0.049	0.066
MiDAS 3.0 Large	0.611	4.416	0.159	0.037	0.050	0.066
GLPN	0.222	13.254	0.270	0.033	0.072	0.121
ZoeDepth	0.279	11.320	0.270	0.033	0.065	0.086
DepthAnything Small	0.641	3.962	0.153	0.035	0.045	0.060
DepthAnything Base	0.642	3.956	0.152	0.035	0.044	0.059
DepthAnything Large	0.640	3.974	0.153	0.035	0.059	0.099
DepthAnything Outdoor Small	0.296	10.813	0.257	0.019	0.056	0.079
DepthAnything Outdoor Base	0.300	10.726	0.254	0.0249	0.061	0.083
DepthAnything Outdoor Large	0.295	11.002	0.261	0.041	0.072	0.089

Table 8: Normalized comparison of depth estimation models on UrbanSyn dataset (n=150)

The results, normalized to the [0,1] range, showed that models originally trained on outdoor data (GLPN, ZoeDepth) outperformed those optimized for indoor scenes. Nonetheless, even these outdoor-specialized architectures exhibited lower-than-anticipated accuracy on UrbanSyn. This underperformance likely arises from UrbanSyn's unique depth characteristics—particularly the vast range introduced by distant elements such as buildings and sky—which can disproportionately influence the normalization process and skew depth predictions. These findings underscore the inherent difficulty of monocular depth estimation in outdoor settings, where models must reconcile both near-field and far-field depth cues under variable lighting and scene complexity.

To diagnose the reduced performance on UrbanSyn, we examined the distributions of ground-truth and predicted depths. Distant scene elements—such as building facades and the sky—exhibit exceptionally large depth values; when these values are min–max normalized, they compress the representational range available for foreground regions. As a consequence, the model allocates insufficient dynamic range to near-field depth variation and struggles to resolve subtle differences at close distances. This compression phenomenon highlights the necessity of normalization schemes or architectural refinements that preserve depth resolution across the entire scale spectrum characteristic of outdoor imagery. Figure 5 illustrates the substantial qualitative improvement achieved with the proposed adjustment.



(a) Ground-Truth Depth (b) Predicted Depth Map

Fig. 5: Comparison of UrbanSyn ground-truth and predicted depth maps.

To mitigate these challenges, we applied a thresholding strategy to the UrbanSyn ground-truth depths. Depth values exceeding 0.004 were capped at 1.0, thereby isolating extreme outliers from the remainder of the distribution. The residual values were then min–max rescaled to the [0,1] interval, enhancing relative depth resolution for proximate objects. Predicted depths were first normalized to [0,1] and subsequently subjected to a logarithmic transformation to correct residual skewness. Specifically, predictions above 0.98 were set to 1.0, and remaining values were transformed according to $\log(\text{depth} + 1)$. This log transformation attenuates the influence of large outliers and increases sensitivity to subtle depth variations at shorter ranges.

Table 9 summarizes model performance following the logarithmic transformation. By attenuating the influence of distant, high-depth regions, this normalization enhanced the capacity of outdoor-trained models to capture relative depth gradients from foreground to background. Consequently, predictive accu-

Model	RMSE	PSNR	SSIM	Accuracy @ 1.25	Accuracy @ 1.25ˆ2	Accuracy @ 1.25ˆ3
MiDAS 3.0 Trans	0.397	8.225	0.291	0.052	0.089	0.111
MiDAS Large Trans	0.403	8.122	0.248	0.055	0.091	0.111
GLPN Trans	0.210	13.739	0.445	0.096	0.187	0.279
ZoeDepth Trans	0.210	13.783	0.458	0.056	0.112	0.165
DepthAnything Small Trans	0.425	7.623	0.275	0.058	0.082	0.099
DepthAnything Base Trans	0.421	7.690	0.280	0.060	0.085	0.104
DepthAnything Large Trans	0.420	7.704	0.253	0.060	0.082	0.102
DA Large Outdoor Trans	0.208	14.063	0.420	0.057	0.106	0.154
DA Base Outdoor Trans	0.208	14.063	0.420	0.057	0.106	0.154
DA Small Outdoor Trans	0.210	13.883	0.450	0.047	0.096	0.151

Table 9: Comparison of log transformation of depth estimation models on UrbanSyn dataset (n=150)

racy improved across key metrics, yielding more reliable estimates of interindividual distances in our scene analysis.

For the Israel street-view corpus, our principal requirement is accurate *depth ordering*—that is, a reliable ranking from near to far—rather than precise metric depth. Any model that preserves this ordinal structure satisfies our analytic criteria. Although GLPN, ZoeDepth, and all DepthAnything-Outdoor variants achieve acceptable ordering, GLPN and ZoeDepth underperform on the SUN RGB-D benchmark and require roughly three times the inference time of the DepthAnything-Outdoor *Base* model. We therefore adopt the DepthAnything-Outdoor Base configuration as the preferred estimator, as it offers the most favorable trade-off between computational efficiency and accuracy across both indoor and outdoor test sets. Figure 6 visualizes representative depth estimations for four Israeli cities generated with the selected model.

6 Distance Estimation

Quantifying interpersonal spacing is central to computational proxemics, yet estimating physical distance from a single two–dimensional (2-D) frame is inherently ambiguous: a monocular image projects a three-dimensional (3-D) scene onto a 2-D plane, discarding explicit depth cues. Our dataset provides neither multi-view imagery nor fixed reference objects; consequently, precise metric depth cannot be recovered directly. Instead, we infer *relative* depth using the DepthAnything Outdoor Base model (Section 5.1) and convert these estimates into pairwise Euclidean distances—a strategy widely adopted in recent social-distancing and interaction studies (e.g., (1; 37; 7)).



Fig. 6: Sample images (middle column) vs. predicted DepthAnything outdoor base model outputs (right column) for four cities.

6.1 Euclidean Distance in 2-D and 3-D Space

Let $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ denote two points in the image plane. Their 2-D Euclidean distance is

$$d_{2D} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{1}$$

Equation (1) ignores depth and thus cannot capture true spatial separation when objects are located at different distances from the camera.

After monocular depth estimation, each pixel acquires a scale–invariant depth value z. Extending Equation (1) to 3-D yields

$$d_{3D} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}, \tag{2}$$

where $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ represent corresponding points in camera-centered 3-D space. Relative depth suffices for Equation (2) because the unknown global scale cancels when comparing distances across a single frame (32; 46).

6.2 Pipeline for Interpersonal Distance

Our processing pipeline proceeds as follows:

1. **Person detection.** Each frame is processed with Coexist-Net to obtain bounding boxes for all visible individuals.
2. **Depth estimation.** The same frame is fed to the DepthAnything Outdoor Base network, producing a dense depth map $D(x, y)$.
3. **Coordinate extraction.** For every bounding box, we record the center pixel (x_i, y_i) and its normalized depth $z_i = D(x_i, y_i)$, yielding a set of 3-D points $\{P_i(x_i, y_i, z_i)\}_{i=1}^{N}$.
4. **Distance matrix.** Pairwise distances are computed with Equation (2), forming an $N \times N$ symmetric distance matrix suitable for proxemic analysis.

This procedure is computationally lightweight, requiring only a single forward pass through the depth network and simple arithmetic for distance computation—advantages noted by earlier monocular social-distancing systems (1; 37).

6.3 Rationale for Using Relative Distances

Absolute metric reconstruction from a lone image demands either camera calibration or reference objects, neither of which is available in our heterogeneous street-view corpus. Relative distances, by contrast, (i) align with the theoretical focus of proxemics on *comparative* spacing between individuals (23; 15) and (ii) avoid the scale-ambiguity problem endemic to monocular depth (32; 46). Empirical studies confirm that relative depth is both more robust across scenes and sufficient for detecting social behaviors such as avoidance, grouping, or interaction (7; 27). Accordingly, we treat the outputs of Equation (2) as *scale-invariant* measures of interpersonal proximity, enabling comparative analysis across all frames regardless of camera configuration.

7 Empirical Approach

Our analysis asks whether the physical spacing between pedestrians— measured as a scale–free, frame-normalized distance—systematically varies with their visible gender and religiosity. For every video frame containing at least three classifiable pedestrians, we calculate all Euclidean pairwise distances δ_{ij} using Eq. 2. Each set of within-frame distances is then min–max–normalized,

$$\tilde{\delta}_{ij} = \frac{\delta_{ij} - \min_j \delta_{ij}}{\max_j \delta_{ij} - \min_j \delta_{ij}} \in [0, 1],$$

so that the closest pair receives a value of 0 and the farthest a value of 1, thereby eliminating camera-scale heterogeneity (37). We label every $\tilde{\delta}_{ij}$ with a single unordered indicator $D_{ij}^{(k)}$ that records the gender (Male or Female) and religiosity (Religious or Non-Religious) of the two pedestrians involved.

Equation (3) reports the core specification.

$$\tilde{\delta}_{ij} = \alpha + \boldsymbol{\beta}^{\top} \mathbf{D}_{ij}^{(-b)} + \gamma_{\text{district}[i]} + \gamma_{\text{night}[i]} + \gamma_{\text{month}[i]} + \varepsilon_{ij}, \tag{3}$$

where $\mathbf{D}_{ij}^{(-b)}$ contains all category–pair dummies except a scheme-specific baseline b. The fixed effects γ_{district}, γ_{night} (equal to 1 for night-time scenes), and γ_{month} (twelve monthly indicators) absorb spatial and temporal heterogeneity. Robust standard errors are clustered by video to account for within-clip dependence.

To demonstrate that our conclusions do not depend on how finely the classes are defined, we estimate Eq. (3) under four alternative taxonomies that range from the most granular to the most parsimonious. In the "All Classes" model we keep the full gender-by-religiosity cross and use the reciprocal pair *religious male Jew versus religious female Arab* (and its inverse) as the reference category. When we collapse Jewish and Arab identities into a single religious category, the baseline becomes *religious female versus religious male*. When we instead merge gender, the benchmark is *religious Arab versus religious Jew*. Finally, when we merge both gender and religiosity into a single "religious" versus "non-religious" split, that very contrast forms the baseline.

Each aggregation scheme is implemented with two independent object detectors, namely *YOLOv9c*, which has a compact backbone, and *YOLOv9e*, which features an extended backbone. In the context of convolutional neural networks for object detection, the *backbone* refers to the core feature-extraction network—the series of convolutional and pooling layers that transform raw image pixels into rich, hierarchical representations of edges, textures, and higher-level visual patterns. A compact backbone typically comprises fewer layers or narrower channel widths, trading off some representational capacity for faster inference and lower memory requirements, while an extended backbone increases network depth or channel dimensionality to capture more detailed and diverse features at the cost of greater computational and storage demands. Consequently, we estimate eight regression models in total—four aggregation schemes each paired with two detectors. For each of these eight models, we also conduct two-sample t-tests that compare the mean normalized distance in every aggregation–detector combination against its baseline counterpart, thereby assessing whether differences in spatial accuracy are statistically significant.

Across all specifications—whether estimated in levels, with YOLOv9c or YOLOv9e, or by t-test—the sign, magnitude, and statistical significance of the key coefficients remain stable. This concordance shows that our substantive conclusions about interpersonal spacing are not an artifact of baseline choice, class aggregation, or detector architecture (the full numerical results are reported in Section 8).

8 Results

Throughout all model specifications, the reference dyad is a visibly very religious Jewish man passing a visibly very religious Muslim (Arab) woman, or the converse.[5] Intercepts therefore capture the largest empirically observed inter-pedestrian gap, ranging from roughly 0.40 to 0.50 on the normalized distance scale, with p-values below 0.01. A positive coefficient indicates that the focal dyad is farther apart than this benchmark, whereas a negative coefficient indicates closer proximity.

Across eight model families—YOLOv9c and YOLOv9e, each estimated under four aggregation schemes—two identity dimensions consistently structure spacing: ethnonational affiliation and gender. All regressions yield three stable findings:

1. Co-ethnic and co-religious pairs stand closer than cross-group pairs.
2. Encounters that mix a visibly religious or very religious pedestrian with a neutral one exhibit pronounced avoidance.
3. Dyads composed entirely of men maintain the largest interpersonal buffers.

These regularities persist across all fixed-effect structures, covariate sets, and detection architectures.

8.1 Ethnonational Identity and In-Group Proximity (H1)

Co-ethnic dyads stand closer. With gender merged and religion disaggregated (Table 13 and Table 10), two religious Arabs (RA–RA) stand approximately 0.22–0.32 normalized units closer than the reference dyad ($p < 0.001$), whereas religious Jews (RJ–RJ) reduce spacing by about 0.07–0.10 units ($p < 0.01$). Fully disaggregated models (Tables 1–2) confirm this pattern: the pairing `RFA vs RFA` is 0.21–0.23 units closer than the benchmark, while `RFJ vs RFJ` is roughly 0.04–0.05 units closer.

There is asymmetry across communities. Religious Arabs exhibit the most pronounced in-group proximity (`RA vs RA`, `VRFA vs VRFA`); corresponding Jewish coefficients are smaller. Among secular pedestrians, ethnonational homophily largely vanishes: neutral Jew–Jew and neutral Arab–Arab dyads do not differ significantly from the cross-ethnic benchmark.

Pairings that mix a neutral pedestrian with a visibly *very religious* partner increase spacing by approximately 0.45–0.60 units (`NF vs VRFA`, `NM vs VRFA`; Table 11 and Table 12. Even the most aggregated specification—which collapses gender and religiosity (Tables 7–8)—shows that any religious cue paired with a non-religious partner adds 0.21–0.27 units ($p < 0.001$).

[5] When the *very-religious* categories are unavailable, the corresponding *religious* categories are substituted.

8.2 Gender Dynamics (H2)

Male-male dyads maintain the greatest distance. Male-only pairs (`NM vs NM`) add roughly 0.23 units in YOLOv9c and 0.31 units in YOLOv9e ($p < 0.001$). Female-only pairs (`NF vs NF`) show modest, often insignificant, increases of 0.03–0.11 units. Mixed-gender encounters fall between these extremes.

There is limited evidence for female exceptionalism. Once ethnonational and religious cues are controlled, women do not consistently stand closer to one another; observed gender effects are driven primarily by male-male avoidance rather than female-female attraction. Below you can find a more structured evaluation of the hypotheses propesed:

- **H1 (Ethnonational homophily):** Supported. Co-ethnic dyads (RA–RA, RJ–RJ) approach up to 0.32 units closer than the cross-ethnic benchmark.
- **H1a/H1b (Asymmetry):** Partially supported. Religious and very religious Arabs show the strongest in-group proximity; secular subgroups in both communities show minimal bias.
- **H2a (Same-gender proximity):** Not supported. Male-male pairs increase distance; female-female effects are weak.
- **H2b (Stronger effect for women):** Not supported. Gender differences arise from male avoidance rather than female closeness.

The specifications that merge gender provide the central summary of the paper's quantitative findings. As demonstrated in Table 10 and Figure 7, homophily effects are most pronounced along dimensions of ethnoreligious identity.

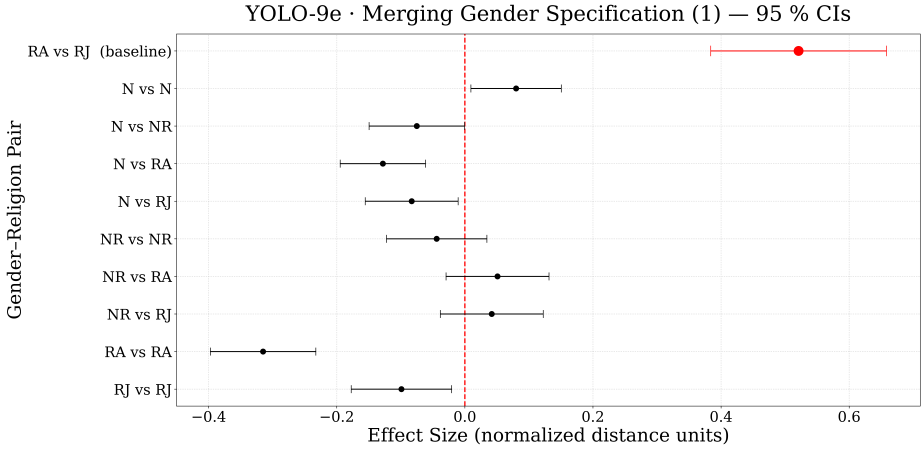


Fig. 7: Merging Gender, Normalized Distances

8.3 Robustness Checks

Effect directions, magnitudes, and significance levels remain stable under both detection models and after collapsing the full sixteen-category taxonomy to two categories. Adding location, video, month, or night controls alters coefficients by less than ± 0.02 units and never reverses a sign, indicating that the results reflect behavior rather than modeling artifacts.

Table 10: YOLO 9e - Merging Gender

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.521***	0.521***	0.531***	0.531***	0.512***	0.511***
N vs N	0.080**	0.081**	0.083**	0.083**	0.088**	0.088**
N vs NR	-0.075*	-0.075*	-0.073*	-0.073*	-0.069*	-0.069*
N vs RA	-0.128***	-0.128***	-0.127***	-0.127***	-0.128***	-0.128***
N vs RJ	-0.083**	-0.082**	-0.082**	-0.082**	-0.079**	-0.079**
NR vs NR	-0.044	-0.044	-0.042	-0.042	-0.039	-0.039
NR vs RA	0.051	0.051	0.052	0.052	0.053	0.053
NR vs RJ	0.042	0.042	0.044	0.044	0.046	0.046
RA vs RA	-0.315***	-0.314***	-0.314***	-0.314***	-0.323***	-0.323***
RJ vs RJ	-0.099**	-0.099**	-0.098**	-0.098**	-0.099**	-0.099**
Controls						
month		0.000		0.000	0.001	0.001
night		-0.003		-0.002	-0.003	-0.003
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.025	0.025	0.024	0.024	0.005	0.005
Observations	35515	35515	35515	35515	35515	35515

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)
Acronym Legend:
N: Neutral, NR: Non-Religious, RA: Religious Arab, RJ: Religious Jew.

8.4 Spatial Behavior and Political Polarization

Pedestrian spacing reproduces Israel's broader cleavages. Ultra-Orthodox Jewish men and visibly religious and very religious Muslim Arab women seldom occupy close physical space, whereas co-ethnic and co-religious pairs—especially religious and very religious Arabs—walk noticeably nearer to one another. Nonetheless, polarization is heterogeneous. In mixed cities such as Haifa and Tel Aviv, secular Jews and Arabs often pass with only modest additional distance. Female pedestrians likewise tend to approach one another more readily than men, even across ethnonational lines. These instances of spatial convergence demonstrate that polarization, though salient, is contingent upon visible cues and local context.

Overall, the findings provide a fine-grained behavioral perspective on Israel's political geography: visible identity cues—particularly religious symbols—predict systematic variation in interpersonal spacing on urban sidewalks.

9 Discussion

Using more than three million street-level images from every Israeli municipality, our computer-vision pipeline documents a pervasive pattern of ethnonational homophily in pedestrian spacing. Across Jerusalem, Haifa, Tel Aviv–Yafo, and small towns from Kiryat Shmona to Eilat, Jewish–Jewish and Arab–Arab dyads stand materially closer together than mixed Jewish–Arab pairs. Regression estimates show that co-ethnic Arab pairs narrow the mean gap by roughly 0.22–0.32 normalized units, and co-ethnic Jewish pairs by 0.07–0.10, relative to a mixed, visibly very-religious benchmark.

The effect is not confined to large cities. It appears in ultra-Orthodox Bnei Brak, Arab Umm al-Fahm, and secular Herzliya alike, echoing classic work on residential and educational segmentation (23). Arab-Arab dyads are typically the closest; Middle-Eastern proxemics research reports a cultural comfort with near physical space, and surveys find Arab Israelis more willing than Jews to engage in close contact (4). Our visual data accord with both strands.

Gender differences are secondary. Male–male dyads keep slightly larger buffers (0.23–0.31 units) than female–female or mixed-gender pairs, replicating earlier findings that men prefer larger personal zones. Visible religious markers—kippot, keffiyehs, niqabs, shtreimels—serve as boundary cues: in mixed dyads, adding such a marker widens spacing by about 0.45–0.60 units, consistent with work showing that outward religiosity accentuates perceived group boundaries (31).

Distancing softens but does not disappear in mixed municipalities. Haifa and Tel Aviv–Yafo display the narrowest Jewish–Arab gaps, whereas mono-ethnic towns show the widest. The pattern aligns with evidence that civic engagement and protest participation are higher in mixed settings (4).

Sidewalk geometry mirrors Israel's partisan "tribes." During the 2023–24 judicial-overhaul crisis, surveys recorded record distrust across ethnic and ideological lines (9). Our proxemic measurements reveal that this distrust materializes as additional distance. The same crisis—framed as a clash between majoritarianism and liberal democracy (40)—thus leaves literal footprints in daily spatial behavior.

The nationwide dataset offers a baseline for monitoring change. Future studies can test whether joint civic projects (for example, a Jewish–Arab youth center in Lod) narrow local gaps or whether new conflicts widen them. Combining distance with gaze, mutual orientation and conversational clustering could extend proxemic analysis from co-presence to interaction and enable comparisons with other divided societies.

Inter-group distance can serve as an evidence-based behavioral indicator. Peace-building NGOs or municipal integration offices could track whether interventions reduce the gap. However, others can take advantage of the same system to support surveillance, do ethnic profiling and social scoring (12). Responsible deployment demands privacy-by-design, continuous impact assessment and strict secondary-use limits (14). Transparent model cards, publication of limitations and participatory ethics boards should accompany any future technical advance

to ensure that behavioral indicators illuminate social division without becoming tools of control.

Acknowledgments

The research team reflects a broad range of perspectives: C. Caliskan (Turkey), Y. Li (People's Republic of China), and T. Loftus (United States). We believe that our cultural and disciplinary diversity strengthened every stage of the project—from framing the research question to interpreting the results—by ensuring that no single national viewpoint dominated the analysis.

Bibliography

[1] Aghaei, M., Poggi, M., Tosi, F., Mattoccia, S.: Single image human proxemics estimation for visual social distancing. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 608–618 (2021)

[2] Agnus, O.M.: Proxemics: The study of space. IRWLE **8**(1), 1–7 (2012)

[3] Ayer, L., Venkatesh, B., Stewart, R., Mandel, D., Stein, B., Schoenbaum, M.: Psychological aspects of the israeli–palestinian conflict: A systematic review. Trauma, Violence, & Abuse **18**(3), 322–338 (2017)

[4] Bassan-Nygate, L., Weiss, C.M.: Party competition and cooperation shape affective polarization: Evidence from natural and survey experiments in israel. Comparative Political Studies **55**(2), 287–318 (2022). https://doi.org/10.1177/00104140211019764

[5] Bassan-Nygate, L., Weiss, C.M.: Party competition and cooperation shape affective polarization: evidence from natural and survey experiments in israel. Comparative Political Studies **55**(2), 287–318 (2022)

[6] Bermeo, N.: On democratic backsliding. Journal of Democracy **27**(1), 5–19 (2016)

[7] Bertoni, L., Kreiss, S., Alahi, A.: Perceiving humans: From monocular 3d localization to social distancing. IEEE Transactions on Intelligent Transportation Systems pp. 1–13 (2021). https://doi.org/10.1109/TITS.2021.3087495

[8] Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)

[9] Bronstein, P.: Polarization and the hidden israeli center. Sapir: A Journal of Jewish Conversations **16** (March 2025)

[10] Bucy, E.P.: Politics through machine eyes: What computer vision allows us to see. Journal of Visual Political Communication **10**(1), 59–68 (2023)

[11] Caliskan, C., Ke, L., Li, Y., Shimazaki, Y., Zhong, M.: Veiled metrics: Measuring political polarization through visible interactions between religious and non-religious citizens. Available at SSRN 4907046 (2024)

[12] Citron, D.K., Pasquale, F.A.: The scored society: Due process for automated predictions. Washington Law Review **89**, 1–33 (2014)

[13] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)

[14] Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)

[15] Cristani, M., Bazzani, L., Farenzena, M., Murino, V.: Towards computational proxemics: Inferring social relations from inter-personal distances. In: IEEE International Conference on Social Computing. pp. 290–297 (2011)

[16] Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V.: Towards computational proxemics: Inferring social relations from interpersonal distances. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. pp. 290–297. IEEE (2011)
[17] Deloitte: Research into the financial cost of the northern ireland sectarian divide (2007)
[18] Deshmukh, I., Tripathi, V., Gangodkar, D.: Enhanced 3d model generation from a single image using glpn depth estimation. In: 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC). pp. 984–989. IEEE (2023)
[19] Dietrich, B.J.: Using motion detection to measure social polarization in the us house of representatives. Political Analysis **29**(2), 250–259 (2021)
[20] Dietrich, B.J., Sands, M.L.: Seeing racial avoidance on new york city streets. Nature human behaviour **7**(8), 1275–1281 (2023)
[21] Ehret, T.: Monocular depth estimation: a review of the 2022 state of the art. Image Processing On Line **13**, 38–56 (2023)
[22] Halbfinger, D.M.: Tourism to israel dropped nearly 80% in january, compared to same time last year. Jewish Insider (February 2024), `https://jewishinsider.com/2024/02/israel-tourism-gaza-war-hotels/`, accessed: 2025-06-14
[23] Hall, E.T.: The Hidden Dimension. Doubleday, New York (1966)
[24] Hall, P.A., Soskice, D. (eds.): Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford (2001)
[25] Hamanaka, S.: Polarization, policy preferences, and democratic decay in israel?
[26] Huntington, S.P.: The clash of civilizations? Foreign Affairs **72**(3), 22–49 (1993)
[27] Jiménez-Velasco, A., Gutiérrez, I., Escalera, S.: Proxemics-net++: Recognizing human interactions in images through touch and proxemic cues. Pattern Analysis and Applications pp. 1–18 (2024). https://doi.org/10.1007/s10044-024-01234-5
[28] Joo, J., Bucy, E.P., Seidel, C.: Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning (2019)
[29] Joseph, J.: Monitor social distancing using python, yolov5, opencv (January 2022)
[30] Khan, F., Salahuddin, S., Javidnia, H.: Deep learning-based monocular depth estimation methods—a state-of-the-art review. Sensors **20**(8), 2272 (2020)
[31] Krogstad, A.: A political history of visual display. The Poster **4**(1–2), 7–29 (2017)
[32] Lee, S., Cho, S., Lin, S., Lee, K.M.: Depth map from camera lens autofocus. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 222–231 (2019)

[33] Levitsky, S., Ziblatt, D.: How Democracies Die. Crown, New York (2018)
[34] Malhotra, P., Garg, E.: Object detection techniques: a comparison. In: 2020 7th international conference on smart structures and systems (ICSSS). pp. 1–4. IEEE (2020)
[35] McCall, C.: Mapping social interactions: the science of proxemics. Social Behavior from Rodents to Humans: Neural Foundations and Clinical Implications pp. 295–308 (2017)
[36] McCall, C., Blascovich, J., Young, A., Persky, S.: Proxemic behaviors as predictors of aggression towards black (but not white) males in an immersive virtual environment. Social Influence **4**(2), 138–154 (2009)
[37] Mingozzi, A., Srinivas, C., Neri, M.: Monitoring social distancing with single-image depth estimation. arXiv preprint arXiv:2203.07031 (2022)
[38] Mock, S., Obeidi, A., Zeleznikow, J.: A brief outline of the israel–palestinian conflict. Group Decision and Negotiation **23**, 1245–1262 (2014)
[39] Moïsi, D.: Israel: Four elections later, deadlocked yet again. Institut Montaigne – Expressions (April 2021), `https://www.institutmontaigne.org/en/expressions/israel-four-elections-later-deadlocked-yet-again`
[40] Mussel, J.D., Shugart, M.S.: Majoritarian politics in a consensus democracy: How judicialization of politics helped lay bare the contradictions of israel's constitution. Frontiers in Political Science **7** (2025). https://doi.org/10.3389/fpos.2025.1554064
[41] Nwodo, J.: Israel's judicial reforms: A potential danger to democracy? Democratic Erosion Blog (April 2025)
[42] Peng, Y.: Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. Journal of Communication **68**(5), 920–941 (2018)
[43] Porat, I.: Political polarisation and the constitutional crisis in israel. Israel Law Review **56**(3), 369–384 (2023)
[44] Pullan, W., Baillie, H.: Locating urban conflict: Ethnicity, violence and the right to the city in jerusalem and belfast. Urban Studies **50**(14), 2823–2840 (2013)
[45] Ramzy, S.: Have netanyahu's judicial reforms opened an existential can of worms? Al Habtoor Research Centre – Early Warning Program (March 2023)
[46] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), miDaS project paper
[47] Robinson, K.: What to know about the arab citizens of israel. Council on Foreign Relations Backgrounder (October 2023), available at `https://www.cfr.org/backgrounder/what-know-about-arab-citizens-israel`
[48] Sanchez, S., Romero, H., Morales, A.: A review: Comparison of performance metrics of pretrained models for object detection using the tensorflow framework. In: IOP conference series: materials science and engineering. vol. 844, p. 012024. IOP Publishing (2020)

[49] Staff, T.J.P.: Tourism to israel hits all-time high in 2019. The Jerusalem Post (2019), https://www.jpost.com/israel-news/tourism-to-israel-hits-all-time-high-in-2019-611975, accessed: 2025-06-14
[50] Varshney, A.: Ethnic Conflict and Civic Life: Hindus and Muslims in India. Yale University Press, New Haven (2002)
[51] Wang, Y., Feng, Y., Luo, J.: Gender politics in the 2016 us presidential election: a computer vision approach. In: Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10. pp. 35–45. Springer (2017)
[52] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
[53] Yiftachel, O.: Democracy or ethnocracy? territory and settler politics in israel/palestine. Middle East Report (207), 8–13 (1998)

Appendix

A Object Detection Metrics Formulas

This appendix provides definitions and formulas for the evaluation metrics used in this study to compare pre-trained object detection models to ground-truth annotations.

Precision

Definition: Precision measures the proportion of true positive predictions among all positive predictions, indicating the model's ability to minimize false positives.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall

Definition: Recall evaluates the proportion of true positives that were correctly identified among all actual positive instances, reflecting the model's effectiveness in detecting all relevant objects.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Intersection over Union (IoU)

Definition: IoU measures the overlap between the predicted bounding box (B_p) and the ground-truth bounding box (B_t).

$$\text{IoU} = \frac{\text{Area of Overlap } (B_p \cap B_t)}{\text{Area of Union } (B_p \cup B_t)}$$

Mean Average Precision at 50% IoU (mAP50)

Definition: mAP50 calculates the mean Average Precision (AP) across multiple detection thresholds (confidence levels), with a fixed IoU threshold of 50%.

- **Average Precision (AP)** is computed as the area under the precision-recall curve.
- **mAP50** is the mean of AP values over all object classes in the dataset.

Mean Average Precision at IoU from 50% to 95% (mAP50-95)

Definition: mAP50-95 extends the mAP50 metric by averaging AP across IoU thresholds from 0.50 to 0.95 (in 0.05 increments). This metric highlights the model's ability to accurately localize objects.

$$\text{mAP}_{50:95} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_{\text{IoU}=0.50:0.05:0.95}$$

Where N is the total number of IoU thresholds considered.

F1 Score

Definition: The F1 score combines precision and recall into a single metric, serving as the harmonic mean. It balances the trade-off between precision and recall, particularly useful for imbalanced datasets.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Number of Predicted People vs. Actual People

Definition: This comparison directly counts the number of people predicted by the model versus the number of people in the ground truth. It helps identify tendencies in predictions, such as over-prediction or under-prediction.

- **Over-prediction**: Predicted Count > Actual Count
- **Under-prediction**: Predicted Count < Actual Count

B Depth Estimation Metrics Formulas

Root Mean Square Error (RMSE)

Definition: Root Mean Square Error (RMSE) measures the square root of the average squared differences between estimated depth values and ground truth, providing an indication of overall prediction accuracy. Lower RMSE values signify better performance.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Where: - y_i is the ground truth depth value for the i-th sample, - $\hat{y}_i$ is the predicted depth value for the i-th sample, - N is the total number of samples.

Peak Signal-to-Noise Ratio (PSNR)

Definition: Peak Signal-to-Noise Ratio (PSNR) evaluates the ratio between the maximum possible power of a signal (in this case, the depth map) and the power of corrupting noise. A higher PSNR indicates a more accurate reconstruction of the depth information.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{R^2}{\text{MSE}}\right)$$

Where: - R is the maximum possible pixel value of the image (e.g., 255 for an 8-bit image), - MSE is the Mean Squared Error, defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Structural Similarity Index (SSIM)

Definition: Structural Similarity Index (SSIM) assesses the similarity between two images by considering changes in structural information, luminance, and contrast. SSIM values range from -1 to 1, where values closer to 1 indicate high similarity, reflecting how well the predicted depth map preserves structural information compared to the ground truth.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Where: - μ_x and μ_y are the average intensities of the two images x and y, - σ_x^2 and σ_y^2 are the variances of the images, - σ_{xy} is the covariance between the images, - c_1 and c_2 are small constants to avoid instability when the denominator is close to zero.

Accuracy @ 1.25, Accuracy @ 1.25^2, and Accuracy @ 1.25^3

Definition: These metrics quantify the proportion of estimated depth values within a specific threshold of the ground truth. Specifically, they measure the percentage of predictions that are within a factor of 1.25, 1.25^2, and 1.25^3 of the actual depth values, respectively. These metrics help evaluate the robustness of depth predictions across varying scales.

$$\text{Accuracy @ } 1.25 = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{y_i}{1.25} \leq \hat{y}_i \leq 1.25 y_i\right)$$

$$\text{Accuracy @ } 1.25^2 = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{y_i}{1.25^2} \leq \hat{y}_i \leq 1.25^2 y_i\right)$$

$$\text{Accuracy @ } 1.25^3 = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{y_i}{1.25^3} \leq \hat{y}_i \leq 1.25^3 y_i\right)$$

Where: - y_i is the ground truth depth for the i-th sample, - $\hat{y}_i$ is the predicted depth for the i-th sample, - I is an indicator function that equals 1 if the condition is true and 0 otherwise.

These formulas provide a detailed overview of how each metric is calculated for evaluating depth estimation models. By assessing RMSE, PSNR, SSIM, and the accuracy metrics, we can comprehensively gauge the quality and performance of depth map predictions.

C Added Regression Specifications

C.1 All Classes - Yolo9c and YOLO9e

C.2 Merging Gender - Yolo9c

C.3 Merging Religion - Yolo9c and 9e

C.4 Merging Gender and Religion - Yolo9c and 9e

Table 11: YOLO9c - All Classes

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.383***	0.387***	0.373***	0.377***	9.76e+09	1.26e+10*
NF vs NF	0.029	0.028	0.027	0.026	0.026	0.026
NF vs NM	0.075	0.074	0.073	0.072	0.073	0.074
NF vs NRF	-0.033	-0.034	-0.036	-0.037	-0.039	-0.038
NF vs NRM	0.181**	0.180**	0.179**	0.178**	0.178**	0.179**
NF vs OM	0.250**	0.250**	0.251**	0.251**	0.247**	0.247**
NF vs RFA	-0.084	-0.085	-0.087	-0.088	-0.093	-0.092
NF vs RFJ	0.103	0.101	0.098	0.097	0.098	0.099
NF vs RMA	0.381	0.381	0.383	0.383	0.389	0.387
NF vs RMJ	0.193**	0.192**	0.193**	0.192**	0.194**	0.193**
NF vs SM	0.136	0.135	0.139	0.138	0.138	0.139
NF vs VRFA	0.224**	0.224**	0.224**	0.224**	0.218**	0.219**
NF vs VRFJ	0.240	0.236	0.246	0.243	0.283	0.284
NF vs VRMJ	0.238***	0.237***	0.236***	0.236***	0.240***	0.240***
NM vs NM	0.232***	0.232***	0.231***	0.230***	0.232***	0.232***
NM vs NRF	0.197**	0.196**	0.195**	0.194**	0.196**	0.197**
NM vs NRM	0.102	0.101	0.099	0.098	0.102	0.105
NM vs OM	-0.029	-0.029	-0.030	-0.030	-0.038	-0.037
NM vs RFA	0.128	0.127	0.125	0.124	0.125	0.127
NM vs RFJ	0.195**	0.194**	0.192**	0.191**	0.195**	0.196**
NM vs RMA	0.160	0.160	0.165	0.165	0.152	0.152
NM vs RMJ	0.049	0.048	0.048	0.048	0.047	0.047
NM vs SM	-0.073	-0.074	-0.072	-0.072	-0.077	-0.076
NM vs VRFA	0.244***	0.244***	0.244***	0.244***	0.239***	0.237***
NM vs VRFJ	0.213	0.211	0.215	0.213	0.223	0.223
NM vs VRMJ	0.183**	0.182**	0.181**	0.181**	0.184**	0.185**
NRF vs NRF	0.082	0.081	0.079	0.078	0.075	0.075
NRF vs NRM	0.025	0.023	0.021	0.020	0.020	0.021
NRF vs OM	0.307**	0.308**	0.307**	0.308**	0.304**	0.304**
NRF vs RFA	0.138	0.137	0.135	0.135	0.132	0.133
NRF vs RFJ	-0.060	-0.061	-0.064	-0.065	-0.067	-0.067
NRF vs RMA	0.029	0.032	0.031	0.034	0.013	0.011
NRF vs RMJ	0.161*	0.160*	0.160*	0.160*	0.160*	0.160*
NRF vs SM	0.219**	0.219**	0.221**	0.221**	0.214**	0.214**
NRF vs VRFA	0.279**	0.278**	0.276**	0.275**	0.276**	0.276**
NRF vs VRFJ	0.617	0.611	0.605	0.601	0.601	0.599
NRF vs VRMJ	0.309***	0.308***	0.306***	0.305***	0.309***	0.309***
NRM vs NRM	0.136	0.134	0.132	0.131	0.129	0.130
NRM vs OM	0.433***	0.432***	0.432***	0.431***	0.431***	0.431***
NRM vs RFA	0.204**	0.203**	0.201**	0.200**	0.198**	0.198**
NRM vs RFJ	0.130	0.129	0.128	0.127	0.125	0.126
NRM vs RMJ	0.136	0.135	0.135	0.135	0.134	0.134
NRM vs SM	0.154	0.153	0.154	0.154	0.149	0.150
NRM vs VRFA	0.380**	0.379**	0.382**	0.380**	0.377**	0.377**
NRM vs VRMJ	0.261***	0.260***	0.258***	0.258***	0.260***	0.261***
OM vs OM	-0.030	-0.030	-0.027	-0.027	-0.041	-0.040
OM vs RFA	0.146	0.147	0.145	0.145	0.130	0.130
OM vs RFJ	0.617	0.618	0.624	0.624	0.627	0.626
OM vs RMJ	0.347**	0.347**	0.347**	0.347**	0.348**	0.348**
OM vs SM	-0.236*	-0.235*	-0.241*	-0.240*	-0.288**	-0.287**
OM vs VRFA	-0.221	-0.221	-0.218	-0.218	-0.234	-0.233
OM vs VRMJ	0.098	0.097	0.098	0.098	0.085	0.085
RFA vs RFA	-0.001	-0.001	-0.004	-0.005	-0.011	-0.011
RFA vs RFJ	0.043	0.041	0.032	0.031	0.026	0.026
RFA vs RMJ	0.261**	0.260**	0.259**	0.258**	0.255**	0.255**
RFA vs SM	0.412***	0.410***	0.415***	0.415***	0.408***	0.408***
RFA vs VRFA	-0.213**	-0.214**	-0.213**	-0.214**	-0.225**	-0.225**
RFA vs VRMJ	0.163*	0.162	0.158	0.158	0.155	0.155
RFJ vs RFJ	-0.040	-0.041	-0.050	-0.050	-0.053	-0.053
RFJ vs RMJ	0.018	0.017	0.016	0.015	0.012	0.012
RFJ vs SM	-0.043	-0.044	-0.041	-0.042	-0.046	-0.046
RFJ vs VRFA	-0.106	-0.107	-0.112	-0.112	-0.105	-0.105
RFJ vs VRFJ	-0.383	-0.386	-0.390	-0.392	-0.398	-0.398
RFJ vs VRMJ	0.038	0.037	0.031	0.031	0.028	0.028
RMA vs RMA	0.617	0.612	0.619	0.615	0.627	0.627
RMA vs RMJ	-0.383	-0.388	-0.381	-0.385	-0.373	-0.373
RMA vs VRMJ	-0.383	-0.381	-0.373	-0.372	-0.416	-0.416
RMJ vs RMJ	0.021	0.020	0.021	0.020	0.018	0.017
RMJ vs SM	0.071	0.069	0.074	0.073	0.069	0.069
RMJ vs VRFA	0.218	0.218	0.217	0.217	0.211	0.211
RMJ vs VRFJ	-0.059	-0.057	-0.060	-0.058	-0.040	-0.039
RMJ vs VRMJ	-0.042	-0.043	-0.045	-0.046	-0.058	-0.058
SM vs SM	-0.219*	-0.220*	-0.217*	-0.218*	-0.235**	-0.235**
SM vs VRFA	0.531	0.526	0.537	0.535	0.531	0.532
SM vs VRFJ	-0.090	-0.087	-0.088	-0.085	-0.117	-0.118
SM vs VRMJ	0.330**	0.330**	0.332**	0.333**	0.321**	0.321**
VRFA vs VRFA	-0.233*	-0.233*	-0.229	-0.229	-0.244*	-0.244*
VRFJ vs VRMJ	-0.098	-0.099	-0.098	-0.098	-0.088	-0.088
VRMJ vs VRMJ	-0.033	-0.034	-0.036	-0.037	-0.044	-0.044
Controls						
month		0.000		0.000	−1.26e+09	−7.64e+08
night		-0.007**		-0.006*	-0.003	-0.003
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.046	0.046	0.045	0.045	0.038	0.038
Observations	70379	70379	70379	70379	70379	70379

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Table 12: YOLO9e - All Classes

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.313***	0.318***	0.317***	0.321***	-1.865e+09	-2.601e+09
NF vs NF	0.113	0.113	0.111	0.111	0.123*	0.123*
NF vs NM	0.135*	0.135*	0.134*	0.134*	0.147**	0.147**
NF vs NRF	0.049	0.049	0.048	0.048	0.061	0.061
NF vs NRM	0.255***	0.255***	0.254***	0.254***	0.270***	0.270***
NF vs OM	0.247**	0.248**	0.246**	0.248**	0.255**	0.255**
NF vs RFA	-0.029	-0.029	-0.029	-0.029	-0.024	-0.024
NF vs RFJ	0.146*	0.146*	0.146*	0.146*	0.161**	0.161**
NF vs RMA	0.181	0.176	0.172	0.169	0.157	0.158
NF vs RMJ	0.286***	0.285***	0.286***	0.286***	0.303***	0.303***
NF vs SM	0.184**	0.183**	0.182**	0.182**	0.195**	0.195**
NF vs VRFA	0.459***	0.460***	0.461***	0.461***	0.461***	0.461***
NF vs VRFJ	0.601	0.604	0.607	0.609	0.600	0.600
NF vs VRMJ	0.276***	0.275***	0.276***	0.276***	0.291***	0.291***
NM vs NM	0.294***	0.294***	0.293***	0.293***	0.310***	0.310***
NM vs NRF	0.252***	0.252***	0.251***	0.251***	0.269***	0.268***
NM vs NRM	0.196***	0.196***	0.194***	0.194***	0.213***	0.213***
NM vs OM	0.005	0.006	0.002	0.003	0.014	0.014
NM vs RFA	0.216***	0.217***	0.217***	0.217***	0.224***	0.223***
NM vs RFJ	0.271***	0.271***	0.271***	0.271***	0.288***	0.288***
NM vs RMA	0.187	0.186	0.181	0.180	0.170	0.170
NM vs RMJ	0.122*	0.122*	0.124*	0.124*	0.141*	0.142*
NM vs SM	-0.013	-0.014	-0.015	-0.015	-0.005	-0.005
NM vs VRFA	0.323***	0.324***	0.322***	0.323***	0.324***	0.324***
NM vs VRFJ	0.319*	0.319*	0.320*	0.320*	0.321*	0.320*
NM vs VRMJ	0.285***	0.285***	0.286***	0.286***	0.305***	0.305***
NRF vs NRF	0.151**	0.150**	0.148**	0.148**	0.165**	0.164**
NRF vs NRM	0.096	0.096	0.093	0.093	0.109	0.108
NRF vs OM	0.387***	0.388***	0.387***	0.388***	0.396***	0.396***
NRF vs RFA	0.248***	0.248***	0.247***	0.247***	0.262***	0.262***
NRF vs RFJ	0.027	0.026	0.026	0.026	0.044	0.044
NRF vs RMJ	0.246***	0.246***	0.249***	0.249***	0.271***	0.271***
NRF vs SM	0.239***	0.239***	0.238***	0.238***	0.257***	0.257***
NRF vs VRFA	0.401***	0.402***	0.399***	0.400***	0.418***	0.418***
NRF vs VRFJ	0.311	0.310	0.311	0.311	0.317	0.317
NRF vs VRMJ	0.340***	0.340***	0.340***	0.340***	0.358***	0.358***
NRM vs NRM	0.210***	0.210***	0.207***	0.207***	0.224***	0.224***
NRM vs OM	0.687***	0.685***	0.674***	0.673***	0.701***	0.701***
NRM vs RFA	0.247***	0.246***	0.245***	0.245***	0.260***	0.260***
NRM vs RFJ	0.189**	0.189**	0.190**	0.190**	0.210***	0.210***
NRM vs RMJ	0.222***	0.222***	0.226***	0.226***	0.248***	0.248***
NRM vs SM	0.226***	0.226***	0.226***	0.226***	0.241***	0.241***
NRM vs VRFA	0.319**	0.320**	0.318**	0.319**	0.332**	0.332**
NRM vs VRMJ	0.315***	0.314***	0.315***	0.315***	0.334***	0.334***
OM vs RFA	-0.115	-0.114	-0.113	-0.112	-0.109	-0.109
OM vs RFJ	-0.124	-0.121	-0.117	-0.115	-0.105	-0.105
OM vs RMJ	0.609***	0.611***	0.610***	0.612***	0.622***	0.622***
OM vs SM	-0.285*	-0.286*	-0.290*	-0.290*	-0.271*	-0.271*
OM vs VRFA	-0.057	-0.057	-0.060	-0.060	-0.056	-0.056
OM vs VRMJ	0.157	0.158	0.153	0.154	0.162	0.162
RFA vs RFA	0.043	0.043	0.043	0.043	0.049	0.049
RFA vs RFJ	0.004	0.005	0.004	0.005	0.014	0.014
RFA vs RMA	-0.313	-0.310	-0.317	-0.316	-0.324	-0.324
RFA vs RMJ	0.312***	0.312***	0.314***	0.314***	0.326***	0.326***
RFA vs SM	0.394***	0.392***	0.395***	0.394***	0.410***	0.410***
RFA vs VRFA	-0.234***	-0.234***	-0.236***	-0.236***	-0.241***	-0.241***
RFA vs VRFJ	-0.313	-0.318	-0.307	-0.310	-0.316	-0.316
RFA vs VRMJ	0.274***	0.274***	0.277***	0.277***	0.288***	0.288***
RFJ vs RFJ	0.036	0.036	0.035	0.035	0.047	0.047
RFJ vs RMJ	0.114	0.114	0.115	0.115	0.134*	0.134*
RFJ vs SM	0.227**	0.226**	0.226**	0.225**	0.238**	0.238**
RFJ vs VRFA	-0.006	-0.006	-0.014	-0.013	-0.006	-0.006
RFJ vs VRFJ	-0.306	-0.306	-0.306	-0.305	-0.300	-0.300
RFJ vs VRMJ	0.160**	0.160**	0.159**	0.159**	0.173**	0.173**
RMJ vs RMJ	0.111	0.111	0.113	0.113	0.131*	0.131*
RMJ vs SM	0.150*	0.149*	0.149*	0.149*	0.162*	0.162*
RMJ vs VRFA	0.318**	0.318**	0.316**	0.316**	0.322**	0.322**
RMJ vs VRFJ	0.687*	0.682*	0.693*	0.690*	0.701*	0.701*
RMJ vs VRMJ	0.053	0.053	0.053	0.053	0.062	0.062
SM vs SM	-0.087	-0.088	-0.093	-0.093	-0.086	-0.086
SM vs VRFA	0.275*	0.271*	0.267*	0.265*	0.277*	0.277*
SM vs VRMJ	0.082	0.081	0.084	0.084	0.092	0.091
VRFA vs VRFA	-0.052	-0.053	-0.051	-0.051	-0.050	-0.050
VRFJ vs VRFJ	-0.242	-0.239	-0.236	-0.233	-0.254	-0.254
VRFJ vs VRMJ	0.123	0.124	0.120	0.121	0.128	0.128
VRMJ vs VRMJ	0.074	0.074	0.076	0.076	0.088	0.088
Controls						
month		0.000		0.000	2.275e+08	1.568e+08
night		-0.007**		-0.005*	-0.003	-0.003
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.042	0.042	0.041	0.041	0.035	0.035
Observations	76604	76604	76604	76604	76604	76604

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Table 13: YOLO9c - Merging Gender

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.478***	0.481***	0.496***	0.498***	0.481***	0.480***
N vs N	0.132***	0.132***	0.131***	0.131***	0.137***	0.137***
N vs NR	-0.050	-0.050	-0.053	-0.053	-0.049	-0.049
N vs RA	-0.081**	-0.081**	-0.082**	-0.082**	-0.082**	-0.082**
N vs RJ	-0.019	-0.019	-0.020	-0.019	-0.016	-0.016
NR vs NR	0.003	0.003	0.000	0.000	0.003	0.003
NR vs RA	0.053	0.053	0.051	0.051	0.051	0.051
NR vs RJ	0.105***	0.105***	0.105***	0.106***	0.107***	0.107***
RA vs RA	-0.216***	-0.216***	-0.218***	-0.218***	-0.226***	-0.226***
RJ vs RJ	-0.073**	-0.073**	-0.074**	-0.073**	-0.072**	-0.072**
Controls						
month		0.000		0.000	0.001	0.001
night		-0.005		-0.005*	-0.003	-0.003
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.030	0.030	0.028	0.028	0.010	0.010
Observations	36200	36200	36200	36200	36200	36200

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Acronym Legend:

N: Neutral, NR: Non-Religious, RA: Religious Arab, RJ: Religious Jew.

Table 14: YOLO9c - Merging Religion

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.130	0.137	0.140	0.147	0.136	0.139
NF vs NF	0.294**	0.294**	0.291**	0.291**	0.309**	0.309**
NF vs NM	0.330**	0.330**	0.327**	0.328**	0.346**	0.346**
NF vs NRF	0.231	0.231	0.227	0.227	0.242*	0.242*
NF vs NRM	0.441***	0.441***	0.437***	0.437***	0.455***	0.455***
NF vs RF	0.250*	0.250*	0.248*	0.248*	0.260*	0.260*
NF vs RM	0.459***	0.460***	0.462***	0.462***	0.481***	0.481***
NF vs VRF	0.583***	0.582***	0.583***	0.583***	0.595***	0.594***
NF vs VRM	0.483***	0.484***	0.486***	0.486***	0.501***	0.501***
NM vs NM	0.487***	0.487***	0.486***	0.486***	0.506***	0.506***
NM vs NRF	0.436***	0.436***	0.433***	0.433***	0.449***	0.449***
NM vs NRM	0.370***	0.369***	0.365***	0.366***	0.383***	0.384***
NM vs RF	0.432***	0.431***	0.430***	0.431***	0.446***	0.446***
NM vs RM	0.264*	0.264*	0.265*	0.266*	0.284*	0.283*
NM vs VRF	0.538***	0.538***	0.538***	0.538***	0.553***	0.550***
NM vs VRM	0.465***	0.465***	0.466***	0.466***	0.481***	0.481***
NRF vs NRF	0.302**	0.302**	0.297**	0.297**	0.311**	0.311**
NRF vs NRM	0.277*	0.277*	0.271*	0.271*	0.285*	0.285*
NRF vs RF	0.262*	0.262*	0.259*	0.259*	0.273*	0.273*
NRF vs RM	0.463***	0.464***	0.467***	0.467***	0.485***	0.485***
NRF vs VRF	0.470***	0.469***	0.469***	0.469***	0.493***	0.491***
NRF vs VRM	0.593***	0.594***	0.595***	0.596***	0.613***	0.614***
NRM vs NRM	0.406***	0.405***	0.398***	0.398***	0.413***	0.413***
NRM vs RF	0.414***	0.414***	0.411***	0.411***	0.428***	0.428***
NRM vs RM	0.447***	0.447***	0.450***	0.450***	0.472***	0.472***
NRM vs VRF	0.758***	0.759***	0.753***	0.755***	0.796***	0.794***
NRM vs VRM	0.549***	0.549***	0.549***	0.549***	0.568***	0.568***
RF vs RF	0.262*	0.261*	0.258*	0.258*	0.273*	0.272*
RF vs RM	0.363**	0.363**	0.364**	0.364**	0.379**	0.379**
RF vs VRF	-0.052	-0.052	-0.053	-0.053	-0.042	-0.045
RF vs VRM	0.342**	0.342**	0.340**	0.340**	0.354**	0.354**
RM vs RM	0.361**	0.361**	0.363**	0.363**	0.383**	0.382**
RM vs VRF	0.430**	0.429**	0.433**	0.432**	0.438**	0.438**
RM vs VRM	0.213	0.214	0.213	0.214	0.225	0.225
VRF vs VRF	0.152	0.154	0.141	0.144	0.181	0.165
VRM vs VRM	0.150	0.150	0.151	0.151	0.161	0.163
Controls						
month		0.000		0.000	-0.004	-0.004
night		-0.008**		-0.007*	-0.004	-0.004
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.042	0.042	0.042	0.042	0.033	0.033
Observations	57472	57472	57472	57472	57472	57472

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Acronym Legend:

N: Neutral, NR: Non-Religious, R: Religious, VR: Very Religious, F: Female, M: Male.

Table 15: YOLO9e - Merging Religion

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.393***	0.396***	0.401***	0.404***	0.375***	0.377***
NF vs NF	0.023	0.023	0.018	0.018	0.028	0.031
NF vs NM	0.058	0.058	0.053	0.053	0.064	0.067
NF vs NRF	-0.026	-0.025	-0.032	-0.031	-0.023	-0.020
NF vs NRM	0.177*	0.178*	0.171*	0.172*	0.183**	0.186**
NF vs RF	-0.014	-0.013	-0.019	-0.019	-0.013	-0.010
NF vs RM	0.209**	0.210**	0.209**	0.210**	0.221**	0.225**
NF vs VRF	0.344***	0.344***	0.338***	0.338***	0.339***	0.339***
NF vs VRM	0.171*	0.171*	0.170*	0.170*	0.182*	0.186**
NM vs NM	0.217**	0.217**	0.213**	0.213**	0.227**	0.230**
NM vs NRF	0.172*	0.172*	0.166*	0.167*	0.178*	0.181*
NM vs NRM	0.098	0.098	0.091	0.092	0.104	0.107
NM vs RF	0.151	0.151	0.146	0.146	0.156*	0.159*
NM vs RM	0.036	0.036	0.034	0.035	0.050	0.053
NM vs VRF	0.310***	0.311***	0.306***	0.306***	0.313***	0.312***
NM vs VRM	0.186**	0.186**	0.183**	0.184**	0.195**	0.197**
NRF vs NRF	0.056	0.057	0.050	0.051	0.061	0.064
NRF vs NRM	0.014	0.014	0.007	0.007	0.017	0.020
NRF vs RF	-0.014	-0.014	-0.021	-0.021	-0.012	-0.009
NRF vs RM	0.165*	0.166*	0.164*	0.165*	0.179*	0.183*
NRF vs VRF	0.279**	0.279**	0.272**	0.273**	0.275**	0.278**
NRF vs VRM	0.337***	0.337***	0.333***	0.334***	0.347***	0.351***
NRM vs NRM	0.158*	0.158*	0.150*	0.150*	0.162*	0.165*
NRM vs RF	0.142	0.143	0.137	0.138	0.149	0.151
NRM vs RM	0.179*	0.179*	0.178*	0.179*	0.196**	0.199**
NRM vs VRF	0.316**	0.316**	0.308**	0.309**	0.321**	0.320**
NRM vs VRM	0.235**	0.236**	0.231**	0.231**	0.244**	0.246**
RF vs RF	0.004	0.005	-0.003	-0.002	0.010	0.013
RF vs RM	0.106	0.107	0.104	0.105	0.118	0.121
RF vs VRF	-0.272***	-0.272***	-0.277***	-0.276***	-0.274***	-0.274***
RF vs VRM	0.155	0.155	0.149	0.150	0.159	0.161
RM vs RM	0.043	0.044	0.042	0.042	0.055	0.057
RM vs VRF	0.393**	0.394**	0.391**	0.392**	0.392**	0.395**
RM vs VRM	0.007	0.007	0.005	0.005	0.011	0.014
VRF vs VRF	-0.223*	-0.222*	-0.225*	-0.225*	-0.209	-0.215*
VRM vs VRM	-0.029	-0.029	-0.029	-0.029	-0.022	-0.018
Controls						
month		0.000		0.000	-0.002	-0.002
night		-0.003		-0.002	-0.005	-0.005
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.038	0.038	0.038	0.038	0.030	0.030
Observations	59419	59419	59419	59419	59419	59419

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Acronym Legend:

N: Neutral, NR: Non-Religious, R: Religious, VR: Very Religious, F: Female, M: Male.

Table 16: YOLO9c - Merging Gender and Religion

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.393***	0.390***	0.390***	0.390***	0.389***	0.393***
NR vs NR	0.262***	0.262***	0.262***	0.262***	0.265***	0.265***
R vs R	0.026**	0.026**	0.024**	0.024**	0.017	0.017
Controls						
month		0.000		0.000	0.001	0.001
night		0.001		0.001	-0.005	-0.004
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.068	0.068	0.063	0.063	0.025	0.025
Observations	13337	13337	13337	13337	13337	13337

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)

Acronym Legend:

NR: Non-Religious, R: Religious.

Table 17: YOLO9e - Merging Gender and Religion

Specification	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Pair Categories						
Intercept	0.421***	0.423***	0.426***	0.426***	0.413***	0.422***
NR vs NR	0.214***	0.215***	0.215***	0.215***	0.217***	0.217***
R vs R	-0.027**	-0.027**	-0.030***	-0.030***	-0.036***	-0.036***
Controls						
month		-0.000		-0.000	0.001	0.001
night		0.001		0.003	0.001	0.001
Fixed Effects						
Location FE	No	No	Yes	Yes	No	Yes
Video ID FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.053	0.052	0.048	0.048	0.009	0.009
Observations	13212	13212	13212	13212	13212	13212

$p < 0.10$ (*), $p < 0.05$ (**) , $p < 0.01$ (***)
Acronym Legend:
NR: Non-Religious, R: Religious.