Grenze International Journal of Engineering and Technology, June Issue

ICIS: A Vision - Language Dataset of Indian Streets for Culturally Aware Image Captioning and Benchmarking

J Bhuvana[1], Sandhya Giribabu[2], Shinigdapriya Sathish[3] and Vidarshanaa Saravanavel[4]
[1-4]Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
Email: bhuvanaj@ssn.edu.in, sandhya2370021@ssn.edu.in, shinigdapriya2370021@ssn.edu.in, vidarshanaa2370047@ssn.edu.in

***Abstract*— Existing large-scale Vision-Language (V+L) datasets are predominantly focused on Western urban environments, resulting in a geographical bias that hinders model performance in visually complex and culturally diverse settings like urban India. To bridge this gap, we introduce Images and its Captions of Indian Streets (ICIS), a novel dataset comprising 1,350 high-resolution images capturing heterogeneous traffic, crowded marketplaces, and traditional attire. An analysis of 5,400 captions reveals a remarkably low inter-annotator agreement score, which validates the high subjectivity and diversity of human interpretation inherent in these scenes. Benchmarking with the state-of-the-art BLIP model demonstrates a significant performance gap; while semantic metrics suggest general thematic understanding, lexical overlap and multimodal alignment scores remain notably low. Experimental evaluation using BLIP, BLIP-2, and GIT observes high semantic similarity (BERTScore ≈ 0.89) but extremely low lexical overlap (BLEU-4 < 0.06), showing the linguistic diversity and visual complexity of Indian street scenes. These results indicate that current V+L models struggle with the culturally nuanced descriptions required for unstructured environments. ICIS provides a challenging benchmark to advance the development of globally robust and culturally aware computer vision systems for applications in smart cities and autonomous navigation.**

***Index Terms*— Vision-Language Datasets, ICIS, Culturally-aware AI, Benchmarking, Urban India, Dataset Bias.**

I. INTRODUCTION

Modern Vision-Language (V+L) models perform well within the scope of their training data, but their effectiveness drops significantly when applied in visually and culturally different environments. This shortcoming arises from a basic bias in current large public datasets, which are predominantly focused on the West. Indian urban settings, in particular, offer a complicated and disorganized visual pattern that is seldom represented in these materials. In India, streetscapes feature bustling pedestrian movement, varied traffic, lively markets, and a blend of contemporary and historic buildings. Unique cultural items, like traditional clothing worn every day, add even more to this visual diversity.

This lack of representation greatly restricts the capacity of modern AI to generalize, making V+L systems ineffective at understanding the visual signals, cultural elements, and complicated, unstructured traffic situations that are specific to Indian cities.

Grenze ID: 01.GIJET.12.2.530

To address this key limitation in generalized AI, we present Images and Their Captions of Indian Streets (ICIS), a carefully selected set of 1,350 high-resolution images along with detailed captions created by humans, specifically intended to showcase the complete visual diversity of urban life in India.

The ICIS dataset is created to support a wide range of research and real-world applications focused on developing AI systems with stronger global capabilities: Scene Captioning and Retrieval: Acting as a difficult benchmark for image captioning [13] and multimodal comprehension in visually and culturally unique settings. Cultural AI: Supporting initiatives that record and examine local architecture and traditional clothing, aiding in the protection of cultural heritage. Smart Cities: Developing computer-vision systems for urban surveillance and infrastructure planning provisions to Indian conditions. Autonomous Driving: Developing perception systems capable of recognizing various road users and managing complex, unpredictable traffic situations.

This paper presents the following main contributions:

- A novel, domain-specific dataset comprising 1,350 high-resolution images with 5,400 captions was created as Images and its Captions of Indian Streets (ICIS) dataset.
- A comprehensive data analysis, highlighting the dataset's thematic diversity and emphasizing the high subjectivity of human captions (low inter-annotator agreement) as a key feature of scene complexity was performed.
- We benchmark a state-of-the-art model (BLIP) and demonstrate that while semantic understanding is high, the dataset poses significant challenges for precise lexical and visual grounding.
- We provide a public release of the dataset to foster research in culturally-aware vision-language models and urban computing.

II. Related Work

Several large-scale datasets have driven progress in image captioning [10] and urban-scene understanding. The COCO (Common Objects in Context) dataset [1] is a large-scale benchmark containing over 300,000 images with dense object annotations and five human-written captions per image. Similarly, Flickr30k [2] comprises 31,000 images focused on natural-language descriptions of everyday activities, supporting research on image-sentence retrieval. In the domain of urban environments, Cityscapes [3] provides pixel-level semantic segmentation for 5,000 images, though it is primarily limited to European street scenes. For the Indian context, the India Driving Dataset (IDD) [4] captures diverse traffic scenarios with fine-grained labels for road objects and lane markings. While these resources have enabled advances in multimodal learning and urban analytics [20], they primarily capture Western contexts or are limited to non-linguistic tasks.

Though they are large, these datasets provide notable limitations for vision-language [9] research in the Indian context. Some of these limitations are; Lack of Language-Based Descriptions: The Cityscapes Dataset and IDD are designed for detection and segmentation tasks and lack natural-language descriptions, making them unsuitable. Geographic and Cultural Bias: The datasets are mostly from Western culture and contain images of Western architecture, signs, and fashion. Hence, they are not very useful in understanding the unique visual characteristics of Indian cities. Underrepresented Scene Complexity: Scenes containing heavy footfall, unpredictable vehicle movement, and dynamic activity are underrepresented. This makes it difficult for a model to handle chaotic scenes. The ICIS dataset does not pose any such problems because it contains images of Indian streets in high resolution with detailed captions written by humans that describe India's architectural, social, and cultural diversity.

III. The ICIS Dataset

A. Collection Methodology

The ICIS dataset was created using a two-step method aimed at enhancing visual variety and ensuring relevance to different viewpoints. First, we carried out on-campus "walking tours," recording continuous video with smartphones held at eye level to mimic a pedestrian's perspective. This egocentric vision method is essential because it reflects the visual input of mobile AI systems and human agents, ensuring the data is highly relevant for navigation and assistive technologies. Still images were taken from these recordings at regular intervals, with one image captured every ten seconds. To expand the range of environmental diversity beyond campus areas, we selected publicly available footage of people walking on streets in Indian urban settings, including busy marketplaces and streets with mixed traffic during both day and night. The same frame-extraction method was used to process these recordings, resulting in a well-balanced set of 1,350 high-resolution real-world scenes.

B. Ethical Considerations and Copyright Compliance

The ICIS collection strictly followed ethical research practices and copyright guidelines. The authors confirm the following: Data Ownership: The researchers took all the original images. For selected secondary footage, the extraction process adhered to fair-use guidelines applicable to academic research. Privacy and Anonymity: The data gathering was carried out in a way that ensured no personal identifying information (PII) or confidential material was captured. The dataset does not include any personal information, ensuring complete anonymity. Institutional Standards: The gathering of data in public settings followed the ethical guidelines set by the institution. The dataset is meant solely for research use and does not violate any existing intellectual property rights.

C. Dataset Statistics

The ICIS dataset includes 1,350 high-resolution images, each accompanied by several captions written by humans. These images were gathered from a variety of environments and devices to guarantee a wide range of differences in visual quality and content. The comprehensive technical details are presented in Table 1 and Table 2. Captions were generated under shared annotation guidelines to maintain factual consistency while preserving the linguistic diversity introduced by multiple annotators. This variety in human perspective makes the dataset uniquely suited for benchmarking multimodal grounding [19] in complex environments.

The ICIS dataset spans multiple scene types to provide a rich representation of the Indian urban landscape. These categories, illustrated in Fig. 1, ensure that models are exposed to the structural and cultural nuances of the region; Heterogeneous Traffic: Urban roads featuring mixed traffic (cars, motorcycles, bicycles) and informal road behaviours typical of Indian cities. Market Dynamics: Crowded streets and open-air markets featuring local vendors, dense signage, and vibrant commercial activity. Landmarks and Infrastructure: Architectural structures such as temples, bus stops, and specific street furniture providing context for urban navigation. Residential Environments: Quieter neighbourhoods showcasing diverse housing layouts and local vegetation. Cultural Cues: Specific visual markers including traditional attire (e.g., sarees, dhotis) and unique transport modes (e.g., auto-rickshaws, cycle rickshaws).

TABLE I: IMAGE SPECIFICATIONS AND CAPTION STATISTICS

Property	Details	Property	Details
Total Images	1,350	Total Captions	≈5400
Resolution Range	720 x 1280 to 1080 x 1920	Captions per Image	3 to 5
Image Type	High-resolution RGB images	Contributors	Three annotators (approx. 450 images each)
Source Variation	Diverse capture devices and lighting conditions	Process	Independent captioning and stylistic review

Figure 1. Representative samples from the ICIS dataset having heterogeneous traffic, Market scenes and Cultural attire.

Captioning Guidelines: To ensure descriptive richness, annotators followed a structured protocol: (i) provide detailed, natural-language descriptions of objects and activities; (ii) prioritize culturally relevant elements (e.g., traditional attire, local transport); (iii) maintain strictly objective visual descriptions; and (iv) ensure grammatical precision. Annotator Background: The dataset was collaboratively annotated by three contributors intimately familiar with the Indian urban context. Each annotator handled a unique subset of 450 images, ensuring context-aware descriptions of marketplaces and cultural artifacts. Inter-Annotator Agreement: A validation subset of 100 images was independently annotated by all contributors.

The average inter-annotator agreement was calculated at 0.103. While statistically low, this indicates a high degree of linguistic variety and subjective interpretation, providing a rich training signal for generative AI models to learn diverse, valid descriptions for complex scenes.

D. Exploratory Data Analysis (EDA)

The ICIS dataset follows a structured directory format to facilitate integration with deep learning frameworks. It is organized into two primary repositories: Reference frames, containing 1,350 high-resolution images, and Captions, containing corresponding text files with at least three annotations per image. Annotation Density and Consistency: The distribution of captions per image is illustrated in Fig. 2 and summarized in Table 2. The results indicate that most images include either three or five captions. Images with exactly three captions appear most frequently in the dataset, while other captions count occur only rarely as shown in figure 2. Quantitative Image Quality Analysis: The visual quality of the dataset using brightness, contrast, sharpness, and colorfulness metrics (Fig. 3–4 and Table 2) are evaluated. Brightness, contrast, and colorfulness show distributions that are close to normal. In contrast, sharpness is strongly skewed. Most images exhibit moderate sharpness values, while a small number of outliers display extremely high values, likely caused by variations in lighting conditions and motion within the scenes.

Scene Type and Categorical Breadth: The categories "Traffic/road" and "Residential/houses" appear most frequently in the dataset. These are followed by "Nature parks" and "Public transport." Less common categories include "Landmarks/Monuments" and "Food/Vendors." A review of the dataset also confirms a 100% caption completion rate, indicating strong annotation consistency and overall dataset quality. The distribution of different urban scene categories in ICIS is illustrated Fig. 5. Inter-Annotator Agreement: The average inter-annotator agreement was measured at 0.1027. This value reflects the variety of perspectives captured in the dataset. Different annotators often describe different elements of the same scene. For generative AI systems, this diversity is beneficial because it encourages models to generate multiple valid interpretations rather than relying on a single fixed description. Lexical and Semantic Frequency: The linguistic characteristics of the captions are examined using word frequency analysis (Fig. 6) [15]. The lexical analysis reveals a high occurrence of common prepositions and articles such as "a," "of," and "on." However, the primary semantic meaning is conveyed through nouns like "street," "motorcycle," and "person," along with action-related words such as "wearing" and "walking." Color terms including "yellow" and "white" also appear frequently, reflecting the visually vibrant nature of Indian street environments. Qualitative Analysis: Success and Failure Cases: To further evaluate the BLIP model on the ICIS dataset, we conducted a qualitative analysis of the generated captions. The examples are grouped into success and failure cases, as shown in Table 4. Dataset Samples and Ground Truth: Figures 7 and 8 provide a comprehensive cross-section of the ICIS dataset and the corresponding multi-annotator outputs.

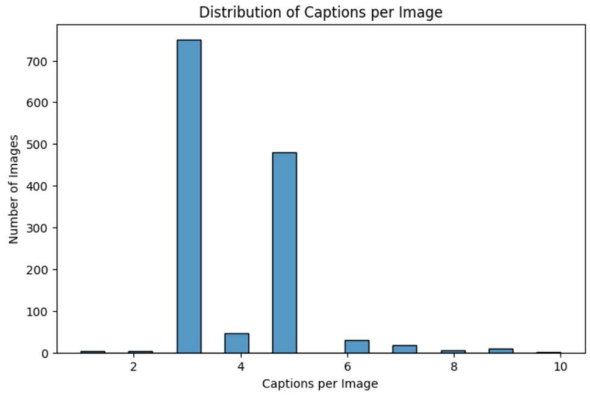

Figure 2. Distribution of Captions per image with Statistics:

Mean :3.93, Std Dev: 1.199, Min: 1.0; Median (50%):3.0, Max: 10.0

TABLE II: DETAILED IMAGE QUALITY METRICS

Stat	Brightness	Contrast	Sharpness	Color
Mean	103.036	62.175	699.496	31.554
Std Dev	18.188	8.509	675.711	11.046
Max	167.868	89.119	7721.447	104.083

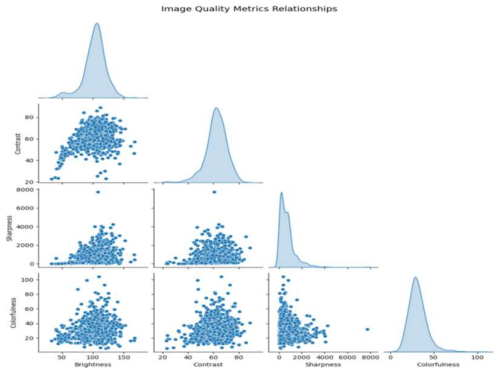

Figure 3. Relationship Between Image Quality Metrics.

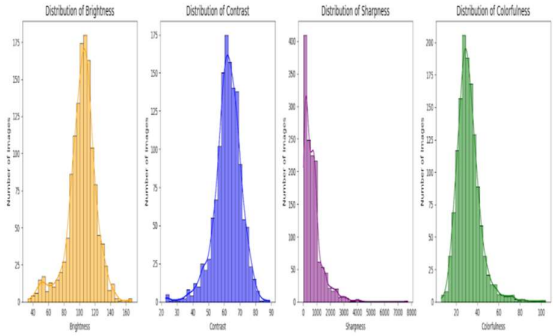

Figure 4. Individual Distribution of Image Quality Metrics.

IV. BENCHMARKING

To establish baseline results for the ICIS dataset, we evaluated three state-of-the-art vision–language models: BLIP (Bootstrapping Language-Image Pretraining): Uses Image-Text Matching (ITM) together with a Caption-Filtering (CapFilt) strategy to learn strong multimodal representations [4], [5]. BLIP-2: An improved and more efficient architecture that connects frozen vision encoders [11] with Large Language Models (LLMs) through a lightweight Querying Transformer (Q-Former) [12]. GIT (Generative Image-to-text Transformer): A unified transformer-based framework designed for end-to-end image-to-text generation tasks [14]. The models were evaluated on the task of Image Captioning. Performance was assessed using both semantic [15], [16]and lexical metrics: BERT Score (F1) for semantic similarity [17], CLIP Score for visual–text alignment, and METEOR together with BLEU-1 to BLEU-4 for n-gram precision and linguistic similarity (Table 3).

TABLE III: COMPARATIVE BENCHMARKING RESULTS: BLIP VS. BLIP-2 VS. GIT

Metric	BLIP	BLIP-2	GIT	Interpretation
BERT Score (F1)	0.887	0.895	0.892	High semantic alignment: captions convey identical thematic meaning.
CLIP Score	0.267	0.225	0.240	Moderate visual-textual alignment; room for improved visual grounding.
METEOR	0.543	0.279	0.194	Captures synonyms and stemming; BLIP shows superior meaningful matching.
BLEU-1	0.159	0.321	0.374	Moderate to high single-word overlaps across models.
BLEU-2	0.007	0.135	0.107	Low bigram overlap; indicates limited phrase-level matching.
BLEU-4	0.002	0.055	0.064	Extremely low four-gram overlap due to dataset complexity.

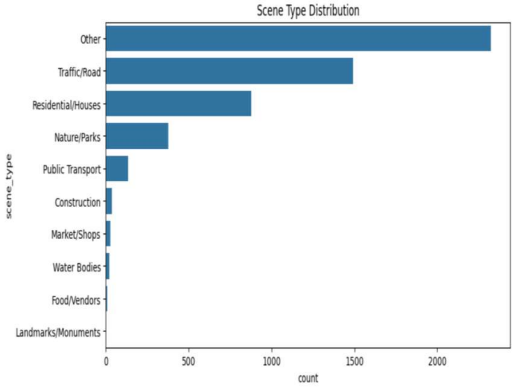

Figure 5. Scene Type Distribution

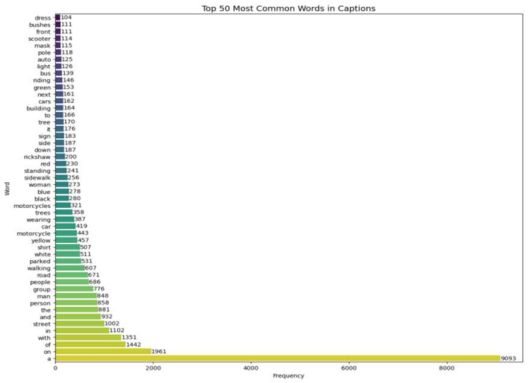

Figure 6. Frequency of Top 50 Common Words

Performance Gaps and Linguistic Diversity: One notable observation from the benchmarking results is the difference between semantic-level metrics (BERT Score and METEOR) and n-gram-based metrics (BLEU). While the high BERT Score values indicate that the models generally capture the overall context of Indian street scenes, the near-zero BLEU-4 scores reveal very limited exact phrase overlap.

This outcome reflects the inherent linguistic diversity of the ICIS dataset. In visually complex street environments, annotators often focus on different elements within the same image. For example, one caption

may describe a rickshaw while another highlights a pedestrian's clothing. As a result, a generated caption may still be accurate even if it does not exactly match a particular reference phrase.
Visual Grounding Challenges: The relatively low CLIP Score values across all evaluated models suggest that current vision–language systems still face difficulties in achieving precise visual grounding in complex Indian urban environments [18]. This further highlight the importance of ICIS as a challenging benchmark for developing culturally aware AI systems capable of interpreting dense and visually diverse scenes.
Applications and Use Cases: The ICIS dataset supports several practical applications aimed at developing more globally capable AI systems: Urban Monitoring: Automated analysis of urban environments for infrastructure planning and public safety. Captioning models can help identify overcrowded areas and assist in monitoring city surveillance systems. Traffic Analysis: Understanding patterns of vehicle and pedestrian movement to support congestion management. Cultural Heritage Documentation: Documenting historic landmarks with descriptive captions for digital archives, supporting tourism, education, and cultural preservation efforts [7].

V. Ethics and Limitations

Although ICIS provides an important domain-specific data set, its size of 1,350 images is smaller than widely used benchmarks such as MS COCO. Therefore, results should be interpreted with caution, as the dataset size may limit the evaluation of extremely large models. Protecting personal information in streetscape imagery is an important concern. Before the dataset is publicly released, all identifiable faces and vehicle license plates will be processed through an automated blurring pipeline to ensure compliance with global privacy and data protection standards

TABLE IV: QUALITATIVE ANALYSIS OF MODEL-GENERATED CAPTIONS

Image and Generated Caption	Performance Rationalization
	Caption: A woman in a purple dress, a woman in a green and pink saree The model accurately identifies distinct clothing and color attributes. Recognition of the saree demonstrates a successful shift beyond Western-centric object detection.
	Caption: A rickshaw driving down the street The model correctly identifies region-specific transport, which is frequently misclassified in standard datasets.
	Caption: A group of trees in a forest Failure: The model overgeneralizes the context as a "forest," failing to recognize the urban streetscape.

Figure 7. Representative samples from the ICIS dataset

- A man on a motorcycle
- A cow with a harness on its back
- A motorcycle parked in front of a building
- A tree

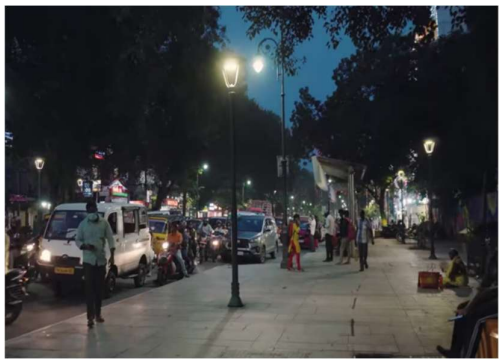

Figure 8. Representative samples from the ICIS dataset

- A street with cars and people on it
- A man wearing a mask standing in front of a white van
- A street light at night

While ICIS helps reduce Western-centric bias [21], it still has several inherent limitations: Geographical Scope: The dataset focuses exclusively on urban regions in India and does not represent rural environments or other South Asian areas. Annotator Perspective: The captions reflect the perspectives of three annotators with similar cultural backgrounds, which may influence the visual elements they choose to emphasize. Linguistic Mitigation: A manual review process was carried out to remove stereotypes or culturally insensitive descriptions in order to ensure fair representation [22].

VI. Conclusion and Future Work

We introduce Images and its Captions of Indian Streets (ICIS), a dataset designed to capture the visual complexity of Indian urban environments. Our analysis highlights the thematic diversity of the dataset and demonstrates how the subjectivity of human captions reflects the richness of real-world street scenes.

Benchmarking with state-of-the-art models like BLIP, BLIP-2, and GIT demonstrates that while semantic understanding is high [6], [8], the dataset poses significant challenges for precise lexical and visual grounding.

Plans for expansion include increasing the dataset size to include more diverse regional scenes; Adding object-level annotations such as bounding boxes and segmentation masks for Visual Question Answering (VQA) tasks; and extending the linguistic breadth by including multilingual captions in regional Indian languages.

References

[1] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 740–755.

[2] B. A. Plummer et al., “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 2641–2649.

[3] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3213–3223.

[4] Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar, “IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2019, pp. 1743–1751.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in Int. Conf. Mach. Learn. (ICML), 2023, pp. 19730–19742.

[6] W. Xie et al., “RA-CLIP: Retrieval augmented contrastive language-image pre-training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 19265–19274.

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] S. Paul and P. Y. Chen, “Vision transformers are robust learners,” in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 2, 2022, pp. 2071–2081.

[9] W. Su et al., “VL-BERT: Pre-training of generic visual-linguistic representations,” arXiv preprint arXiv:1908.08530, 2019.
[10] X. Li et al., “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 121–137.
[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.
[12] Yan et al., “Task-adaptive attention for image captioning,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 1, pp. 43–51, 2021.
[13] O. Ondeng, H. Ouma, and P. Akuon, “A review of transformer-based approaches for image captioning,” Appl. Sci., vol. 13, no. 19, p. 11103, 2023.
[14] N. Carion et al., “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 213–229.
[15] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 382–398.
[16] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 4566–4575.
[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” arXiv preprint arXiv:1904.09675, 2019.
[18] Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, “The Mapillary Vistas dataset for semantic understanding of street scenes,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 4990–4999.
[19] B. A. Paranjape and A. A. Naik, “DATS_2022: a versatile Indian dataset for object detection in unstructured traffic conditions,” Data Brief, vol. 43, p. 108470, 2022.
[20] Caesar et al., “nuScenes: A multimodal dataset for autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 11621–11631.
[21] Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, “Undoing the damage of dataset bias,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2012, pp. 158–171.
[22] Karkkainen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2021, pp. 1548–1558.