# Introduction to the Special Issue on Trustworthy Multimedia Computing and Applications in Urban Scenes

The fast spread of urbanization has significantly improved how people live, but it has also created important issues such as traffic jams, high energy usage, crime, and pollution. Nowadays, multiple multimedia sensing technologies, coupled with large-scale computing infrastructures, are producing a broad variety of big, multi-modality data within urban spaces at an unprecedented rate. These data provide valuable information about a city, helping us address these challenges. As a result, the research field of intelligent urban computing is gaining more and more attention from academia and industry.

Among other machine learning and artificial intelligence technologies, deep learning has made significant advancements in various aspects of intelligent urban computing. It has been particularly effective in areas like identifying and recognizing faces, detecting and tracking individuals, analyzing crowds, understanding behavior and events, capturing human–object interactions, recognizing vehicles, and so on. However, despite the practical success of multimedia computing and recognition in urban environments, there are concerns that these technologies are often seen as black-box systems, making them vulnerable and potentially unfair. As a result, researchers in the fields of multimedia, urban computing, and artificial intelligence are working hard to develop artificial intelligence systems that are trustworthy. These systems not only focus on performance metrics like accuracy, robustness, and efficiency but also emphasize fairness, accountability, transparency, interpretability, and accessibility in the context of multimedia computing and recognition in urban scenarios.

The goal of this special issue (SI) is to encourage a collaborative endeavor to explore the possibilities and obstacles associated with trustworthy multimedia analysis. We aim to gather inventive approaches, systems, and applications focused on trustworthy multimedia computing and its applications in urban environments. Of the 18 submissions received for this SI, only 9 of them have been accepted based on feedback from peer reviewers. Below is a brief summary of these accepted articles.

Attack, defense, and anti-spoofing are important topics in the field of trustworthy multimedia computing and applications in urban scenes. In the article "Improving Face Anti-Spoofing via Advanced Multi-Perspective Feature Learning," Wang et al. present a novel Advanced Multi-Perspective Feature Learning network (AMPFL) for Face Anti-Spoofing that is a crucial aspect of face recognition security. Previous methods have primarily failed to consider the diversity

**ACM Reference format:**
Wu Liu, Hailin Shi, Yunchao Wei, Dan Zeng, Nicu Sebe, and Jiebo Luo. 2023. Introduction to the Special Issue on Trustworthy Multimedia Computing and Applications in Urban Scenes. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 211 (July 2023), 4 pages.
https://doi.org/10.1145/3603534

1551-6857/2023/07-ART211 $15.00
https://doi.org/10.1145/3603534

among various attack types and the resemblances between some attacks and authentic faces. To address this, AMPFL leverages multiple perspectives to learn discriminative features and enhance the performance of face anti-spoofing. The network captures universal cues and perspective-specific cues, combines them, and further refines the features to effectively detect and differentiate between genuine and spoof faces. Evaluations on public datasets show that AMPFL achieves promising results and effectively addresses the shortcomings of single-perspective methods.

Besides the face anti-spoofing, to distinguish generated Deepfake media in visual, in the article "TCSD: Triple Complementary Streams Detector for Comprehensive Deepfake Detection," Liu et al. introduce a novel Triple Complementary Streams Detector (TCSD). TCSD extracts and integrates multi-view information from visual content. First, it incorporates a newly developed depth estimator to derive depth information (DI) not previously used in Deepfake detection. Second, it identifies inconsistencies between the image's foreground and background information (FBI) and discrepancies between local and global information (LGI). The process also involves an attention-based multi-scale feature extraction module that captures additional complementary features from DI, FBI, and LGI. Last, two attention-based feature fusion modules merge the gathered data adaptively. The experimental results demonstrate that TCSD performs exceptionally well in Deepfake detection, surpassing existing methods.

Hidden information within images shared online is also an increasing concern given the ubiquity of social networks and growing urbanization. In the article "A Siamese Inverted Residuals Network Image Steganalysis Scheme Based on Deep Learning," Li et al. propose a novel method for image steganalysis based on the Inverse Residuals structured Siamese network. This network consists of three stages: preprocessing, feature extraction, and classification. The preprocessing layer uses high-pass filters and depthwise separable convolution to accurately capture the correlation of residuals between feature channels. The feature extraction layer is based on the Inverse Residuals structure, enhancing the model's ability to obtain residual features. The final stage uses a fully connected layer to classify between cover and stego images. Through extensive experiments using three datasets (BossBase-1.01, BOWS2, and ALASKA#2), the authors demonstrate that their proposed method underlines superior adaptability to multi-source and arbitrary-size images.

Semantic segmentation of remote sensing images plays a vital role in urban scenes, involving urban planning, natural disaster monitoring, and land resource management. The article "Cross-scale Graph Interaction Network for Semantic Segmentation of Remote Sensing Images" proposes a Cross-scale Graph Interaction Network (CGIN) for semantic segmentation of complex, low-resolution remote sensing images. The CGIN integrates a semantic branch and a boundary branch. The semantic branch employs atrous convolution and a novel cross-scale graph interaction module to capture multi-scale semantic features. In contrast, the boundary branch uses a multi-scale boundary feature extraction module for multi-scale boundary features. To address sparse boundary pixels during fusion, a multi-scale similarity-guided aggregation module is introduced. The proposed CGIN demonstrated superior performance compared to existing methods in numerical experiments using two benchmark remote sensing datasets.

Vehicle re-identification is one of the emerging application topics in the urban sense. In the article "SSR-Net: A Spatial Structural Relation Network for Vehicle Re-identification," Xu et al. present a Spatial Structural Relation Network (SSR-Net) for vehicle re-identification, addressing the limitations of existing feature learning-based approaches. SSR-Net employs a Graph Convolution Network to model spatial structural relationships, incorporating both local and global features, improving robustness and discriminative capacity. Performance is enhanced by combining classification and metric learning. Results on VehicleID and VeRi-776 datasets demonstrate the effectiveness of SSR-Net compared to recent works.

Crowd localization is another important application in the urban scenes. The article "Dilated Convolution-based Feature Refinement Network for Crowd Localization" proposes the Dilated Convolution-based Feature Refinement Network (DFRNet) to solve the challenges of continuous scale variations in complex crowd scenes, particularly with small individuals at the periphery of images. DFRNet employs three branches and incorporates a Feature Perception Module to capture long-range contextual information at various scales, facilitating the detection of small individuals at image edges. Additionally, a Feature Refinement Module refines features at different scales, enhancing multi-scale contextual information representation. Experimental results across multiple datasets showcase DFRNet's superior performance.

Temporal Sentence Grounding in Videos (TSGV) seeks to align complex human activities in untrimmed videos of urban scenes with natural language sentences. In the article "A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach," Lan et al. try to address the biases in TSGV benchmarks by proposing a reorganization of datasets for out-of-distribution testing and introducing a new evaluation metric, "$dR@n, Iou = m$." They also present a causality-based multi-branch deconfounding debiasing framework for unbiased moment prediction, utilizing a multi-branch deconfounder to eliminate confounding effects. Enhanced feature encoding, including fine-grained textual features and decomposed visual positional information, improves the alignment between sentence queries and video moments. Experimental results reveal the approach outperforms the state-of-the-art models.

Interpretability is another important research topic in the field of trustworthy multimedia computing and applications in urban scenes. In the article "Temporal Dynamic Concept Modeling Network for Explainable Video Event Recognition," Zhang et al. propose to explore temporal concept receptive fields to identify crucial event concepts in varying time frames. The study introduces the temporal dynamic convolution (TDC) and the cross-domain temporal dynamic convolution (CrTDC) methods, which are designed to dynamically model these receptive fields and enhance concept representation. TDC and CrTDC form the basis of the temporal dynamic concept modeling network (TDCMN) developed for explainable video event recognition. Tested on large-scale datasets (FCVID, ActivityNet, CCV), TDCMN showed significant improvements in event recognition performance and provided new insights for constructing more explainable models.

Also for interpretability, the article "Sim2Word: Explaining Similarity with Representative Attribute Words via Counterfactual Explanations" proposes a novel interpretation method for image similarity models, which employs salience maps and attribute words to address concerns regarding explainability in deep neural networks. The proposed model generates visual salience maps and counterfactual explanations, utilizing two branches for salience map generation, i.e., the global identity relevant region discovery and multiattribute semantic region discovery. These branches identify visual evidence and semantic regions affecting similarity scores. The model also generates attribute words that best explain similarity via an erasing model. Evaluated on VGG-Face2 and Celeb-A benchmarks, the proposed approach offers convincing, interpretable explanations for similarity and is applicable to evidential learning cases.

In conclusion, we express our sincere appreciation to all the authors who have made noteworthy contributions to this SI, as well as the reviewers, whose insightful feedback has been invaluable. We are also grateful to the Editor-in-Chief and the dedicated publication staff, who have worked closely with us throughout the entire process. Our hope is that this SI will serve as an inspiration for future research and applications in trustworthy multimedia computing within urban environments.

Wu Liu
Explore Academy of JD.com, China
Hailin Shi
NIO, China
Yunchao Wei
Beijing Jiaotong University, China
Dan Zeng
Shanghai University, China
Nicu Sebe
University of Trento, Italy
Jiebo Luo
University of Rochester, USA
*Guest Editors*