Guest Editorial: Special issue on The British Machine Vision Conference 2022

Guang Yang[1], Angelica Aviles-Rivero[2], Yingying Fang[3], Zhenhua Feng[4], Gianluigi Ciocca[5], Yulia Hicks[6], Constantino Carlos Reyes-Aldasoro[7]

Introduction

This special issue in the International Journal of Computer Vision is dedicated to the 33rd British Machine Vision Conference, held from the 21st to the 24th of November 2022 in London (the Kia Oval, Cricket Ground), UK. The articles included in this issue have undergone rigorous peer-review, adhering to the International Journal of Computer Vision's highest standards.

Author Contributions

Comprising 8 papers, this issue focuses on machine vision, machine and deep learning, and a broad spectrum of applications.

The first article, by Jin et al., introduces a novel unsupervised multi-frame denoising strategy named One-Pot Denoising (OPD). This approach extends traditional pairwise or self-supervision to involve supervision among multiple noisy frames. The paper presents two specific algorithms, OPD-RC and OPD-AL, for data allocation and loss design. OPD performs well across denoising tasks and even outperforms some supervised methods. It exhibits robustness in handling inter-frame shifts and variations in noise levels. Besides, OPD shows excellent adaptability to different network architectures and loss compositions. Traditional denoising methods and deep learning-

[1] Guang Yang (g.yang@imperial.ac.uk) is with the Bioengineering Department and Imperial-X, Imperial College London, London W12 7SL, UK, the National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK, and the School of Biomedical Engineering & Imaging Sciences, King's College London, London WC2R 2LS, UK. Guang Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, NIHR Imperial Biomedical Research Centre (RDA01), Wellcome Leap Dynamic Resilience, and the UKRI Future Leaders Fellowship (MR/V023799/1).

[2] Angelica Aviles-Rivero (ai323@cam.ac.uk) is with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK.

[3] Yingying Fang (y.fang@imperial.ac.uk) is with the National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK.

[4] Zhenhua Feng (z.feng@surrey.ac.uk) is with the School of Computer Science and Electronic Engineering, the University of Surrey, Guildford, Surrey GU2 7XH, UK.

[5] Gianluigi Ciocca (gianluigi.ciocca@unimib.it) is with the Department of Informatics, Systems and Communication, University of Milan-Bicocca, 20126 Milano MI, Italy.

[6] Yulia Hicks (hicksya@cardiff.ac.uk) is with the School of Engineering, Cardiff University, Cardiff CF10 3AT, UK.

[7] Constantino Carlos Reyes-Aldasoro (Constantino-Carlos.Reyes-Aldasoro@city.ac.uk) is with the School of Mathematics, Computer Science and Engineering, City University of London, London EC1V 0HB, UK.

based methods are discussed, highlighting the limitations of detail loss and manual noise estimation. The difficulty of obtaining clean images has led to the development of unsupervised methods, with multi-frame denoising (MFD) being a key approach. While existing MFD methods are supervised, the paper's OPD method is the first unsupervised multi-frame denoising strategy. OPD employs mutual supervision among multiple frames, demonstrating superior performance in various denoising tasks, including challenging scenarios like mixed-blind denoising. The paper presents two implementations of OPD, OPD-random coupling and OPD-alienation loss, and provides source code and pre-trained models for further exploration. Experimental results indicate that OPD achieves state-of-the-art performance in denoising tasks, making it a promising approach for unsupervised multi-frame denoising.

The second paper, by Jain et al., proposes two extensions of Neural Radiance Fields (NeRF) for scene representation, specifically addressing challenges in neural image-based rendering from unposed images. These approaches aim to synthesize realistic images from novel viewpoints using unposed images while recovering accurate camera parameters, crucial for accurate rendering. The proposed methods leverage concepts from multi-view geometry, NeRF representation, and camera pose estimation. The first extension involves a multi-scale neural scene representation and single-image depth prediction to model object content at different resolutions, handling camera motion effectively. The camera parameters are made learnable within a neural fields-based modelling framework, and relative pose estimation between frames contributes to accurate absolute camera pose estimation. Robust cameras pose estimation is emphasized for accurate multi-scale neural scene representation. The second extension focuses on addressing practical issues in multi-view images, allowing the modelling of randomly captured image sets without relying on third-party software. It introduces a graph-neural network-based multiple motion averaging for camera pose estimation, enhancing the absolute camera pose solution. The paper demonstrates the effectiveness of these approaches through extensive experiments on benchmark datasets. Future directions for research include extending these methods to different cameras and scenes with specular objects. The paper emphasizes the importance of robust camera pose estimation for accurate scene representation and image synthesis, making significant contributions to the field of neural image-based rendering.

The third paper from Rao et al. presents a novel approach, referred to as VoRF, for editing and relighting human heads from a single image. The method represents human heads as a continuous volumetric field with disentangled latent spaces for identity and illumination. It employs a face prior model learned in an auto-decoder manner, and a reflectance Multi-Layer Perceptron (MLP) predicts One-Light-at-A-Time (OLAT) images for target lighting conditions. The proposed method demonstrates photorealistic and view-consistent results, outperforming existing works in the field. VoRF addresses the challenges of portrait editing, which is crucial in various applications such as virtual reality (VR), augmented reality (AR), movies, and photography. The paper emphasizes that current methods often struggle with 3D

modeling and encounter difficulties in handling viewpoint and illumination editing simultaneously. VoRF, by representing human heads as a continuous volumetric field and learning a prior model, is capable of generalizing to novel test identities and synthesizing novel illuminations effectively. The method combines OLAT images with target environment maps for relighting and is demonstrated to be effective through extensive qualitative and quantitative evaluations. The paper provides an application of VoRF for relighting faces using textual input and extends previous work by offering a more in-depth evaluation and ablative studies. The literature review highlights the vast landscape of portrait editing methods, including NeRF-based methodologies for general scene relighting. The paper positions VoRF as a unique approach, differing from traditional NeRF by representing multiple scenes, using HDR environment maps for lighting manipulation, and addressing specific challenges related to human head editing. However, the paper acknowledges some limitations, such as challenges in synthesizing eyes and facial expressions during relighting, difficulty in reproducing certain details, and potential generation of regions that do not exist in reality. Despite these limitations, the proposed VoRF approach stands out for its ability to synthesize novel lighting conditions and views, making it a promising solution for portrait editing and relighting tasks.

In the fourth work by Lai et al. develops a transformer-based architecture designed for egocentric gaze estimation, introducing a novel task of predicting gaze saccade/fixation from egocentric videos. The model incorporates a Global-Local Correlation Module to enhance representation learning by modelling correlations between global and local tokens. It achieves competitive performance on action recognition without additional design modifications, showcasing its versatility. This transformer-based model is the first of its kind for egocentric gaze estimation and outperforms the previous state-of-the-art model significantly. The proposed model demonstrates strong generalization capabilities, successfully applied to gaze saccade/fixation prediction and action recognition tasks. It can be easily integrated into other transformer-based architectures, making it an accessible plug-in for various applications. The evaluation is conducted on EGTEA Gaze+ and Ego4D datasets, and the paper provides visualizations supporting the claim of the effectiveness of the Global-Local Correlation Module. The paper contextualizes its work within the larger field of computational analysis of human gaze behaviour, emphasizing the novelty of predicting gaze targets from egocentric videos. Prior works on egocentric gaze estimation and transformer-based video representation learning are discussed to provide a comprehensive background. Despite the model's success, potential limitations are acknowledged, including the larger computational cost associated with transformer-based models, which may not be feasible for on-device computing. The paper also highlights potential challenges in learning gaze distribution when trained to predict action labels. In summary, the paper contributes a novel transformer-based architecture for egocentric gaze estimation, introduces a new task of gaze saccade/fixation prediction from egocentric videos, demonstrates improved action recognition performance without additional design, and showcases strong generalization capabilities. The model is presented as an easy-to-use

plug-in for other transformer-based architectures, making it a valuable addition to the field of egocentric gaze estimation and related tasks.
In the fifth paper, a novel approach for non-rigid reconstruction using event-based cameras is proposed by Xue et al. This method surpasses state-of-the-art event-based non-rigid reconstruction approaches and demonstrates robustness to noisy events and initial parameter estimates. The proposed technique employs a probabilistic optimization framework for estimating non-rigid object deformations, associating events with mesh faces on the object contour. The results show superior performance, particularly in reconstructing the motion of human hands. The paper highlights the advantages of event cameras over conventional cameras in computer vision tasks and notes the limited research on non-rigid reconstruction with event cameras. The presented algorithm takes event streams as input and outputs reconstructed object pose parameters, modelling event measurements at contours in a probabilistic manner. The evaluation encompasses both synthetic and real data sequences, showcasing improvements over existing methods, especially in hand reconstruction. The related work section provides an overview of various approaches, including scene reconstruction, rigid tracking, non-rigid shape reconstruction with frame-based cameras, and recent attention to non-rigid reconstruction and tracking with event-based cameras. It discusses a differentiable event stream simulator for non-rigid motion tracking, a deep neural network trained on synthetic event streams for deformation estimation, and proposes geometric contour alignment within a probabilistic optimization framework. The challenges and limitations of the proposed approach are acknowledged, including the loose coupling of frames and events in the optimization process, potential issues with self-occlusions within the hand, and the need for sufficient contour events for accurate deformation estimation. The paper suggests further investigation into challenging settings like 6D pose estimation or scenarios involving crossing hands. The evaluation and results section presents an ablation study on the data likelihood formulation and provides quantitative results of variants. The proposed approach outperforms existing event-based non-rigid reconstruction methods, highlighting its robustness and potential for combining with texture-based reconstruction. The paper suggests future work focusing on improving runtime efficiency by efficiently searching for correspondences.

Paper number six discusses ConMH, a one-stage self-supervised video hashing method for generating short binary representations of videos without ground truth supervision (by Wang et al.). ConMH incorporates video semantic information and understanding of video similarity relationships. It utilizes an encoder-decoder structure to reconstruct videos from temporally-masked frames, with a higher masking ratio found to be beneficial for video understanding. The method maximizes agreement between two augmented views of a video to generate more discriminative and robust hash codes. ConMH achieves state-of-the-art results on three large-scale video datasets, surpassing existing self-supervised video hashing methods. The proposed ConMH addresses the limitations of traditional Self-Supervised Video Hashing (SSVH) models, which often use a two-stage training pipeline. ConMH is designed as a one-stage SSVH method,

eliminating the need for a two-stage process. It combines video semantic information and video similarity relationship understanding, employing an encoder-decoder structure for video reconstruction from temporally-masked frames. The approach maximizes agreement between augmented views, contributing to the generation of more discriminative and robust hash codes. The results demonstrate the effectiveness of ConMH, achieving state-of-the-art performance on three large-scale video datasets. The paper contextualizes its work within the evolution of video hashing methods, highlighting the shift from early image hashing techniques to recent deep neural network-based video hashing approaches. It notes the prevalence of neighbourhood-preserving methods using a two-stage training strategy and mentions the popularity of contrastive learning and masking operations in self-supervised learning. In summary, ConMH is presented as a one-stage framework for self-supervised video hashing that combines video semantic understanding and similarity relationship exploitation in a single stage. It employs random temporal masking as a data augmentation technique for contrastive learning, achieving state-of-the-art results on three large-scale video datasets.

The paper seven by Lukezic et al. has two significant contributions: the Trans2k dataset and the DiTra tracker, both aimed at advancing transparent object tracking. The Trans2k dataset, the first of its kind, comprises over 2,000 sequences, providing 104,343 images for transparent object tracking. Standard trackers trained on Trans2k consistently show performance improvements of up to 16%. The DiTra tracker, a novel distractor-aware transparent object tracker, achieves state-of-the-art performance in transparent object tracking and demonstrates generalization capabilities to opaque objects. Both the Trans2k dataset and the DiTra tracker will be publicly released. Transparent object tracking poses challenges due to appearance dependence on the background, and existing trackers trained on opaque objects tend to perform poorly in this context. The Trans2k dataset addresses this gap by providing a comprehensive training dataset for transparent object tracking, leading to consistent improvements in the performance of standard trackers. The DiTra tracker introduces a distractor-aware formulation, treating localization accuracy and target identification as separate tasks. This innovative architecture sets a new state-of-the-art in transparent object tracking and generalizes well to opaque objects. The paper contextualizes its contributions within the broader landscape of visual object tracking, highlighting the limitations of existing trackers in handling transparent objects. The advancements made by DiTra and Trans2k fill this void in transparent object tracking research. The Trans2k dataset generation engine enables the creation of sequences with specific challenges, and its rendering engine has potential applications in training data for 6-DoF video pose estimation. In summary, the Trans2k dataset and the DiTra tracker are introduced as groundbreaking contributions to transparent object tracking. The Trans2k dataset enhances the performance of standard trackers, while the DiTra tracker, with its distractor-aware formulation, achieves state-of-the-art results in both transparent and opaque object tracking. The paper emphasizes the release of these resources to the public and suggests potential applications and inspirations for future research,

including the use of the Trans2k rendering engine in training data for 6-DoF video pose estimation.

The final study carried out by Bounareli demonstrates a novel approach to neural head/face re-enactment, leveraging a 3D shape model in combination with pretrained GANs for high-quality results. The method focuses on controlling facial pose and expression variations by discovering directions in the latent space of GANs. This enables the generation of realistic re-enacted faces, achieving successful outcomes for both single-source image re-enactment and cross-person re-enactment. The proposed approach surpasses state-of-the-art methods in producing higher-quality re-enacted faces. The paper emphasizes the limitations of previous methods in disentangling identity and pose and presents a simple pipeline that utilizes a 3D shape model to learn interpretable directions in the latent GAN space. The method is demonstrated to be effective for re-enacting real-world faces, showcasing capabilities for one-shot re-enactment and cross-person re-enactment. Despite the successes, the paper acknowledges potential challenges and visual artifacts, particularly in extreme head poses, where source and target faces are on the outskirts of the distribution. Large distances between source and target head poses can result in visual artifacts that affect source identity preservation. Inverted latent codes in extreme head poses are noted to be less editable. The proposed approach is positioned within the broader context of face re-enactment, drawing inspiration from previous works in discovering disentangled directions in GAN's latent space. It contrasts with other methods for explicit controllable facial image editing and reviews GAN inversion techniques for encoding real images into latent space. In summary, the paper introduces a novel neural head/face re-enactment approach, utilizing a 3D shape model and pretrained GANs for high-quality results. The method demonstrates superior performance in various applications, such as art and video conferencing, while acknowledging potential dangers associated with face re-enactment.

Conclusions

In conclusion, the 8 papers presented in this special issue collectively delve into a diverse range of research topics within machine vision, machine learning, and deep learning. With a focus on contributing to both expert practitioners and those seeking a comprehensive snapshot of current computer vision research, these papers underscore the multifaceted nature of ongoing investigations in the field: (1) Unsupervised Multi-Frame Denoising (OPD): Jin et al. propose a ground-breaking unsupervised multi-frame denoising strategy, One-Pot Denoising (OPD). This method outperforms traditional denoising approaches, demonstrating robustness in handling inter-frame shifts and noise level variations; (2) Extensions of Neural Radiance Fields (NeRF): Jain et al. address challenges in neural image-based rendering from unposed images by proposing two NeRF extensions. These approaches leverage multi-view geometry, NeRF representation, and camera pose estimation to achieve accurate scene representation and rendering; (3) VoRF for Editing and Relighting Human Heads: Rao et

al. introduce VoRF, a novel approach for editing and relighting human heads from a single image. Utilizing a continuous volumetric field representation and a disentangled latent space, VoRF achieves photorealistic and view-consistent results, demonstrating its efficacy in portrait editing; (4) Transformer-Based Architecture for Egocentric Gaze Estimation: Lai et al. present a transformer-based model for egocentric gaze estimation, achieving competitive performance on action recognition. The model's novel Global-Local Correlation Module enhances representation learning and exhibits strong generalization capabilities; (5) Novel Non-Rigid Reconstruction with Event Cameras: Xue et al. propose a novel non-rigid reconstruction approach for event cameras, demonstrating superior performance in reconstructing non-rigid object deformations. The method addresses limitations in existing event-based reconstruction approaches and shows robustness to noisy events; (6) ConMH for Self-Supervised Video Hashing: Wang et al. introduce ConMH, a one-stage self-supervised video hashing method that outperforms existing models. ConMH incorporates video semantic information, utilizes an encoder-decoder structure, and achieves state-of-the-art results on large-scale video datasets; (7) Trans2k Dataset and DiTra Tracker for Transparent Object Tracking: Lukezic et al. contribute the Trans2k dataset and DiTra tracker, advancing transparent object tracking. The dataset enhances tracker performance, while DiTra achieves state-of-the-art results by treating localization accuracy and target identification as separate tasks; and (8) Neural Head/Face Reenactment with a 3D Shape Model: Bounareli presents a novel approach to neural head/face reenactment, leveraging a 3D shape model and pretrained GANs. The method achieves high-quality results for reenacting faces, demonstrating effectiveness in single-source and cross-person reenactment scenarios. Collectively, these contributions showcase the evolving landscape of machine vision and computer vision research, offering innovative solutions to challenges and pushing the boundaries of what is possible in the field.

Dr Guang Yang on behalf of all co-guest editors (01-Feb-2024 in London, UK)

Senior Lecturer
UKRI Future Leaders Fellow
Imperial-X and Department of Bioengineering
Imperial College London
SW7 2AZ, London, UK
Email: g.yang@imperial.ac.uk

Acknowledgement

The British Machine Vision Conference is organised by The British Machine Vision Association and Society for Pattern Recognition for the purposes of the scholarly advancement of education and research in machine vision, pattern recognition and associated academic research areas including the application of such scholarly research within industry. The 33rd British Machine Vision Conference was also sponsored by ROKE, VIVO, Woven Planet, Intel and the Institution of Engineering and Technology.