Computer vision per sistemi di trasporto intelligenti: il progetto S.Pa.Ce.

Giuseppe Riccardo Leone, Andrea Carboni, Simone Nardi, Davide Moroni

Signals and Images Lab, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", ISTI-CNR
Eikontech srl, Pisa, Italy

giuseppericcardo.leone@cnr.it, andrea.carboni@isti.cnr.it, davide.moroni@cnr.it,
simone.nardi@eikontech.it

Abstract

Lo Smart Passenger Center (SPaCe) è una piattaforma integrata che mira a superare la complessità della gestione centralizzata delle infrastrutture di trasporto pubblico e dei veicoli. Il motore di intelligenza artificiale esamina i flussi quotidiani di persone, correla dati ed eventi diversi, prevede minacce ed eventi critici e propone contromisure. Questa enorme mole di dati proviene da una rete pervasiva di telecamere intelligenti che monitora costantemente le attività in stazioni, treni, autobus e altri luoghi di interesse. In questo lavoro, presentiamo il sottosistema distribuito di visione artificiale, lo stato dell'arte delle tecniche adottate e le funzionalità avanzate che questo sistema di sorveglianza intelligente offre ai livelli superiori di SPaCe. Tutto è sviluppato seguendo il paradigma della privacy-by-design: nessuna immagine viene registrata o trasmessa, ma tutte le elaborazioni avvengono sui nodi periferici del sistema.

1 Introduzione

Con il trasporto pubblico accessibile a un numero sempre maggiore di persone, gli operatori che lo gestiscono si trovano ad affrontare la duplice sfida di soddisfare la crescente domanda di flussi sempre maggiori di passeggeri garantendo allo stesso tempo elevati livelli di qualità e sicurezza. Oggi le aree di interesse sono completamente monitorate con sistemi video a circuito chiuso utilizzati principalmente dalle forze dell'ordine per fini di sicurezza. Tuttavia, le potenzialità di un sistema di controllo a circuito chiuso per ambienti confinati, come un veicolo, sono molteplici se analizzate attraverso moderni sistemi automatici di rilevamento e classificazione. A conoscenza dell'autore, ad oggi ci sono due esempi degni di nota che si propongono di applicare tecniche di AI alla videosorveglianza nel trasporto pubblico: il sistema NAIA [Thalesgroup2022] e il sistema MASTRIA [Alstom2022]. Entrambi sono soluzioni sperimentali che condividono gli stessi obiettivi: ottimizzare le risorse del trasporto pubblico, reagire agli incidenti stradali, incorporare tutte le parti interessate nel processo decisionale e nell'esecuzione, fornire viaggi più agevoli prevedendo le variazioni dei passeggeri.

Il sistema Smart Passenger Center (SPaCe) [Mermec2022] si inserisce in questo settore con l'ambizioso obiettivo di permettere agli operatori del trasporto pubblico di superare la complessità della gestione centralizzata delle infrastrutture e dei veicoli e di soddisfare i requisiti di sicurezza e protezione di tutte le parti interessate. Grazie al motore di intelligenza artificiale della piattaforma di integrazione SPaCe, i gestori dei servizi potranno migliorare l'efficienza delle proprie operazioni riducendo drasticamente l'impatto degli eventi di sicurezza. SPaCe consentirà notevoli risparmi sui costi fornendo agli operatori la possibilità di concentrarsi su eventi critici piuttosto che sulla microgestione. In questo lavoro, ci concentriamo sul sottosistema di computer vision di SPaCe che monitora costantemente le attività in stazioni, treni, autobus e altri luoghi di interesse. Esso è composto da una rete pervasiva di telecamere intelligenti il cui compito principale è il conteggio delle persone e il monitoraggio del viaggio dei passeggeri oltre ad una serie di funzionalità ausiliarie come il rilevamento di oggetti incustoditi, danni, sporcizia e la presenza di fuoco o fumo. Nelle sezioni seguenti, presentiamo lo stato dell'arte delle tecniche di computer vision di base coinvolte e le funzionalità avanzate che il sistema offre agli livelli superiori del sistema.

2 Architettura e funzionalità

L'architettura del sistema segue il paradigma "privacy-by-design": nessuna immagine viene registrata o trasmessa, ma tutte le elaborazioni avvengono nei sensori intelligenti del sistema. Pertanto ci concentriamo sulla pipeline di visione artificiale utilizzata su ognuno di questi sensori (fig. 1). Il primo passo è classificare e localizzare tutti gli elementi nell'immagine con *Object detection*, che è fondamentale per localizzare persone e oggetti nella scena. Per l'identificazione dei passeggeri ci affidiamo al *Riconoscimento facciale* che è anche utile per costruire il modello della persona utilizzato per il *Tracciamento delle persone*. Per stime e misure migliori, utilizziamo telecamere stereo o sensori di profondità in combinazione con telecamere a colori. Vediamo una breve rassegna delle tecniche fondamentali di visione artificiale ampiamente utilizzate nella pipeline di elaborazione delle immagini.

Object detection

I migliori risultati per questo compito si ottengono con le reti neurali convoluzionali (CNN) con modelli di apprendi-

mento supervisionato. R-CNN (Region-Based Convolutional Network) [Girshick *et al.*2014] permette di localizzare oggetti addestrando un modello usando una piccola quantità di dati annotati; SPP-net (Spatial Pyramid Pooling) [He *et al.*2015] mira ad eliminare la necessità di fornire immagini a dimensione fissa al classificatore e quindi ad evitare il ritaglio o il ridimensionamento delle immagini; Fast R-CNN (Fast Region-Based Convolutional Network) [Girshick2015] garantisce una migliore precisione media rispetto ai precedenti con apprendimento a stadio singolo. Faster R-CNN [Ren *et al.*2017] genera proposte regionali in un modo meno costoso di R-CNN e Fast R-CNN. R-FCN (Rete completamente convoluzionale basata sulla regione) [Dai *et al.*2016] a differenza di altri metodi basati su RPN, sostituisce i livelli FC dopo il pooling del ROI con operazioni di calcolo semplici e a basso costo. SSD (Single Shot Detector) [Liu *et al.*2016] è un metodo di rilevamento di oggetti che utilizza un'unica rete neurale profonda, garantendo una velocità adatta all'elaborazione video in tempo reale. YOLO (You Only Look Once) [Redmon *et al.*2016] come SSD è un metodo basato sulla regressione, che è computazionalmente adatto per il tempo reale (può raggiungere anche 155 fps) ma si comporta peggio di SSD su piccoli oggetti o vicino a oggetti tra di loro.

Riconoscimento facciale

La tecnica principale per identificare le persone si basa sul riconoscimento facciale (FR) [Whitelam *et al.*2017]; più precisamente nel sistema SPaCe ci interessa l'Identificazione del Volto: data l'immagine di un volto, viene generata una rappresentazione numerica, che serve per cercare all'interno di un database se il soggetto è noto. Gli algoritmi di Deep Learning utilizzati per FR vengono generalmente addestrati tramite set di dati di grandi dimensioni come LFW [Huang *et al.*2008] e VGGface2 [Cao *et al.*2018]. Ci sono molti problemi con la FR: errori di quantizzazione [Wu *et al.*2020], sbilanciamento delle classi nel dataset [Liu *et al.*2019, Yin *et al.*2019] e, soprattutto, le diverse dimensioni delle immagini [Massoli *et al.*2020]. Quest'ultimo perché i dataset utilizzati in fase di training sono composti da immagini ad alta risoluzione mentre, nei sistemi di sorveglianza, si possono avere immagini con risoluzioni molto basse (fino a 8x8 pixel). Per affrontare questo problema troviamo le tecniche di SuperResolution [Singh *et al.*2018, Kolouri e Rohde2015], le proiezioni in uno spazio comune [Li *et al.*2019] e il training di Cross-Resolution [Massoli *et al.*2020].

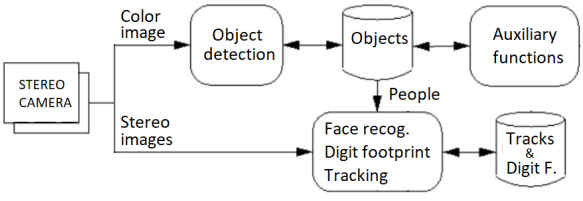

Figura 1: La pipeline di visione: i risultati dello object detector sono usati sia per il conteggio ed il tracciamento delle persone che per identificare danni, sporcizia e oggetti incustoditi.

Tracciamento delle persone

Tracciare le persone significa assegnare un numero di identificazione (ID) a tutte le persone presenti in una determinata immagine e riconoscere gli stessi soggetti in quelle successive riportando gli ID assegnati. Se qualcuno non è più presente, l'ID viene inserito nella lista *scomparsi*; ogni volta che compare una nuova persona, viene assegnata una nuova identificazione solo dopo aver verificato se la persona non è presente in questa lista. Il problema appena descritto va sotto il nome di reidentificazione della persona e in letteratura sono stati proposti molti approcci [Vezzani *et al.*2013]. Il Deep Learning ha permesso di ottenere ottime performance anche in questo settore [Ciaparrone *et al.*2020, Wojke *et al.*2017]. Un sito di riferimento che mostra i risultati dei migliori algoritmi è [MOT2022]: il vincitore del concorso CVPR19 proposto in [Mykheievskyi *et al.*2020] è ancora oggi l'algoritmo migliore. Utilizza un rilevamento basato su Faster R-CNN e una stima di somiglianza basata su caratteristiche di dimensione 128 prodotte dalle patch dell'immagine convertita nello spazio colore HSV. In secondo posizione troviamo [Zhou *et al.*2020] il cui codice in formato open-source è liberamente utilizzabile e modificabile.

Basandosi sulle precedenti tecniche di visione artificiale, il sistema implementa le seguenti funzionalità: conteggio dei passeggeri, riconoscimento di sporco e rifiuti, segnalazione di oggetti incustoditi, rilevamento di danni, allarme per la presenza di fuoco o fumo, tracciamento di persone con telecamere multiple. Tali funzionalità saranno descritte più in dettaglio nei prossimi paragrafi.

Conteggio delle persone

Una prima implementazione di questa funzionalità si ottiene contando le istanze di persone trovate da object detector. Questo approccio funziona bene per spazi di piccole e medie dimensioni in cui le persone sono distinte. Nel caso di luoghi grandi e affollati, invece, le tecniche basate sulla stima della densità di istanze [Zhang *et al.*2016, Boominathan *et al.*2016] ottengono prestazioni superiori rispetto al conteggio. Recentemente sono stati proposti in letteratura metodi basati su strategie self-supervised e unsupervised. In particolare, in [Liu *et al.*2018] gli autori hanno utilizzato un approccio self-supervised sfruttando il paradigma *learning-to-rank*, mentre in [Sam *et al.*2019] una tecnica basata su *winner-takes-all* applicato alle rappresentazioni prodotte da un DCNN dando vita ad un framework non supervisionato. Diversamente, gli autori in [Xiong *et al.*2019] hanno proposto un approccio self-supervised basato sull'idea che il conteggio può essere scomposto insieme alle componenti spaziali dell'immagine agendo indipendentemente sulle loro partizioni. In [Shen *et al.*2018] viene utilizzato un approccio "adversarial learning" per ridurre gli effetti di sfocatura tipici dei metodi di stima che utilizzano mappe di densità, sempre utilizzando una perdita che garantisce la coerenza spaziale del conteggio. Recentemente, alcuni articoli come [Lian *et al.*2019, Sam *et al.*2021, Yuting *et al.*2020] hanno tentato con successo di combinare approcci di regressione con approcci di rilevamento nel tentativo di sfruttare al meglio i vantaggi di entrambe le metodologie.

Rifiuti, danni, oggetti mancanti o incustoditi
Un object detector efficiente e preciso è la base per l'implementazione di tutte queste funzionalità insieme ad altre informazioni spazio-temporali. Per quanto riguarda l'identificazione di sporco o oggetti abbandonati, tali controlli verranno effettuati in assenza di persone e solo in momenti precisi come al termine del viaggio; in questo contesto, la corretta classificazione è già il risultato desiderato. I danni possono anche essere identificati più facilmente quando il veicolo è vuoto utilizzando un'immagine di riferimento. Sarà verificata l'efficacia di queste funzioni anche mentre i passeggeri sono a bordo.

Rilevamento di fuoco e fumo
Nel libro [Cetin *et al.*2016] viene presentata una rassegna completa dei metodi tradizionali di visione artificiale, comprese alcune considerazioni sul rilevamento mediante sistemi multi-view e multimodali. I metodi basati sul deep learning si sono rivelati utili nel rilevamento degli incendi, soprattutto per prevenire i limiti dei metodi tradizionali dovuti all'alto tasso di falsi allarmi e alla bassa precisione nel distinguere il fumo dalla foschia isolata. Per una revisione recente e completa, rimandiamo a [Gaur *et al.*2020, Khan *et al.*2021]. In [Yin *et al.*2017] viene proposta una rete profonda con normalizzazione (DNCNN) con 14 livelli per implementare l'estrazione e la classificazione automatiche delle caratteristiche. I risultati sperimentali mostrano che il metodo proposto raggiunge tassi di falsi allarmi inferiori allo 0,60% e tassi di rilevamento superiori al 96,37% su un set di dati di riferimento non liberamente disponibile. Alcuni set di dati utili per il benchmarking degli algoritmi di rilevamento del fumo sono [Toreyin e Cetin2009] e [Firesense2021].

Tracciameto di passeggeri rispettando il GDPR
Il sistema SPACe deve essere in grado di ricostruire il percorso completo delle persone durante i loro spostamenti quotidiani, dalla partenza all'arrivo. Pertanto, il sistema implementa un processo di tracciamento in tempo reale delle persone attraverso la rete di telecamere montate sui veicoli e nelle aree di parcheggio e transito. Tutto ciò deve essere realizzato nel rispetto della normativa vigente in materia di privacy, ovvero nessuna immagine con persone riconoscibili deve essere memorizzata su supporti permanenti né trasmessa in rete. Variazioni di luce, ambienti affollati e le conseguenti occlusioni parziali o totali, diversi punti di vista e angolazioni delle telecamere rendono questo compito particolarmente difficile rispetto ad un piccolo spazio interno. Per confrontare le persone in diversi frame di sensori diversi, abbiamo bisogno di un modello di aspetto che consenta una correlazione efficace. Con una buona immagine facciale si possono dedurre le caratteristiche fenotipiche del passeggero come sesso ed età; il volto, però, non è sempre ben visibile; è necessario considerare l'intera figura estraendo caratteristiche uniche per creare un modello di aspetto robusto: dall'analisi della silhouette si può stimare l'altezza e la corporatura della persona osservata, e si possono anche cogliere le caratteristiche estetiche associate all'abbigliamento o agli accessori presi in considerazione (es. colore dei vestiti, occhiali, trolley, valigie, zaini). In base al viso, alle caratteristiche fisiche e all'aspetto, il sistema SPACe costruisce una Orma Digitale (OD) della persona osservata, ovvero un modello matematico associato a quella persona, utile per il suo riconoscimento (reidentificazione) ogni volta che viene visto da una telecamera. La probabilità di successo di questo riconoscimento è fortemente influenzata dalla nitidezza dell'immagine, dalla prospettiva dell'inquadratura e dalle eventuali occlusioni presenti. In situazioni favorevoli di inquadratura frontale, il sistema garantisce una chiara identificazione delle persone osservate.

Un punto di partenza interessante per la costruzione di un Orma Digitale veloce e affidabile è quello basato sulle *deep features* che troviamo descritte in [Wojke *et al.*2017]. Per calcolare queste caratteristiche viene utilizzato un modello addestrato su milioni di immagini umane; è un vettore di dimensione 128 per ogni riquadro di delimitazione del classificatore che rappresenta le caratteristiche principali del riquadro. Il limite principale di questo approccio è il caso di un riquadro di delimitazione molto grande che include troppi background o informazioni non relative alla persona.

3 Conclusioni

In questo articolo abbiamo presentato l'idea generale dello Smart Passenger Center (SPaCe), una piattaforma di integrazione per la gestione e il supporto degli operatori dei sistemi di trasporto pubblico. Ci siamo concentrati sul sottosistema di Computer Vision che si basa su una rete di telecamere intelligenti per offrire servizi avanzati in modo solido, scalabile e salvaguardando la privacy. Ancora in fase di progettazione, il lavoro in corso esplorerà l'interazione tra visione artificiale e tecnologie pervasive e affronterà aspetti legati alla comunicazione e all'orchestrazione tra nodi-sensore. I risultati saranno dimostrati in una configurazione sperimentale realistica utilizzando un'installazione su vagoni di test per simulare le problematiche del mondo reale.

Riferimenti bibliografici

[Alstom, 2022] Alstom. The MASTRIA homepage. https://bit.ly/mastria_alstom, 2022.

[Boominathan *et al.*, 2016] L. Boominathan, S. Kruthiventi, e R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM Int. Conf. on Multimedia*, 2016.

[Cao *et al.*, 2018] Q. Cao, L. Shen, W. Xie, O. Parkhi, e A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, 2018.

[Cetin *et al.*, 2016] A. E. Cetin, B. Merci, O. Günay, B. U. Töreyin, e S. Verstockt. *Methods and techniques for fire detection: signal, image and video processing perspectives*. Academic Press, 2016.

[Ciaparrone *et al.*, 2020] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, e F. Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381, 2020.

[Dai *et al.*, 2016] J. Dai, Y. Li, K. He, e J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *ACM NeurIPS*, 2016.

[Firesense, 2021] Firesense. Firesense project homepage. https://cordis.europa.eu/project/id/244088, 2021.

[Gaur *et al.*, 2020] A. Gaur, A. Singh, A. Kumar, A. Kumar, e K. Kapoor. Video flame and smoke based fire detection algorithms: A literature review. *Fire Technology*, 56, 2020.

[Girshick *et al.*, 2014] R. Girshick, J. Donahue, T. Darrell, e J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[Girshick, 2015] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[He *et al.*, 2015] K. He, X. Zhang, S. Ren, e J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI*, 37(9), 2015.

[Huang *et al.*, 2008] G. B. Huang, M. Mattar, T. Berg, e E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[Khan *et al.*, 2021] S. Khan, K. Muhammad, T. Hussain, J. Del Ser, F. Cuzzolin, e S. Bhattacharyya. Deepsmoke: Deep learning model for smoke detection and segmentation in outdoor environments. *Expert Systems with Applications*, 182, 2021.

[Kolouri e Rohde, 2015] S. Kolouri e G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *CVPR*, 2015.

[Li *et al.*, 2019] P. Li, L. Prieto, D. Mery, e P. J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Trans. on Information Forensics and Security*, 14(8), 2019.

[Lian *et al.*, 2019] D. Lian, J. Li, J. Zheng, W. Luo, e S. Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *CVPR*, 2019.

[Liu *et al.*, 2016] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, e A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[Liu *et al.*, 2018] X. Liu, J. van de Weijer, e A. D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, 2018.

[Liu *et al.*, 2019] H. Liu, X. Zhu, Z. Lei, e S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *CVPR*, 2019.

[Massoli *et al.*, 2020] F. V. Massoli, G. Amato, e F. Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99, 2020.

[Mermec, 2022] Mermec. The Smart PAssenger CEnter. https://bit.ly/mermec_space, 2022.

[MOT, 2022] MOT. Multiple object tracking benchmark. https://motchallenge.net, 2022.

[Mykheievskyi *et al.*, 2020] D. Mykheievskyi, D. Borysenko, e V. Porokhonskyy. Learning local feature descriptors for multiple object tracking. In *ACCV*, 2020.

[Redmon *et al.*, 2016] J. Redmon, S. Divvala, R. Girshick, e A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[Ren *et al.*, 2017] S. Ren, K. He, R. Girshick, e J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6), 2017.

[Sam *et al.*, 2019] D. B. Sam, N. N. Sajjan, H. Maurya, e R. V. Babu. Almost unsupervised learning for dense crowd counting. In *33rd AAAI Conf. on Art. Intelligence*, 2019.

[Sam *et al.*, 2021] D. Sam, S. Peri, M. Sundararaman, A. Kamath, e R. Babu. Locate, size, and count: Accurately resolving people in dense crowds via detection. *IEEE TPAMI*, 43(8), 2021.

[Shen *et al.*, 2018] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, e X. Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *CVPR*, 2018.

[Singh *et al.*, 2018] M. Singh, S. Nagpal, M. Vatsa, R. Singh, e A. Majumdar. Identity aware synthesis for cross resolution face recognition. In *CVPR Workshops*, 2018.

[Thalesgroup, 2022] Thalesgroup. Tha NAIA homepage. https://bit.ly/naia_thalesgroup, 2022.

[Toreyin e Cetin, 2009] U. Toreyin e E. Cetin. Sample fire and smoke video clips. http://signal.ee.bilkent.edu.tr/VisiFire/Demo/SampleClips.html, 2009.

[Vezzani *et al.*, 2013] R. Vezzani, D. Baltieri, e R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Survey*, 46-2(29), 2013.

[Whitelam *et al.*, 2017] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, e P. Grother. Iarpa janus benchmark-b face dataset. In *CVPR Workshops*, 2017.

[Wojke *et al.*, 2017] N. Wojke, A. Bewley, e D. Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

[Wu *et al.*, 2020] Y. Wu, Y. Wu, R. Gong, Y. Lv, K. Chen, e D. Liang. Rotation consistent margin loss for efficient low-bit face recognition. In *CVPR*, 2020.

[Xiong *et al.*, 2019] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, e C. Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *ICCV*, 2019.

[Yin *et al.*, 2017] Z. Yin, B. Wan, F. Yuan, X. Xia, e J. Shi. A deep normalization and convolutional neural network for image smoke detection. *IEEE Access*, 5, 2017.

[Yin *et al.*, 2019] X. Yin, X. Yu, K. Sohn, X. Liu, e M. Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019.

[Yuting *et al.*, 2020] L. Yuting, W. Zheng, S. Miaojing, S. Shin'ichi, Z. Qijun, e Y. Hongyu. Towards unsupervised crowd counting via regression-detection bi-knowledge transfer. In *ACM Int. Conf. on Multimedia*, 2020.

[Zhang *et al.*, 2016] Y. Zhang, D. Zhou, S. Chen, S. Gao, e Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.

[Zhou *et al.*, 2020] X. Zhou, V. Koltun, e P. Krähenbühl. Tracking object as points. In *ECCV*, 2020.