2025 14th Mediterranean Conference on Embedded Computing (MECO)

including

13th International Conference on Cyber-Physical Systems and Internet-of-Things (CPS&IOT'2025)

Editors

Radovan Stojanović

Budimir Lutovac

Lech Jóźwiak

Proceedings – Research Monograph

June 10th — 14th, 2025, Budva, Montenegro

MECO 2025 covers:

    – **MECO 2025 Conference**
    – **CPS&IOT'2025 (13th International Conference on Cyber-Physical Systems and Internet-of-Things)**

For difficulty accessing documents, please contact the conference organizers.
Email: meco.conference@gmail.com

Transforming Faces Into Video Stories-VideoFace2.0

Branko Brkljač[1§*], Vladimir Kalušev[2§], Branislav Popović[1], Milan Sečujski[1]

[1]*Department of Power, Electronic and Telecommunication Engineering*, *Faculty of Technical Sciences, University of Novi Sad*
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
{brkljacb*, bpopovic, secujski}@uns.ac.rs
[2]*Visual Computing & Perception Group*, *The Institute for Artificial Intelligence Research and Development of Serbia*
Fruškogorska 1, 21000 Novi Sad, Serbia
vladimir.kalusev@ivi.ac.rs

Abstract—**Face detection and face recognition have been in the focus of vision community since the very beginnings. Inspired by the success of the original Videoface digitizer, a pioneering device that allowed users to capture video signals from any source, we have designed an advanced video analytics tool to efficiently create structured video stories, i.e. identity-based information catalogs. VideoFace2.0 is the name of the developed system for spatial and temporal localization of each unique face in the input video, i.e. face re-identification (ReID), which also allows their cataloging, characterization and creation of structured video outputs for later downstream tasks. Developed near real-time solution is primarily designed to be utilized in application scenarios involving TV production, media analysis, and as an efficient tool for creating large video datasets necessary for training machine learning (ML) models in challenging vision tasks such as lip reading and multimodal speech recognition. Conducted experiments confirm applicability of the proposed face ReID algorithm that is combining the concepts of face detection, face recognition and passive tracking-by-detection in order to achieve robust and efficient face ReID. The system is envisioned as a compact and modular extensions of the existing video production equipment. Presented results are based on test implementation that achieves between 18–25 fps on consumer type notebook. Ablation experiments also confirmed that the proposed algorithm brings relative gain in the reduction of number of false identities in the range of 73%–93%.**

Index Terms—**VideoFace2.0, face detection, face re-identification (ReID), video analysis, multi-modal datasets**

I. INTRODUCTION

Video content has become the prevailing type of global internet traffic [1]. Similarly, media productions that include news, podcasts, and interviews had been steadily growing in the new internet formats. Due to the large amount of available data, different types of advanced machine learning (ML) vision tasks involving human speakers had also come into the focus of research community. However, availability of adequate video analysis tools represents a technological burden for more effective mastering of the respective tasks. Inspired by the historical success of the original Videoface digitizer [2] we have developed VideoFace2.0, Fig. 1. The main assumption is that analyzed video content, regardless of source and application, contains appearance of multiple persons or speakers that randomly enter the scene during the video, with face orientation towards the camera. The main goal of the system is to efficiently perform spatial and temporal localization of each unique face in the input video, i.e. solve the
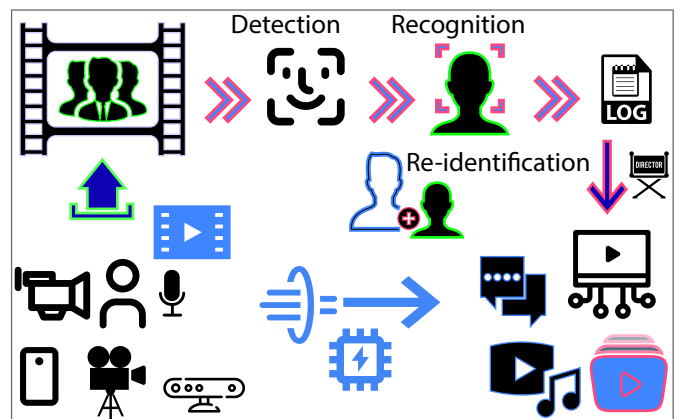
§ equal contribution

Fig. 1: VideoFace 2.0 processing workflow and applications.

face re-identification (ReID) task and provide functionalities for cataloging, characterization and production of structured video outputs or video stories, Fig. 1. It is assumed that neither the identities nor the number of people are known in advance.

Input video from recorded file, network stream or camera feed is uploaded into the system and analyzed frame-by-frame in order to detect human faces (spatial and temporal localization), catalog their identity (create a new entry, i.e. video story or re-identify the person if it was already present in the video before). It can also generate additional personal characteristics with specific time stamps (e.g. facial landmarks, facial dynamics, age, gender and audio/speech). The system is designed to operate in near real-time mode and allows for standalone embedded hardware implementation.

VideoFace2.0 is primarily designed to be utilized in application scenarios that include TV production, media analysis and creative industries, where video stories generated by the system would be an efficient way for fast and easier video montage of scenes in which specific persons of interest appear. E.g. editing of interviews or reportages brought by the field reporters to TV studios, talk-shows, podcasts and other formats that include multiple speakers or participants.

Another important application field includes production of custom video-based datasets that are required in different ML tasks involving multi-modal inputs like speech, text and image. Challenging vision tasks like lip reading and multi-modal speech recognition require large-scale language and speaker specific video datasets necessary for ML model training. Thus, one of the main goals of VideoFace2.0 and the proposed functionality and designed elements was also to facilitate creation

of high-quality multi-modal datasets involving human speakers. The main contribution of the paper is the generic and modular algorithm for face ReID, which is combining the concepts of face detection, recognition and passive tracking-by-detection.

The rest of the paper is organized as follows. In Section II we introduce the main vision tasks behind the proposed processing workflow and discuss current state-of-the-art in the field. Next, in Section III we present the main technical and algorithmic elements of the proposed solution, including proposed person re-identification improvements. Section IV brings results of conducted experiments, performance evaluation and discussion of possible limitations. Finally, in Section V we reexamine the strategies for further improvements and provide a reference for future work and planned applications. Code implementation is available at: https://github.com/brkljac/VideoFace2.0, while the results of test experiments are also available on the following link: https://youtube.com/@kalusev.

II. Related work

The proposed processing workflow, Fig. 1, is based on four vision tasks that have been extensively studied and analyzed in the literature. These are face detection [3], [4], face recognition [5], [6], face re-identification (ReID) [7] and object tracking. Since each of these tasks is characterized by specific performance metrics and comprehensive evaluation benchmarks that go beyond the scope of this presentation, in this paper we have decided to rely on off-the-shelf solutions that have been properly tested and proven as competitive to current state-of-the-art approaches in practice. In terms of applied algorithms and chosen pre-trained models we have followed the usual paradigm in which face detection results are used as a prerequisite for face characterization using learned face embeddings and followed by face identity matching in the corresponding high dimensional feature space during the face ReID phase. Nevertheless, we provide a brief overview of related works that illustrate design challenges and characteristics of the proposed VideoFace2.0.

The main goal of face detection is to provide localization (bounding box) and probability (decision confidence score) of face appearance in the image. Since design requirements almost always are that face detection algorithm are fast and lightweight in terms of computational and memory resources, detection is usually performed over still images or individual video frames without the use of temporal information. It is a special case of object detection task, and as such depending on application requires the choice of optimal working point on the detector's receiver operating characteristic [8]. The trade-off between face detection sensitivity and the level of false alarms (precision) is determined by detector's role in processing workflow. Similarly to other object detectors [9], traditionally face detection was also designed as a two stage process involving face localization (e.g. sliding window search or region proposal networks [10]) and subsequent binary classification of region of interest (ROI). However, this changed with introduction of more efficient single stage detectors [11], [12].

Open-set face ReID task is a challenging vision problem that is attracting significant attention of research community. It can be regarded as an extension of face recognition task, in which face identity is determined by matching the query image of the face with collected face samples in the dynamically changing gallery or database of face identities. ReID or recognition can be based on metric learning, i.e. query association with the closest identity in the gallery based on nonlinear distance between their vector embeddings. Embedding computation is usually preceded by face alignment [13], which is also the case in this paper. Since pre-trained feature extraction models can be obtained using different learning strategies, it creates possibility to impose various additional requirements on embedding invariance to different signal perturbations or distortions.

A closely related problem to face ReID is also object tracking in video. In case of active tracker, after initialization generally there is no need for subsequent object detection or recognition in following frames, but the matching of ROIs can still be based on described nonlinear image embeddings. However, in specific setups that are characterized by efficient face detection models, active tracking can be easily replaced by equally successful passive tracker [14], which is significantly less computationally expensive. This is exactly the design approach that was pursued in the proposed VideoFace2.0, where the tracker component is introduced in order to alleviate possible identity mismatch. E.g. when the person that was present in the previous frame is not correctly recognized (paired with generated face embedding of any available identity in the gallery during the recognition phase), or not present (tracker confirms the new identity).

III. Method and implementation

During the design process we have relied on gradual system improvement and synergistic effect of combining different vision components described in Section II. The primary requirements were robust face based identification of all persons in the video and at the same time efficient implementation with near real-time processing. The main elements of the proposed face re-identification method are summarized in Algorithm 1.

Algorithm 1

1: **procedure** VIDEOFACE2.0 FACE REID
2:    *Initialization:* $I(t)$, *video frame at time t; open-set face gallery $\mathcal{G}$; face detector $\mathcal{D}$ and confidence threshold σ_h; face embedding $\mathcal{R}$; distance $d(\cdot,)$ and threshold τ_d; passive tracker $\mathcal{T}$ and IoU threshold τ; post filtering period t_{min};*
3:    Detect new faces: $\mathcal{F} = \mathcal{D}(I(t)) = \{F_1, ..., F_N\}$
4:    *Determine identities:*
5:    Discard unreliable F_i: $\mathcal{F} = \mathcal{F} \setminus \{\sigma(F_i) < \sigma_h\}, i \in 1..N$
6:    Recognition:
7:    *# 1:* compute $\mathcal{R}_\mathcal{F} = \mathcal{R}(\mathcal{F}) = \{\mathcal{R}_{F_1}, ..., \mathcal{R}_{F_N}\}$
8:    *# 2:* $\forall \mathcal{R}_{F_i} \in \mathcal{R}_\mathcal{F} \wedge \mathcal{R}_\mathcal{G} = \{\mathcal{R}_{G_1}, ..., \mathcal{R}_{G_K}\}$ **do:**
9:    *# 3: if:* $\min_{j=1..K} d(\mathcal{R}_{F_i}, \mathcal{R}_{G_j}) < \tau_d$
10:       *then: new detection i is re-identified as an existing person* $j^* = \arg\min_{j=1..K} d(\mathcal{R}_{F_i}, \mathcal{R}_{G_j})$
11:       *# 4: else: new detection "i" is not significantly similar to any of K existing identities in gallery $\mathcal{G} \Rightarrow$ possibly new identity (candidate for inclusion in gallery $\mathcal{G}$):* **call** $\mathcal{T}$
12:       *# 4.1: for candidate "i" call tracker $\mathcal{T}$ and check IoU between bounding box of the closest detection $G_c \in \mathcal{G}$ and bounding box of F_i.* **if:** $\mathcal{T}_{IoU}(F_i, G_c) < \tau$ **then:** *new candidate identity is not valid,* **else:** *add F_i as new identity to $\mathcal{G}$: $\mathcal{G} = \mathcal{G} \cup F_i \wedge \mathcal{R}_\mathcal{G} = \mathcal{R}_\mathcal{G} \cup \mathcal{R}(F_i)$, $|\mathcal{G}| = K + 1$*
13:    *Post filtering: If $t_{min} > 0$, next t_{min} frames count number of appearances of all newly added identities at step t, and only after that make them active in the gallery (i.e. put them on hold for t_{min} to confirm their validity). Otherwise discard them from the gallery.*
14: **end procedure**

Starting assumption of the adopted fast prototyping approach was that each of the face detection, recognition and tracking algorithms in Section II could be very efficient and stand out individually, but that on their own they would not provide the desired level of face identification functionality. This was

confirmed through preliminary tests, where despite the high level of sophistication and performance each of the above-mentioned components individually failed to provide consistent and robust extraction of faces from the video. E.g. face detector performance used to vary depending on sensitivity to input video signal characteristics. Face recognition that relied on pre-trained face embedding models was exhibiting high level of mismatches due to the fact that the embeddings were learned on unoccluded or less diverse face images compared to those present in the open-set of real-world human face appearances. Similarly, tracker component had difficulty in maintaining precise position and extent of the face over the entire video, with unresolved interruptions in the event that a person goes out of frame. On the other hand, design approaches that have proven the concept of combining highly efficient individual components in order to achieve functionality that goes beyond their individual performance, like the passive tracker in [14], were the main motivation for the proposed face ReID method.

The main concept of Algorithm 1 is that in terms of performance and efficient implementation face detection is the most mature vision technology in Fig. 1, and as such should be the core of the proposed system. Thus, the underlying assumption was that the chosen detector $\mathcal{D}$ will work in high sensitivity mode with possibly high false positives rate per frame $I(t)$, but very unlikely miss of the face F_i at time t if detection i has confidence score $\sigma(F_i) > \sigma_h$. On the other hand, matching of face inside the bounding box of F_i with one of the already present open-set identities in gallery $\mathcal{G} = \{G_1, ..., G_K\}$ is based on computation of similarity or distance score $d(\cdot, \cdot)$ defined by eq. (1). It operates on high-dimensional feature vectors $\mathbb{R}^d$ generated by nonlinear embedding $\mathcal{R}$ that is pre-trained on face recognition task.

$$\cos \angle(\mathcal{R}_{F_i}, \mathcal{R}_{G_j}) = \frac{\mathcal{R}_{F_i} \cdot \mathcal{R}_{G_j}}{\|\mathcal{R}_{F_i}\| \|\mathcal{R}_{G_j}\|},$$
$$d(\mathcal{R}_{F_i}, \mathcal{R}_{G_j}) = 1 - \cos \angle(\mathcal{R}_{F_i}, \mathcal{R}_{G_j}). \tag{1}$$

Thus, if the matching of $\mathcal{R}(F_i) = \mathcal{R}_{F_i}$ with any of K known feature vectors $\mathcal{R}_{G_j} \in \mathbb{R}^d, j = 1..K$ in gallery fails, it is an indicator of the possibly new face identity. However, since it is assumed that the adopted embedding $\mathcal{R}$ is pre-trained on some large-scale but standard face recognition task, despite being highly discriminative it will be exhibiting poor identity matching performance, i.e. high identity rejection rate, under considered open world setting. Namely, the presence of varying facial expressions, head pose, illumination, face occlusions (by person's hands or other objects like microphone or glasses) will lead to situations in which embeddings $\mathcal{R}_{F_i}, \mathcal{R}_{G_j} \in \mathbb{R}^d$, despite being very informative (with high discriminative power), will lead to undesirable rejections of face detections F_i belonging to already known identity j^* in the gallery. It will happen due to relatively simple decision rule, which is based on combining with cost efficient distance functions like cosine similarity distance d in (1). Therefore, in order to avoid possible false rejections of known identities, ReID threshold τ_d for distance based test statistic in Algorithm 1 was intentionally raised.

In order to further improve ReID performance, additional passive tracker $\mathcal{T}$ was also included in Algorithm 1, however

in a careful way that does not increase processing latency. It means that it was designed not to be constantly active, but only when needed, in addition to being computationally light (tracking-by-detection). Namely, since the detector and recognizer are successfully complementing each other most of the time, the only situations in which the tracker is invoked are when additional confirmation of the detection of new identity is necessary. In such cases, standard Intersection-over-Union (IoU) metric and the chosen IoU threshold τ are utilized to confirm that new identity has entered the video, step 4.1 in Algorithm 1. We note that such setting was possible since in the planned application scenarios, Fig. 1, faces are expected to enter the scene gradually, and it is highly unlikely that a new face will appear immediately over the same position of some face from the previous frame.

In addition to the described algorithmic elements, there is also a final, post filtering step where the deliberately introduced delay t_{min} can be traded for additional protection against fake new identities. However, introduced t_{min} delay in new identity approval only affects new identities when they appear for the first time, and could be replaced by a more complex ReID rule.

IV. RESULTS AND DISCUSSION

Extensive experiments of the proposed open-set face ReID procedure were conducted based on the code implementation provided in Section I. Experiments included visual assessment of ReID results on created visuals and video stories, statistical analysis of processing log files, and tests on three representative video sequences with varying number of people, dynamics and shooting conditions, Fig. 2. These also included side-by-side visual comparisons of the ablation experiments, Fig. 2d, in which the impact of each of the described algorithmic components was analyzed, Table I. For more information about ablation experiments and test sequences please see the descriptions in the provided code repository. We note that the given implementation is modular and allows for independent choice of methods for each of the components in Algorithm 1.

Regarding the implementation we would like to highlight the following capabilities. First, the ability to choose between near real-time generation of specified output video or efficient off-line post production of different types of video output based on detailed log of initial video analysis that is conducted only once. The log file contains information about the temporal and spatial localization of each valid identity in the gallery, their face embeddings and, depending on the settings, other data (gender and age estimates, distinctive face landmarks for each detection, and other statistics, Fig. 2e). Control flags and parameters allow fine-tuning of functions and their operation. Predefined video stories, with and without sound, for the user-selected identity include: a) ReID visualizations (overlays of various processing results over original video), Fig. 2d; b) cropped and optionally scaled face videos, Fig. 2a, or c) mouth region videos, Fig. 2b.

TABLE I: Reduction of false identities brought by Algorithm 1 - ablation study

#identities	exp 1	exp 2	exp 3	exp 4	true	γ[a] [%]	[m:s][b]
testvideo1	50	42	30	7	4	83	02:44
testvideo2	421	378	263	23	13	93	07:25
testvideo3	39	37	25	9	6	73	18:45

[a] Alg. 1 relative gain: $\gamma = (1 - \exp 4/(\sum_{i=1}^{3}(\exp i)/3)) \times 100\%$; [b] duration
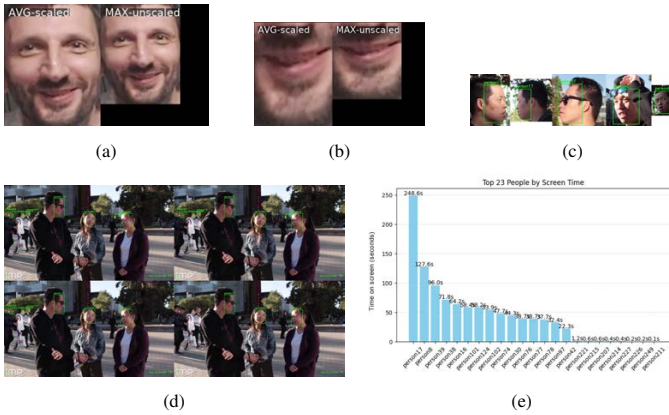
Fig. 2: Video stories and face ReID analyses: (a) face and (b) mouth region videos; (c) identity mismatch, (d) ablation experiments on testVideo2 [25], (e) on screen presence of all 23 identities found by the full Algorithm 1 in *testVideo2* - the best face ReID shown in the lower-right part of image in (d).

Note that generated log files contain only anonymized data (information about face appearances), and as such are suitable for sharing even if the original videos require specific licenses.

Experimental setup was based on SCRFD face detector [15] and ArcFace embeddings [16], which have been trained on the subset of WebFace dataset [17] and kindly provided by InsightFace project [18], i.e. buffalo_l model pack [19], with overall model size of 325 MB. VideoFace2.0 was tested on Intel i-7 notebook with 16GB of RAM and 4GB NVIDIA RTX3050 GPU. For efficient inference we have utilized ONNX runtime [20] with CUDA support, instead of possibly more efficient TensorRT [21], and achieved a processing speed between 18-25 fps. Relatively small model size should allow for implementation on embedded platforms like Jetson Nano [22], which is planned future work. For image processing and video coding we have relied on OpenCV [23] and FFmpeg [24] libraries. In all experiments involving full Algorithm 1, lower-right image in Fig. 2d, parameters were pre-set to $\sigma_h = 0.6$, ReID similarity score threshold of 0.4, i.e. $\tau_d = 0.6$, $\tau = 0.8$, and $t_{min} = 60$ frames, based on extensive pre-testing. We note that these are dependant on the specific choice of components in Algorithm 1.

V. CONCLUSIONS

In this paper we have presented a novel video analysis tool for face ReID and creation of structured identity-based video outputs or video stories. The proposed algorithm proved to be robust on various challenges brought by open-set face ReID and tests corresponding to the considered real-world application scenarios. Regarding limitations, we would like to point out that the system sometimes creates multiple identities associated with the same person, as a consequence of the proposed system design. Namely, the recognition stage in Algorithm 1 is relying on relatively simple identity association rule, as well as on face embeddings that do not account for occlusions and face image variations during the video. Possible solutions would be a more complex gallery updates and more robust association rules, which are planned for future work.

ACKNOWLEDGMENT

REFERENCES

[1] Cisco VNI. (2019) Cisco visual networking index: Forecast and trends, 2017–2022 white paper. [Online]. Available: https://www.futuretimeline.net/data-trends/pdfs/cisco-2017-2022.pdf (Accessed 2025-03-30).

[2] F. Frey, "Videoface to face," *Crash*, vol. 1987, no. 37, pp. 86–87, 1987. [Online]. Available: https://archive.org/details/crash-magazine-37/page/n86/mode/1up?view=theater

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[4] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Underst.*, vol. 138, pp. 1–24, 2015.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2008.

[6] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, and Y. Guo, "A survey of face recognition – OPPO Research Institute,," in *Int. Conf. on Automatic Face and Gesture Recognition*, 2023, pp. 1–59.

[7] Z. Cheng *et al.*, "Face re-identification challenge: Are face recognition models good enough?" *Pattern Recognit.*, vol. 107, p. 107422, 2020.

[8] A. Whalen, *Detection of Signals in Noise*. Academic press (HBJ), 1971.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2011.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[12] J. Deng, J. Guo, E. Ververas *et al.*, "RetinaFace: Single-shot multi-level face localisation in the wild," in *CVPR*, 2020, pp. 5203–5212.

[13] E. Onaran, E. Sarıtaş, and H. K. Ekenel, "Impact of face alignment on face image quality," *arXiv preprint arXiv:2412.11779*, 2024.

[14] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *14th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*. IEEE, 2017, pp. 1–6.

[15] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," in *ICLR*, 2022.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[17] Z. Zhu, G. Huang, J. Deng, Y. Ye *et al.*, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *CVPR*, 2021, pp. 10 492–10 502.

[18] J. Guo, J. D. Deng *et al.* (2025) InsightFace: 2D and 3D face analysis project. [Online]. Available: https://github.com/deepinsight/insightface (Accessed 2025-04-02).

[19] J. Guo, J. D. Deng *et al.* (2025) buffalo_l: InsightFace model pack. [Online]. Available: https://github.com/deepinsight/insightface/tree/master/python-package (Accessed 2025-04-02).

[20] (2025) Open Neural Network Exchange - ONNX. [Online]. Available: https://onnx.ai (Accessed 2025-04-02).

[21] (2025) NVIDIA corporation. TensorRT SDK for high-performance DL inference. [Online]. Available: https://developer.nvidia.com/tensorrt (Accessed 2025-04-02).

[22] (2025) NVIDIA corporation. Jetson Nano. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano (Accessed 2025-04-02).

[23] G. Bradski *et al.* (2025) OpenCV: Open Computer Vision library. [Online]. Available: https://opencv.org/ (Accessed 2025-04-02).

[24] F. Bellard, B. Bingham *et al.* (2025) FFmpeg: Fast Forward mpeg. [Online]. Available: https://ffmpeg.org/ (Accessed 2025-04-02).

[25] Joe and Mario. (2017) "Vancouver Talks" . Creative Commons Attribution license (reuse allowed). [Online]. Available: https://www.youtube.com/watch?v=ixhZadXXABg (Accessed 2025-04-02).