James Tompkin | Research Statement

jamestompkin.com • December 18, 2023

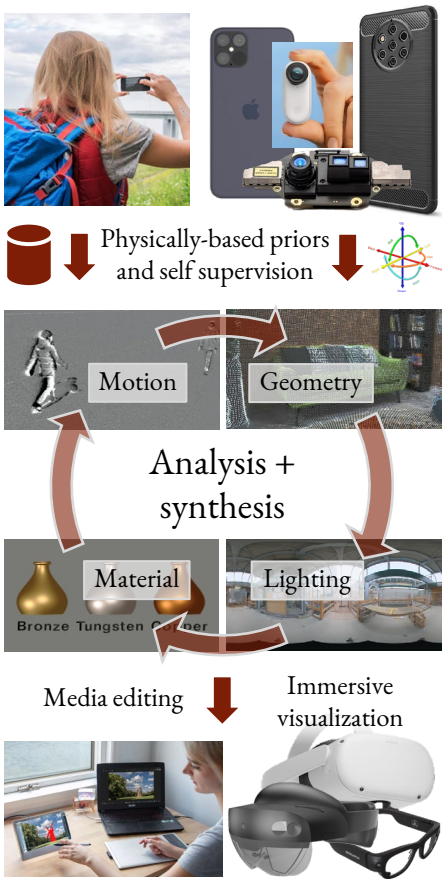
I am a visual computing researcher. I contribute to the disciplines of computer vision, computer graphics, and human-computer interaction to investigate how humans and computers can *make* and *make sense of* images. Much of my work concerns camera-captured images and videos. However, the tools to extract information from images often require expert human skill or use error-prone AI—this is because our computational methods lack real-world understanding. For instance, smartphones let anyone capture images and videos of our world, but editing and interacting with imagery with precise control and high quality is still difficult. As such, my lab develops algorithms to better computationally understand the world in images; in the process, we create new image and video generation and editing capabilities, new methods for virtual and augmented reality, and new techniques for applications like immersive teleoperation of robots.

I use different analogies across conceptual levels to describe my work:

Academic: As 'proof by reconstruction'. We evaluate our computational ability to understand images via a *cycle* of analysis by computer vision and synthesis by computer graphics: Can we reconstruct what we capture while still inferring useful and principled models of the visual world?

Technological: Defining the future smartphone. My algorithms create and improve the capabilities of camera systems comprising many sensors, lenses, and illuminators. My work must consider the trade-offs within design spaces of software and hardware to be accurate, efficient, and practical.

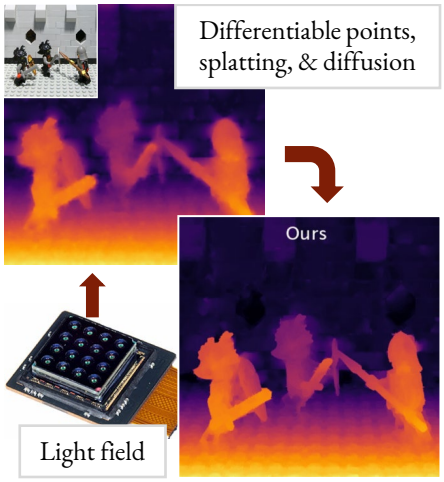
Personal: For human creativity. While new sensing capabilities might apply to many tasks, a guiding principle for me is how to give humans new interactive tools to experience the world and tell stories with images. Thus, questions of perception, visualization, and interaction are essential to me.



Approach overview. As analysis and synthesis methods, vision and graphics form a cycle of mutual support that lets us build useful models of the world. My research helps to realize this cycle.

Reconstructing the Real World from Images

One grand challenge of visual computing is how to use cameras to digitally reconstruct the physical properties of dynamic real world scenes. The task is to model complex geometry, motion, lighting, and material properties and infer them from the visual appearance captured as pixel values. This task contains many ill-posed problems—e.g., the 3D scene is projected to 2D, motion may be occluded, and different lightings and materials map to the same appearance. But, solving them has inherent value as these properties conform to the fundamental science of the natural world. This provides robustness to downstream applications as they can reason about physically-meaningful representations rather than pixel values. Thus, if we can complete the task accurately, quickly, and in practical capture scenarios, then this capability has broad impact across our work and play lives.

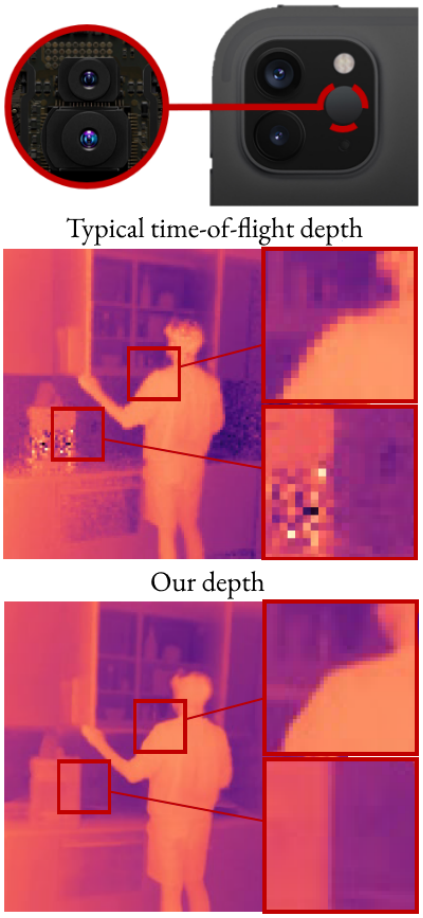


Computers must estimate depth to reason about the real world (more orange = closer). My work creates optimization approaches based on more-accurate physical reasoning of the sensing process.

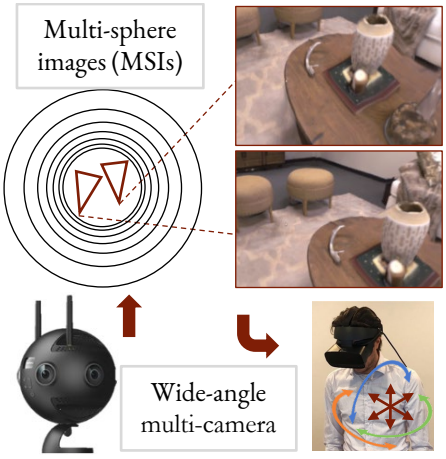
Physically-based Methods for Depth Sensing In reconstruction, the first step is often to estimate the depth of a scene for each camera pixel. Existing methods assume a simplified world or use supervised deep learning and data to overcome the need to explain appearance. But, these produce erroneous estimates for cases that lie outside the assumptions or data, such as dark or untextured surfaces, or materials that transmit or reflect light like glass. We showed that integrating physically-plausible image formation models into depth estimation can significantly reduce error: for camera systems that use many sensors or many lenses for robustness in dense/sparse and regular/irregular structures (e.g., light field, Raytrix; 1.4–8.4× less error) [9, 8], for camera systems that measure the time that emitted light takes to return but are susceptible to light bounce errors (e.g., Samsung Galaxy back camera; 2× less) [2], or for camera systems that emit light in structured patterns using projectors (e.g., Apple iPhone front FaceID camera; 3× less) [17]. Low light leads to significant sensor noise, so we also showed how to use the same models to learn robust noise estimates for specific sensors (2× less) [15]. Beyond better measurement, our better depth can be directly used in applications like photo editing, where compositing virtual graphics into a scene requires accurate depth edges and correct occlusion.

Reconstructing Scenes Quickly and with High Quality Moving our camera to capture different locations let us reconstruct larger-scale scenes like an apartment. This requires efficiency in how we capture and in how we computationally model and infer scene properties. One promising way to capture scenes quickly is with 360° cameras. We showed that a scene representation composed of concentric spheres can be inferred in real time and is sufficient to add missing motion parallax and binocular stereo perceptual cues to VR video [3]. For high quality, a DSLR can take hundreds of images of a scene. To handle the large data, we showed how to spend computational effort wisely and skip empty space by using a coarse initial geometry estimate. This lets us split the scene into independent tiles to be refined in parallel [22], and lets us model large outdoor scenes via distributed processing [21]. Further, we automatically separate complex lighting effects like reflections from polished wooden floors and mirrors, so that they cannot confuse geometric reconstruction. Then, we add them back afterwards to create photo-real interactive navigation of the real world [22, 21].

New Underlying Techniques Such approaches require new differentiable rendering and neural field techniques that find plausible solutions to ill-posed problems. Three years ago, this research area hardly existed; now, it has exploded in activity as the techniques achieve a step-change in quality. I helped to define this space with an influential state-of-the-art report [23], community webpage, and two courses at major conferences. Differentiable approaches that my lab has helped develop are 3D point splatting with radiative transport [8, 12], rasterization for 360° cameras [26], active illumination models for volume rendering [2], mesh refinement with retopologization [4], and material rendering for appearance acquisition [4, 27, 11].



Approaches to fuse colour and time-of-flight images are often heuristic. We show that physically-plausible integration leads to a significant reduction in depth errors and super-resolved details.



Our fast scene reconstruction method for 360° video lets us reproduce motion parallax and binocular stereo perceptual cues in VR ('6DoF').

Time, Meaning, and Space One persistent problem that I am currently tackling is how to reconstruct and represent dynamic scenes accurately and efficiently. For instance, some depth sensing techniques combine measurements taken at different times, and these fail under scene motion. I investigate how to overcome this by jointly reconstructing scene geometry and motion through self-supervision [2]: the scene and the sensing process itself must have internal physical consistency for what went where. Sometimes, pixel intensities alone really are insufficient to reconstruct scene geometry and motion, e.g., when using a single moving colour camera to capture dynamic objects. Humans can exploit semantics to disambiguate objects; our early work shows how to integrate low-level or 'bottom-up' cues like motion with high-level or 'top-down' semantic cues from unsupervised deep learned features to produce precise 4D object boundaries [13].

At the application level, rendering these scene representations produces images that are often indistinguishable from a photo yet provide interactive virtual navigation. Further, their accurate geometry and appearance reconstruction provide good scene measurement. These properties give our methods potential to improve robot teleoperation, where situational awareness is key to efficient control [16, 18]. We will explore this in a new NASA grant, where remote robots in the upcoming Artemis moon missions will send imagery from their roving back to Earth-based operators.

For indoor scenes, our reconstructions from multiple photographs allow high quality interactive navigation of real world spaces. All images above are from virtual cameras at uncaptured positions.

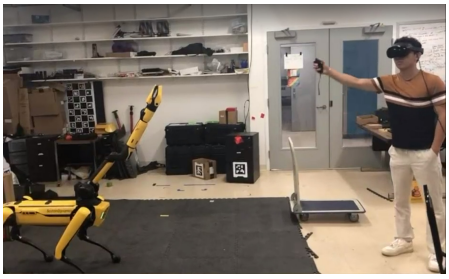
Defining the detailed dynamic boundaries of objects within scenes automatically is difficult from a single video, but we show that combining low-level motion cues with high-level semantic information lets us precisely decompose the scene.

Insights from Human Perception

Computational estimation of depth and motion often finds inspiration in biological vision. Cases where images do not provide sufficient information for reconstruction are paralleled in humans, where ambiguous stimuli in psychophysical studies reveal the successes and failures of our vision system. With my colleagues in cognitive science and in visualization, we have investigated when artificial neural networks share similar patterns of success and failure as humans [6, 25, 24]. For instance, in the estimation of depth from texture when observing Polka dot patterns on slanted surfaces, four human biases are reproduced by unsupervised convolutional neural networks [20]. Such experiments show that even basic problems in perception require more than just naïve deep learning if we want accurate results.

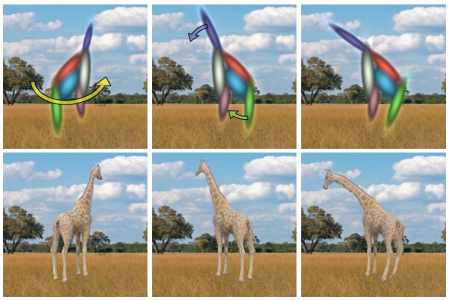
Generating Images to Model the Real World

Within the cycle of vision and graphics, I consider how to exploit large databases of images to infer models of the world. My lab uses these data to generate priors that ease ill-posed problems within physically-based algorithms, e.g., using deep learning to describe the distribution of a measurable physical quantity like the variations in skin reflectivity or in surrounding scenes that might illuminate a face [27]. With my students and colleagues, I have also explored whether changing the workings (or *inductive biases*) of deep learning models can help us exploit physical reasoning.



With my Brown colleagues in robotics, we are developing reconstruction algorithms to create VR teleoperation interfaces for Spot that improve human situational awareness. This is also useful for training robots from demonstration.

For instance, we proposed a popular method to embed a simple background / foreground image formation model into deep learning within the task of image translation [5, 1]. This induced the ability to automatically segment foreground objects from different classes by a statistical fortune: over a database of photos, the set of foreground objects are more similar to each other than to the set of backgrounds. Further, within a database of photos of, say, giraffe, the subjects are seen under varying viewpoints and each with a different body pose (walking, bending down, etc.). We showed that embedding 3D camera reasoning, the idea that subjects are composed of parts under motion, and the idea that all parts must align by their motion to a canonical subject as self-supervision is sufficient to automatically recover a controllable 'mannequin' for giraffe from image data alone. This enables interactive controllable generation of objects for image editing [14].

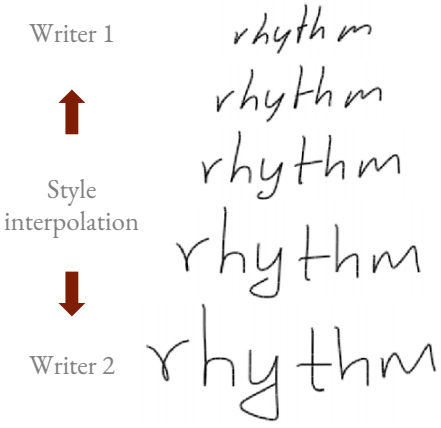


From a database of images of giraffe under different viewpoints and body poses, we automatically recover a parts-based generative model for giraffe images. This allows people to interactively make new images by posing a camera and a giraffe.

However, deep learning methods suffer from the problem of *spurious correlations* in data, where properties that we know to vary separately become entangled. For instance, for human faces, a deep-learned image generator might unwittingly encode 'frowning' and 'cropped hairstyles' together. One way to overcome this is to constrain known physical properties via neural network *conditioning*, but this was thought to impose an inevitable quality tax. My lab's more-rigorous analysis led to an approach that incurs no quality tax and creates realistic images with subtle control over face identity, expression, appearance, and lighting [7]. One of my motivations to achieve highly-accurate image modeling is to be able to detect and explain image manipulations (e.g., deep fakes) with even subtle changes [19].

Generative models of images entangle independent factors that happen to be correlated in data, so control of a smile might also change hairstyle or lighting. Our method correctly conditions the network so that these factors are independent, leading to subtle control and high quality images.

Beyond images, with one of my undergraduates who has a passion for script, we investigated how to represent the space of handwriting styles for the Latin alphabet. For this, we collected a new dataset of time-varying vector strokes drawn with digital pens. Then, accounting for pen lifts, temporal bar and dot ordering, and both writer and character styles, let us learn to represent handwriting and allow free interpolation of different styles [10].



We pose the computer a learning problem of how to model handwriting formed as time-varying vector strokes. The learned representation can smoothly vary between distinct styles such as the cursive 't' bar (Writer 1) and flicked 'y' (Writer 2).

Outlook

To computationally model the visual world, my research has added better physical reasoning through differentiable rendering and learning methods, aiding the rapid progress in visual computing over the past seven years. Yet, this pace belies a complex set of representational trade-offs—the *many* variations in the real world do not imply that 'one algorithm fits all'—and our current gains are wrought from considerable computation. Future progress will be born from hybrid approaches that know how best to model what is in the scene, and from hierarchical approaches that can abstract detail across spacetime. I will continue to explore how such approaches can be useful to *people* in creating practical interactive systems for everyday reconstruction, editing, and immersive visualization. Ultimately, I aim to advance the cycle of analysis and synthesis and help realize visual computing's promise of real-world computational understanding.

References

[1] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised Attention-guided Image-to-image Translation. *Neural Information Processing Systems (NeurIPS)*, 2018.

[2] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. TöRF: Time-of-Flight Radiance Fields for Dynamic Scene View Synthesis. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[3] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. MatryODShka: Real-time 6DoF Video View Synthesis using Multi-sphere Images. In *European Conference on Computer Vision (ECCV)*, 2020.

[4] Purvi Goel, Loudon Cohen, James Guesman, Vikas Thamizharasan, James Tompkin, and Daniel Ritchie. Shape from Tracing: Towards Reconstructing 3D Object Geometry and SVBRDF Material from Images via Differentiable Path Tracing. In *International Conference on 3D Vision (3DV)*, 2020.

[5] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving Shape Deformation in Unsupervised Image-to-image Translation. In *European Conference on Computer Vision (ECCV)*, 2018.

[6] Daniel Haehn, James Tompkin, and Hanspeter Pfister. Evaluating 'Graphical Perception' with CNNs. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2018.

[7] Yiwen Huang, Zhiqiu Yu, Xinjie Yi, Yue Wang, and James Tompkin. Removing the Quality Tax from Controllale Face Generation. *Winter Conference on Computer Vision (WACV)*, 2024.

[8] Numair Khan, Min H Kim, and James Tompkin. Differentiable Diffusion for Dense Depth Estimation from Multi-view Images. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Numair Khan, Qian Zhang, Lucas Kasser, Henry Stone, Min H Kim, and James Tompkin. View-consistent 4D Light Field Superpixel Segmentation. In *International Conference on Computer Vision (ICCV)*, 2019.

[10] Atsunobu Kotani, Stefanie Tellex, and James Tompkin. Generating Handwriting via Decoupled Style Descriptors. In *European Conference on Computer Vision (ECCV)*, 2020.

[11] Hyun Jin Ku, Hyunho Hat, Joo Ho Lee, Dahyun Kang, James Tompkin, and Min H Kim. Differentiable Appearance Acquisition from a Flash/No-flash RGB-D Pair. In *International Conference on Computational Photography (ICCP)*, 2022.

[12] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. GauFRe: Gaussian Deformation Fields for Real-time Dynamic Novel View Synthesis. In *In review*, 2024.

[13] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic Attention Flow Fields for Monocular Dynamic Scene Decomposition. In *International Conference on Computer Vision (ICCV)*, 2023.

[14] Youssef A Mejjati, Isa Milefchik, Aaron Gokaslan, Oliver Wang, Kwang In Kim, and James Tompkin. GaussiGAN: Controllable Image Synthesis with 3D Gaussians from Unposed Silhouettes. In *British Machine Vision Conference (BMVC)*. Presented at the CVPR 2021 Workshop on AI for Content Creation, 2021.

[15] Andreas Meuleman, Hakyeong Kim, James Tompkin, and Min H Kim. FloatingFusion: Depth from ToF and Image-stabilized Stereo Cameras. In *European Conference on Computer Vision (ECCV)*, 2022.

[16] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. Communicating and Controlling Robot Arm Motion Intent through MR Head-mounted Displays. *International Journal of Robotics Research (IJRR)*, 2019.

[17] Aarrushi Shandilya, Benjamin Attal, Christian Richardt, James Tompkin, and Matthew O'Toole. Neural Fields for Structured Light. In *International Conference on Computer Vision (ICCV)*, 2023.

[18] Austin Sumigray, Eliot Laidlaw, James Tompkin, and Stefanie Tellex. Improving Remote Environment Visualization through 360 6DoF Multi-sensor Fusion for VR Telerobotics. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021.

[19] Eleanor Tursman, Marilyn George, Seny Kamara, and James Tompkin. Towards Untrusted Social Video Verification to Combat Deepfakes via Face Geometry Consistency. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[20] Yuanhao Wang, Qian Zhang, Celine Aubuchon, Jovan Kemp, Fulvio Domini, and James Tompkin. On Human-like Biases in Convolutional Neural Networks for the Perception of Slant from Texture. *ACM Transactions on Applied Perception (TAP)*, 2023.

[21] Xiuchao Wu, Jiamin Xu, Xin Zhang, Hujun Bao, Qixing Huang, Yujun Shen, James Tompkin, and Weiwei Xu. ScaNeRF: Scalable Bundle-adjusting Neural Radiance Fields for Large-scale Scene Rendering. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023.

[22] Xiuchao Wu, Jiamin Xu, Zihan Zhu, Hujun Bao, Qixing Huang, James Tompkin, and Weiwei Xu. Scalable Neural Indoor Scene Rendering. *ACM Transactions on Graphics (SIGGRAPH)*, 2022.

[23] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural Fields in Visual Computing and Beyond. In *Computer Graphics Forum (EG STAR)*, 2022.

[24] Fumeng Yang, Yuxin Ma, Lane Harrison, James Tompkin, and David H Laidlaw. How Can Deep Neural Networks Aid Visualization Perception Research? Three Studies on Correlation Judgments in Scatterplots. In *Human Factors in Computing Systems (CHI)*, 2023.

[25] Fumeng Yang, James Tompkin, Lane Harrison, and David H Laidlaw. Visual Cue Effects on a Classification Accuracy Estimation Task in Immersive Scatterplots. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2022.

[26] Qian Zhang, Min H. Kim, and James Tompkin. Learning Physically-based Material and Lighting Decompositions for 360 Images with Differentiable Rasterization. In *preparation*, 2023.

[27] Qian Zhang, Vikas Thamizharasan, and James Tompkin. Learning Physically-based Material and Lighting Decompositions for Face Editing. In *Computational Visual Media (CVM)*, 2022.

Appendix: Software and Datasets

Many of our projects release open-source code and data to support the academic community in validating and building upon our work (as we benefit from the sharing of others); below is a selection.

Software

Semantic Attention Flow Fields 4D semantic scene reconstruction from monocular video, using coarse-to-fine preprocessing for large deep semantic features and including new labeled video segmentation data [13].
https://github.com/brownvc/saff

MatryODShka 6DoF VR Video Real-time scene reconstruction into multi-sphere image representation from 360° stereo cameras, including a stereo 360° (ODS) renderer for synthetic training data generation [3].
https://github.com/brownvc/matryodshka
https://github.com/brownvc/matryodshka-replica360

DiffDiffDepth Reconstruction from structured and unstructured light fields. Differentiable 3D Gaussian splatting via radiative transport and a locally-adaptive hierarchical basis preconditioner for fast diffusion [8].
https://github.com/brownvc/diffdiffdepth

Light Field Superpixels Low-level segmentation of light field data via piece-wise-planar depth reconstruction and scene clustering, including epipolar plane image edge bipartite matching for correct occlusion [9].
https://github.com/brownvc/lightfieldsuperpixels

Learned Material Decomposition Physically-based learned priors for decomposing face images into normals, SVBRDF material, and environment map lighting using a full high-dynamic-range pipeline [27].
https://github.com/brownvc/phaced

GaussiGAN Inferring parts-based mannequins of objects seen from different views and under different articulated poses, including interactive image generator application with control of pose and camera [14].
https://github.com/brownvc/gaussigan

GANimorph Inferring how to translate between two image classes with large object deformations in an unsupervised way [5].
https://github.com/brownvc/ganimorph

Datasets

Brown University Stylus Handwriting (BRUSH) 27,649 online (temporal) cursive handwriting samples in the Latin alphabet using English words from 170 writers. Vector stroke format collected with digital pen on tablet. Superseded existing datasets by containing a) temporal information for each stroke, b) character-level segmentation and labeling, c) writer repetition to model individual writer style variation [10].
https://github.com/brownvc/decoupled-style-descriptors

Brown University Social Video Verification Six-camera multi-view video of 27 participants giving speeches, along with deep-faked equivalents of the participants giving other speeches. This was the first multi-view video dataset for deep fake analysis [19].
https://github.com/brownvc/social-video-verification