

Computer Vision and Machine Learning for Computer Graphics

Saarland University - Summer Term 2022

Marc Habermann

Agenda

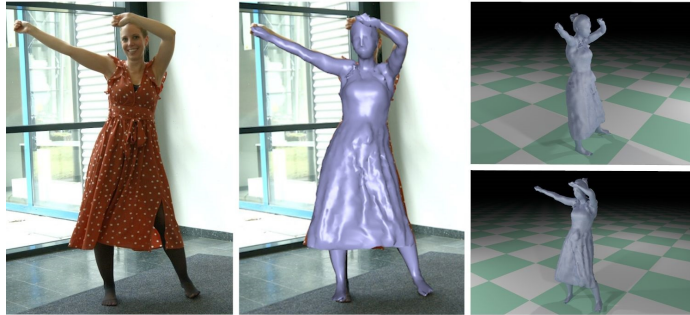
- About myself ←
- Introduction of participants
- How to read a scientific paper
- How to give a good scientific talk
- Questions and answers

About Myself

- **Marc Habermann**, Dr.-Ing.
- **Research group leader**
 - Graphics and Vision for Digital Humans Group
- **Max Planck Institute for Informatics**
 - Visual Computing and Artificial Intelligence Department
- **Research areas**
 - Computer Vision and Graphics
 - Human performance capture
 - Surface tracking
 - Neural rendering
- **Website:** <https://people.mpi-inf.mpg.de/~mhaberma/>
- **Group page:** <https://gvdh.mpi-inf.mpg.de>



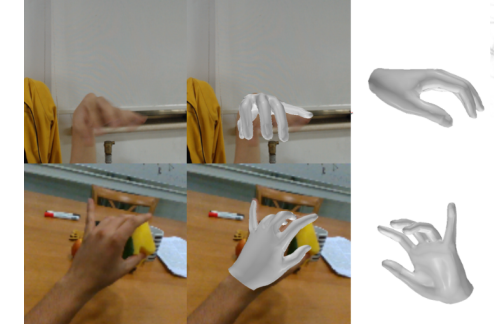
Research Interests



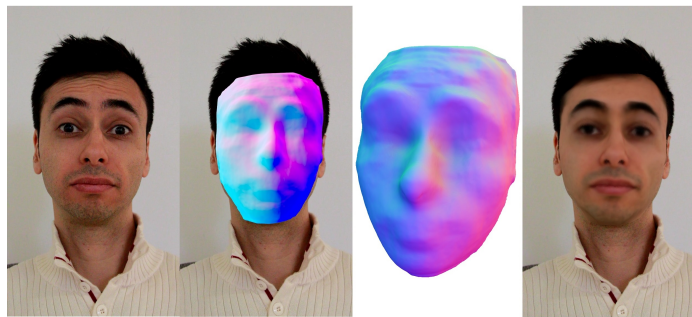
Human performance capture



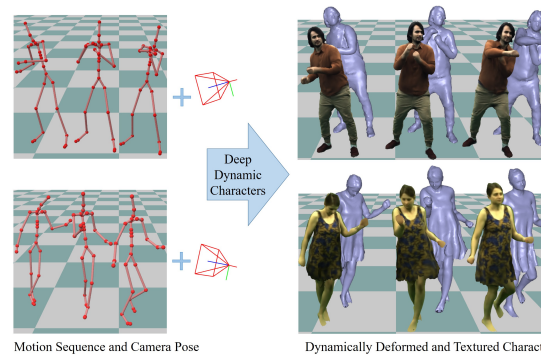
Performance capture



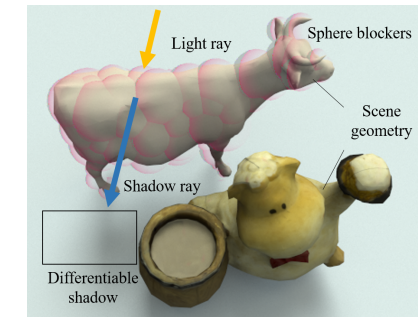
Hand Tracking



Surface Tracking




**Human Synthesis /
Neural rendering**



Differentiable Rendering

Agenda

- About myself
- Introduction of participants 
- How to read a scientific paper
- How to give a good scientific talk
- Questions and answers

Introduction of participants

Who are you?

Research interests?

Previous lectures?

What do you expect from this seminar?

**Seminar Topics are assigned.
Check the seminar webpage**

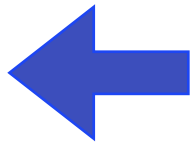
Additional session for the last topic:

21.07.2022 17:00-19:00

or

28.07.2022 14:00-16:00

Agenda

- About myself
- Introduction of participants
- How to read a scientific paper 
- How to give a good scientific talk
- Questions and answers

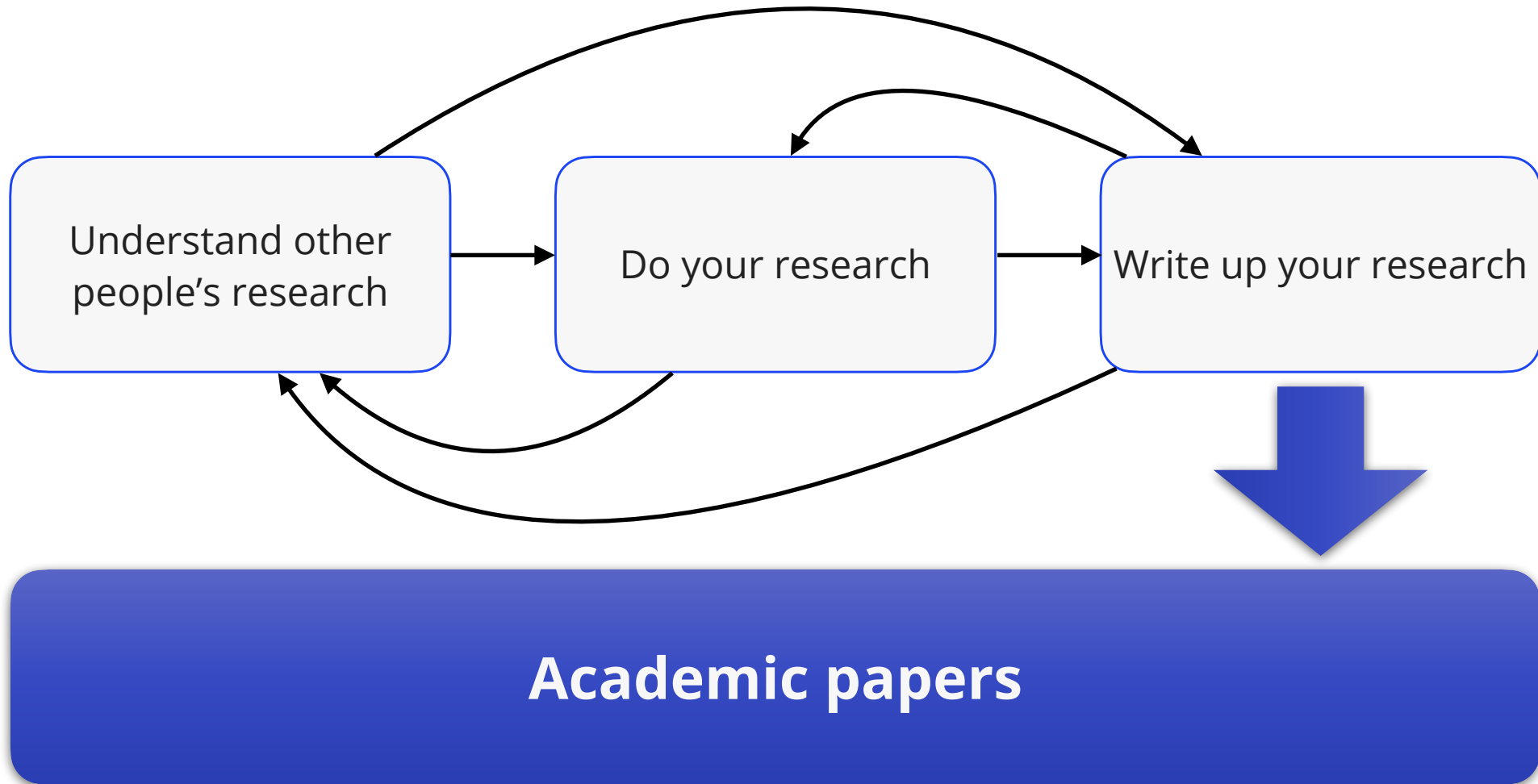
How to read a scientific paper

- The **research process**
- **Why** do we read academic papers?
- **What** is the nature of academic papers?
- **How** to read papers?

How to read a scientific paper

- The **research process**
- **Why** do we read academic papers?
- **What** is the nature of academic papers?
- **How** to read papers?

The Research Process

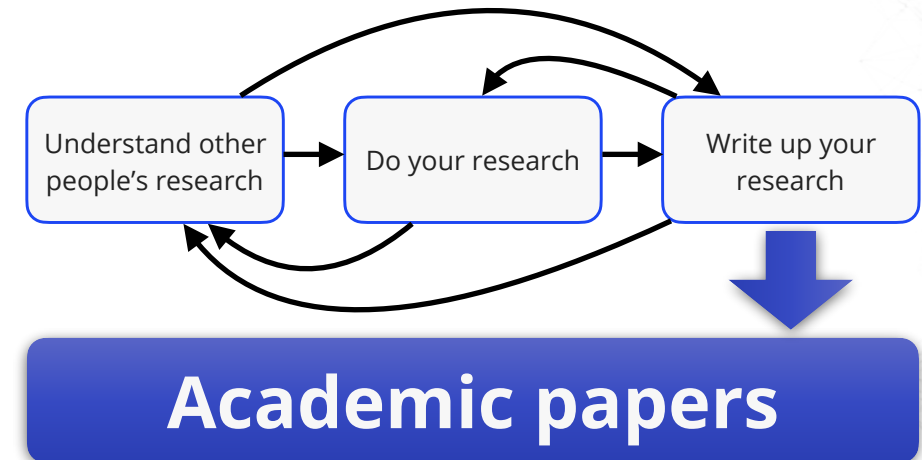


How to read a scientific paper

- The **research process**
- **Why** do we read academic papers?
- **What** is the nature of academic papers?
- **How** to read papers?

Why read papers?

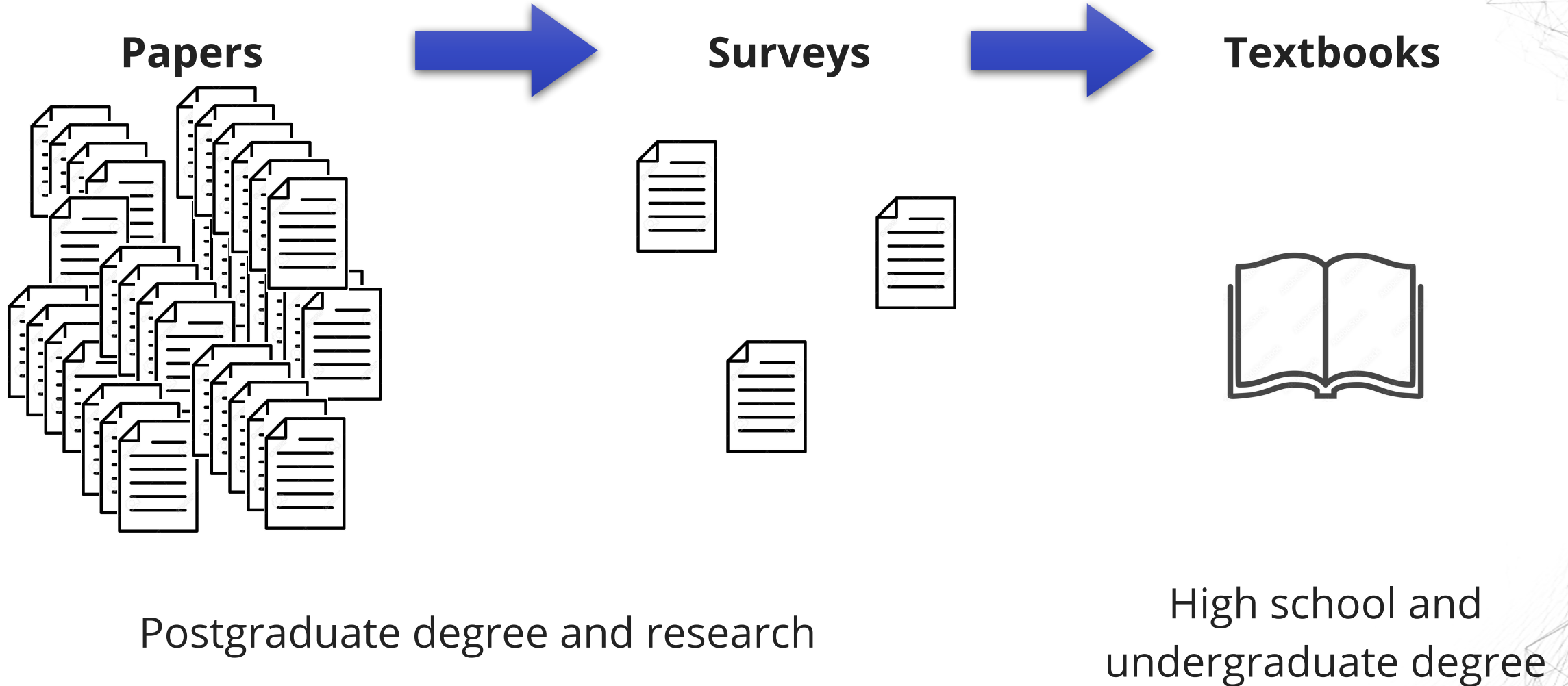
- **Understand other people's research**
 - Understand the context of a research area
 - Keep up-to-date with a field
 - Learn techniques used in a particular research area
- **Do your research**
 - Inspire your ideas
 - Help formulate your own research problems
 - Solve specific problem
- **Write up your research**
 - See good/bad writing and good/bad research
 - Related works/references



How to read a scientific paper

- The **research process**
- **Why** do we read academic papers?
- **What** is the nature of academic papers?
- **How** to read papers?

The Nature of Academic Writing



The Nature of Papers

Good research

Correct

Important

Well written

VS.

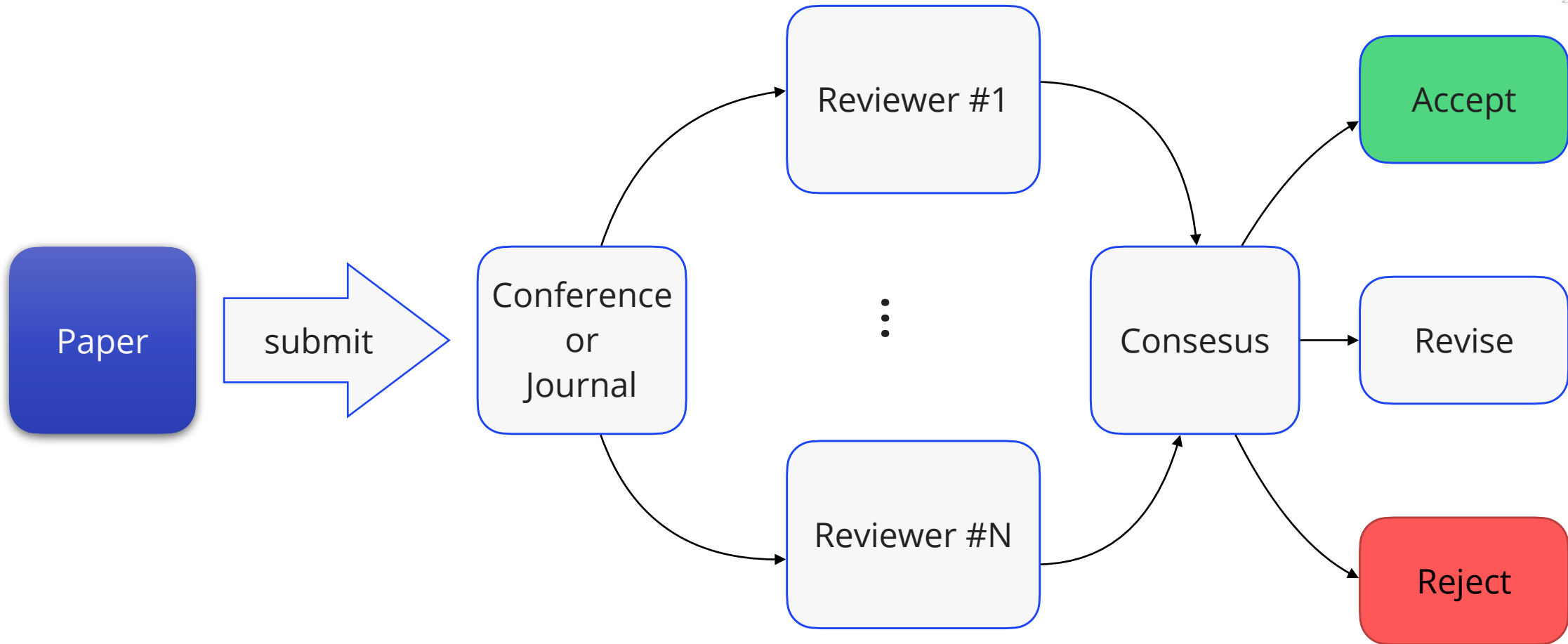
Poor research

Wrong

Unimportant

Incomprehensible

The Peer-review Process



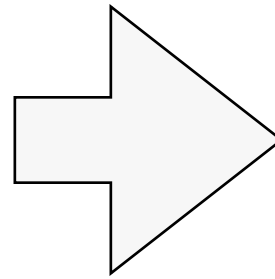
When you read a paper ... be a reviewer

Good research

Correct

Important

Well written



- Identify interesting concepts
- Acknowledge novel ideas

but also ...

- Apply critical judgement
- Ask questions as you read

Questions to ask

- What are the researchers trying to find out?
- Why is the research important?
- What things were measured?
- What were the results?
- What do the authors conclude and why?
- Can I accept the findings as true?



Why publish?

- **Primarily to communicate:**
 - New ideas and theories
 - Solutions to existing and new problems
 - Combinations of existing and new components (systems)
 - Organise works on some topic (surveys, text books)
- **But also:**
 - For (a sense of) achievement
 - To travel to new places and meet new people
 - To further one's academic career
 - Get well known for your work



Publication Venues

- Conference papers
- Journal articles
- Posters
- Workshop papers
- ArXiv
- Technical reports
- Dissertations
- Book chapters
- Text books



Where to find papers

- Google / Google Scholar
 - ArXiv
 - CiteSeerX
 - DBLP
 - CVF website (CVPR, ICCV)
 - Ke-Sen Huang's website
 - Authors' websites
 - Institutional repository
- **Digital libraries:**
 - ACM Digital Library (SIGGRAPH, TOG ...)
 - IEEE Explore (ICCV, CVPR, PAMI...)
 - SpringerLink(ECCV, IJCV...)
 - Wiley Online Library, Elsevier ScienceDirect,
 - **Traditional libraries:**
 - Campus-Bibliothek für Informatik und Mathematik
 - Saarländische Universitäts-und Landesbibliothek (SULB)
 - Deutsche Nationalbibliothek
 - Google Books

How to read a scientific paper

- The **research process**
- **Why** do we read academic papers?
- **What** is the nature of academic papers?
- **How** to read papers?

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

1.

DeepCap: Monocular Human Performance Capture Using Weak Supervision

2.

Marc Habermann^{1,2} Weipeng Xu^{1,2} Michael Zollhoefer³ Gerard Pons-Moll^{1,2} Christian Theobalt^{1,2}

3.

¹Max Planck Institute for Informatics, ²Saarland Informatics Campus, ³Stanford University

4.

Abstract

5.

Human performance capture is a highly important computer vision problem with many applications in movie production and virtual/augmented reality. Many previous performance capture approaches either required expensive multi-view setups or did not recover dense space-time coherent geometry with frame-to-frame correspondences. We propose a novel deep learning approach for monocular dense human performance capture. Our method is trained in a weakly supervised manner based on multi-view supervision completely removing the need for training data with 3D ground truth annotations. The network architecture is based on two separate networks that disentangle the task into a pose estimation and a non-rigid surface deformation step. Extensive qualitative and quantitative evaluations show that our approach outperforms the state of the art in terms of quality and robustness.

6.

1. Introduction

Human performance capture, i.e. the space-time coherent 4D capture of full pose and non-rigid surface deformation of people in general clothing, revolutionized the film and gaming industry in recent years. Apart from visual effects, it has many use cases in generating personalized dynamic virtual avatars for telepresence, virtual try-on, mixed reality, and many other areas. In particular for the latter applications, being able to performance capture humans from *monocular video* would be a game changer. The capability of established monocular methods only captures limited motion (including hands or sparse facial expressions at most). However, the monocular tracking of dense full-body deformations of skin and clothing, in addition to articulated pose, which play an important role in producing realistic virtual characters, is still at its infancy.

In literature, multi-view marker-less methods [13, 14, 15, 17, 24, 29, 50, 55, 81, 82, 86, 64, 65] have shown compelling results. However, these approaches rely on well-controlled multi-camera studios (typically with green screen), which prohibits them from being used for location shootings of films and telepresence in living spaces.

Recent monocular human modeling approaches have shown compelling reconstructions of humans, including clothing, hair and facial details [70, 99, 2, 3, 9, 60, 52].

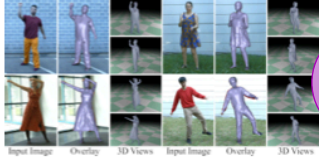


Figure 1. We present the first learning-based approach for dense monocular human performance capture using weak multi-view supervision that not only predicts the pose but also the space-time coherent non-rigid deformations of the model surface.

Some directly regress voxels [28, 99] or the continuous occupancy of the surface [70]. Since predictions are pixel aligned, reconstructions have nice detail, but limbs are often missing, especially for difficult poses. Moreover, the recovered motion is not factorized into articulation and non-rigid deformation, which prevents the computer-graphics style control over the reconstructions that is required in many of the aforementioned applications. Importantly, surface vertices are not tracked over time, so no space-time coherent model is captured. Another line of work predicts deformations or displacements to an articulated template, which prevents missing limbs and allows more control [2, 9, 5, 67]. However, these works do not capture motion and the surface deformations.

The state-of-the-art monocular human performance capture methods [89, 32] densely track the deformation of the input view. They leverage deep learning-based sparse keypoint detections and perform an expensive template fitting afterwards. In consequence, they can only non-rigidly fit to the input view and suffer from instability. By contrast, we present the first learning-based method that jointly infers the articulated and non-rigid 3D deformation parameters in a single feed-forward pass at much higher performance, accuracy and robustness. The core of our method is a CNN model which integrates a fully differentiable *mesh* template parameterized with *pose* and an *embedded deformation graph*. From a single image, our network predicts the skeletal pose, and the rotation and translation parameters for each node in the deformation graph. In stark contrast to implicit representations [70, 99, 22], our mesh-based

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

method tracks the surface vertices over time, which is crucial for adding semantics, and for texturing and rendering in graphics. Further, by virtue of our parameterization, our model always produces a human surface *without missing limbs*, even during occlusions and out-of-plane motions.

While previous methods [70, 99, 2, 9] rely on 3D ground truth for training, our method is weakly supervised from multi-view images. To this end, we propose a fully differentiable architecture which is trained in an analysis-by-synthesis fashion, without explicitly using any 3D ground truth annotation. Specifically, during training, our method only requires a personalized template mesh of the actor and a multi-view video sequence of the actor performing various motions. Then, our network learns to predict 3D pose and dense non-rigidly deformed surface shape by comparing its single image feed-forward predictions in a differentiable manner against the multi-view 2D observations. At test time, our method only requires a single-view image as input and produces a deformed template matching the actor's non-rigid motion in the image. In summary, the main technical contributions of our work are:

- A learning-based 3D human performance capture approach that jointly tracks the skeletal pose and the non-rigid surface deformations from monocular images.
- A new differentiable representation of deforming human surfaces which enables training from multi-view video footage directly.

Our new model achieves high quality dense human performance capture results on our new challenging dataset, demonstrating, qualitatively and quantitatively, the advantages of our approach over previous work. We experimentally show that our method produces reconstructions of higher accuracy and 3D stability, in particular in depth, than related work, also under difficult poses.

2. Related Work

In the following, we focus on related work in the field of dense 3D human performance capture and do not review work on sparse 2D pose estimation.

Capture using Parametric Models. Monocular human performance capture is an ill-posed problem due to its high dimensionality and ambiguity. Low-dimensional parametric models can be employed as shape and deformation priors. First, model-based approaches leverage a set of simple geometric primitives [63, 74, 71, 54]. Recent methods employ detailed statistical models learned from thousands of high-quality 3D scans [6, 33, 59, 65, 51, 41, 45, 85, 35, 97, 10]. Deep learning is widely used to obtain 2D and/or 3D joint detections or 3D vertex positions that can be used to inform model fitting [37, 48, 53, 11, 46]. An alternative is to regress model parameters directly [42, 62, 43]. Beyond

body shape and pose, recent models also include facial expressions and hand motion [61, 88, 40, 69] leading to very expressive reconstruction results. Since parametric body models do not represent garments, variation in clothing cannot be reconstructed, and therefore many methods recover the naked body shape under clothing [8, 7, 95, 90]. The full geometry of the actor can be reconstructed by non-rigidly deforming the base parametric model to better fit the observations [68, 3, 4]. But they can only model tight clothes such as T-shirts and pants, but not loose apparel which has a different topology than the body model, such as skirts. To overcome this problem, ClothCap [64] captures the body and clothing separately, but requires active multi-view setups. Physics based simulations have recently been leveraged to constrain tracking (SimulCap [78]), or to learn a model of clothing on top of SMPL (TailorNet [60]). Instead, our method is based on person-specific templates including clothes and employs deep learning to predict clothing deformation based on monocular video directly.

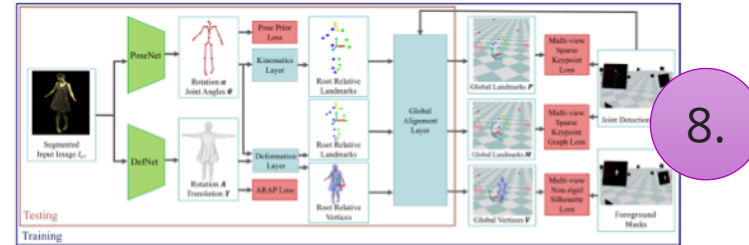
Depth-based Template-free Capture. Most approaches based on parametric models ignore clothing. The other side of the spectrum are prior-free approaches based on one or multiple depth sensors. Capturing general non-rigidly deforming scenes [73, 31], even at real-time frame rates [57, 39, 31], is feasible, but only works reliably for small, controlled, and slow motions. Higher robustness can be achieved by using higher frame rate sensors [30, 47] or multi-view setups [91, 27, 58, 26, 96]. Techniques that are specifically tailored to humans increase robustness [93, 94, 92] by integrating a skeletal motion prior [93] or a parametric model [94, 84]. HybridFusion [98] additionally incorporates a sparse set of inertial measurement units. These fusion-style volumetric capture techniques [36, 1, 49, 23, 66] achieve impressive results, but do not establish a set of dense correspondences between all frames. In addition, such depth-based methods do not directly generalize to our monocular setting, have a high power consumption, and typically do not work well under sunlight.

Monocular Template-free Capture. Quite recently, fueled by the progress in deep learning, many template-free monocular reconstruction approaches have been proposed. Their regular structure, many implicit reconstruction methods [80, 99] make use of uniform voxel grids. DeepFusion [99] combines a coarse scale volumetric reconstruction with a refinement network to add high-frequency details. Multi-view CNNs can map 2D images to 3D volumetric fields enabling reconstruction of a clothed human body at arbitrary resolution [38]. SiCloPe [56] reconstructs a complete textured 3D model, including cloth, from a single image. PIPu [70] regresses an implicit surface representation that locally aligns pixels with the global context of the corresponding 3D object. Unlike voxel-based representations, this implicit per-pixel representation is more memory

7.

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material



8.

Figure 2. Overview of our approach. Our method takes a single segmented image as input. First, our pose network, *PoseNet*, is trained to predict the joint angles and the camera relative rotation using sparse multi-view 2D joint detections as weak supervision. Second, the deformation network, *DefNet*, is trained to regress embedded graph rotation and translation parameters to account for non-rigid deformations. To train *DefNet*, multi-view 2D joint detections and silhouettes are used for supervision.

need to transform $\mathbf{P}_{c'}^T$ to the world coordinate system:

$$\mathbf{P}_m = \mathbf{R}_{c'}^T \mathbf{P}_{c',m} + \mathbf{t}, \quad (1)$$

where $\mathbf{R}_{c'}$ is the rotation matrix of the input camera c' and \mathbf{t} is the global translation of the skeleton.

Global Alignment Layer. To obtain the global translation \mathbf{t} , we propose a global alignment layer that is attached to the kinematics layer. It localizes our skeleton model in the world space, such that the globally rotated landmarks $\mathbf{R}_{c'}^T \mathbf{P}_{c',m}$ project onto the corresponding detections in all camera views. This is done by minimizing the distance between the rotated landmarks $\mathbf{R}_{c'}^T \mathbf{P}_{c',m}$ and the corresponding rays cast from the camera origin \mathbf{o}_c to the 2D joint detections:

$$\sum_c \sum_m \sigma_{c,m} \|(\mathbf{R}_{c'}^T \mathbf{P}_{c',m} + \mathbf{t} - \mathbf{o}_c) \times \mathbf{d}_{c,m}\|^2, \quad (2)$$

where $\mathbf{d}_{c,m}$ is the direction of a ray from camera c to the 2D joint detection $\mathbf{p}_{c,m}$ corresponding to landmark m :

$$\mathbf{d}_{c,m} = \frac{(\mathbf{E}_c^{-1} \hat{\mathbf{p}}_{c,m})_{xyz} - \mathbf{o}_c}{\|(\mathbf{E}_c^{-1} \hat{\mathbf{p}}_{c,m})_{xyz} - \mathbf{o}_c\|}. \quad (3)$$

Here, $\mathbf{E}_c \in \mathbb{R}^{4 \times 4}$ is the projection matrix of camera c and $\hat{\mathbf{p}}_{c,m} = (p_{c,m}, 1, 1)^T$. Each point-to-line distance is weighted by the joint detection confidence $\sigma_{c,m}$, which is set to zero if below 0.4. The minimization problem of Eq. 2 can be solved in closed form:

$$\mathbf{t} = \mathbf{W}^{-1} \sum_{c,m} \mathbf{D}_{c,m} (\mathbf{R}_{c'}^T \mathbf{P}_{c',m} - \mathbf{o}_c) + \mathbf{o}_c - \mathbf{R}_{c'}^T \mathbf{P}_{c',m}, \quad (4)$$

where

$$\mathbf{W} = \sum_{c,m} \mathbf{I} - \mathbf{D}_{c,m}. \quad (5)$$

Here, \mathbf{I} is the 3×3 identity matrix and $\mathbf{D}_{c,m} = \mathbf{d}_{c,m} \mathbf{d}_{c,m}^T$. Note that the operation in Eq. 4 is differentiable with respect to the landmark position $\mathbf{P}_{c'}$.

Sparse Keypoint Loss. Our 2D sparse keypoint loss for the *PoseNet* can be expressed as

$$\mathcal{L}_{kp}(\mathbf{P}) = \sum_c \sum_m \lambda_m \sigma_{c,m} \|\pi_c(\mathbf{P}_m) - \mathbf{p}_{c,m}\|^2, \quad (6)$$

which ensures that each landmark projects onto the corresponding 2D joint detections $\mathbf{p}_{c,m}$ in all camera views. Here, π_c is the projection function of camera c and $\sigma_{c,m}$ is the same as in Eq. 2. λ_m is a kinematic chain-based hierarchical weight which varies during training for better convergence (see the supplementary material for details).

Pose Prior Loss. To avoid unnatural poses, we impose a pose prior loss on the joint angles

$$\mathcal{L}_{limit}(\boldsymbol{\theta}) = \sum_{i=1}^{27} \Psi(\theta_i) \quad (7)$$

$$\Psi(x) = \begin{cases} (x - \theta_{max,i})^2, & \text{if } x > \theta_{max,i} \\ (\theta_{min,i} - x)^2, & \text{if } x < \theta_{min,i} \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

that encourages that each joint angle θ_i stays in a range $[\theta_{min,i}, \theta_{max,i}]$ depending on the anatomic constraints.

3.3. Deformation Network

With the skeletal pose from *PoseNet* alone, the non-rigid deformation of the skin and clothes cannot be fully explained. Therefore, we disentangle the non-rigid deformation and the articulated skeletal motion. *DefNet* takes the

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

efficient. These approaches have not been demonstrated to generalize well to strong articulation. Furthermore, implicit approaches do not recover frame-to-frame correspondences which are of paramount importance for downstream applications, e.g., in augmented reality and video editing. In contrast, our method is based on a mesh representation and can explicitly obtain the per-vertex correspondences over time while being slightly less general.

Template-based Capture. An interesting trade-off between being template-free and relying on parametric models are approaches that only employ a template mesh as prior. Historically, template-based human performance capture techniques exploit multi-view geometry to track the motion of a person [76]. Some systems also jointly reconstruct and obtain a foreground segmentation [13, 15, 50, 87]. Given a sufficient number of multi-view images as input, some approaches [21, 17, 24] align a personalized template model to the observations using non-rigid registration. All the aforementioned methods require expensive multi-view setups and are not practical for consumer use. Depth-based techniques enable template tracking from less cameras [100, 91] and reduced motion models [86, 29, 81, 50] increase tracking robustness. Recently, capturing 3D dense human body deformation just with a single RGB camera has been enabled [89] and real-time performance has been achieved [32]. However, their methods rely on expensive optimization leading either to very long per-frame computation times [89] or the need for two graphics cards [32]. Similar to them, our approach also employs a person-specific template mesh. But differently, our method directly learns to predict the skeletal pose and the non-rigid surface deformations. As shown by our experimental results, benefiting from our multi-view based self-supervision, our reconstruction accuracy significantly outperforms the existing methods.

3. Method

Given a single RGB video of a moving human in general clothing, our goal is to capture the dense deforming surface of the full body. This is achieved by training a neural network consisting of two components: As illustrated in Fig. 1, our pose network, *PoseNet*, estimates the skeletal pose of the actor in the form of joint angles from a monocular video (Sec. 3.2). Next, our deformation network, *DeformNet*, regresses the non-rigid deformation of the dense surface, which cannot be modeled by the skeletal motion, in the embedded deformation graph representation (Sec. 3.3). To avoid generating dense 3D ground truth annotation, our network is trained in a weakly supervised manner. To this end, we propose a fully differentiable human deformation and rendering model, which allows us to compare the rendering of the human body model to the 2D image evidence and back-propagate the losses. For training, we first capture a video sequence in a calibrated multi-camera green screen

studio (Sec. 3.1). Note that our multi-view video is only used during training. At test time we only require a single RGB video to perform dense non-rigid tracking.

3.1. Template and Data Acquisition

Character Model. Our method relies on a person-specific 3D template model. We first scan the actor with a 3D scanner [79] to obtain the textured mesh. Then, it is automatically rigged to a kinematic skeleton, which is parameterized with joint angles $\theta \in \mathbb{R}^{27}$, the camera relative rotation $\alpha \in \mathbb{R}^3$ and translation $t \in \mathbb{R}^3$. To model the non-rigid surface deformation, we automatically build an embedded deformation graph \mathcal{G} with K nodes following [77]. The nodes are parameterized with Euler angles $\mathbf{A} \in \mathbb{R}^{K \times 3}$ and translations $\mathbf{T} \in \mathbb{R}^{K \times 3}$. Similar to [32], we segment the mesh into different non-rigidity classes resulting in per-vertex rigidity weights s_i . This allows us to model varying deformation behaviors of different surface materials, e.g. skin deforms less than clothing (see Eq. 13).

Training Data. To acquire the training data, we record a multi-view video of the actor doing various actions in a calibrated multi-camera studio with green screen. To provide weak supervision for the training, we first perform 2D pose detection on the sequences using OpenPose [19, 18, 72, 83] and apply temporal filtering. Then, we generate the foreground mask using color keying and compute the corresponding distance transform image $D_{f,c}$ [12], where $f \in [0, F]$ and $c \in [0, C]$ denote the frame index and camera index, respectively. During training, we randomly sample one camera view c' and frame f' for which we crop the recorded image with a bounding box, based on the 2D joint detections. The final training input image $I_{f',c'} \in \mathbb{R}^{256 \times 256 \times 3}$ is obtained by removing the background and augmenting the foreground with random brightness, hue, contrast and saturation changes. For simplicity, we describe the operation on frame f' and omit the subscript f' in following equations.

3.2. Pose Network

Our *PoseNet*, we use ResNet50 [34] pretrained on ImageNet [25] as backbone and modify the last fully connected layer to output a vector containing the joint angles θ and the camera relative root rotation α , given the input image $I_{f,c}$. Since generating the ground truth for θ and α is a non-trivial task, we propose weakly supervised training based on fitting the skeleton to multi-view 2D joint detections.

Kinematics Layer. To this end, we introduce a kinematics layer as the differentiable function that takes the joint angles θ and the camera relative rotation α and computes the positions $\mathbf{P}_{c'} \in \mathbb{R}^{M \times 3}$ of the M 3D landmarks attached to the skeleton (17 body joints and 4 face landmarks). Note that $\mathbf{P}_{c'}$ lives in a camera-root-relative coordinate system. In order to project the landmarks to other camera views, we

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material



Figure 3. Qualitative results. Each row shows results for a different person with varying types of apparel. We visualize input frames and our reconstruction overlaid to the corresponding frame. Note that our results precisely overlay to the input. Further, we show our reconstructions from a virtual 3D viewpoint. Note that they also look plausible in 3D.

input during testing to crop the frames and to obtain the 3D global translation with our global alignment layer. Finally, we temporally smooth the output mesh vertices with a Gaussian kernel of size 5 frames.

Dataset. We evaluate our approach on 4 subjects ($S1$ to $S4$) with varying types of apparel. For qualitative evaluation, we recorded 13 in-the-wild sequences in different indoor and outdoor environments shown in Fig. 3. For quantitative evaluation, we captured 4 sequences in a calibrated multi-camera green screen studio (see Fig. 4), for which we computed the ground truth 3D joint locations using the multi-view motion capture software, The Capture [20], and we use a color keying algorithm for ground truth foreground segmentation. All sequences contain a large variety of motions, ranging from simple ones like walking up to more difficult ones like fast dancing or baseball pitching. We will release the dataset for future research.

Qualitative Comparisons. Fig. 3 shows our qualitative

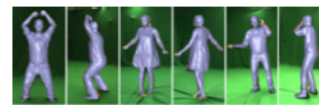


Figure 4. Results on our evaluation sequences where input views (IV) and reference views (RV) are available. Note that our reconstruction also precisely overlays on RV even though they are not used for tracking.

results on in-the-wild test sequences with various clothing styles, poses and environments. Our reconstructions not only precisely overlay with the input images, but also look plausible from arbitrary 3D view points. In Fig. 5, we qualitatively compare our approach to the related human capture and reconstruction methods [42, 32, 70, 99]. In terms of the shape representation, our method is most

AMVtoU, RVtoU, and SVtoU (in %) on $S1$ sequence			
Method	AMVtoU \uparrow	RVtoU \uparrow	SVtoU \uparrow
IDMR [42]	62.25	61.7	68.85
IDMR [43]	65.98	65.58	70.77
LiveCap [32]	56.02	54.21	77.75
DeepHuman [99]			94.87
Ours	87.2	87.03	89.26
MVBL	91.74	91.72	92.02

AMVtoU, RVtoU, and SVtoU (in %) on $S4$ sequence			
Method	AMVtoU \uparrow	RVtoU \uparrow	SVtoU \uparrow
IDMR [42]	45.1	44.66	70.84
IDMR [43]	63.79	63.29	70.23
LiveCap [32]	59.96	59.02	72.16
DeepHuman [99]			84.15
Ours	82.53	82.22	86.66
MVBL	88.14	88.03	89.66

Table 2. Surface deformation accuracy. We outperform all other monocular methods and are even close to the multi-view baseline.

shape on $S1$ and $S4$ for every 100th frame. We evaluate the IoU on all views, on all views expect the input view, and on the input view which we refer to as AMVtoU, RVtoU and SVtoU, respectively. To factor out the errors in global localization, we apply the ground truth translation to the reconstructed geometries. For DeepHuman [99] and PiFu [70], we cannot report the AMVtoU and RVtoU, since we cannot overlay their results on reference views as discussed before. Further, PiFu [70] by design achieves perfect overlay on the input view, since they regress the depth for each foreground pixel. However, their reconstruction does not reflect the true 3D geometry (see Fig. 5). Therefore, it is meaningless to report their SVtoU. Similarly, DeepHuman [99] achieves high SVtoU, due to their volumetric representation. But their results are often wrong, when looking from side views. In contrast, our method consistently outperforms all other approaches in terms of AMVtoU and RVtoU, which shows the high accuracy of our method in recovering the 3D geometry. Further, we are again close to the multi-view baseline.

Ablation Study. To evaluate the importance of the number of cameras, the number of training images, and our *DefNet*, we performed an ablation study on $S4$ in Tab. 3. 1) In the first group of Tab. 3, we train our networks with supervision using 1 to 7 views. We can see that adding more views consistently improves the quality of the estimated poses and deformations. The most significant improvement is from one to two cameras. This is not surprising, since the single camera settings is inherently ambiguous. 2) In the second group of Tab. 3, we reduce the training data to 1/2 and 1/4. We can see that the more frames with different poses and deformations are seen during training, the better the reconstruction quality is. This is expected since a larger number of frames may better sample the possible space of poses and deformations. 3) In the third group of Tab. 3, we evaluate the AMVtoU on the template mesh animated with the results of *PoseNet*, which we refer to as *PoseNet-only*. One can see that on average, the AMVtoU is improved by around 4%. Since most non-rigid deformations rather happen locally,

3DPCK and AMVtoU (in %) on $S4$ sequence		
Method	3DPCK \uparrow	AMVtoU \uparrow
1 camera view	62.11	65.11
2 camera views	93.52	78.64
3 camera views	94.70	79.75
7 camera views	95.95	81.73
8500 Frames	85.19	73.41
13000 frames	92.35	78.97
PoseNet-only	96.74	78.51
Ours(14 views, 26000 frames)	96.74	82.53

Table 3. Ablation study. We evaluate the number of cameras and the number of frames used during training in terms of the 3DPCK and AMVtoU metrics. Adding more cameras and frames consistently improves the quality of reconstruction. Further, *DefNet* improves the AMVtoU compared to pure pose estimation.



Figure 6. *PoseNet* + *DefNet* vs. *PoseNet-only*. *DefNet* can deform the template to accurately match the input, especially for loose clothing. In addition, *DefNet* also corrects slight errors in the pose and typical skinning artifacts.

the difference is visually even more significant as shown in Fig. 6. Especially, the skirt is correctly deformed according to the input image whereas the *PoseNet-only* result cannot fit the input due to the limitation of skinning.

5. Conclusion

We have presented a learning-based approach for monocular dense human performance capture using only weak multi-view supervision. In contrast to existing methods, our approach directly regresses poses and surface deformations from neural networks, produces temporal surface correspondences, preserves the skeletal structure of the human body, and can handle loose clothes. Our qualitative and quantitative results in different scenarios show that our method produces more accurate 3D reconstruction of pose and non-rigid deformation than existing methods. In the future, we plan to incorporate hands and the face to our mesh representation to enable joint tracking of body, facial expressions and hand gestures. We are also interested in physically more correct multi-layered representations to model the garments even more realistically. **Acknowledgements.** This work was funded by the ERC Consolidator Grant 4DRePLY (770784) and the Deutsche Forschungsgemeinschaft (Project Nr. 409792180, Emmy Noether Programme, project: Real Virtual Humans).

10.

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

AMViU, RViU, and SViU (in %) on S1 sequence			
Method	AMViU \uparrow	RViU \uparrow	SViU \uparrow
HMR [43]	62.25	61.7	68.85
HMMR [45]	65.98	65.58	70.77
LiveCap [52]	56.02	54.21	77.75
DeepHuman [99]	-	-	91.87
Ours	87.2	87.03	89.26
MVBL	91.74	91.72	92.02

AMViU, RViU, and SViU (in %) on S4 sequence			
Method	AMViU \uparrow	RViU \uparrow	SViU \uparrow
HMR [43]	65.1	64.66	70.84
HMMR [45]	63.79	63.29	70.23
LiveCap [52]	59.96	59.02	72.16
DeepHuman [99]	-	-	84.15
Ours	82.53	82.22	86.66
MVBL	88.14	88.03	89.66

Table 2. Surface deformation accuracy. We outperform all other monocular methods and are even close to the multi-view baseline.

shape on $S1$ and $S4$ for every 100th frame. We evaluate the IoU on all views, on all views except the input view, and on the input view which we refer to as $AMViU$, $RViU$ and $SViU$, respectively. To factor out the errors in global localization, we apply the ground truth translation to the reconstructed geometries. For DeepHuman [99] and PiFu [70], we cannot report the $AMViU$ and $RViU$, since we cannot overlay their results on reference views as discussed before. Further, PiFu [70] by design achieves perfect overlay on the input view, since they regress the depth for each foreground pixel. However, their reconstruction does not reflect the true 3D geometry (see Fig. 5). Therefore, it is meaningless to report their $SViU$. Similarly, DeepHuman [99] achieves high $SViU$, due to their volumetric representation. But their results are often wrong, when looking from side views. In contrast, our method consistently outperforms all other approaches in terms of $AMViU$ and $RViU$, which shows the high accuracy of our method in recovering the 3D geometry. Further, we are again close to the multi-view baseline.

Ablation Study. To evaluate the importance of the number of cameras, the number of training images, and our *DefNet*, we performed an ablation study on $S4$ in Tab. 3. 1) In the first group of Tab. 3, we train our networks with supervision using 1 to 7 views. We can see that adding more views consistently improves the quality of the estimated poses and deformations. The most significant improvement is from one to two cameras. This is not surprising, since the single camera settings is inherently ambiguous. 2) In the second group of Tab. 3, we reduce the training data to 1/2 and 1/4. We can see that the more frames with different poses and deformations are seen during training, the better the reconstruction quality is. This is expected since a larger number of frames may better sample the possible space of poses and deformations. 3) In the third group of Tab. 3, we evaluate the $AMViU$ on the template mesh animated with the results of *PoseNet*, which we refer to as *PoseNet-only*. One can see that on average, the $AMViU$ is improved by around 4%. Since most non-rigid deformations rather happen locally,

3DPCK and AMViU (in %) on S4 sequence		
Method	3DPCK \uparrow	AMViU \uparrow
1 camera view	62.11	65.11
2 camera views	93.52	78.44
3 camera views	94.70	79.75
7 camera views	95.85	81.73
6500 frames	85.19	73.41
13000 frames	92.25	78.97
PoseNet-only	96.74	78.31
Ours(14 views, 26000 frames)	96.74	82.53

Table 3. Ablation study. We evaluate the number of cameras and the number of frames used during training in terms of the $3DPCK$ and $AMViU$ metrics. Adding more cameras and frames consistently improves the quality of reconstruction. Further, *DefNet* improves the $AMViU$ compared to pure pose estimation.



Figure 6. *PoseNet + DefNet* vs. *PoseNet-only*. *DefNet* can deform the template to accurately match the input, especially for loose clothing. In addition, *DefNet* also corrects slight errors in the pose and typical skinning artifacts.

the difference is visually even more significant as shown in Fig. 6. Especially, the skirt is correctly deformed according to the input image whereas the *PoseNet-only* result cannot fit the input due to the limited flexibility.

5. Conclusion

We have presented a learning-based approach for monocular dense human performance capture under weak multi-view supervision. In contrast to existing methods, our approach directly regresses poses and surface deformations from neural networks, produces temporal surface correspondences, preserves the skeletal structure of the human body, and can handle loose clothes. Our qualitative and quantitative results in different scenarios show that our method produces more accurate 3D reconstruction of pose and non-rigid deformation than existing methods. In the future, we plan to incorporate hands and the face to our mesh representation to enable joint tracking of body, facial expressions and hand gestures. We are also interested in physically more correct multi-layered representations to model the garments even more realistically.

Acknowledgements. This work was funded by the ERC Consolidator Grant 4DReLy (770784) and the Deutsche Forschungsgemeinschaft (Project Nr. 409792180, Emmy Noether Programme, project: Real Virtual Humans).

Parts of a Paper

1. Title
2. Author list
3. Affiliations
4. Teaser
5. Abstract
6. Introduction
7. Related work
8. Overview
9. Method
10. Results
11. Discussion
12. Conclusion
13. References
14. Appendices
15. Supplemental material

References

- [1] B. Allain, J.-S. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In *CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, June 2015. IEEE. 2
- [2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1175–1186, Jun 2019. 1, 2
- [3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR Spotlight Paper*. 1, 2
- [4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, Sep 2018. 2
- [5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1
- [6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 2
- [7] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. 2
- [8] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 2
- [9] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2
- [10] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *International Conference on Computer Vision (ICCV)*, pages 2300–2308, Dec. 2015. 2
- [11] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 2
- [12] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986. 3, 5
- [13] M. Bray, P. Kohli, and P. H. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *European conference on computer vision*, pages 642–655. Springer, 2006. 1, 3
- [14] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-d pose tracking: exploiting contour and flow based constraints. In *European Conference on Computer Vision*, pages 98–111. Springer, 2006. 1
- [15] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region and motion-based 3d tracking of rigid and articulated objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):402–415, 2010. 1, 3
- [16] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5
- [17] C. Cagniat, E. Boyer, and S. Ilic. Free-form mesh tracking: a patch-based approach. In *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on, pages 1339–1346. IEEE, 2010. 1, 3
- [18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 3, 5
- [19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3, 5
- [20] The Capture. <http://www.thecapture.com/>. 6, 7
- [21] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3), July 2003. 3
- [22] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1
- [23] M. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 2
- [24] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM, 2008. 1, 3
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3
- [26] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6):246:1–246:16, Nov. 2017. 2
- [27] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusions4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016. 2
- [28] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2019. 1
- [29] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pat-*

13.

How to read a paper (by S. Keshav)

- Suggested approach for efficient reading
- Make up to three passes over the paper:
 - **1. Quick pass:**
 - Get general idea about the paper
 - **2. Content pass:**
 - Grasp paper contents, but skip details
 - **3. Details pass:**
 - Understand the paper in depth

How to Read a Paper

David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

S. Keshav

ABSTRACT

Researchers spend a great deal of time reading research papers. However, this skill is rarely taught, leading to much wasted effort. This article outlines a practical and efficient three-pass method for reading research papers. I also describe how to use this method to do a literature survey.

Categories and Subject Descriptors:

A.1 [Introductory and Survey]

General Terms: Documentation.
Keywords: Paper, Reading, Hints.

1. INTRODUCTION

Researchers must read papers for several reasons: to review them for a conference or a class, to keep current in their field, or for a literature survey of a new field. A typical researcher will likely spend hundreds of hours every year reading papers.

Learning to efficiently read a paper is a critical but rarely taught skill. Beginning graduate students, therefore, must learn on their own using trial and error. Students waste much effort in the process and are frequently driven to frustration.

For many years I have used a simple approach to efficiently read papers. This paper describes the 'three-pass' approach and its use in doing a literature survey.

2. THE THREE-PASS APPROACH

The key idea is that you should read the paper in up to three passes, instead of starting at the beginning and plowing your way to the end. Each pass accomplishes specific goals and builds upon the previous pass. The first pass gives you a general idea about the paper. The second pass lets you grasp the paper's content, but not its details. The third pass helps you understand the paper in depth.

2.1 The first pass

The first pass is a quick scan to get a bird's-eye view of the paper. You can also decide whether you need to do any more passes. This pass should take about five to ten minutes and consists of the following steps:

1. Carefully read the title, abstract, and introduction
2. Read the section and sub-section headings, but ignore everything else
3. Read the conclusions

4. Glance over the references, mentally ticking off the ones you've already read

At the end of the first pass, you should be able to answer the five Cs:

1. *Category:* What type of paper is this? A measurement paper? An analysis of an existing system? A description of a research prototype?
2. *Context:* Which other papers is it related to? Which theoretical bases were used to analyze the problem?
3. *Correctness:* Do the assumptions appear to be valid?
4. *Contributions:* What are the paper's main contributions?
5. *Clarity:* Is the paper well written?

Using this information, you may choose not to read further. This could be because the paper doesn't interest you, or you don't know enough about the area to understand the paper, or that the authors make invalid assumptions. The first pass is adequate for papers that aren't in your research area, but may someday prove relevant. Incidentally, when you write a paper, you can expect most reviewers (and readers) to make only one pass over it. Take care to choose coherent section and sub-section titles and to write concise and comprehensive abstracts. If a reviewer cannot understand the gist after one pass, the paper will likely be rejected; if a reader cannot understand the highlights of the paper after five minutes, the paper will never be read.

2.2 The second pass

In the second pass, read the paper with greater care, but ignore details such as proofs. It helps to jot down the key points, or to make comments in the margins, as you read.

1. Look carefully at the figures, diagrams and other illustrations in the paper. Pay special attention to graphs. Are the axes properly labeled? Are results shown with error bars, so that conclusions are statistically significant? Common mistakes like these will separate rushed, shoddy work from the truly excellent.

2. Remember to mark relevant unread references for further reading (this is a good way to learn more about the background of the paper).

How to read a paper - Pass 1

- **Quick scan** to get a bird's-eye view of the paper
- Decide whether you need to do any more passes
- Should take about **5–10 minutes**:
 - Carefully read title, abstract and introduction
 - Read headings, but ignore everything else
 - Look at the maths (if any)
 - Read conclusion
 - Glance over the references
- **Tip:** Read the figures (teaser, method overview, results, tables..)

How to read a paper - Pass 2

- **Read the paper with greater care, but ignore details (1h)**
 - It helps to make notes in the margins as you read
 - Look carefully at figures, diagrams and other illustrations
- Appropriate for an interesting paper **outside your research speciality**
- If you still **don't understand** a paper, you can **choose to:**
 - Set the paper aside
 - Return to the paper later
 - Go on to the third pass

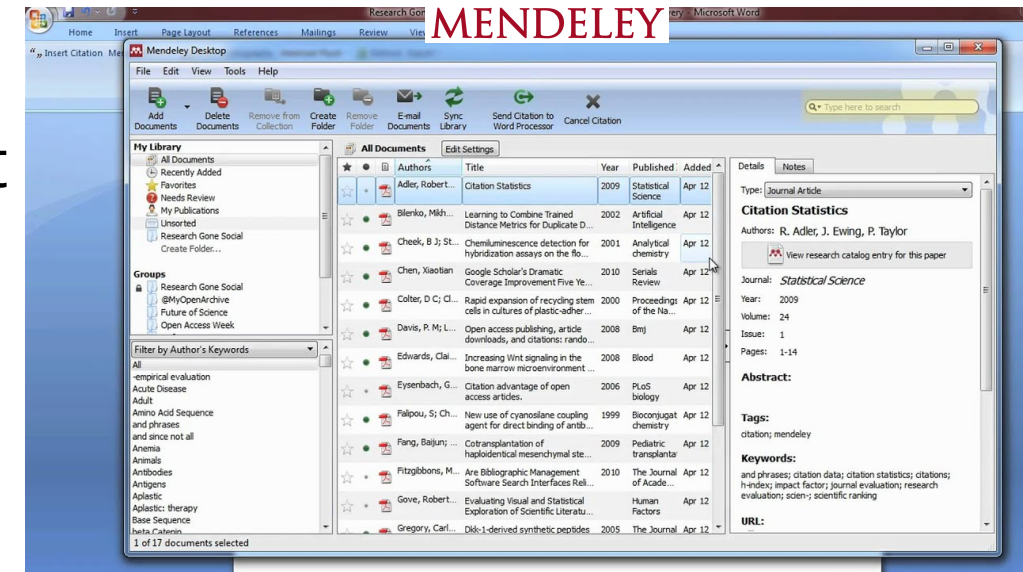
How to read a paper - Pass 3

- The key is to attempt to **virtually re-implement** the paper:
 - Make the same assumptions as the authors, re-create the work.
 - Compare your re-creation with the actual paper
- This pass requires **great attention** to detail
- **Identify and challenge** every assumption
- **Identify strong and weak points:**
 - Implicit assumptions
 - Missing citations to relevant work
 - Potential issues with experimental or analytical techniques



Remember what you read

- Organise papers to keep track of them:
 - Mendeley: Free online reference manager with social network
 - Notion
 - BibTeX file
- Minimum paper details:
 - Authors, title, venue, year, keywords, abstract
- Write a brief summary:
 - Problem, solution, results, future work



How to read a paper - Conclusion

- Papers are used to communicate research
- Don't expect all papers to be totally correct and well written
- 3 pass manner
- Think when reading
- Don't get frustrated if you don't understand anything



Agenda

- About myself
- Introduction of participants
- How to read a scientific paper
- How to give a good scientific talk ←
- Questions and answers

How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- Concluding your presentation
- Questions and answers

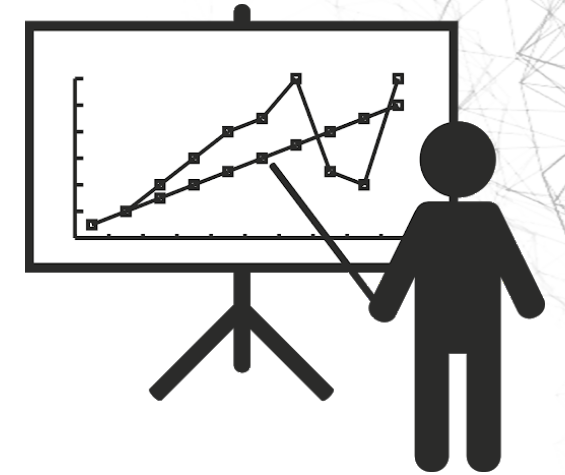
How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- Concluding your presentation
- Questions and answers

Presentation Structure: Basic Rule

- Say what you are going to say (introduction)
- Say it (give the core talk)
- Say what you said (summarise and conclude)

**This is about scientific findings and implications:
Do not try building suspense and then unveiling a surprise ending.**



VS



Exemplary Structure of the Presentation

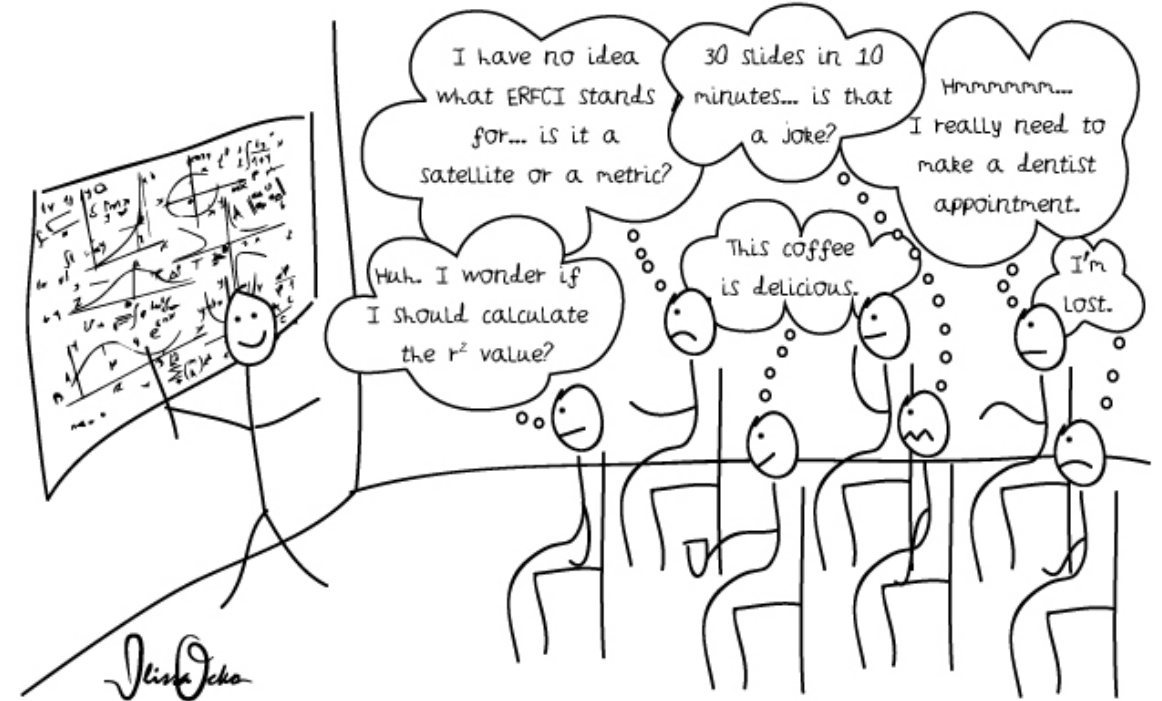
- **Title page** (title, date, authors, venue, acknowledgements)
- **Seminar specifics:** Recap of the previous topic
- **Introduction / Motivation** (including an overview and related works)
- **Approach** (technical details of the method, maths)
- **Experimental Results** (including evaluation methodology, interpretation of the results and discussion)
- **Conclusion** (summary and core implications)

Audience

- **Why** are you giving this presentation?
- **To whom** are you giving this presentation?
- What are **your expectations** from that talk?
- What are the **expectations of the audience**?
- Is the presentation **live or online**?
- How much **time** do I have?



- **Keep that in mind while preparing the talk**
- **Edit / adjust the slides**



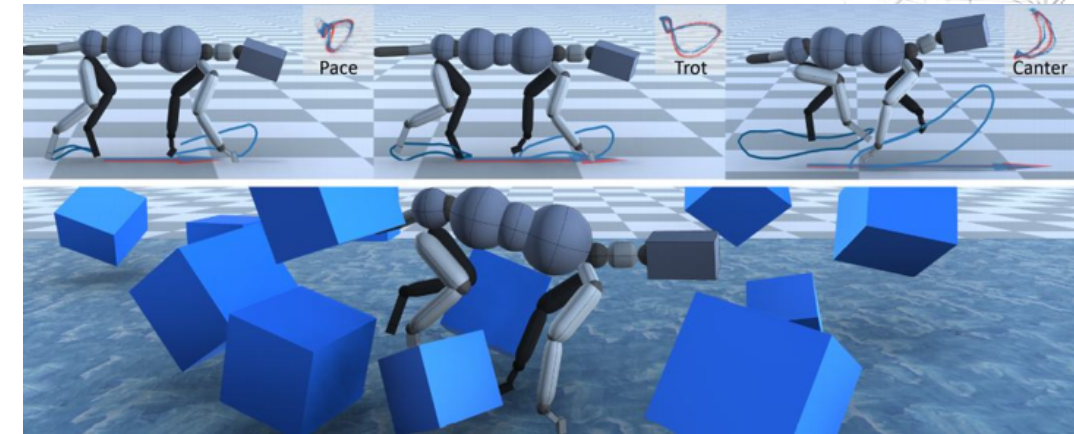
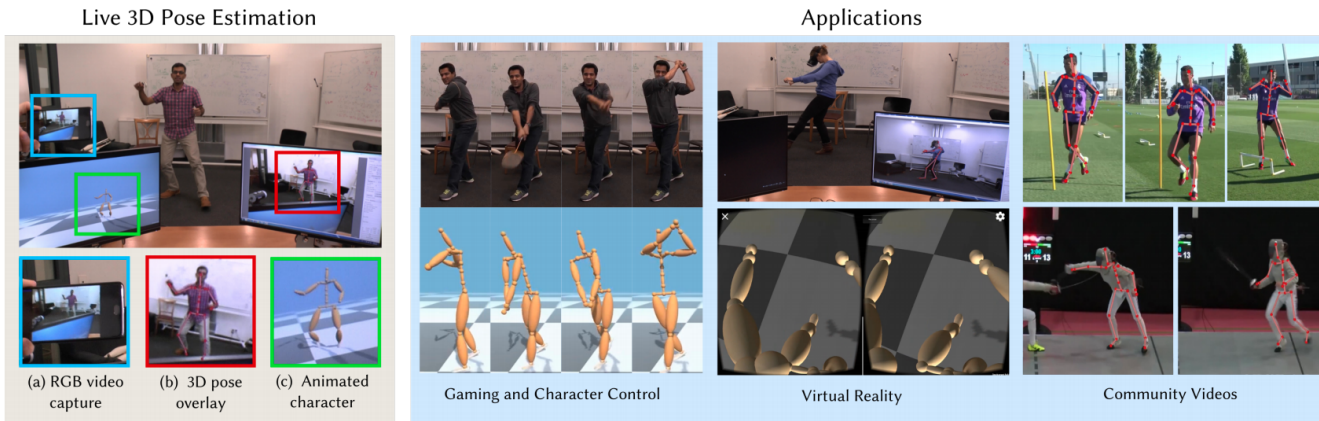
Audience: University Seminar

- Audience with broad technical background
- Many topics: Provide an overview of state of the art
- **Message:**
 - Why the problem is important?
 - Why the proposed solution is novel and impactful?
 - What are the main ideas and insights?
 - “Being a graduate student”: discussion, ideas for improvement
 - To include a slide or not:
 - How important is it for the story?
 - Will the audience understand and value the point?

How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- Concluding your presentation
- Questions and answers

Preparing the Talk: Overview Figures

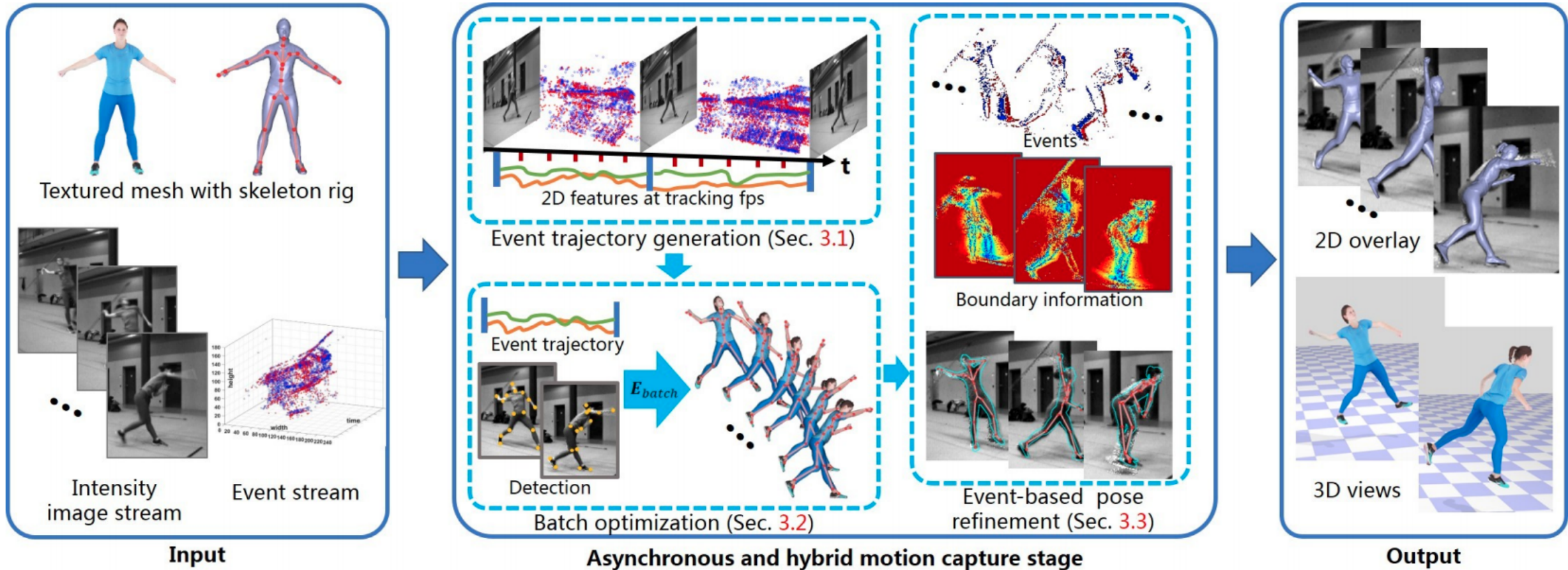


Mehta *et al.*, SIGGRAPH 2017.

Luo *et al.*, SIGGRAPH 2020.

- A figure with a summary of findings
- Overview of the method, problem or a core concept
- Helps to motivate why the problem is important
- If you use web sources, reference the source

Preparing the Talk: Overview Figures



Xu *et al.*, CVPR 2020.

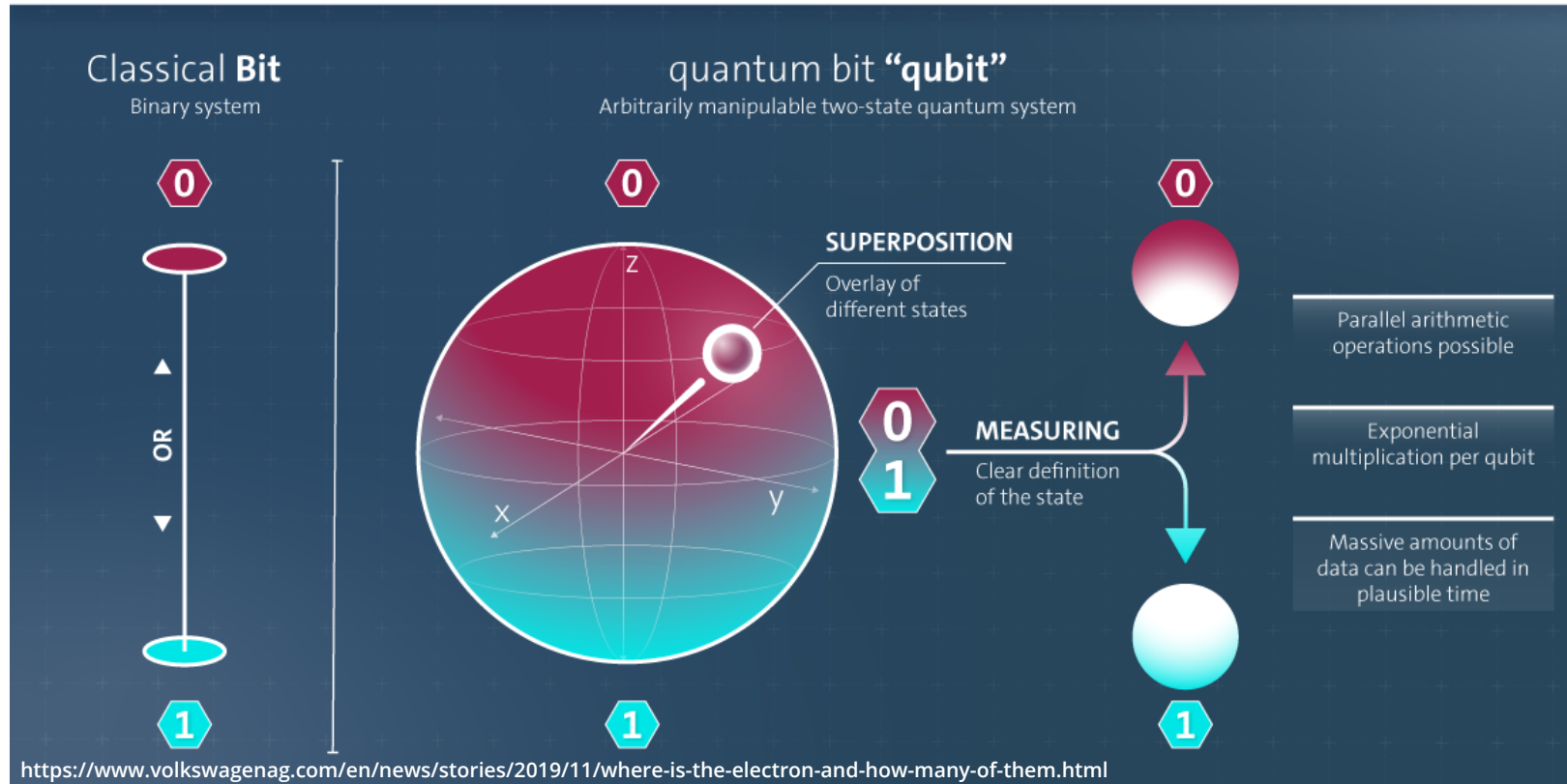
Example: What is a Qubit?

Qubit. Quantum computing encompasses tasks which can be performed on quantum-mechanical systems [53]. Quantum *superposition* and *entanglement* are two forms of parallelism evidenced in quantum computers. A *qubit* is a quantum-mechanical equivalent of a classical bit. A qubit $|\phi\rangle$ — written in the *Dirac* notation — can be in the state $|0\rangle$, $|1\rangle$ or an arbitrary *superposition of both states* denoted by $|\phi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are the (generally, complex) probability amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$. In quantum computing, the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$ denoted by $|+\rangle$ is often used for initialisation of a qubit register. The state of a qubit remains hidden during the entire computation and reveals when measured. If qubits are *entangled*, measuring one of them influences the measurement outcome of the other one [59]. During the measurement, the qubit's state irreversibly collapses to one of the basis states $|0\rangle$ or $|1\rangle$. Efficient physical realisation of a qubit demand very low temperatures. Otherwise, thermal fluctuations will destroy it and lead to arbitrary changes of the measured qubit state.

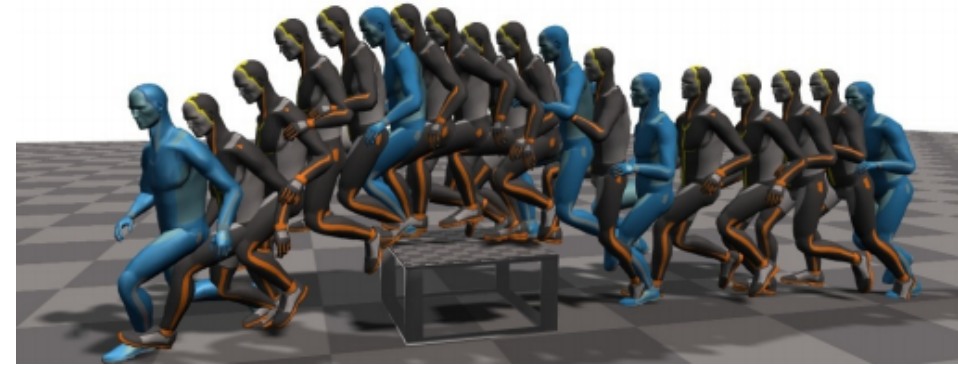
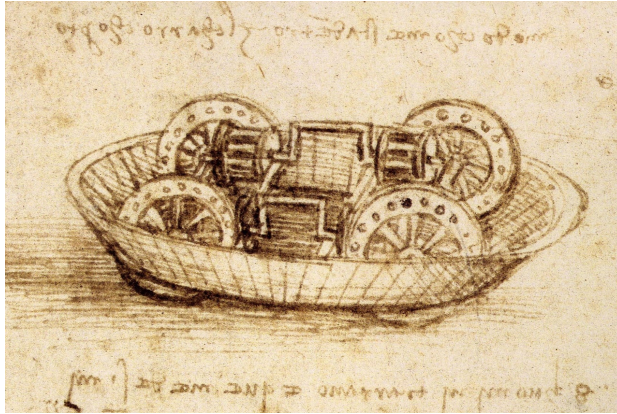
Example: What is a Qubit?

HOW A QUANTUM COMPUTER WORKS

Principle of superposition allows parallelism in the calculations

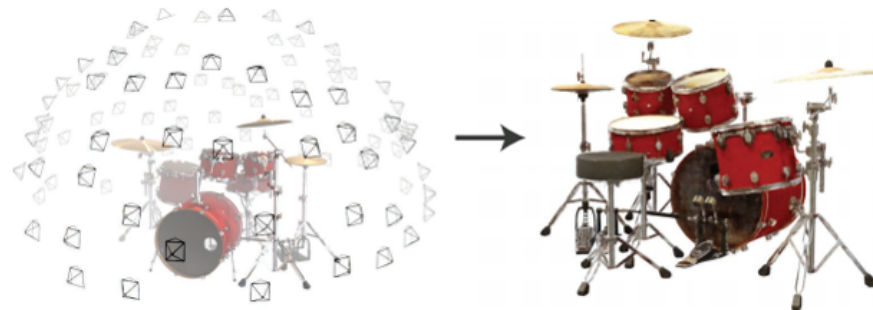


Example: Overview Figures



Technical Drawings of da Vinci.

Harvey *et al.*, SIGGRAPH 2020.



Mildenhall *et al.*, ECCV, 2020.

Using Tables

date	discharge (cf/s)	precipitation (in/day)
-------------	-----------------------------	-----------------------------------

1-Nov	631	0
2-Nov	808	0
3-Nov	794	0.08
4-Nov	826	0
5-Nov	1060	1.09
6-Nov	1080	0.48
7-Nov	1040	0.28
8-Nov	779	0
9-Nov	686	0
10-Nov	670	0
11-Nov	696	0.53
12-Nov	831	0.23
13-Nov	985	0.45
14-Nov	1080	0.14
15-Nov	1350	0.65
16-Nov	1430	0
17-Nov	2440	1.6
18-Nov	2280	0
19-Nov	2040	0
20-Nov	1830	0.55
21-Nov	1650	0
22-Nov	1560	0
23-Nov	1520	0.39
24-Nov	1410	0
25-Nov	1320	0
26-Nov	1310	0.11
27-Nov	1450	0.78
28-Nov	1560	0.22
29-Nov	1550	0.45
30-Nov	1480	0

date	discharge (cf/s)	precipitation (in/day)
-------------	-----------------------------	-----------------------------------

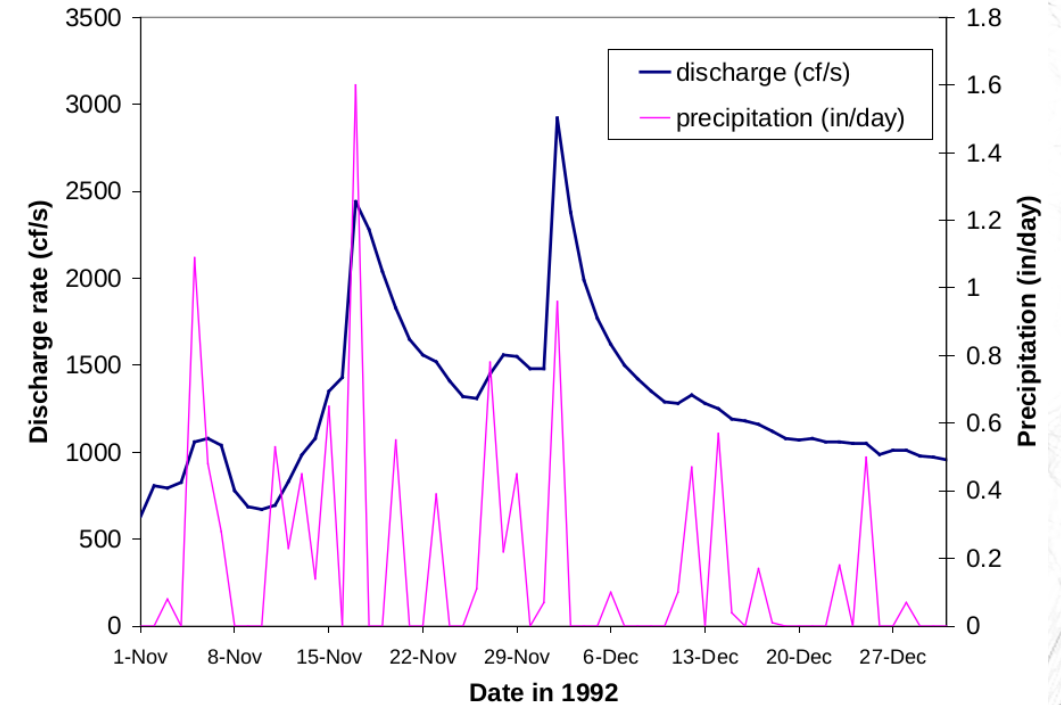
1-Dec	1480	0.07
2-Dec	2920	0.96
3-Dec	2380	0
4-Dec	1990	0
5-Dec	1770	0
6-Dec	1620	0.1
7-Dec	1500	0
8-Dec	1420	0
9-Dec	1350	0
10-Dec	1290	0
11-Dec	1280	0.1
12-Dec	1330	0.47
13-Dec	1280	0
14-Dec	1250	0.57
15-Dec	1190	0.04
16-Dec	1180	0
17-Dec	1160	0.17
18-Dec	1120	0.01
19-Dec	1080	0
20-Dec	1070	0
21-Dec	1080	0
22-Dec	1060	0
23-Dec	1060	0.18
24-Dec	1050	0
25-Dec	1050	0.5
26-Dec	986	0
27-Dec	1010	0
28-Dec	1010	0.07
29-Dec	977	0
30-Dec	972	0
31-Dec	957	0

Using Tables

date	discharge (cf/s)	precipitation (in/day)
1-Nov	631	0
2-Nov	808	0
3-Nov	794	0.08
4-Nov	826	0
5-Nov	1060	1.09
6-Nov	1080	0.48
7-Nov	1040	0.28
8-Nov	779	0
9-Nov	686	0
10-Nov	670	0
11-Nov	696	0.53
12-Nov	831	0.23
13-Nov	985	0.45
14-Nov	1080	0.14
15-Nov	1350	0.65
16-Nov	1430	0
17-Nov	2440	1.6
18-Nov	2280	0
19-Nov	2040	0
20-Nov	1830	0.55
21-Nov	1650	0
22-Nov	1560	0
23-Nov	1520	0.39
24-Nov	1410	0
25-Nov	1320	0
26-Nov	1310	0.11
27-Nov	1450	0.78
28-Nov	1560	0.22
29-Nov	1550	0.45
30-Nov	1480	0

date	discharge (cf/s)	precipitation (in/day)
1-Dec	1480	0.07
2-Dec	2920	0.96
3-Dec	2380	0
4-Dec	1990	0
5-Dec	1770	0
6-Dec	1620	0.1
7-Dec	1500	0
8-Dec	1420	0
9-Dec	1350	0
10-Dec	1290	0
11-Dec	1280	0.1
12-Dec	1330	0.47
13-Dec	1280	0
14-Dec	1250	0.57
15-Dec	1190	0.04
16-Dec	1180	0
17-Dec	1160	0.17
18-Dec	1120	0.01
19-Dec	1080	0
20-Dec	1070	0
21-Dec	1080	0
22-Dec	1060	0
23-Dec	1060	0.18
24-Dec	1050	0
25-Dec	1050	0.5
26-Dec	986	0
27-Dec	1010	0
28-Dec	1010	0.07
29-Dec	977	0
30-Dec	972	0
31-Dec	957	0

Vs.

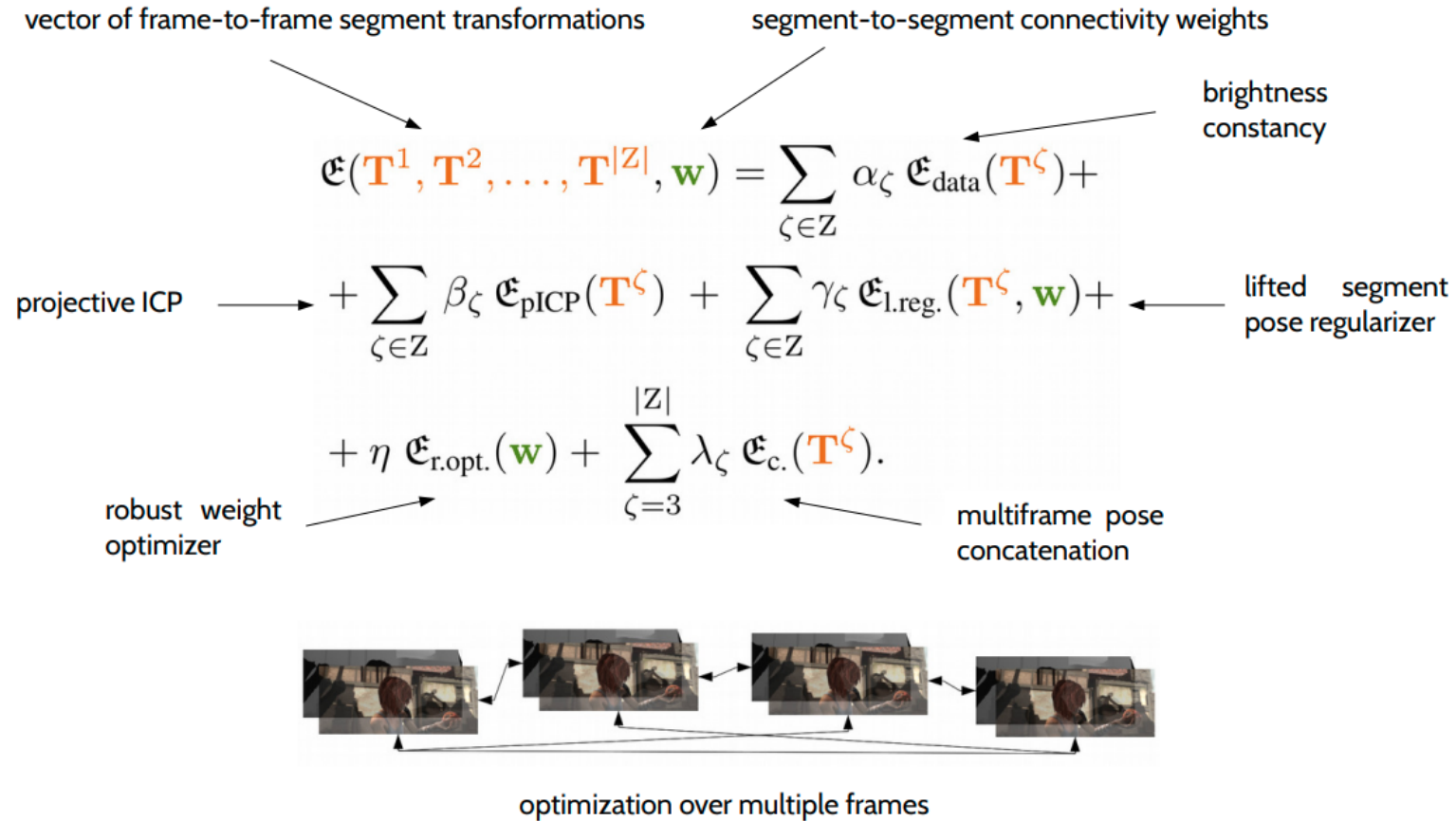


Using Math

$$\begin{aligned} \mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \dots, \mathbf{T}^{|\mathcal{Z}|}, \mathbf{w}) &= \sum_{\zeta \in \mathcal{Z}} \alpha_{\zeta} \mathfrak{E}_{\text{data}}(\mathbf{T}^{\zeta}) + \\ &+ \sum_{\zeta \in \mathcal{Z}} \beta_{\zeta} \mathfrak{E}_{\text{pICP}}(\mathbf{T}^{\zeta}) + \gamma_{\zeta} \sum_{\zeta \in \mathcal{Z}} \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^{\zeta}, \mathbf{w}) + \\ &+ \eta \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|\mathcal{Z}|} \lambda_{\zeta} \mathfrak{E}_{\text{c.}}(\mathbf{T}^{\zeta}). \end{aligned}$$

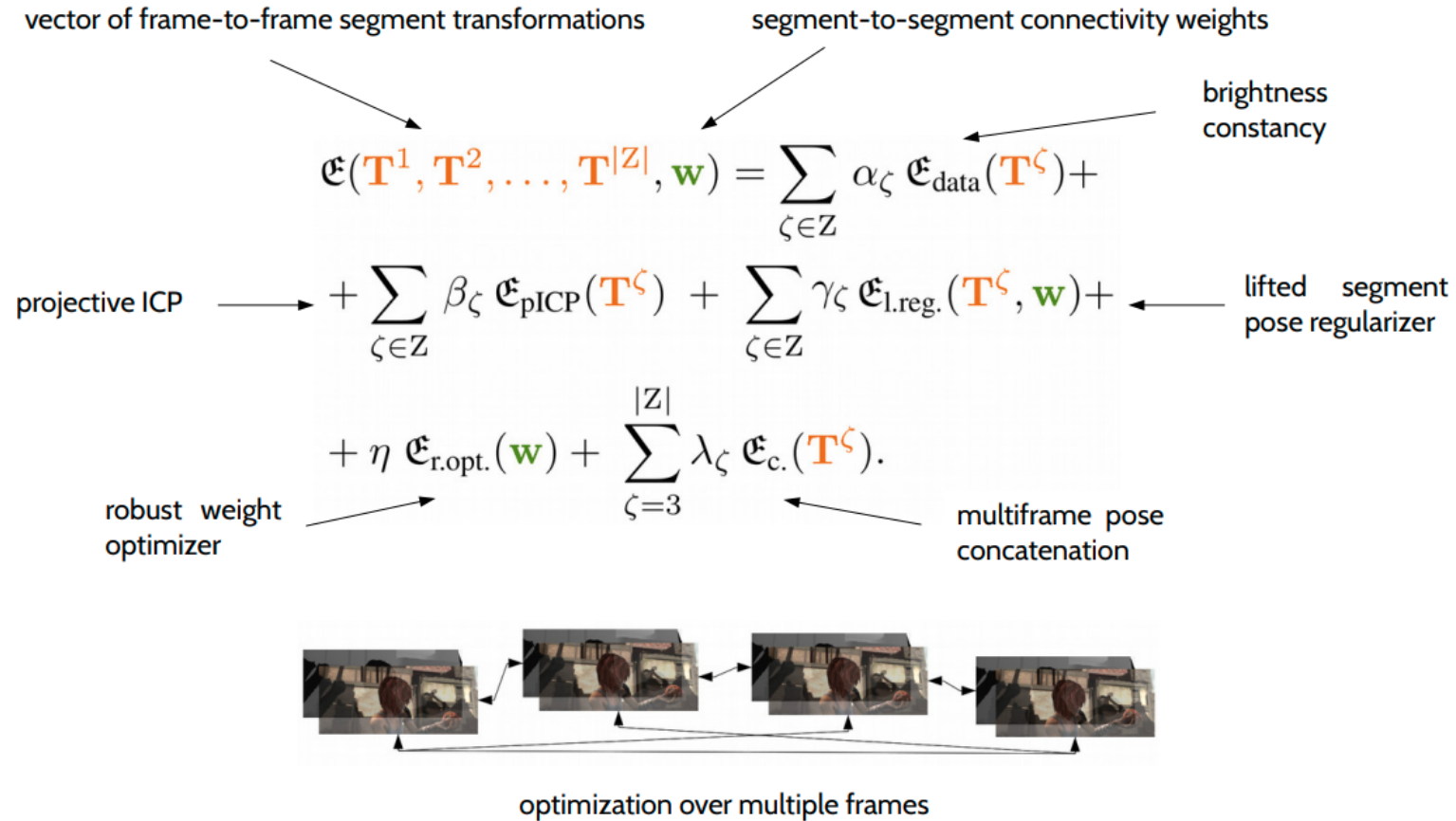
Using Math

$$\begin{aligned} \mathcal{E}(\mathbf{T}^1, \mathbf{T}^2, \dots, \mathbf{T}^{|\mathcal{Z}|}, \mathbf{w}) = & \sum_{\zeta \in \mathcal{Z}} \alpha_{\zeta} \mathcal{E}_{\text{data}}(\mathbf{T}^{\zeta}) + \\ & + \sum_{\zeta \in \mathcal{Z}} \beta_{\zeta} \mathcal{E}_{\text{pICP}}(\mathbf{T}^{\zeta}) + \gamma_{\zeta} \sum_{\zeta \in \mathcal{Z}} \mathcal{E}_{\text{l.reg.}}(\mathbf{T}^{\zeta}, \mathbf{w}) + \\ & + \eta \mathcal{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|\mathcal{Z}|} \lambda_{\zeta} \mathcal{E}_{\text{c.}}(\mathbf{T}^{\zeta}). \end{aligned}$$



Using Math

$$\begin{aligned} \mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \dots, \mathbf{T}^{|\mathcal{Z}|}, \mathbf{w}) = & \sum_{\zeta \in \mathcal{Z}} \alpha_{\zeta} \mathfrak{E}_{\text{data}}(\mathbf{T}^{\zeta}) + \\ & + \sum_{\zeta \in \mathcal{Z}} \beta_{\zeta} \mathfrak{E}_{\text{pICP}}(\mathbf{T}^{\zeta}) + \gamma_{\zeta} \sum_{\zeta \in \mathcal{Z}} \mathfrak{E}_{\text{l.reg.}}(\mathbf{T}^{\zeta}, \mathbf{w}) + \\ & + \eta \mathfrak{E}_{\text{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|\mathcal{Z}|} \lambda_{\zeta} \mathfrak{E}_{\text{c.}}(\mathbf{T}^{\zeta}). \end{aligned}$$



Use equations as little as possible and as much as needed

How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- Concluding your presentation
- Questions and answers

General Rule: Presenting Methodology

- A scientific talk is always about

HOW and WHY

- Explain what you do
- What is new and innovative
- **AND** motivate why this is the way to go

General Rule: Presenting Methodology

- A scientific talk is always about

HOW and WHY

- Explain what you do
- What is new and innovative
- **AND** motivate why this is the way to go

THIS INFLUENCES THE STORY

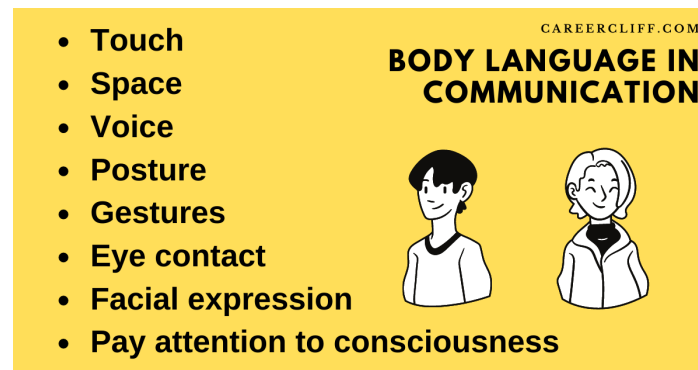
Preparing and Polishing Presentation

- Use **3-7 bullets** per page
 - Avoid complete sentences
- No more than **one minute** per slide on average
- Check the slide appearance **consistency**
- **No sound** unless it is part of results
- **Videos** are often results in visual computing
- **Spelling and writing style**
 - Use the same font (or a few fonts)
 - Check the text for typos; check the grammar
 - Decide between British and American English



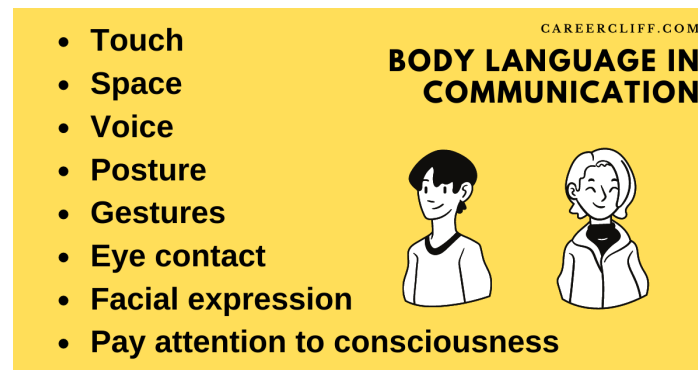
Preparing Yourself

- The way how you present yourself is as important as your slides
- Immerse yourself in what you are going to say
- Make sure that you are **familiar** with (remote) conference software, check your equipment (microphone, projector, etc.)



Preparing Yourself

- **Online format:**
- Perception of gestures and body language is limited:
 - Use other tools of expressiveness
 - There is no eye contact with the audience, you do not see other participants
 - Use intonation in combination with the visual tools (e.g., colours)
- **Rehearsing is very important!**
 - Be on time, know what you want to say, prepare transitions between the slides/papers



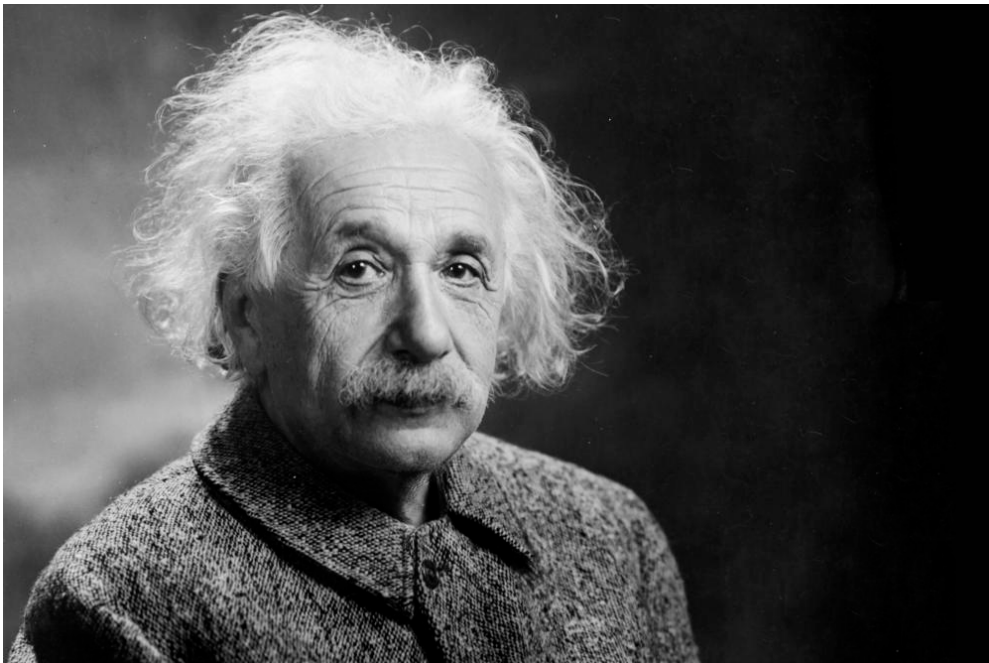
Rehearsing

- Practice – actually stand up and say the words out loud
 - Discover what you do not understand and develop a natural flow
- Do not memorise the talk, do not over-rehearse
- Stay within the time limit
- The Feynman Technique: *A mental model and a breakdown of the thought process to convey information using concise thoughts and simple language [1].*



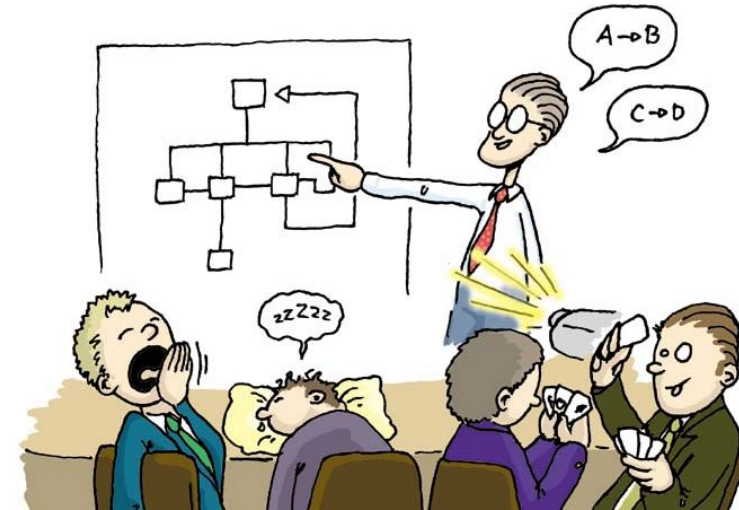
Rehearsing

If you can't explain it simply, you don't understand it well enough.



A. Einstein.

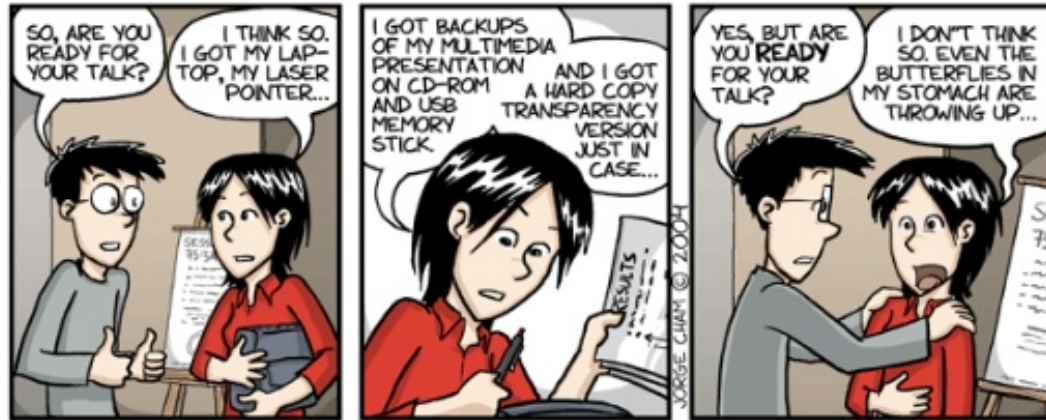
Presenting



- Make yourself comfortable, speak freely, be enthusiastic but do not rush
- Ensure that people can hear you well and see your shared screen
- Seminar specifics: Switch on your camera

Presenting

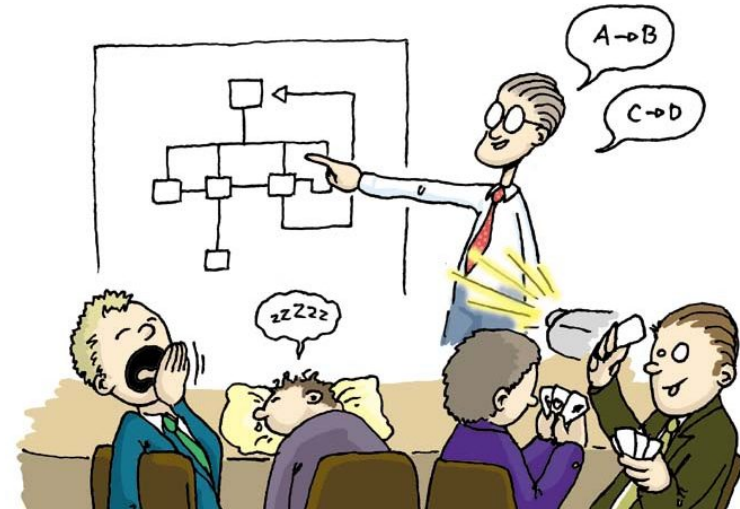
Piled Higher and Deeper by Jorge Cham



title: "Conference" - originally published 8/25/2004

www.phdcomics.com

www.phdcomics.com



- Starting is the most difficult part
 - Memorise the first lines
- Nervousness is normal, don't worry about stopping to think

How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- **Concluding your presentation**
- Questions and answers

Concluding the Presentation

- **Announce the ending** so that people are prepared
- Have only a few **concluding statements** (the core points)
- Come back to the **big picture** and **summarize the significance** of your work
- Open up **new perspective** (could be another slide)
 - Describe future work
 - Raise questions and potential implications
- Think carefully about the **final words** (which people tend to memorise)

How to give a good scientific talk

- Structuring your story
- Preparing your data and information
- Preparing and giving the presentation
- Concluding your presentation
- Questions and answers

Questions and Answers

- **Difficult questions** can help **improving your skills**, writing, and research
 - Identifies parts the audience did not understand
 - Focuses and adds an additional dimension to your analysis
- You can **repeat the question** using your own words
 - This gives you time to think
 - Helps in understanding the question by more people
 - Presents an opportunity for clarification
- **Be concise** in your answers, do not drift away
- **Anticipate questions**, prepare backup slides if required
- **Do not say that the question is bad** or it has been already addressed
- **Never demean the question** or questioner



How to give a good scientific talk - Conclusion

- **Structure your content** in a way that is comfortable for you and your audience
- **Filter** out core aspects and **build convincing story**
- Use figures, videos and maths appropriately
- **Rehearse** and present within the **time limit**
- **Online format:** Using body language in communication is difficult
- **Be prepared** for questions

Materials Used

This talk is a revised version of

How to Give a Good Scientific Talk by V. Golyanik, 2021.

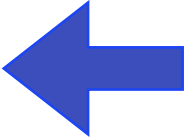
How to Give a Good Scientific Talk by C. Theobalt, 2017.

Some ideas are from

How to Give a Good Talk by S. Pfirman (Cornell University) and

How to give Scientific Presentations by T. Williams (Texas A&M University).

Agenda

- About myself
- Introduction of participants
- How to read a scientific paper
- How to give a good scientific talk
- Questions and answers 



Thank You!

Bonus: 12 Rules for a Bad Talk

Best Presentation-Ever Bingo

Didn't pre-load the presentation	Over-ran time	Used as many bullet points as humanly possible	
	Apologized for unreadable slides	Acted as if had never used PowerPoint	Embraced Obfuscation
Used incredibly complex plots		Used as many slides as humanly possible	Crammed as much as possible onto each slide
Included a video fail	Didn't check the presentation worked beforehand		Used tables with more data than any sane person could read

Bonus: Moderating the Discussion

- You will be assigned as a moderator and get a set of questions one day before the appointment
- Most probably, some questions will be already addressed; all questions cannot be addressed due to time limits
 - 2-4 questions to each paper, up to 2 questions to both papers
 - You decide which questions are the most relevant and engaging
- Prepare a set of points to discuss
 - Weaknesses / Limitations of the methods
 - Comparisons between the papers
 - Ask other participants about their ideas
 - Build bridges to other talks in the seminar
 - Points you were unclear about while reading the papers