SKOLKOVO INSTITUTE OF SCIENCE AND TECHNOLOGY

As a manuscript

Egor A. Burkov

LEARNING FROM DATA FOR HUMAN MODELING AND TRACKING

PhD dissertation summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Academic supervisor:
Candidate of Sciences Viktor S. Lempitsky

Moscow — 2024

Contents

1 Dissertation Topic

Overview. This dissertation deals with human modeling and tracking, collectively termed *human capture* – a set of problems with the end goal of automatically understanding the appearance and the pose of a person. The meaning of ”understanding” and the problems here vary widely depending on the technologies involved, the sensors used, the desired level of detail, and, importantly, the target applications.

Example applications of human capture algorithms include VR/AR telepresence, action recognition for surveillance, automated sports highlight generation, or automatically turning oneself into a 3D video game character. The research problems range from gesture classification and gait recognition to full 3D shape estimation of body, hair, or clothing.

Relevance. Human capture, i.e. the understanding of pose and appearance of people by the means of computer vision, underlies dozens of useful applications. These applications make human life safer (smart security alarm; emergency systems for monitoring elderly people), more convenient (virtual fitness coach; virtual fitting room; device control via hand gestures), and more fun (cartoon avatars for telepresence; dance and sport simulators or games). At the same time, nowadays almost no human capture problem can be considered fully solved. Many of the human capture algorithms are yet to improve, and usually have to be at least carefully fine-tuned for their applications. For instance, current algorithms for 3D human digitization from one photo usually lack realism and require manual refinement from an artist; some applications like gesture control are just generally inhibited by the flaws of current hand pose estimation algorithms (low accuracy, temporal shakiness, vulnerability to occlusions and poor lighting, and low speed).

This dissertation aims to solve some of the above issues of the existing human capture algorithms, and therefore is both relevant from the scientific perspective (a multitude of human capture challenges is now attracting huge numbers of researchers) and directly important for practical applications.

Goals. Our inspiration is that, for a human brain, it is usually sufficient to rely only on visual cues to perform ”human capture” (e.g. imagine how someone might look like from different views). We argue that this mental inference is possible because people ”learn from data”, having seen many other people before. In contrast, many previous state-of-the-art human capture algorithms seem to underuse the prior knowledge about human bodies, using e.g. hand-engineered machine learning features or complex pipelines with manual refinement.

Hence, **our overall goal** is to improve some of current state-of-the-art algorithms for human capture in several scenarios where learning from data has previously been particularly overlooked by promoting rich datasets, end-to-end-learnable models, and paradigms such as self-supervised learning and meta-learning. Namely, we narrow our scope to **solve these human capture problems**:

- 3D body pose estimation with telepresence as the primary application;
- the estimation of head pose and facial expression with cross-person reenactment and telepresence as the primary applications;
- 3D head shape reconstruction.

Specifically, these translate into the following **tasks**:

1. develop an algorithm for realistic and temporally smooth estimation of 3D coordinates of full body keypoints from a single-view RGB video;
2. improve the temporal smoothness and occlusion robustness of 3D pose estimation in a multi-camera scenario;
3. construct an entirely different pose representation (for head and face) that, unlike keypoints, is person-agnostic, i.e. doesn't contain information about the identity;
4. test the above pose representation in a cross-person reenactment pipeline;
5. devise a trainable algorithm to reconstruct the 3D shape of a human head from a single or few RGB images.

2 Key Results

The main outcomes of the thesis are:

1. An existing telepresence algorithm has been augmented with free-viewpoint capability thanks to a new algorithm for full 3D pose estimation (including face, hands, and feet) from monocular RGB video.
2. Two more precise, multi-view 3D pose estimation algorithms have been developed, significantly improving the state-of-the-art on the common datasets and boosting occlusion robustness. The algorithms have brought conceptual novelty in that they are end-to-end learnable using ground truth 3D coordinates directly, unlike previous multi-view pose estimation methods.

3. A latent representation of facial expression and head pose that is as descriptive as other popular representations (like keypoints or 3DMM) but contains much less information about the person's identity and doesn't require a manually labeled dataset since it is learned in a self-supervised way.

4. A head-and-shoulders reenactment system based on the above representation that allows cross-person reenactment without giving away driver's identity.

5. A method that predicts the textured 3D shape of a head based on one or few images (such as a selfie or a painting). Compared to the most similar method in the literature, the proposed one is trained on a much simpler dataset (100 smartphone videos instead of 10.000 3D scans).

The novelty of these outcomes can be summarized as follows:

- Unlike previous methods, both proposed algorithms for multi-view pose triangulation are end-to-end differentiable and optimize the target metric directly during training. The algorithms have significantly improved state-of-the-art accuracy on two popular multi-view datasets.

- The latent pose representation is both learned in a self-supervised way (and can potentially learn arbitrary levels of detail) and is provably person-agnostic.

- The head reenactment system allows cross-person driving while providing arguably simpler pipeline and better quality than the available systems at the time of publication.

- Two original methods for fitting neural implicit functions to multiple objects simultaneously are introduced.

- Compared to the previously best 3D head reconstruction algorithm of the same family, the proposed one achieves better results while being trained on a vastly simpler dataset.

The practical significance of the results stems directly from our original motivation for addressing the above challenges. The proposed 3D pose estimation algorithms – both single-view and multi-view – have been demonstrated to successfully augment an existing telepresence pipeline with a free-viewpoint capability. This enables its use e.g. in AR or VR. Person-agnostic pose descriptors allow training neural renderers for cross-person reenactment without having to deal with the driver's identity "leaking" into the rendered avatar. Such reenactment systems can be useful for character animation in movie production or for entertainment. The proposed single-view 3D reconstruction algorithm shortens the practical gap to several applications, such as multimedia aids in historical museums or putting oneself into a video game.

The thesis' results carry some indirect practical value too. Precise and smooth 3D pose estimation without physical wearable markers can make a number of products much cheaper (VR quest rooms; motion capture in movie production; hands-free VR headsets without depth cameras). Person-agnostic pose descriptors could be useful in privacy-preserving applications that do pose estimation.

Personal contribution. All the results of the dissertation were obtained personally by the applicant or with his direct involvement. The entire pipelines for 3D pose estimation from video, latent head pose and expression (except pose augmentations), and 3D head reconstruction were designed and implemented by the author. In these parts of the thesis, the author also came up with the original ideas for the methods; besides, the related work was studied, analyzed, and compiled by the author. As for the learnable triangulation, the author implemented full handling of one of the datasets (Human3.6M), validated the proposed algorithm on it, as well as launched some of the relevant training experiments. All visualization tools were implemented by the author.

To summarize, **the provisions to be defended** are:

1. Two methods for learnable multi-view triangulation.
2. A latent pose representation learned in a self-supervised way.
3. A cross-person head reenactment system based on the above pose representation.
4. An algorithm for 3D head reconstruction from a single image with known camera parameters.

3 Publications and Approbation of Research

First-tier publications

1. E. Burkov, I. Pasechnik, A. Grigorev, V. Lempitsky. **Neural Head Reenactment with Latent Pose Descriptors**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13786-13795, 2020. [Indexed in SCOPUS, CORE A*]

2. E. Burkov, R. Rakhimov, A. Safin, E. Burnaev, V. Lempitsky. **Multi-NeuS: 3D Head Portraits from Single Image with Neural Implicit Functions**. *IEEE Access*, vol. 11, pp. 95681-95691, 2023. [Indexed in SCOPUS, Q1]

3. K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov. **Learnable Triangulation of Human Pose**. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 7718-7727, 2019. [Indexed in SCOPUS, CORE A*]

4. A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, V. Lempitsky. **Textured Neural Avatars**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2387-2397, 2019. [Indexed in SCOPUS, CORE A*]

5. E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky. **Few-Shot Adversarial Learning of Realistic Neural Talking Head Models**. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9459-9468, 2019. [Indexed in SCOPUS, CORE A*]

Reports at conferences and seminars

1. ”Textured Neural Avatars”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 16–20, 2019.

2. ”Learnable Triangulation of Human Pose”. IEEE International Conference on Computer Vision (ICCV). October 27 – November 2, 2019.

3. ”Few-Shot Adversarial Learning of Realistic Neural Talking Head Models”. IEEE International Conference on Computer Vision (ICCV). October 27 – November 2, 2019.

4. ”Neural Head Reenactment with Latent Pose Descriptors”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 14–19, 2020.

The author has also contributed to the following publications:

1. E. Burkov, V. Lempitsky. **Deep Neural Networks with Box Convolutions**. *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 6214-6224, 2018. [Indexed in SCOPUS, CORE A*]

4 Contents

The main part of the thesis consists of 3 chapters, each summarized below. The first chapter highlights the research outcomes from publication #4 first and then from #3, the second chapter is derived from publication #1 and its predecessor #5, and the third chapter matches #2.

4.1 3D-Rotatable Pose for Human Body

This chapter is inspired by a target application which is a full body telepresence system. We consider an implementation where the neural renderer generates an avatar given a pose as 2D keypoints (Figure 1).

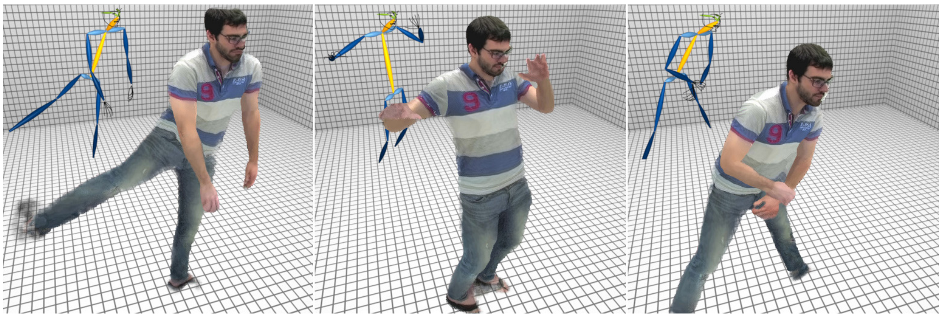

Figure 1: Example inputs and outputs of the full body telepresence system. In each example, at the top left is the driver's pose encoded as keypoint locations (and visualized as a ”stickman”), and the rest is the RGB render of the avatar.

It would be useful if the pose would come as 3D keypoints, because then we could re-project it to any view, thus enabling the free-viewpoint capability, essential for e.g. AR or VR headsets (Figure 2).

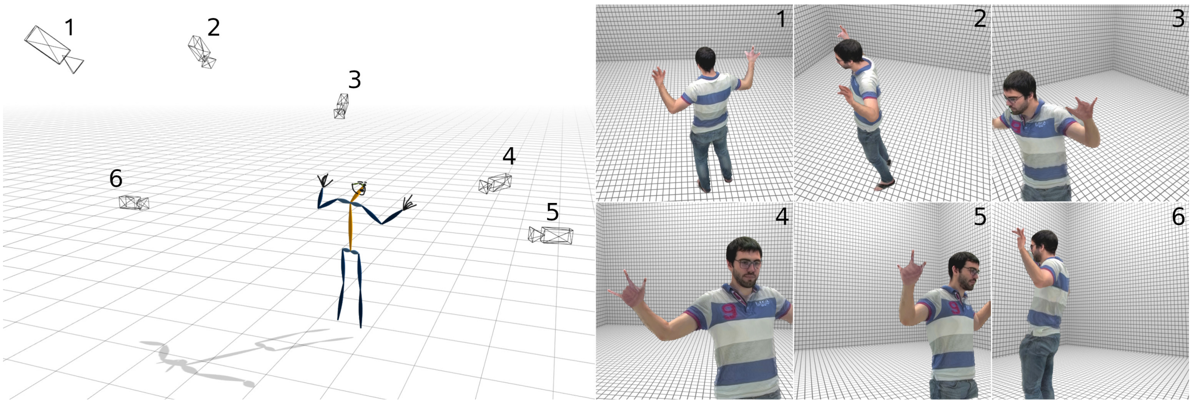

Figure 2: The improvement considered in Chapter 3: 3D pose capture that enables free viewpoint rendering, that is, choosing arbitrary camera. Left: the driver's pose (keypoint locations) and 6 examples of novel views. Right: the corresponding rendered avatars.

4.1.1 Temporal 3D Lifting of 2D Keypoints

First, we consider a simple setup where the driver (the person that drives the avatar, i.e. the source of the pose) is captured by a single RGB camera. We come up with a simple algorithm that predicts their 3D pose based on the motion cues in the 2D poses, detected by an off-the-shelf pose detector OpenPose [2]. It is simple shallow neural network regression but with a few twists (Figure 3). Notable features are the sophisticated normalization and un-normalization procedures, and the keypoint validity masks.

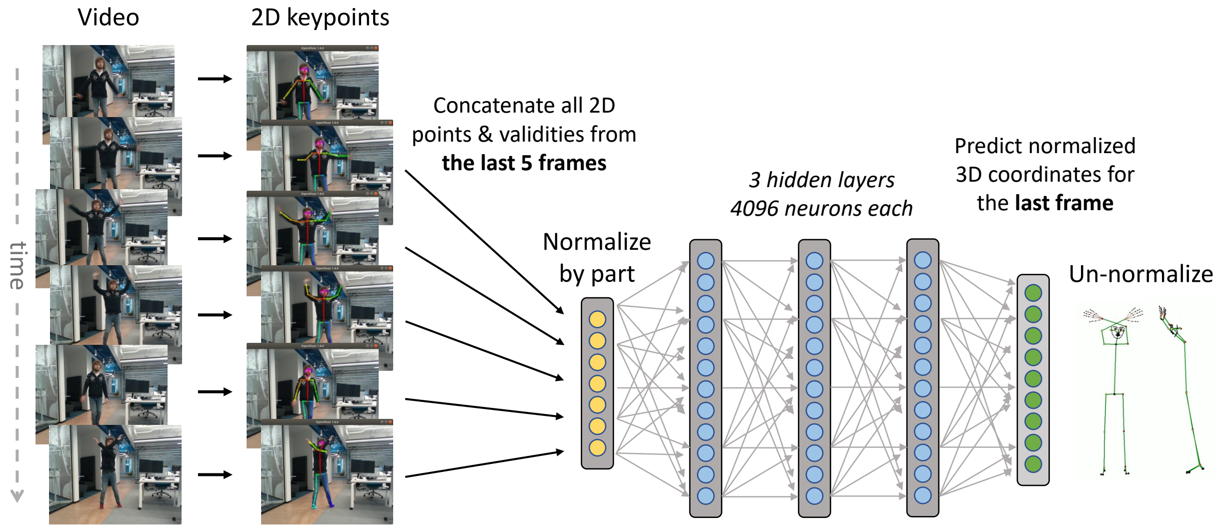
Video
2D keypoints
time
Concatenate all 2D points & validities from **the last 5 frames**
Normalize by part
3 hidden layers 4096 neurons each
Predict normalized 3D coordinates for the **last frame**
Un-normalize

Figure 3: The multi-layer perceptron regression pipeline for 3D lifting of 2D poses in a video.

One inspiration for this algorithm was a large dataset of multi-view videos called CMU Panoptic (Figure 4) we had at hand. We could derive (extract, detect, and triangulate) around 770 000 3D poses in total, allowing our network to learn pose and motion priors from it.

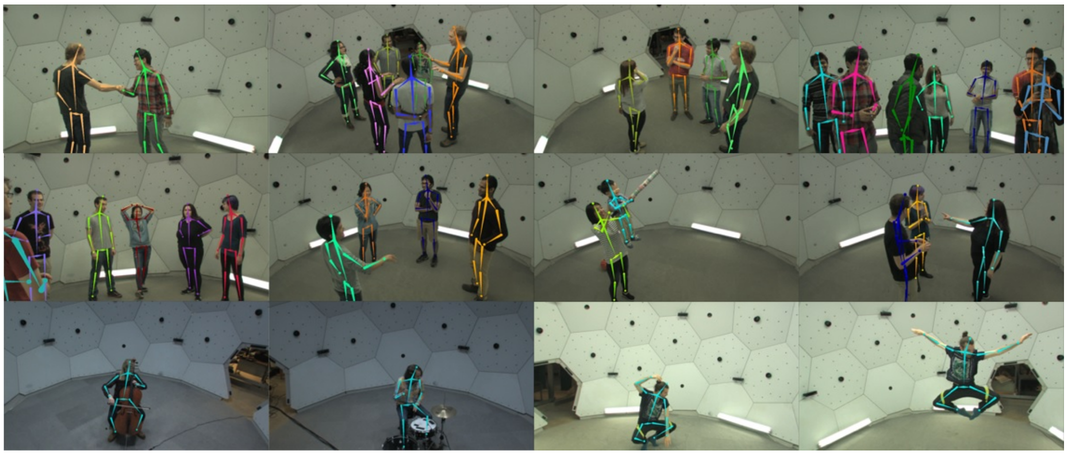

Figure 4: Example frames from the CMU Panoptic dataset with principal body keypoints visualized.

A qualitative demonstration is presented in Figure 5. Our model is able to predict 3D poses even though only 2D coordinates of keypoints are given and no 3D information is available. Although the poses may not be exactly correct, and sometimes even their 2D reprojection to the OpenPose camera may differ from the OpenPose prediction, they are mostly realistic and in general represent the original human pose.

In general, the model yields reasonably smooth predictions thanks to the ”built-in temporal smoothing” induced by the continuous conditioning on the temporal context. Importantly, lifting only adds negligible performance overhead to 2D pose estimation (on the order of 1.5 milliseconds per frame).

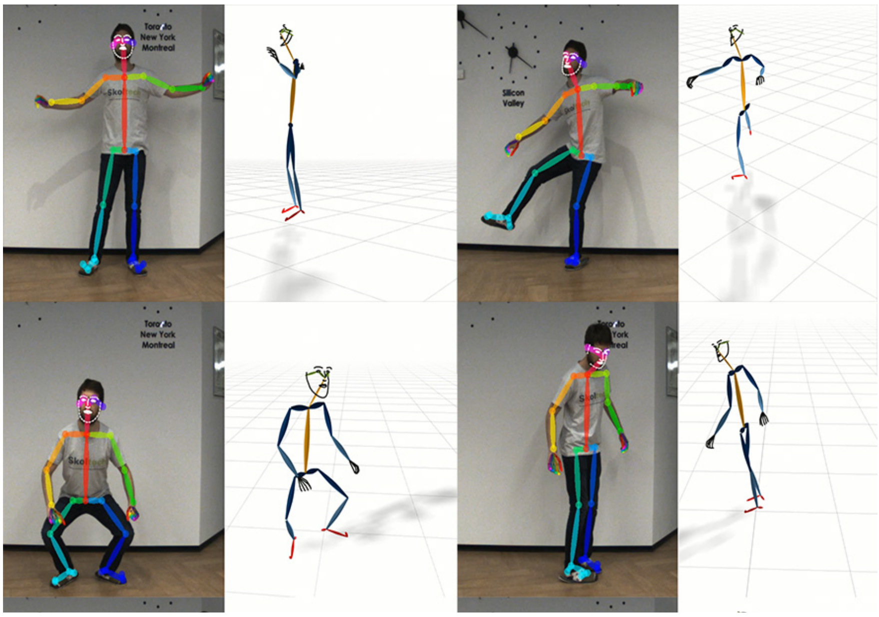

Figure 5: Qualitative examples for the 2D-to-3D pose lifting model. On the left is the last input video frame with the overlay of OpenPose keypoint predictions, on the right is the predicted 3D pose shown from a different angle.

There are many ways to improve this rather simple baseline. Admittedly, the plain binary validity mask as input and the zeroing of features are not the most natural methods for neural networks. Also, the model often has issues with fine motion where it seems to exhibit ”regression-to-mean” behavior; this could be addressed with masked adversarial loss [3]. Finally, while the temporal context helps the model to disambiguate, it has no effect when there is no motion, so shakiness/flickering still kick in when the driver keeps a still ambiguous pose for more than 5 frames; this could be solved by recurrent architectures.

4.1.2 Learnable Triangulation of 3D Keypoints for the Multi-View Case

This section takes a different path towards 3D pose estimation and attempts to reconstruct precise 3D locations of keypoints (while leaning towards pose realism, too). With one camera, precise 3D pose cannot theoretically be recovered due to projection ambiguity. This time, we consider a setup with *multiple* calibrated cameras.

If several RGB cameras are available and all their parameters (poses and projection matrices) are

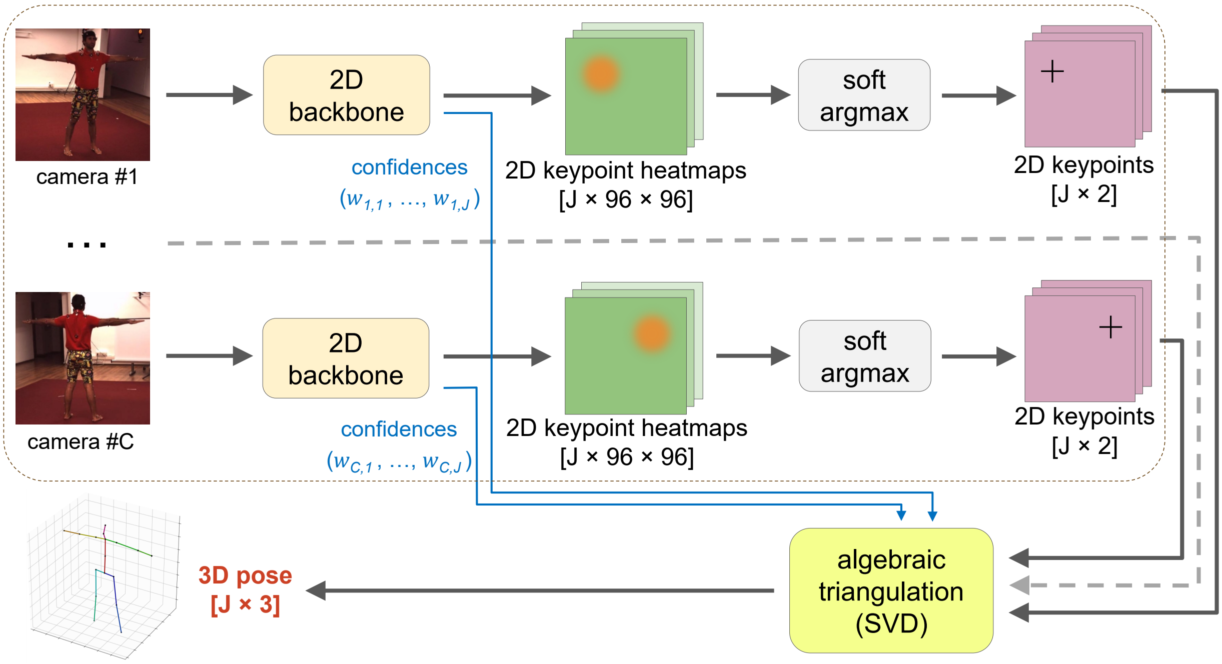

Figure 6: Outline of the ”algebraic” approach.

known, it should be possible to estimate precise 3D keypoints by detecting their 2D projections in each camera image and then *triangulating* them in 3D. In theory, this is possible even just from two cameras; in practice, there are many obstacles hampering triangulation: errors in 2D pose estimation, errors in camera calibration, other kinds of noise, and most importantly, keypoints that are occluded or otherwise unobservable in the camera frame. This is why previous works that used multi-view triangulation for constructing datasets relied on excessive, almost impractical number of views (cameras) to get the 3D ground truth of sufficient quality [8, 24].

Here, we propose two solutions that, compared to previous state of the art, significantly reduce obviously unrealistic poses, inability to handle keypoints unobservable from any camera, and the reliance on large number of views for precise predictions. Behind both of them lies the idea of *learnable* triangulation; during learning, we either use marker based motion capture ground truth or “meta”-ground truth obtained from the excessive number of views. Crucially, both of the proposed solutions are fully differentiable, which permits end-to-end training.

The **”algebraic” approach** resembles classical triangulation methods but is end-to-end differentiable (Figure 6). The 2D convolutional network detects keypoints’ coordinates via heatmaps followed by soft-argmax, and also predicts scalar confidences per view per keypoint. All these predictions are then triangulated using the camera matrices via differentiable operations. The 2D backbone is then trained by backpropagating the loss between the predicted and the true 3D keypoints (a ”soft” version of MSE, mean squared error).

”Volumetric” approach. The main drawback of the baseline algebraic triangulation approach is

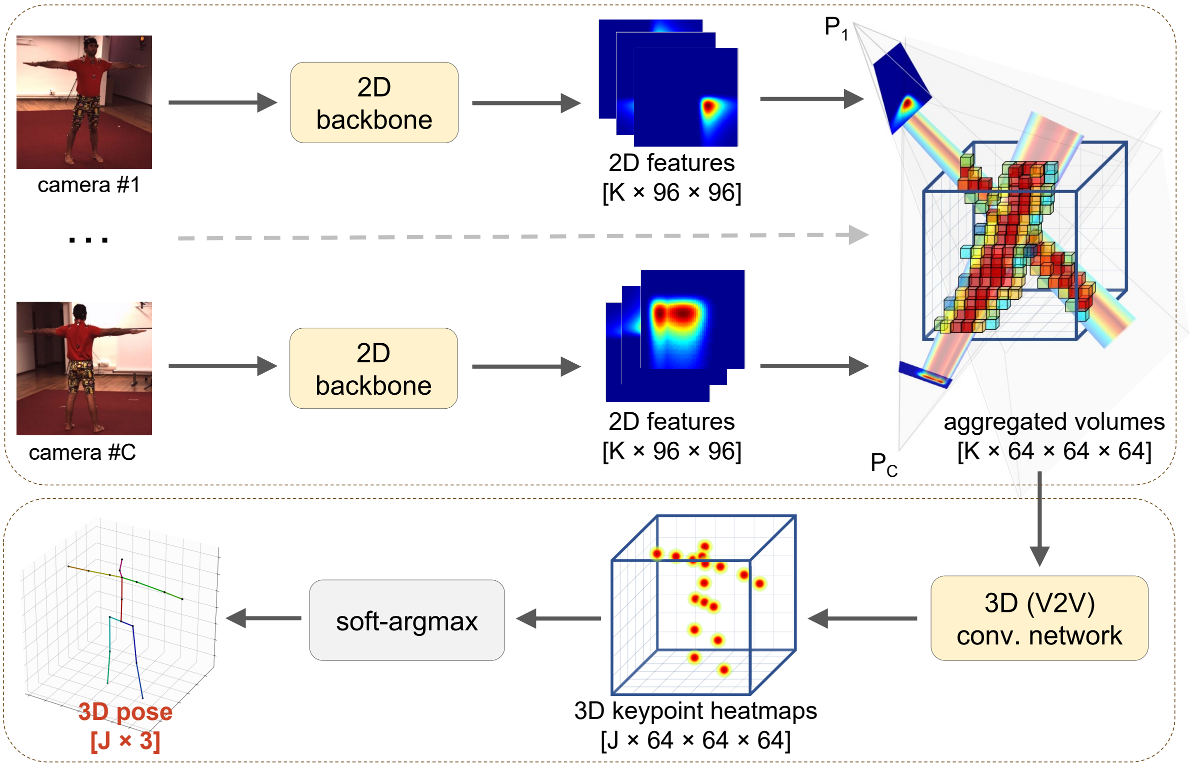

Figure 7: Outline of the ”volumetric” approach.

that the images from different cameras are processed independently from each other, so there is no easy way to add a 3D human pose prior and no way to filter out the cameras with wrong projection matrices.

To solve this, we propose to add another trainable step after aggregating the information from 2D views (Figure 7). Instead of predicting the 2D coordinates, the backbone predicts latent heatmaps, which are then *unprojected* into a volumetric grid using the camera matrices. This grid is processed by a 3D convolutional net to predict interpretable 3D keypoint heatmaps.

Results. We conduct experiments on two large multi-view datasets with available ground-truth 3D pose annotations: Human3.6M [7] and CMU Panoptic [8, 22, 17] (the same used earlier for 2D-to-3D pose lifting). Our methods provably outperform previous ones (Table 1). The qualitative assessment (Figure 8) shows that the algebraic approach is better that the baseline, and the volumetric approach is even more robust at occluded keypoints. We also show that a model trained on CMU Panoptic successfully transfers to Human3.6M, which has a different camera setup. Additionally, we have found out that to reach the same error of 17 mm, the baseline needs 12 cameras, while the algebraic approach needs 4 and the volumetric just 2.

Monocular methods (MPJPE relative to pelvis, mm)	
Martinez et al. [11]	62.9
Sun et al.[18]	**49.6**
Pavllo et al. [15] (∗)	**46.8**
Hossain & Little [5] (∗)	58.3
Ours, volumetric single view (†)	49.9
Multi-view methods (MPJPE relative to pelvis, mm)	
Multi-View Martinez [19]	57.0
Pavlakos et al. [14]	56.9
Tome et al. [19]	52.8
Kadkhodamohammadi & Padoy [9]	49.1
RANSAC (our implementation)	27.4
Ours, algebraic (w/o conf)	26.9
Ours, algebraic	22.6
Ours, volumetric (softmax aggregation)	**20.8**
Ours, volumetric (sum aggregation)	21.3
Ours, volumetric (conf aggregation)	**20.8**

Model	MPJPE, mm
RANSAC	39.5
Ours, algebraic (w/o conf)	33.4
Ours, algebraic	21.3
Ours, volumetric (softmax aggregation)	**13.7**
Ours, volumetric (sum aggregation)	**13.7**
Ours, volumetric (conf aggregation)	14.0

Table 1: The results of evaluation on the Human3.6M dataset (**left**) and CMU Panoptic (**right**). The methods that use temporal information during inference are marked by (∗). Note that our monocular method (labeled by †) is using the approximate position of the pelvis estimated from multiple views.

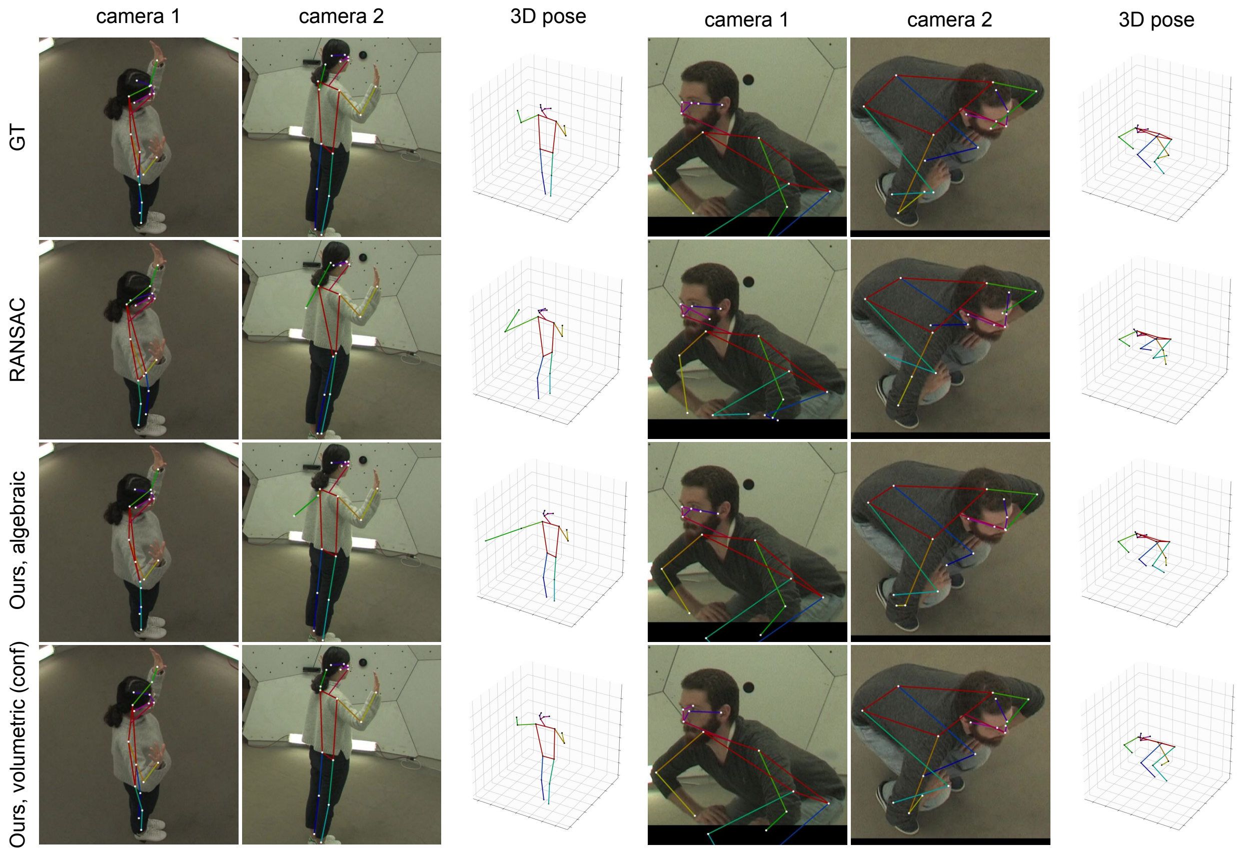

Figure 8: Pose triangulation results on the CMU Panoptic validation set (2 cameras).

4.2 Person-Agnostic Pose for Head and Face

In this chapter, we continue to study pose capture (pose estimation) and our main motivation is again telepresence. While the previous chapter was devoted to better estimation of the keypoints, in this chapter we highlight the fundamental disadvantages of keypoints and try to devise an entirely new pose representation.

We consider a *head reenactment* system, with a reference implementation [25] similar to that in the previous chapter. Since, as before, its main input is rasterized keypoints, it suffers from many disadvantages of keypoints. The most prominent disadvantages here are the identity information contained in keypoints (harms cross-person reenactment), limited descriptiveness (Figure 9), and temporal shakiness.

To overcome these issues, we replace fixed keypoint conditioning with a *pose encoder* network, trained end-to-end with the entire system. The model is trained on a dataset of videos to predict the hold-out frame as shown in the Figure 10.

Intuitively, nothing prevents our system from learning to encode person-specific information into the pose embedding, in the worst case degrading into an auto-encoder (so that the identity encoder becomes useless). Initially, we expected that some form of adversarial training [3] or cycle-consistency [26, 6] would be necessary. Apparently, these 3 techniques turned out to be enough to ensure the *disentanglement* of pose and identity:

1. Pose encoder's capacity is lower than that of the identity encoder (in our case, MobileNetV2 vs ResNeXt-50).

2. Pose augmentations (transformations that preserve person's identity in an image) are applied to the pose source.

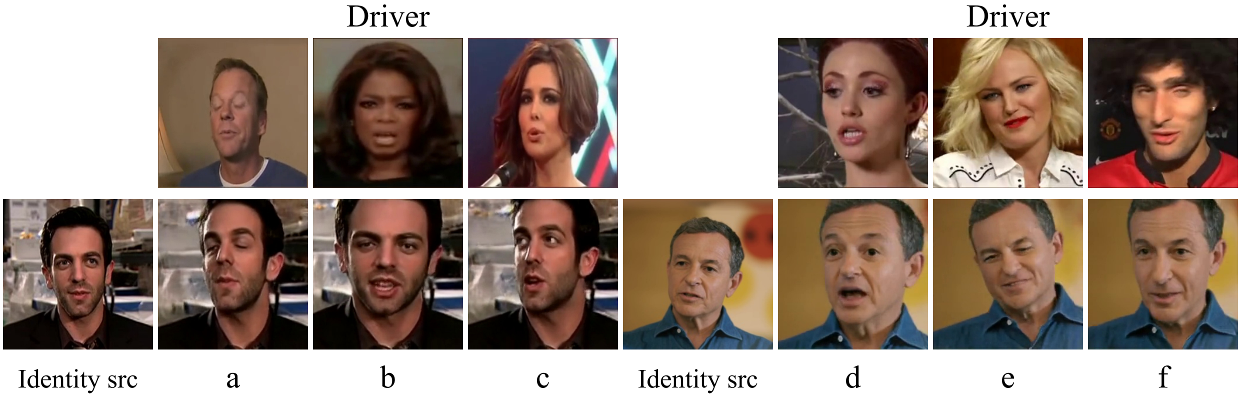

Figure 9: Avatar renderings (bottom row), generated by the keypoint-based system [25] using the pose from the corresponding driving images (top row). The driver's shape "leaks" into the rendering, resulting in a perceptible identity gap (a, b, f), and the pose differs from that of the driver (c, d, e, f).

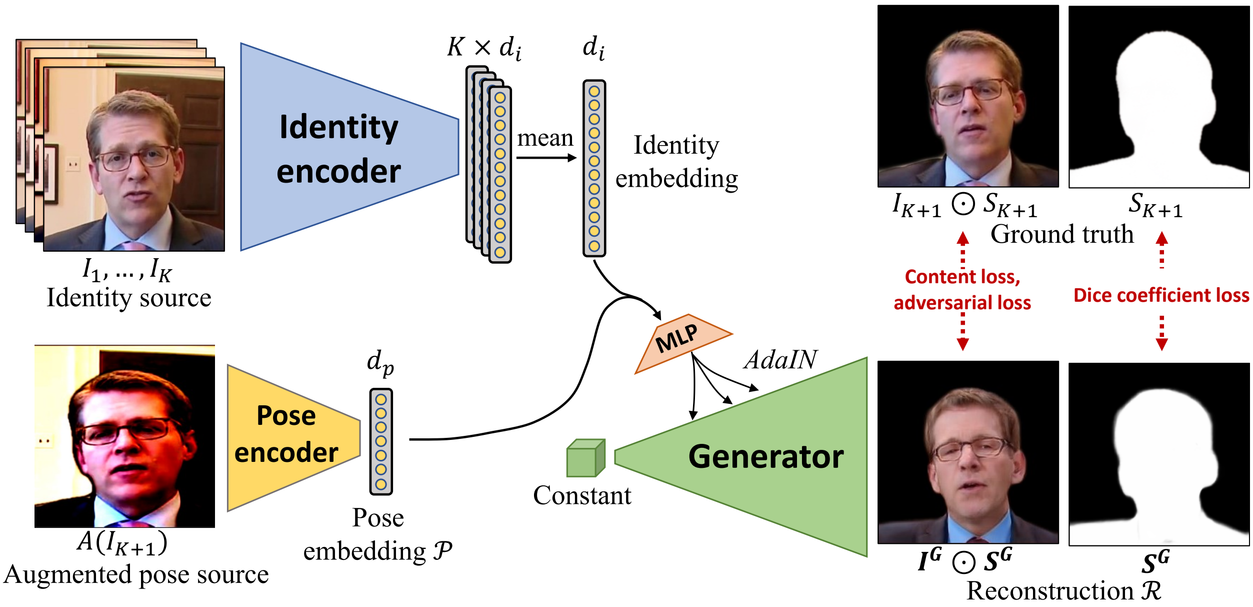

Figure 10: At each step of meta-learning, our system samples a set of frames from a video of a person. The frames are processed by two encoders. The bigger *identity encoder* is applied to several frames of the video, while the smaller *pose encoder* is applied to a hold-out frame. The obtained embeddings are passed to the generator network, whose goal is to reconstruct the last (hold-out) frame. Since the capacity of the pose encoder is limited, and since its input does not exactly match other frames w.r.t. identity (thanks to data augmentation), the system learns to extract all pose-independent information through the identity encoder, and uses the smaller encoder to capture only pose-related information, thus achieving pose-identity disentanglement.

3. Foreground mask is predicted, and reconstruction losses are applied computed with background blacked out.

Results. First, we evaluate the latent pose descriptors (pose embeddings) learned within our system. We compare against other available descriptors in emotion classification on the Multi-PIE dataset [4] and show that our descriptors win at matching different people in the same pose. Then, we train a shallow neural network to regress keypoint locations from our pose and identity embeddings, and achieve better error than FAb-Net [21], though slightly worse than the state-of-the-art for that task. The supplementary video (`https://shrubb.github.io/research/latent-pose-reenactment/`) demonstrates that our descriptors are much smoother in time, and that spherical interpolation between them leads to meaningful pose changes.

Then, we assess our head reenactment system. For quantitative evaluation, we introduce two scores, the *identity error* (tells how well the identity of the avatar is preserved compared to the reference image) and the *pose reconstruction error* (tells how precise the driver's pose and facial expression are replayed). We show that our system is strictly better (or, compared to [25], achieves a better trade-off between the two scores) than the competitor systems (Figure 12). This is confirmed by the qualitative comparison (Figure 11).

Finally, we conduct thorough **ablations** where we study the effects of reducing pose vector dimensionality, increasing pose encoder capacity, keeping the background in images, and removing pose augmentations. We prove quantitatively (Figure 13) and qualitatively that our choices in the final model provide the best trade-off between errors in pose and identity.

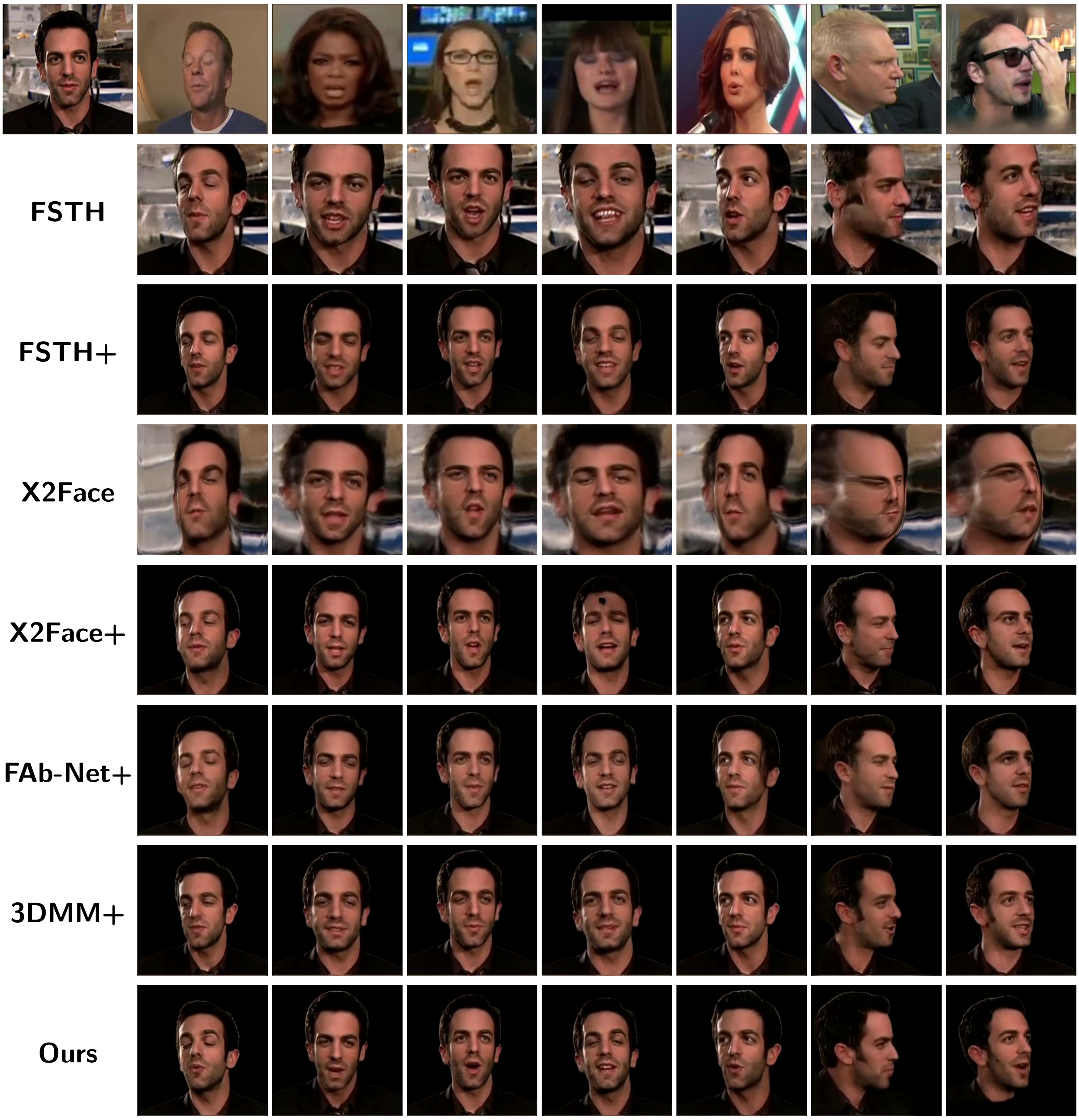

Figure 11: Comparison of cross-person reenactment for several systems on VoxCeleb2 test set. The top left image is one of the 32 identity source frames. The other images in the top row are pose drivers. Our method better preserves the identity of the target person and successfully transfers the mimics from the driver person.

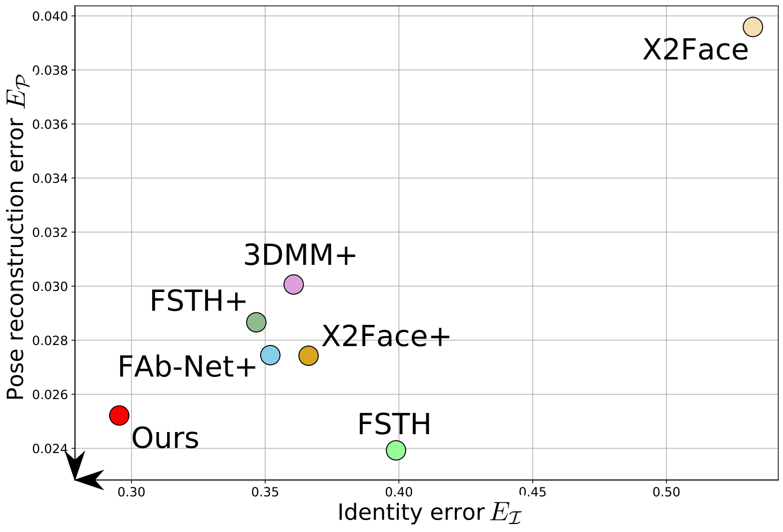

Figure 12: Evaluation of reenactment systems in terms of their ability to represent the driver pose and to preserve reference identity (arrows point towards improvement).

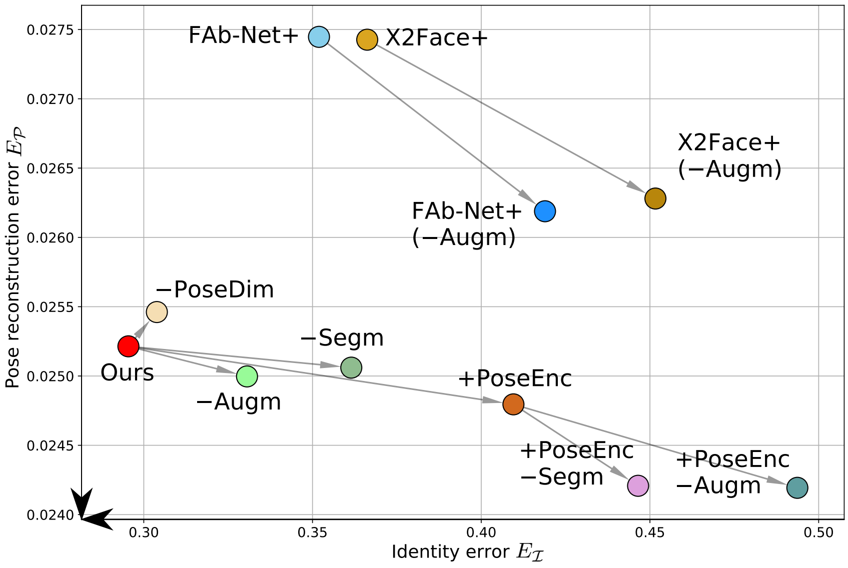

Figure 13: Quantitative evaluation of how ablating several important features of the training setup impacts our system (and two related ones). See the full thesis for legend and discussion.

4.3 Head Geometry Capture from Single RGB Image

While two previous chapters study pose capture, with the latter leaning to identity preservation, this one focuses on identity capture exclusively. We are interested in automatic acquisition of a textured 3D model of a human head from few and, most importantly, from a single RGB image.

As before, we rely heavily on learning from data, and we also want our model to generalize well to few (1-2) novel samples. Therefore, we are looking for a meta-learning architecture in the spirit of those two discussed in the previous chapter. Besides, we would like to avoid complicated datasets such as 3D scans or synthetic 3D models.

For these reasons, we turn to neural implicit functions, which have been shown to successfully reconstruct 3D shapes from posed RGB images only [20, 13, 23] and are based on trainable neural

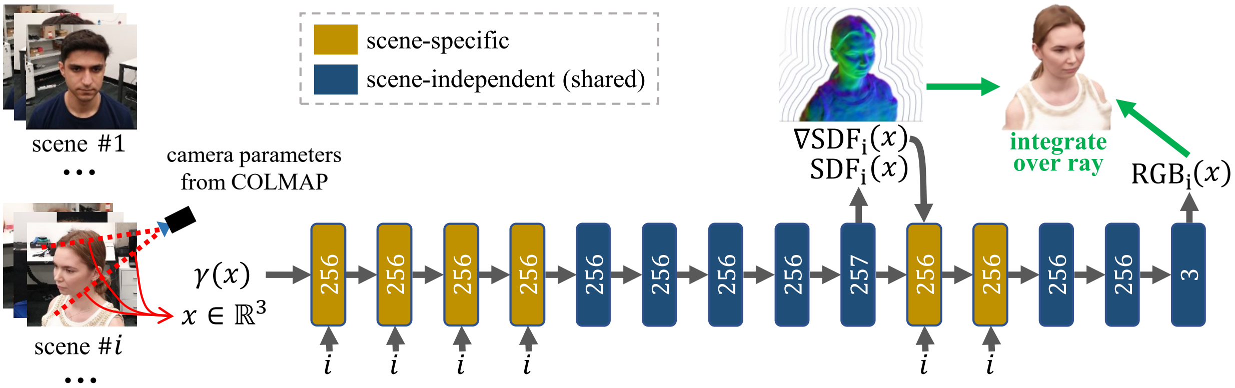

Figure 14: Architecture of Multi-NeuS, a 3D neural implicit function that can represent multiple objects of a class simultaneously (boxes depict fully connected layers and their output dimensionalities; γ is the positional encoding function). Since some layers (blue) are shared between all scenes, they can learn class priors to then transfer knowledge to novel scenes of the same class, enabling few-shot reconstruction. The model is trained via volume rendering and pixelwise loss, just like NeuS [20], but on a dataset of multiple scenes. Afterwards, when fitting to an unseen object, scene-specific layers (yellow) are fitted first, and finally all layers are fine-tuned together.

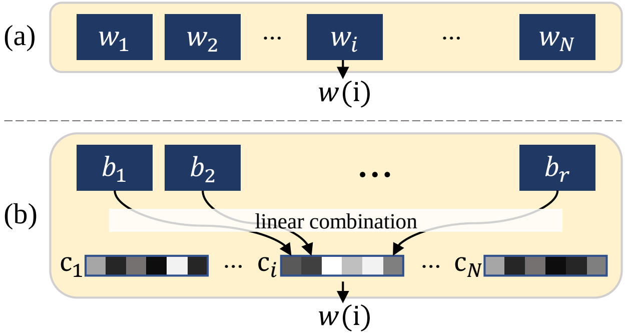

Figure 15: The two options for scene-specific layers we have explored within Multi-NeuS, *independent* (a) and *low-rank* (b). They are fully connected layers whose weights and biases $w(i)$ depend on scene index i. An *independent* layer learns individual weights and biases for each of N scenes, while a *low-rank* layer learns r copies and then linearly combines them with each scene's own learnable coefficients.

networks. Namely, our base architecture is NeuS [20] which is a modification of NeRF [12] for non-transparent objects. It is a shallow regression network that consumes a 3D coordinate and outputs signed distance to the object surface and RGB radiance. The training dataset is pictures of an object and their camera parameters. During training, a random image pixel is picked and points are sampled on an imaginary camera ray casted through it. The network predicts the pixel intensity (computed through the rendering equation), which is compared against the ground truth in the loss function.

Our solution is called Multi-NeuS (Figure 14). We upgrade NeuS to fit N scenes simultaneously by creating N copies of scene-specific NeuS instances that share *some* of the layers, while keeping other layers unshared (*scene-specific*). We then fit these N instances to the scenes simultaneously, while optionally imposing additional structural regularization on scene-specific layers (Figure 15).

First of all, we fit Multi-NeuS to a subset of SmartPortraits [10] which consists of $N = 107$ short smartphone videos of still people. Then, to reconstruct a new person, we add a new $(N + 1)$-st set of scene-specific layers to the model, initialize them with the average of the first N sets, retrain

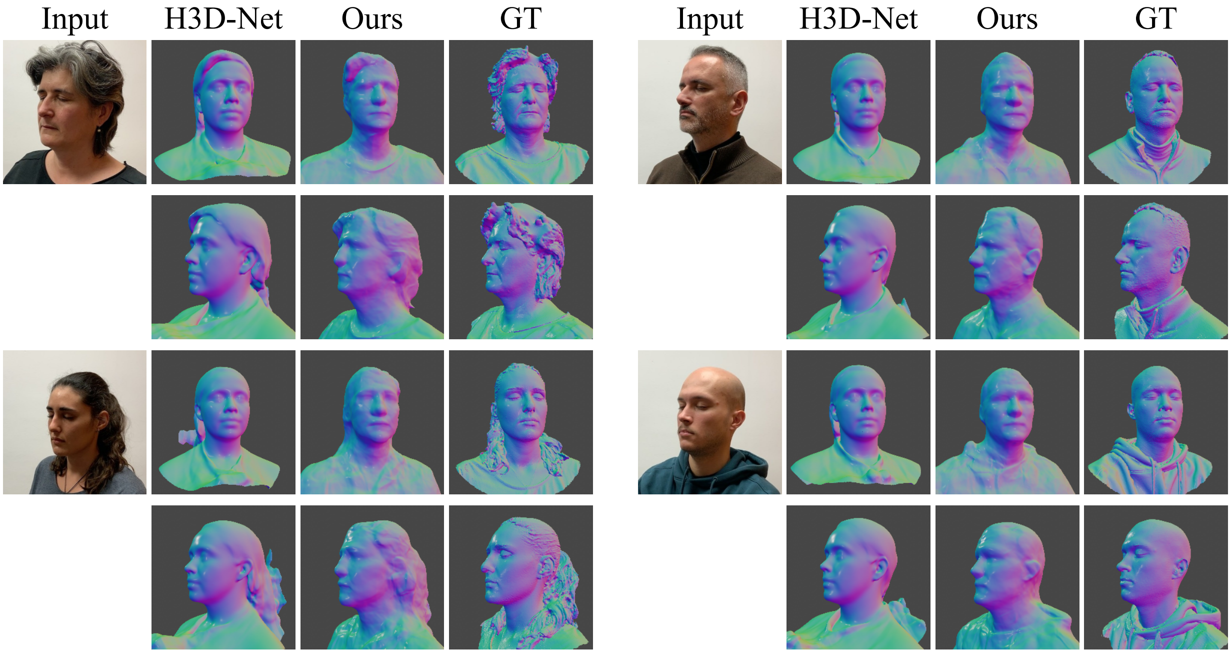

Figure 16: Single-view mesh reconstruction on the first four scenes of the H3DS dataset.

	face				head			
Input view	*F*	*L*	*R*	*mean*	*F*	*L*	*R*	*mean*
H3D-Net 3-view	-	-	-	1.34	-	-	-	10.53
H3D-Net 1-view	**1.82**	1.83	1.91	1.85	13.83	**13.01**	12.51	13.12
Ours 1-view	1.89	**1.77**	**1.86**	**1.84**	**13.00**	13.27	**11.95**	**12.74**

Table 2: Mesh reconstruction error (Chamfer distance) on the H3DS dataset. Lower is better, ”F/L/R” are for ”frontal/left/right”.

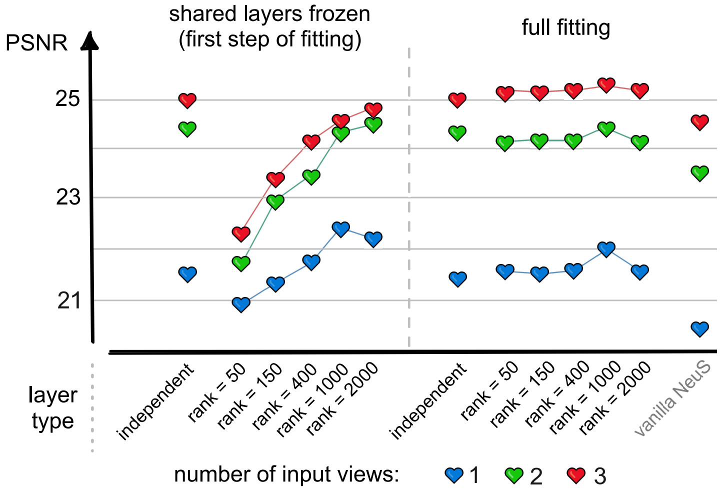

Figure 17: Quality of novel view reconstruction depending on scene-specific layer type. Lower-rank metamodels underfit during the first step of fitting; higher rank models fit better and provide a more convenient initialization for the second step of fitting (fine-tuning of all weights).

just them to fit the image(s) of the new person, and finally unfreeze and fine-tune all layers and the camera parameters. If camera parameters for an in-the-wild photo are not available, we roughly estimate them using an algorithm that predicts 3D facial landmarks [1].

Results. We validate on the H3DS dataset [16] which allows quantitative comparison by providing ground truth 3D scans. Our best model turns out to be no worse than our main rival H3D-Net [16] (Table 2), while being trained on a different and much smaller dataset (1̃00 sets of images vs 1̃0 000 3D scans). In addition, the ”regression-to-mean” effect is smaller (Figure 16).

We also demonstrate additional single-view geometry reconstruction on several in-the-wild photographs and paintings in the figure (available in the full text) and in the supplementary video (`https://shrubb.github.io/research/multi-neus`).

Finally, we conduct extra experiments to validate our architecture choices. There are a quantitative (Figure 17) and a qualitative comparison of the generalization strength for different scene-specific layer settings which reveal that low-rank layer type with $r = 1000$ is optimal for single-view reconstruction. Besides, we run a parameter search to understand which of the NeuS layers we should replace with their scene-specific counterparts (the optimal placement is reflected in Figure 14).

5 Conclusions

This thesis is motivated by numerous challenges in human capture and by the advantages of learning nontrivial patterns from data. Therefore, we have picked 4 challenges where we feel that such data-driven paradigm (e.g. large datasets, end-to-end differentiable neural models, self-supervised learning, meta-learning) is underused in the existing research but could be a good fit.

First, motivated by telepresence systems, we consider the problem of 3D body keypoints estimation. For the single-camera scenario, we propose a simple regression approach, leveraging a large dataset of 3D poses. For the multi-camera scenario, we develop two solutions, each using a convolutional neural network trained to directly minimize the 3D objective for the first time, thus becoming much more robust to occlusions and complex poses, and surpassing previous approaches. Further, we note the inherent disadvantages of body keypoints (as a pose representation in general) revealed by our above research and let a neural network learn a novel *latent* pose representation from an unlabeled dataset of videos. This network is trained within a novel system for cross-person head reenactment which preserves avatar's identity. Finally, we turn to the capture of human appearance, and implement a solution for reconstructing the 3D human head from one or few RGB images. For that, we use a recent self-supervised technique called neural implicit functions which allows our method to be trained on a hundred examples only.

We expect that some of the above results should broaden the practical applications of human capture, for example widen the range of suitable hardware by allowing simple setups, or improve accuracy or prediction quality in general to permit more demanding applications. Not less importantly, our work brings more general and conceptual contributions to the field, such as a simple ”automatic” method for self-supervised feature disentanglement (tailored to human heads) and a meta-learnable neural implicit function. Our experiments have also highlighted some avenues of future work, for instance: enforcing our latent pose to be free of identity by involving more disciplined feature disentanglement methods, e.g. from the specialized literature on generative models; more compact weight sharing mechanisms for faster and more precise learning of multi-scene neural implicit functions.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In *Proc. ICCV*, pages 1021–1030, 2017. 20

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 9

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, 2014. 10, 14

[4] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, September 2008. 15

[5] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, 2018. 13

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. ECCV*, 2018. 14

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 12

[8] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 11, 12

[9] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3D human pose regression. *arXiv*, 1804.10462, apr 2018. 13

[10] Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, and Gonzalo Ferrer. Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis. In *Proc. CVPR*, June 2022. 19

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 13

[12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 19

[13] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. ICCV*, 2021. 18

[14] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1253–1262, 2017. 13

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. *arXiv*, abs/1811.11742, 2018. 13

[16] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proc. ICCV*, 2021. 20

[17] Tomas Simon, Hanbyul Joo, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *CVPR*, 2017. 12

[18] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 13

[19] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking Pose in 3D: Multi-stage Refinement and Recovery for Markerless Motion Capture. In *2018 International Conference on 3D Vision (3DV)*, pages 474–483. IEEE, sep 2018. 13

[20] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proc. NeurIPS*, 2021. 18, 19

[21] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proc. BMVC*, 2018. 15

[22] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *arXiv preprint arXiv:1812.01598*, 2018. 12

[23] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Proc. NeurIPS*, 2021. 18

[24] Zhixuan Yu, Jae Shin Yoon, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi 1.0: Human multiview behavioral imaging dataset. *arXiv preprint arXiv:1812.00281*, 2018. 11

[25] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proc. ICCV*, 2019. 14, 15

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017. 14