We thank the reviewers for their extensive comments. We are particularly grateful for the engagement and effort that clearly went into our reviews, and all reviewers seem happy with the substantial potential impact of our approach, and the importance of the problem we're addressing. As R4 pointed out, there are plenty of applications in which energy minimisation was previously successful but has yet to be combined with modern feature learning techniques, because of the limitation in propagating gradients through their solutions. Our method addresses that specific issue, with a framework general enough to be suited for any kind of energy functions, as R6 noted. That said, some reviewers raised significant objections based on a different interpretations of the paper, and we wish to clarify here.

1. **Where is the novelty (R2+R4) / What is the point of the new proofs (R2)?** R2 and R4 are correct in that we're showing that a stabilised variant of [1] works. However, our primary result is to show *why* it works. To this end we: (i) showed that the result of [1] could be re-derived as a fixed point of the Newton method. (ii) Observed that replacing Newton's method with a more stable trust-region based method gave rise to a more stable fixed-point (line 131), and (iii) characterised the relationship between the two updates using an existing result from the optimisation literature (line 141).

2. **Partial derivatives vs full-derivatives (R2).** We believe the confusion here comes from the use of $H_{x^*}^{-1}$ to compress notation from line 124 onwards. Every sub-scripted x^*, should be understood as a fixed-point arising from the previous use of Newton's method and not as dynamically varying. Given this, partial derivatives and full derivatives coincide. In the final version, the use of partial derivatives in eq. 9, and clarifying remarks at line 119 will make this clear.

3. **'Wiberg optimisation is alternation (see [4]), and an inappropriate description for our work' (R6).** Wiberg methods should be understood as a *replacement* for alternation, whereby some variables are replaced by a function that characterises their minimum value, as a function of the remaining values leading to more informative higher-order gradients and faster convergence. This mischaracterisation by R6 is our fault; we had intended to cite Fitzgibbon's later work [†] which corrected the mischaracterisation of [4], rather than [4] itself.

4. **R6 requested better baselines.** We emphasise that we're modifying a baseline [24] that was published independently as state-of-the art, and as such many obvious tweaks have already been done. In particular: (i) [24], and our baseline, made use of the Human3.6M PPCA basis of [23]; (ii) [24] also performed grid-search over the basic parameters governing the Huber loss, the regularizer coefficient, and the fusion of poses. Much of the benefit to our approach comes from updating the basis shape and regularizer (taken from PPCA) to be consistent with the new joint detections (note the large decrease in performance from [24] to our baseline). We will add an additional baseline to the final version showing this.

5. **Computational bottlenecks and comparisons.** More generally grid-search, although effective, simply does not scale to modern problems defined over large datasets with even moderate numbers of parameters. A single pass over the training set of Human3.6M using the Huber norm takes around 5 hours on a standard CPU. Grid-search using 10 values over k parameters would take $5 \cdot 10^k$ hours. The table below shows the parameters involved in the human pose experiment (‡ were optimised with grid-search in [23] and [24]). The tracking experiments had a total of 964 parameters.

Component	body models	reprojection	cameras	covariance	Huber ‡	fusion ‡	scaling ‡
Number of parameters	3978	92	36	2523	9	6	4

In comparison, our stochastic online approach took approximately 120 core hours, for more than $6,500$ parameters, to converge fully (i.e. substantially faster than searching over 2 parameters). All sets of parameters (as R6 was wondering) are trained simultaneously. This is necessary as most parameters are tightly coupled, and some parameters (Huber loss and poses fusion) are tuned on the distribution of the residuals, and their optimal values change as the basis and cameras alter the residuals.

6. **R4 and R6 requested further applications showing tighter integration with deep learning approaches.** This is an area of on going work, requiring substantial engineering worthy of a future paper, and can not be described alongside current work within the page limit.

7. **Additional comments.** All issues raised by the reviewers will be clarified. We thank R2 for pointing out papers we missed in our literature review, and we will include them in the final version of the paper. As R4 suggested, we will also add more examples of energy based optimisation problems (to which our method is applicable). R6 flagged the use of 'network' instead of 'parameters of the energy function' in the pose experiment; we agree the name should be changed. R1 asked about the 'frames missing at random' tracking experiment: each sequence is randomly split between training and testing set, meaning that both sets contain the same sequences but no frame is presented in both. We will change this notation to talk about sampling the training/test sets at the frame and sequence level, and discuss in the text.

[†] J. H. Hong and A. W. Fitzgibbon. Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In IEEE International Conference on Computer Vision (ICCV), 2015.