We would like to sincerely thank all the reviewers for reading the paper carefully and their very valuable feedback.

General comments:

Visualization of vanishing/exploding gradients: Similarly, as in Fig. 4 from [2], we plotted L_2-norms of the gradients $\frac{\partial \mathcal{L}}{\mathbf{x}_t}$ as a function of time t at the beginning of the optimization and after 100 iterations (first two plots in Fig. 1). We see that the norm of the gradient for ODEtoODE barely changes, while for the NeuralODE it converges to 0 as we backpropagate through time (we observed convergence to 0 also for other methods, yet for the clarify of the picture we did not present additional curves). The plots were created for Humanoid training from the paper, where ODEtoODE was **the only method** that successfully trained the agent.

Presentation: in the final version we will simplify Sections 1-4 and provide definitions of ISO/Gated-ODEtoODE in the Introduction. As suggested, we will improve citations' style. In Section 5.2.3, we will provide more detailed discussion, and incorporate new results (see: Table from Fig. 1). Training ODEtoODEs does not require any additional modules since the parameters governing the evolution of weight matrices are unconstrained and therefore can be handled by standard backpropagation in the supervised setting and standard GD-approach with ES-gradients for the RL one. We do agree that this requires clarification, thus in the final version we will incorporate an algorithmic box on ODEtoODEs in the experimental section for clarity.

Reviewer 1:

ANODEV2: Thank you for pointing this out. The HyperNets networks we compare against in Sec. 5.2 (see: Table 1, 2) are precisely ANODEV2 architectures (as indicated by our citation [56] to ANODEV2 in l.265). This is also the case for the ES experiments, and thus citation [26] (which is a typo) in l.238 of Sec. 5.1 should be replaced by [56].

Pros & cons of orthogonality:

Even though in principle deep architectures leveraging orthogonal matrices can hurt accuracy by restricting model capacity, a rich line of work ([2,30,40]) on orthogonal RNNs and related methods shows that they do not, if designed correctly. In fact, the main motivation behind these architectures are accuracy improvements due to training not suffering from vanishing/exploding gradients - one of the critical problems in training deep machine learning systems. We also demonstrate it exhaustively in the empirical section, covering both: the supervised and RL setup (the latter usually not touched in the papers on the subject). To summarize, ODEtoODEs outperform accuracy-wise other methods (in particular highly competitive ANODEV2) in **11 out of 15** considered supervised tasks, as we clearly state in the paper (l.272). In the RL setting ODEtoODE is **the only method** that manages to train the most difficult Humanoid task and, as we demonstrate in "*General comments*" section above, it is precisely due to its ability to stabilize gradients' lengths.

Theoretical results: We are not aware of **any** results in the literature on neural networks (even in the simpler setting with shallow discrete-architectures & supervised learning) on the convergence to global minimum under weak assumptions from our Theorem 1 and to the best of our knowledge we are the first to show convergence to local minima with depth-independent bounds in the more challenging ES scenario. Given that, we interpret our theoretical results as the strength of the paper. Lemma 4.1 is implicitly used in the proof of our Theorem 1 (see: Eq. 21 in the Appendix) and techniques used to prove it (Sec. 8) play key role in establishing depth-independent bounds. In the final version of the paper we will explicitly refer to it. Lemma 4.1. is also a natural continuous extension of the analogous results for the orthogonal RNNs (see for instance: [2]), so can be used in various analogous continuous variants, not only ODEtoODEs. We will clarify it in the final version.

ResNets in supervised setting: We run these additional experiments and present results in Table from Fig. 1.

Reviewer 3:

Thank you very much for the review.

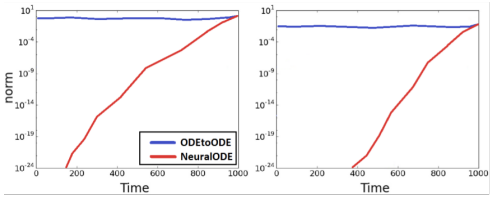
Mathematical notation: we will incorporate suggested improvements and clarify in the final version (in particular, by "strong convergence results/guarantees" we mean "compelling"). **L.102:** This is a generic notation with parameters encapsulated in ψ. We will clarify this in the final version.

Visualization of gradients' norms: See, our paragraph on visualization in "*General comments*" section.

Reviewer 4:

Thank you very much for the review.

Visualization of gradients' norms: See, our paragraph on visualization in "*General comments*" section.



Tasks	Dotted Lines	Spatter	Stripe	Translate	Rotate	Scale	Shear	Motion Blur	Glass Blur	Shot Noise	Identity
ResNet-1	92.89	93.6	30.06	25.97	82.59	58.11	91.86	76.55	91.22	96.15	97.65
ResNet-10	92.9	93.56	27.96	26.1	83.25	63.56	91.44	72.42	89.16	94.31	97.38
ODEtoODE-1	**95.42**	**94.9**	**44.51**	**26.82**	83.9	66.68	**93.37**	**78.63**	**93.91**	**96.91**	**97.94**
ODEtoODE-10	95.22	**94.9**	44.37	26.63	**84.1**	**66.76**	92.93	77.58	93.29	96.71	97.88

Figure 1: Showcasing gradient vanishing & extra experiments with ResNets. First two plots present norm of the gradient $\frac{\partial \mathcal{L}}{\mathbf{x}_t}$ as a function of t for Humanoid training after 0 (first plot) and 100 iterations (second plot). Time $= 1000$ corresponds to $T = 1.0$ from the paper. Table presents additional results on ResNets and comparison with ODEtoODEs. **Bolded blue** correspond to best results and **they are all** for variants of ODEtoODEs.