We thank the reviewers for their detailed comments. We first address a few common queries and then follow with comments specific to individual reviewers.

Linearity and comparisons: Several reviewers comment on a lack of comparison with classical methods. We emphasize that the background wave speed, σ, and, hence, the forward operator A_σ are *unknown* (lines 37-38). The dependence of A_σ on σ is very nonlinear. As pointed out by Reviewer 1, this makes the problem a challenging and novel variant of traditional linear inverse problems, out of reach of the classical approaches. (In practice σ in, say, reflector imaging is inferred separately using, crucially, multiple sources rather than just one as in our paper.) It is truly a problem that was unlocked by machine learning. Unfortunately, most iterative or learning-based solutions assume a known forward operator. This is in particular true for the deep image prior and Tikhonov- and sparsity-regularized inverses. There are thus no fair comparisons that can be made with such methods. Further, the true forward operator in our inverse problems is non-linear; our networks are modeled after a certain linear approximation but are themselves strict non-linear generalizations. The closest papers we refer to, for example, Fan and Ying, assume a smooth *known* background. We compare to the U-Net because the U-Net is incredibly successful on a variety of imaging tasks including the ones we address, regularly outperforming the classical methods. Our main message is that this (if we may say so) *annoying effectiveness* is limited to the training dataset for hard, very non-convolutional wave imaging problems. By modeling the physics via the routing network, the FIONet robustly generalizes out-of-distribution (OOD).

Clarity/reorganizing: Several reviewers state that the paper would benefit from reorganization. Our first strategy was to describe the core novelty and FIO in the main paper and give technical details in appendices. The reviews make it clear that this is unfriendly to people unfamiliar with wave imaging. Thankfully they show a clear path towards improving clarity: move some appendix material to main text (explanations and quant. results), expand the intro on wave imaging, and move some figures and derivations to the appendix. These changes are not hard and we would be pleased to effect them. We thank Reviewer 2 for pointing out issues with notation which we addressed, and to Reviewer 4 for specifying the hard-to-follow parts. We are confident that the reorganization will make things clearer.

Reviewer 1—FIONet outside wave imaging? We reiterate that wave-based imaging encompasses a vast range of modalities in practice today (lines 25-27, 42-44). Therefore, we do not find the focus on wave imaging particularly limiting. That said, the routing network idea applies whenever there is direction-dependent transport, for example in fluid flows or ray tracing. **Figure 7**: Figure 7 shows how wave packets oriented along different directions, ν (red line in the Figure) move over a given background. The two warped grids show two directions of propagation after time T (Appendix C explains the two directions). The learned warping is exactly what the physics dictates. **Renaming**: Thank you for pointing us to the original "routing network". We will add a comment and use "wave packet routing networks".

Reviewer 2—Large filters and the U-Net? The main reason for the discussion of how the U-Net implements large filters is theoretical rather than practical: we argue that the FIONet approximates FIOs by connecting to earlier analytic results. You are right (and we will emphasize this better) that there are many ways to implement large filters. The practical reason we chose the U-Net is that, empirically, it excels at convolutional and to some extent pseudodifferential problems. One may say that the routing network unwarps the input in such a way that it is a simple job for the U-Net. We did try a variety of other architectures including dilated convolutions and direct implementation of filters in the Fourier domain; the U-Net performed the best. **No paired data?** Training with ys alone might be accomplished via self-supervision. Usually A is known, but knowing it up to a class could suffice! This is exciting, thank you! We will try it out asap, but it is out of the scope of this paper which introduces the interpretable geometric architecture. Let us add that strong OOD generalization allows us to train on simulated data and test on real data. We can thus learn to solve the problem "offline" and then apply to real measurements without fear of overfitting.

Reviewer 3—Noisy Measurements: We do have results for noisy measurements in Appendix A.3. We trained on clean data and tested on 10dB noisy data. Note that in our problems the inverse is L^2-stable so additive noise will not cause big problems. Training with noise improves results but not dramatically; we would be happy to add those results. The U-Net is excellent at removing additive noise and artifacts but it cannot handle geometry even without noise. **Artifacts in interferometer imaging:** Indeed FIOs associated with canonical graphs can exhibit artifacts in case of caustics (Appendix E). In fact, a single source in our reflector imaging experiment develops caustics. Interestingly, we show in Appendix E that the FIONet handles it well. The explanation is that our routing network can learn arbitrary warpings, not just diffeos. (The training data was generated with kWave which simulates full wave physics, so we commit no inverse crime.) **Empirical OOD:** Though motivated by theory our OOD findings are indeed empirical. We will emphasize this. **Relevance to ML**: There is a burgeoning interest in physics-inspired deep learning architectures as pointed out by Reviewer #2, as well as in OOD generalization. **Comparison to Fan and Ying**: While refs. [18,19,20,21] are also physics-inspired, they address pseudodifferential operators (contained in FIOs) with simple geometry (singularities do not propagate). Their OOD experiments are combinations of training scenarios (e.g., 2 triangles in training, 4 in testing), whereas we completely change the data. They do not compare to a strong learning baseline like the U-Net which would almost certainly generalize well.