We would first like to thank the reviewers for especially detailed and high quality reviews.

Main points

Notation, the overloaded use of L_s, R1: Thanks for a very detailed break-down of issues with the notation. We wanted to emphasize that the action L_s is very general and so we used it in multiple definitions, as we felt that we did not want to introduce too many symbols. That said, this may have had the opposite effect of being confusing. One way we could resolve this would be to add the domain of the action as a superfix, for example $L_s^X : X \to X$ or $L_s^S : S \to S$ or $L_s^{F(X)} : F(X) \to F(X)$, where $F(X)$ is a space of functions on X. This will also disambiguate the difference between an action in the input space and on the activations. We shall definitely spend more time explaining these subtleties to readers. When it comes to specific examples of actions, as you suggest, it may just be better for us to use specific notation, for instance T_s for a translation or R_s for a rotation.

Examples of actions on whole signals, R1: Indeed scale transformations (with bandlimiting) are actions on the whole signal. Simple blurs are another example of a semigroup action on the whole signal. What we found interesting is that while an action acts on the whole input, the first layer of convolution lifts that input on X onto the semigroup S. The induced action on these lifted activations always acts on just the domain, which as you point out allows us to use pointwise nonlinearities. In our experiments with scale, we first lift to scale-space, which is why in the scale-equivariant correlation that we ended up using, it appears we only ever act on the domain.

Related work, R2: Thank you for drawing these works to our attention. The paper "Multigrid Neural Architectures" (Ke et al., 2017) is indeed close to ours, and we should be able to make a comparison in the camera-ready version of our paper. "Feature Pyramid Networks for Object Detection" (Lin et al., 2017) and "Multiscale Dense Networks for Resource Efficient Image Classification" (Huang et al., 2018) are also very interesting and have now been added to our related work section.

Novelty, R2: By our own estimation and based on the reviews from **R1** and **R3**, we feel that the mathematical exposition we provide contains an appropriate level of mathematical rigor. Furthermore, we feel it neatly dovetails the prior corpus of work on group equivariance. Indeed the requirement for equivariance led to some not-so-obvious conclusions, for instance, we only need to bandlimit once at the input of the network and not at every layer. The use of nonlinearities actually increases the frequency content of each consequent activation, which from a signal-processing perspective would indicate we need bandlimiting at each layer. Without the equivariance perspective, we would never have arrived at our current solution. Indeed, in Dilated Residual Networks (Yu et al., 2017), there is no bandlimiting at all (we presume it it probably learned in the first convolution). Moreover, maybe the use of dilated convolutions in the end "does not seem surprising" as you state, but primarily we were interested in providing mathematically solid and principled arguments behind why we make certain architectural decisions.

Experiments, R2: Of course we can perform more extensive empirical evaluations of the ideas set down in our paper, and we intend for this to be the subject of follow up works. We believe this is essential to establish the utility of equivariant methods in general. For the time being, we felt that one large dataset (Cityscapes, which was used in Dilated Residual Networks), one medium sized dataset (PCam—digital histopathology can naturally benefit from scale), and an introspective experiment were enough for a proof-of-concept. Furthermore, in related works there are no "standard datasets". For instance, in Multigrid Neural Architectures test on CIFAR-100, a toy MNIST segmentation task, and ImageNet, but in Feature Pyramid Networks, the authors look at COCO.

More unstructured input spaces, R3: Thanks for your very positive review, hence why this is the only point directed at you. We do intend to extend the current work to other domains and to other kinds of semigroup action. Indeed scale-spaces on graphs would be an interesting area to pursue, since diffusion equations are often deployed in Graph CNNs.

Minor points

Accuracy vs. training data amount, R1: This is a good idea and we agree it sits inline with the motivations for building in inductive biases, and thus we shall try to add this to the camera-ready version.

Kanazawa et al, R1: Thanks for pointing this out, we have now added it to the related work.

Equation 22, R1: Indeed $L'_s[x]$ is a right-action, we think the notation $R_s[x]$ would resolve this issue.

Pretraining, R1: This would be an interesting experiment to run, we shall try to fit it into the camera-ready paper.

Equation 14, R1: This is our mistake, L_s should have been defined for functions on X, thanks for catching this.

Rotation example, R1: We shall make this example more explicit, to clear up confusion.