To All Reviewers

Main novelty: Existing non-local models can only be sparsely inserted into the original network backbones, because either over-high complexity of the non-local operator (*e.g.*, [Wang et al.2018]) or the lack of multi-scale information (*e.g.*, [Chi et al.2019]). As pointed by R2, the proposed FFC is the first work that implements "a single conv unit which combines local and non-local information". Moreover, the complexity of FFC is comparable to vanilla convolution. These facts collectively enable FFC to directly replace vanilla convolutions in modern deep networks, achieving mixed receptive fields (local / semi-global / global) at each layer.

Cross-scale fusion: We would use empirical results to justify the necessity of cross-scale fusion (or inter-path transitions). For example, on ImageNet, using same parameters (*e.g.*, $\alpha = 0.25$), FFC with all cross-scale fusion achieves top-1 accuracy of 77.6%. Removing global-to-local fusion or local-to-global fusion reduces the accuracy to 76.6%, 76.2% respectively. Removing $f_{l\to g}, f_{g\to l}$ in Fig. 1 only strikes an accuracy of 75.6%. Similar observations are found on other benchmarks. Unfortunately these results were not included in the current draft due to our unwise page space organization. We will surely include the ablation studies in the revision.

To R1

Our responses for the major concerns (difference with [Chi et al.2019], cross-scale fusion, and inter-path transitions) can be found in the Section "to all reviewers" of this rebuttal.

R1 requested "I3D + FFC v.s. I3D + NL". We are sorry that the experimental log of I3D + FFC is not successfully retrieved from our server. Nonetheless, for reference, Table 5 reports the accuracies of both C2D + FFC and C2D + NL, which are 73.5 v.s. 73.8. The corresponding GFLOPs are 20.2 v.s. 30.7. In comparison, the accuracy of original C2D is 71.9. We conclude that FFC and NL are similar in accuracy, but FFC is more efficient. Moreover, FFC / NL are complementary (FFC-C2D + NL -> 74.9).

Thanks for suggesting AA-ResNet (ICCV19). We will include its ResNet-50 results (77.7% v.s. FFC 77.8%) of as suggested in the revision.

To R2

For questions 1 and 2, please refer to the Section "to all reviewers" of this rebuttal. The suggested "channel shuffling" essentially implements the same function to our current design (if we understand this suggestion correctly). However, its efficacy in comparison to FFC is unclear to us at this moment.

Table 3 investigates the final performance with or without LFU under different α. It is observed that FU (global scale) / LFU (semi-global scale) are consistently complementary. We will conduct additional trials with only LFU as suggested in the revision.

R2 suggested "do multiplication at spatial domain directly". This applies to spectral 1x1 conv owing to the convolution theorem. However, it is not the case for spectral ReLU, which has a thresholding step (doing the job of frequency band-passing) that has no spatial correspondence.

FFC does channel splitting by the scheme of "Some groups for local operations and some for global ones". The resultant benefits are two-fold: it can be implemented by build-in group convolution in PyTorch. Moreover, the performance on CIFAR-100 using this scheme is slightly better than other alternatives, which serves as an indicator of better empirical choice.

To R3

For the key novelty of the work and benefit of cross-scale fusion, please check Section "to all reviewers" of this rebuttal. We are not confident about the implication of "intermediate scales" as in the reviewing comment. FFC combines three scales: local (by vanilla convolution), semi-global (via local Fourier unit), and global (via Fourier unit). If "intermediate scales" was referring to the semi-global scale that operates on image patches, Table 3 investigates the effect of LFU.

To R4

We indeed have included the requested comparisons in the submission, including non-local networks [Wang et al.2018], OctConv [Chen et al.2019a] and SRL [Chi et al.2019]. Dilated convolution is typically not chosen as a baseline in the literature of non-local models due to its inferior performance. Please see Tables 5 and 6 for more details. We fully agree with the reviewer that ablation studies are crucial for understanding each part of FFC. Please refer to Section "to all reviewers" of this rebuttal, where we provide detailed experimental results on ImageNet.