Paper ID: 2499 ARMA Nets. We thank all reviewers for their valuable feedback, and we are encouraged that all reviewers found our work important and our method novel. **Major updates of the paper: (1)** To address the concerns from R3 and R4, we extend Theorem 1 to take into account dilated convolutions and non-uniform layer coefficients.

Theorem 1 (ERF of a linear ARMA vs CNN network) *Consider an L-layer linear ARMA network, where the l-th layer computes $y^{(\ell)}[i] - a^{(\ell)} y^{(\ell)}[i-1] = \sum_{p=0}^{K^{(\ell)}-1} (1/K^{(\ell)}) \cdot y^{(\ell-1)}[i - d^{(\ell)} p]$. Suppose $0 \leq a^{(\ell)} < 1, \forall \ell \in [L]$, the radius of its ERF is*

$$r(ERF)_{ARMA} = \sqrt{\sum_{\ell=0}^{L-1} \left\{ \left[(d^{(\ell)} K^{(\ell)})^2 - 1 \right] / 12 + a^{(\ell)} / (1 - a^{(\ell)})^2 \right\}}$$

When $a^{(\ell)} = 0, \forall \ell \in [L]$, the ERF radius of the resulted CNN is $r(ERF)_{CNN} = \sqrt{\sum_{\ell=0}^{L-1} [(d^{(\ell)} K^{(\ell)})^2 - 1]/12}$.

The result reduces to $\sqrt{L} \cdot \sqrt{(K^2-1)/12 + a/(1-a)^2}$ as in the paper if all layers are identical and the dilation is 1. **(2)** We change the baseline for semantic segmentation to *"Non-local U-Net for Biomedical Image Segmentation"* (we add a global aggregating module both at the bottom and up-sampling blocks in U-Net) following R3's suggestion.

Table 1: **Semantic segmentation on ISIC dataset**. For all metrics (ACC, SE, SP, PC, F1 and JS), higher values indicates better performance. The reported numbers are an average of 10 runs with different seeds.

Model	params.	ACC	SE	SP	PC	F1	JS
Non-local U-Net	4.403M	0.945 ± 0.003	0.877 ± 0.017	**0.973 ± 0.004**	0.844 ± 0.014	0.831 ± 0.012	0.741 ± 0.013
ARMA-U-Net	3.455M	0.955 ± 0.003	0.896 ± 0.011	0.972 ± 0.005	**0.873** ± 0.011	0.861 ± 0.007	0.780 ± 0.009
Non-local ARMA-U-Net	4.405M	**0.960** ± 0.002	**0.909** ± 0.009	0.968 ± 0.004	0.870 ± 0.011	**0.870** ± 0.006	**0.790** ± 0.008

R1 - Theoretical and practical analysis of FFT truncation. The direct application of FFT/DFT assumes the input is periodically extended (a.k.a. circularly padded). The boundary artifacts by various padding are numerically analyzed in [r1], and we plan to support reflective padding to mitigate potential artifacts following [r1]. In our applications, we found such artifacts are imperceptible since the boundary pixels are mostly background.

[r1] Aghdasi F, Ward RK. Reduction of boundary artifacts in image restoration. IEEE TIP. 1996 Apr;5(4):611-8.

R1 - Large-scale experiments such as MS-COCO segmentation. We evaluate our approach on an advanced (challenging) but controllable task of video prediction — our method adapts to predicting objects moving at different speeds, which requires adaptive ERF. We agree that segmentation with varying object sizes is an interesting future direction.

R1 - Related works in spatial recurrent neural networks. Most prior works consider nonlinear RNNs, where the activation between recursions prohibits an efficient FFT-based algorithm. On the contrary, ARMA is linear between recursions, thus the recursions reduce to a single deconvolution. We will add this discussion to our related works.

R3 - Confusing statement "ARMA nets are complementary to the aforementioned architectures..." By "complementary", we mean the methods expand the ERF via different ideas. We agree the wording could be confusing and will modify it. We prove *dilated convolution* is complementary to ARMA in the updated Theorem 1, since dilation ($d^{(\ell)}$) and auto-regression ($a^{(\ell)}$) contribute to different terms of the ERF. Furthermore, our lightweight ARMA layers can be used on high-resolution features where *non-local block* could be too expensive in memory and computation.

R3 - Missing citations in Section 7. Section 7 is a short conclusion and discussion section; we will add citations to justify our claims. Specifically, we will add [19, 26] to impulse response filters, [4, 15, 21, 25, 29] to spatial RNNs, and "Miyato, Takeru, et al. *Spectral normalization for generative adversarial networks.*" to spectral normalization.

R3 - Inappropriate baseline for semantic segmentation. We thank the reviewer for suggesting a more appropriate baseline, and we have updated our experiments (with 10 runs) according to the suggestion. As shown in the updated Table 1, our ARMA layer is complementary to non-local block — The non-local ARMA U-Net outperforms both ARMA U-Net and non-local U-Net on most metrics.

R3 - The choice of our video prediction baseline. The suggested baseline [31] by the reviewer is for video classification instead of pixel-level video prediction. To the best of our knowledge, the state-of-the-art video prediction models (such as ContextVP [5], PredRNN++ [32]) rely on a ConvLSTM backbone. As shown in Figure 8 (Appendix A.2), the non-local blocks are inserted between layers, while LSTM mechanism is used between steps: **the LSTM** learns a long-term dynamic through **time**, while **non-local/ARMA layer** captures large receptive field over **spatial domain**. Therefore, the functionalities of LSTM and non-local layer are not overlapping, and our baseline is valid and reasonable.

R4 - Strategies to select autoregressive coefficients. Since autoregressive coefficients are learnable, we can initialize them to zeros and let the networks decide the proper coefficients automatically. In Figure 6 (Line 302), we demonstrate how the networks learn different range of autoregressive coefficients given specific tasks.

R4 - The effect of using different autoregressive coefficients at different layers. We update Theorem 1 as in the beginning of the response — The ERF area (squared radius) is actually a summation of the one for each layer.

R4 - Experiments on regression and generative tasks. Video prediction is indeed a dense regression problem, since a model needs to predict a continuous value for each pixel. Due to space limit, We leave generative task as future work.