Rebuttal for ID 41. We would like to thank the reviewers for their time and thoughtful comments.

[R1] *"The goal of the replacing convolutions with local self-attention is a bit in contradiction..."* Self-attention has several advantages over convolutions, even when restricted locally.

- In contrast with convolutions where each position shares the same kernel weights, multi-head self-attention generates local kernels that can have different weights per position due to the computation depending on content-content interactions.
- Additionally, local self-attention is more parameter efficient than convolutions: using a 7×7 local self-attention layer outperforms using a 3×3 convolution while having $3\times$ fewer parameters. Furthermore, the 7×7 local self-attention layer has $2.4\times$ fewer FLOPs than a 3×3 convolutional layer.

[R1] *"it seems that the parameters of the convolutional layer have been replaced by the parameters ... W_q, W_k and W_v."* We view the attention mechanism as a method for manufacturing convolutional kernels based on the content of a given location. In some sense, this is a relaxation of locally-connected layers by not requiring that the kernels be identical across spatial locations. The relaxation goes further by allowing the weights themselves to depend dynamically on the content of each image.

[R1] *"Why are positional features important for self-attention"* Without positional information, attention will not be sensitive to the ordering of the pixels because it will only use content-content interactions. Convolutions implicitly carry a relative positional encoding by having weights that depend on relative distance.

[R1] *"One could consider that using only the positional interaction is a degenerated form of convolution"* We agree that the importance of the content-relative interaction is surprising and concur in its similarity to traditional convolution, but expect that in future investigations on more challenging tasks than classification the relative importance of content-content interactions will increase.

[R1] *"throrough comparison of CNN and the proposed self-attention from a computational point of view [...] and the expected behaviour and properties.* For kernel size k, channels d, convolution cost scales as k^2d^2 FLOPs per position with k^2d^2 parameters, while self-attention cost scales as $3d^2 + k^2d + kd$ FLOPs per position with d^2 parameters.

[R3] *"Why on Table 1 for ResNet-50 is full attention better than convolution-stem + attention?"* In the cases where full attention outperforms convolutional-stem with attention, the difference is small ($\leq 0.2\%$) and can likely be explained by variance in training runs. In the final version, we will add error bars to capture the variance.

[R3] *"...enlarging the spatial extent k in attention improves performance but plateaus off at 11×11..."* We will experiment with larger k in the final version. We suspect that the effect of changing k is task dependent.

[R3] *"What if you have binary/illusory/sketch images where you may need attention in the first place?"* While this work is focused on demonstrating the attention can be used as a fundamental primitive for building vision models, studying the performance on different input domains is an exciting future direction, as is understanding the relative merits of convolution and attention beyond standard classification and detection tasks. Other study directions include benchmarking performance of convolutional vs. attentional models on transfer and self-supervised settings.

[R3] *"grammatical errors and typos. Also somehow most of the references were missing in the paper."* We apologize for accidentally clipping 4 pages of references section and any grammatical errors. We have addressed all of these issues and will restore the references in the revised manuscript.

[R4] *"downsampling is carried out with average pooling with stride 2... instead of increasing the stride of the self-attention layer"* We tried this downsampling approach in early experimentation and found it slightly underperforms compared to average pooling. However, this experiment was conducted on a preliminary architecture, so we plan on running experiments to benchmark this conceptually simpler approach on our final architecture.

[R4] *"not clear what self-attention can learn with respect to convolution, and what would happen with deeper models"* We agree that a more rigorous study of the modeling and optimization capabilities of attention and convolution would be illuminating. We leave this to future work. However, one clear difference is attention can generate a different kernel per position based on content, while convolution uses the same kernel for every position.