https://arxiv.org/pdf/1805.11604.pdf

BN is considered to have reduced internal covariate shift -- but instead what it really may be doing is:

1. improve the Lipschitzness of the gradients of the loss function
    - this means gradients are more reliable and predictive
    - after taking a step in a gradient direction, you can be more confident that this is the right gradient direction for the next step as well
    - so we can use a larger learning rate
2. improve the Lipschitzness of the loss function itself
    - the loss landscape changes at a smaller rate
    -