We thank the reviewers for their careful reading of the manuscript and their constructive suggestions.

Reviewer-1, Weak baseline (baseline ciphertext is not packed), comparison against MPC+HE baselines: First, our RHECNNS baselines, LoLa [7] and CHET [9], use the ciphertext packing technique shown in Fig.4. LoLa [ICML'19] and CHET [PLDI'19] are the state-of-the-art HE-enabled neural networks. Compared to secure Multi-Party Computation (MPC), HE supports non-interactive operations and greatly reduces the communication cost. Gazelle [13] is described at lines 73 - 75 of our paper. Gazelle is based on interactive MPC+HE, so it suffers from huge communication overhead between clients and server (1.2 GB per CIFAR-10 image). The MPC protocol requires clients to be always online to communicate with servers. In the setting of Machine Learning as a Service, it is difficult for some clients to stay online and share the data by a stable and high speed connection during the entire service. Our Falcon is proposed to reduce the non-interactive HE operations, NOT the MPC+HE hybrid operations.

Reviewer-1, Why not ReLU activation?: (1) Our baselines LoLa [7] and CHET [9] show the square polynomial approximation have competitive accuracy, compared to the original ReLU on MINIST and CIFAR-10. (2) We use the SAME activation approximation as LoLa [7] and CHET [9] for a fair comparison.

Reviewer-2/3, Novelty (Comparison against E2DM [CCS'18]): E2DM [Jiang et.al at CCS'18] is not the state-of-the-art secure inference. One of our baselines CHET [9] claims it has better performance than prior works like E2DM. We would like to point out that we compared our work against state-of-the-art RHECNNs, including CHET and LoLa that obtain better performance than E2DM on CIFAR-10. E2DM is shown effective in matrix multiplications, not in the modern convolution operations which are well-studied by CHET [7] and LoLa [9]. In addition, E2DM is mainly proposed to reduce the number of multiplications, but our Falcon is proposed to reduce the number of rotations that are much more expensive. We will add E2DM into related work and compare it against our Falcon in the revised version.

Reviewer-2, Rotations in Costache's paper [23]: We agree that Costache's paper [23] has not heavy rotations. Instead of [23], homomorphic DFT [25] heavily depends on rotation. In our Falcon, [25] works as the baseline since it is the state-of-the-art homomorphic DFT and outperforms the Costache's work [23]. We show that our Falcon outperforms the state-of-the-art homomorphic DFT [25] in section 3.

Reviewer-3, Novelty (Difference with ENSEI (Bian Song, et al. CVPR'20)): Firstly, we would like to point out that ENSEI and our work Falcon target on different cryptography protocols. ENSEI adapts interactive HE+MPC setting, but Falcon uses non-interactive HE setting. We described the fact that MPC-based ENSEI suffers from high communication overhead in line 73 - 75. **Simply adopting ENSEI in non-interactive HE-based networks is NOT trivial**. This is because cheap DFT and IDFT in plaintext domain CANNOT be performed by clients in the non-interactive HE setting, instead expensive homomorphic DFT and IDFT are required to be performed by servers. As Figure 1 shows, LoLa-S (adopting ENSEI into LoLa) simply using DFT on non-interactive HE setting prolongs the baseline LoLa's latency because of the expensive and essential homomorphic DFT and IDFT operations. Our work Falcon proposes efficient homomorphic DFT in the algorithm 1 to solve the above problem. Second, ENSEI is only shown effective in the convolutional layers, not the fully-connected layers. This is because dot-product operations in the fully connected layers CANNOT be directly applied into convolution theorem. In contrast, Falcon uses block-circulant matrix to support underlying dot-product operations, so both convolution and fully-connected operations are supported well in our work. We will highlight that LoLa-S in figure 1 refers to the spectral-version LoLa using the similar method in ENSEI. Falcon's novelty and contributions can be concluded by three points: 1. We propose efficient homomorphic DFT and IDFT algorithms. 2. We use block-circulant matrix to support efficient spectral convolution and fully-connected operations in encrypted data. 3. Our experiments show that Falcon can be applied into any non-interactive HE networks for reducing expensive HE operations.

Reviewer-2/3, Typos and References formats: Thanks for reviewers' correction. We will fix them in the revised version. Especially thanks for the advice of reviewer 3 on "Changing the name $HReLU$ into $HSquare$ or $HActivation$."