Synopsis of the paper

- The manuscript introduces a residual learning framework for training very deep convolutional neural networks by reformulating target mappings as residual functions added via identity shortcut connections (Eq. (1); Fig. 2; Sec. 3.1–3.2; p.2–3). It presents architectures up to 152 layers on ImageNet and over 1000 layers on CIFAR-10, with bottleneck blocks to control complexity (Fig. 3–5; Sec. 3.3; Table 1; p.3–6). Empirical evaluations show improved optimization behavior and accuracy compared to plain networks, with state-of-the-art results on ImageNet classification (Tables 3–5; Sec. 4.1; p.5–6) and substantial gains in detection/localization tasks (Sec. 4.3; Tables 7–14; p.8–12).

Summary of Review

- The paper addresses the degradation problem in deep networks with a simple, well-motivated residual formulation and rigorous, large-scale experiments (Fig. 1; Sec. 1; Sec. 3.1–3.2; Sec. 4.1; p.1–6).
- Empirical evidence is broad and convincing across ImageNet, CIFAR-10, COCO, and VOC, with clear architecture descriptions and training protocols (Fig. 3–6; Table 1; Sec. 3.4; Sec. 4.1–4.3; p.3–12).
- However, theoretical justification for why residual learning improves optimization is largely intuitive; formal analysis is limited and deferred (Sec. 3.1; statement referencing future work; p.3). No direct evidence found in the manuscript.
- Some implementation details and ablation analyses (e.g., role of BN, shortcut type choices, initialization sensitivity) could be expanded for reproducibility and to isolate causal factors (Sec. 3.4; Table 3 options A/B/C; Fig. 4; p.4–6).

Strengths

- **Clear identification of the degradation problem and motivation**
  - The manuscript documents increased training error with depth in plain nets on CIFAR-10 and ImageNet, motivating the need for better formulations (Fig. 1; Fig. 4 left; Sec. 1; p.1,4). This matters for technical soundness, grounding the contribution in observed failure modes.
  - The constructed-solution argument (identity layers) highlights optimization gaps without overfitting explanations (Sec. 1; p.1–2), improving conceptual clarity.
  - The residual formulation F(x)+x is presented succinctly with minimal changes to standard CNNs (Eq. (1); Fig. 2; Sec. 3.1–3.2; p.2–3), aiding broad impact and usability.

- **Simple, parameter-efficient architectural design with strong empirical gains**
  - Identity shortcuts add no parameters; projections are used only for dimension changes (Eq. (2); Fig. 3; Sec. 3.2–3.3; p.3–4), demonstrating efficiency—important for scalability.
  - Bottleneck blocks (1×1–3×3–1×1) enable depths 50/101/152 while keeping FLOPs below VGG-16/19 (Fig. 5; Table 1 FLOPs; Sec. 3.3; p.5–6), showing practical resource-conscious design.
  - Consistent validation gains as depth increases (Tables 3–4; Fig. 4 right; p.5–6) indicate robustness and impact.

- **Comprehensive large-scale experiments and cross-task generalization**

  - ImageNet classification results: single-model top-5 error 4.49% (ResNet-152) and ensemble 3.57% on test (Tables 4–5; Sec. 4.1; p.6), evidencing state-of-the-art performance.
  - CIFAR-10: successful optimization to 110 and 1202 layers, with training/testing curves and analysis of residual response magnitudes (Fig. 6–7; Table 6; Sec. 4.2; p.7–8), supporting claims of optimization ease and behavior.
  - Detection/localization: substantial mAP improvements on COCO/VOC and top-5 localization error reduction to 9.0% (Tables 9–14; Sec. 4.3; p.8–12), demonstrating transferability and real-world utility.

- **Clarity of network specifications and training protocol**

  - Architecture layouts and downsampling positions are detailed (Fig. 3; Table 1; Sec. 3.3; p.3–5), improving reproducibility.
  - Training settings (batch size, LR schedule, augmentation, BN placement) are documented (Sec. 3.4; p.4), providing practical guidance.

Weaknesses

- **Limited theoretical grounding for optimization improvements**

  - The core hypothesis that residual functions are easier to optimize is argued qualitatively; no formal convergence or landscape analysis is provided (Sec. 3.1; p.3). This matters for technical soundness and generalization beyond reported regimes.
  - The manuscript conjectures exponentially low convergence rates for deep plain nets without empirical diagnostics beyond curves (Sec. 4.1; p.5). No direct evidence found in the manuscript.
  - The relationship to prior shortcut/gated architectures (e.g., highway networks) is descriptive; conditions under which identity shortcuts outperform gates are not theoretically characterized (Sec. 2; p.2–3).

- **Ablation coverage and factor isolation are incomplete**

  - The roles of batch normalization, initialization, and learning-rate warm-up are acknowledged but not isolated via controlled ablations (Sec. 3.4; Sec. 4.2 warm-up note; Fig. 6 middle; p.4,7). This impacts experimental rigor and reproducibility.
  - Shortcut variants A/B/C are compared, but quantitative analysis of where projection shortcuts help most (e.g., layer-wise) is limited (Table 3; Sec. “Identity vs. Projection Shortcuts”; p.6).
  - The claim that identity shortcuts are sufficient for addressing degradation lacks targeted tests removing BN or altering normalization to confirm necessity/sufficiency (Fig. 4; Sec. 3.2; p.4–6).

- **Notation and consistency issues in mathematical formulations**

  - The residual block uses y = F(x;{Wi}) + x (Eq. (1); p.3), but activation placement (”second nonlinearity after the addition”) could be ambiguous across implementations; clearer operator order diagrams would help (Fig. 2; Sec. 3.2; p.3).
  - The explanation of when W_s projections are required versus identity with zero-padding could be clarified with explicit dimensional constraints and stride interactions (Eq. (2); Fig. 3; p.3–4).
  - Discussion of response magnitudes (std after BN) would benefit from explicit definitions and aggregation procedures (Fig. 7 captions; Sec. 4.2; p.8), as current text is concise but not fully

formal.

- **Reproducibility and resource reporting gaps**
  - While FLOPs are listed, memory footprints, training wall-clock, and hardware specs per model/depth are not reported (Table 1; Sec. 3.4; p.4–6), limiting practical adoption planning.
  - Detection/localization pipelines include multiple improvements; some choices (e.g., fixing BN statistics during fine-tuning) could use more justification and sensitivity checks (Appendix A; p.10).
  - For the 1202-layer CIFAR-10 model, overfitting is hypothesized without regularization studies (Table 6; Fig. 6 right; Sec. 4.2; p.7–8).

Suggestions for Improvement

- **Strengthen theoretical framing of residual optimization advantages**
  - Provide a formal analysis or empirical diagnostics (e.g., loss landscape curvature, gradient norms, or Hessian spectra) comparing plain vs. residual blocks across depths and datasets (Sec. 3.1; Fig. 4–6; p.3–7), clarifying mechanisms behind improved convergence.
  - Characterize when identity shortcuts are preferable to gated/projection alternatives with assumptions on activation/normalization, possibly via controlled synthetic studies (Sec. 2–3; p.2–4).
  - Include a discussion connecting residual learning to known preconditioning interpretations, with measurable quantities (No direct evidence found in the manuscript).
- **Expand ablations to isolate critical components**
  - Conduct BN-off/BN-on comparisons, different initialization schemes, and learning-rate warm-up/no warm-up across depths to quantify contributions (Sec. 3.4; Sec. 4.2; p.4,7).
  - Provide layer-wise analyses for options A/B/C showing where projection shortcuts materially change gradients or activations (Table 3; p.6).
  - Test necessity/sufficiency: e.g., plain nets with BN and identity shortcuts selectively removed to validate the attribution of gains to residual formulation (Fig. 4; p.4–5).
- **Clarify mathematical and notation aspects of blocks**
  - Add explicit diagrams or equations specifying activation order (pre-activation vs. post-activation variants), with consistency across figures and text (Fig. 2; Sec. 3.2; p.3).
  - Detail dimensionality constraints and stride behaviors requiring W_s, including zero-padding exact rules and potential artifacts (Eq. (2); Fig. 3; p.3–4).
  - Formalize residual response statistics: define computation points (post-BN/pre-ReLU), aggregation across layers/batches, and report summary tables with confidence intervals (Fig. 7; Sec. 4.2; p.8).
- **Enhance reproducibility and practical reporting**
  - Add memory usage, training time, and hardware details per model; include scalability guidance for different batch sizes and GPUs (Table 1; Sec. 3.4; p.4–6).

  - In detection/localization, justify fixing BN during fine-tuning with ablations (on/off) and discuss effects on mAP/latency (Appendix A; p.10–11).
  - For CIFAR-10 1202-layer model, run regularization ablations (dropout/maxout/weight decay settings) to substantiate overfitting claims and report best practices (Table 6; Fig. 6 right; p.7–8).

References

- [16] Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML, 2015 (appears in manuscript's reference list; cited in Sec. 3.4; p.4).
- [41] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. ICLR, 2015 (appears in manuscript's reference list; used as comparisons; Tables 3–4; p.5–6).
- [44] Szegedy, C., et al. Going deeper with convolutions. CVPR, 2015 (appears in manuscript's reference list; used as comparisons; Table 4; p.6).
- [32] Ren, S., He, K., Girshick, R., & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS, 2015 (appears in manuscript's reference list; cited in Sec. 4.3; p.10–11).