

How Much Does Depth Matter? An Input-Modality Ablation Study of Lightweight Convolutional Grasp Detection on the Cornell Grasping Dataset

Jake Mancuso

AI Engineer & Computer Vision Specialist
Mancuso Intelligence & Defense (MIND) LLC
Dayton, OH, USA
jakeycuso@gmail.com

Abstract—Planar antipodal grasp detection is a core perception capability for robotic manipulation, and modern pipelines increasingly rely on lightweight fully convolutional networks that regress pixel-wise grasp quality, orientation, and gripper width directly from images. These networks are commonly fed depth images, color images, or fused RGB-D input, yet published accuracy figures for different input modalities come from experiments that also vary in architecture, augmentation, and training protocol, making the isolated contribution of depth sensing difficult to judge. In this work we hold a 1.9M-parameter generative residual grasp network and its entire training procedure fixed and vary only the input modality (depth-only, RGB-only, and fused RGB-D) on the Cornell Grasping Dataset, evaluating each configuration with 5-fold cross-validation under the standard rectangle metric. Across 15 training runs, depth-only input reaches $92.3 \pm 0.9\%$ mean grasp-detection accuracy, RGB-only reaches $91.5 \pm 2.2\%$, and fused RGB-D reaches $90.8 \pm 1.9\%$: under an identical, modest training budget no modality difference is statistically significant (paired t -tests across folds, all $p > 0.2$), while depth-only training is the most stable and reproduces the published GR-ConvNet depth result almost exactly. RGB-based configurations learn faster in early epochs but exhibit $2\text{--}3\times$ higher cross-fold variance. Single-image inference at ≈ 11 ms on a consumer Apple M4 GPU (≈ 68 ms CPU) confirms that this class of network remains deployable in closed-loop control regardless of modality. We release code, per-epoch logs, and scripts that regenerate every number and figure in this paper.

Index Terms—robotic grasping, grasp detection, convolutional neural networks, RGB-D perception, ablation study

I. Introduction

Grasping is the gateway skill of robotic manipulation: before a robot can use, sort, or assemble objects it must first pick them up. Four decades of research have moved the field from analytic grasp synthesis built on precise object models and contact mechanics toward data-driven methods that predict grasps directly from sensor data [1], [2]. Within this family, the planar grasp detection problem, predicting an oriented antipodal gripper configuration in the image plane, has become a standard benchmark task, driven by the Cornell Grasping Dataset and the rectangle representation of Jiang et al. [3].

A practical question for anyone building a grasping cell is what sensor the perception stack actually needs. Depth cameras add cost, calibration burden, and failure

modes (glare, transparent and specular objects, ambient IR interference), while plain RGB cameras are cheap and ubiquitous. The literature offers indirect guidance: some influential systems are depth-only [4], [5], others RGB-only [6], [7], and others fuse both [8]–[10]. Published numbers, however, are rarely comparable: they differ in architecture, capacity, augmentation, train/test protocol, and even metric details. Reported modality gaps therefore conflate sensor information content with everything else that differs between papers.

This paper asks the narrow, controlled version of the question: holding architecture and training fixed, how much does depth input contribute to planar grasp detection accuracy? We take a lightweight generative residual convolutional network in the style of GR-ConvNet [10], small enough to train on a desktop-class machine and fast enough for closed-loop use, and train it from scratch under three input configurations that differ only in the number of input channels: depth-only (1), RGB-only (3), and fused RGB-D (4). All other factors (loss, optimizer, augmentation, epochs, folds, evaluation code) are byte-for-byte identical.

Our contributions are:

- a controlled input-modality ablation of a modern lightweight grasp detection network on the Cornell Grasping Dataset, evaluated with 5-fold cross-validation (15 training runs) under the standard rectangle metric;
- an independent reimplement and reproduction study of the GR-ConvNet family, with an honest account of where our numbers agree with and deviate from published results;
- inference-latency measurements on consumer Apple silicon (CPU and Metal GPU), relevant to low-cost robot cells that lack discrete GPUs;
- a fully reproducible pipeline: every number, table, and figure in this paper is regenerated by a script from logged training data.

II. Related Work

A. From analytic to data-driven grasp synthesis

Classical grasp synthesis assumed known object geometry and searched for contact configurations satisfying force- or form-closure; Bohg et al. [1] survey this transition to data-driven methods that generalize to unseen objects by learning from labeled examples, simulation, or self-supervised robot experience. Self-supervised collection at scale was demonstrated by Pinto and Gupta [11] with 50k grasp attempts, and by Levine et al. [12] with an “arm farm” of 14 robots; QT-Opt [13] extended this to closed-loop reinforcement learning. Simulation-heavy approaches such as Dex-Net 2.0 [5] instead train grasp-quality CNNs on millions of synthetic depth images backed by analytic wrench-space metrics. Recent surveys [2], [14], [15] map this now-broad landscape.

B. Planar grasp detection and the rectangle metric

Jiang et al. [3] introduced the oriented rectangle representation and the associated detection metric on the Cornell Grasping Dataset. Lenz et al. [8] were first to apply deep networks, using a two-stage sliding-window cascade over RGB-D features (73.9% image-wise). Redmon and Angelova [6] reformulated detection as single-shot global regression (88.0%), and Kumra and Kanan [9] brought ResNet-scale backbones to the task (89.2%). Anchor-box detectors were adapted by Chu et al. [16] and Zhou et al. [17], while Asif et al. [18] targeted low-power devices and Ainetter and Fraundorfer [7] coupled RGB-only grasp detection with semantic segmentation.

Most relevant here is the generative (fully convolutional) formulation: GG-CNN [4], [19] regresses pixel-wise grasp quality, angle, and width maps from depth images with a network of under 100k parameters, enabling real-time closed-loop grasping; GR-ConvNet [10] scales the same idea to a 1.9M-parameter residual encoder-decoder and reports 97.7% image-wise on Cornell with RGB-D input, 96.6% with RGB, and 93.2% with depth. This per-pixel dense-prediction style descends from fully convolutional segmentation networks [20]. Our study network follows the GR-ConvNet design.

C. Beyond planar grasps and single datasets

Larger datasets reduce Cornell’s small-sample pathologies: Jacquard [21] provides 54k synthetic scenes, and GraspNet-1Billion [22] a large real-world 6-DoF benchmark. Point-cloud and 6-DoF methods [23]–[25] generalize the output space beyond the image plane. We restrict this study to the planar Cornell setting precisely because its rectangle metric and modest scale make a many-run controlled ablation tractable; Section VII discusses the external-validity limits this imposes.

Unlike prior comparisons across papers, and complementing the single-split modality table in [10], our contribution is a same-everything comparison with cross-validated uncertainty estimates.

III. Problem Formulation

A. Grasp representation

Following [4], [10], a planar antipodal grasp in the image frame is

$$\mathbf{g} = (u, v, \theta, w, q), \quad (1)$$

where (u, v) is the grasp center in pixel coordinates, $\theta \in [-\pi/2, \pi/2]$ is the gripper closing direction in the image plane (grasps are symmetric under rotation by π), w is the jaw opening in pixels, and $q \in [0, 1]$ is a scalar quality. Given the camera intrinsics, a calibrated depth value $z(u, v)$, and the hand-eye transform $\mathbf{T}_{RC}$, an image-frame grasp maps to a robot-frame grasp pose

$$\mathbf{g}_R = \mathbf{T}_{RC}(\pi^{-1}(u, v, z(u, v)), \theta, s w), \quad (2)$$

where π^{-1} back-projects the pixel through the intrinsic model and s converts pixels to meters at depth z . This work evaluates in the image frame, as is standard for the Cornell benchmark.

B. Generative grasp maps

Rather than regressing a single tuple (1), the generative formulation predicts a grasp at every pixel. The network is a function

$$f_\phi : \mathbb{R}^{C \times H \times W} \rightarrow (\mathbf{Q}, \boldsymbol{\Theta}_{\cos}, \boldsymbol{\Theta}_{\sin}, \mathbf{W}) \in \mathbb{R}^{4 \times H \times W}, \quad (3)$$

where $C \in \{1, 3, 4\}$ is the number of input channels (the only quantity varied in this study) and the four output maps hold, per pixel, the grasp quality, the angle encoded as $(\cos 2\theta, \sin 2\theta)$, and the normalized jaw opening. The double-angle encoding removes the wrap-around discontinuity at $\pm\pi/2$: it is continuous, unique on $[-\pi/2, \pi/2]$, and decodes as

$$\hat{\theta}(u, v) = \frac{1}{2} \operatorname{atan2}(\boldsymbol{\Theta}_{\sin}(u, v), \boldsymbol{\Theta}_{\cos}(u, v)). \quad (4)$$

The output grasp is read at the quality maximum,

$$(\hat{u}, \hat{v}) = \arg \max_{(u, v)} \tilde{\mathbf{Q}}(u, v), \quad \tilde{\mathbf{Q}} = G_\sigma * \mathbf{Q}, \quad (5)$$

with a small Gaussian filter G_σ suppressing isolated spurious maxima, and $\hat{w} = w_{\max} \mathbf{W}(\hat{u}, \hat{v})$ with $w_{\max} = 150$ px.

C. Ground-truth encoding and loss

Each Cornell grasp rectangle is converted to training targets by marking the central third of the rectangle along the closing direction: inside this region $\mathbf{Q}^* = 1$, $\boldsymbol{\Theta}_{\cos}^* = \cos 2\theta^*$, $\boldsymbol{\Theta}_{\sin}^* = \sin 2\theta^*$, and $\mathbf{W}^* = \min(w^*/w_{\max}, 1)$; elsewhere all targets are zero. The training objective is the smooth- L_1 (Huber) loss summed over the four maps,

$$\mathcal{L}(\phi) = \sum_{\mathbf{M}} \frac{1}{HW} \sum_{u, v} \ell(\mathbf{M}_\phi(u, v) - \mathbf{M}^*(u, v)), \quad (6)$$

$$\ell(e) = \begin{cases} \frac{1}{2}e^2 & |e| < 1, \\ |e| - \frac{1}{2} & \text{otherwise,} \end{cases} \quad (7)$$

which is less sensitive than L_2 to the sparse large residuals that occur at object boundaries [10].

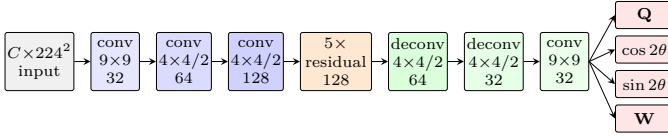


Fig. 1. Study network (GR-ConvNet-style [10]): encoder, five residual blocks, decoder, and four 224×224 output heads. Only the input channel count $C \in \{1, 3, 4\}$ differs between the three modality configurations.

D. Rectangle metric

A predicted grasp rectangle $\hat{r}$ (constructed from $\hat{\mathbf{g}}$ with plate extent $\hat{w}/2$, following [4]) is correct if there exists a ground-truth rectangle r^* such that

$$\frac{|\hat{r} \cap r^*|}{|\hat{r} \cup r^*|} > 0.25 \quad \wedge \quad |\hat{\theta} \ominus \theta^*| < 30^\circ, \quad (8)$$

where $\ominus$ is angular difference modulo π . This is the standard Cornell rectangle metric [3], [8], enabling direct comparison with prior work.

IV. Method

A. Network architecture

We reimplement the generative residual architecture of [10] (Fig. 1). An encoder of three convolutions (a 9×9 stem at stride 1, then two 4×4 stride-2 downsampling stages to 32, 64, and 128 channels) is followed by five residual blocks [26] at 128 channels and 56×56 resolution, then a mirrored decoder of two 4×4 stride-2 transposed convolutions and a 9×9 refinement convolution back to 32 channels at full 224×224 resolution. Four independent 3×3 convolution heads emit the quality, $\cos 2\theta$, $\sin 2\theta$, and width maps. All convolutions are followed by batch normalization [27] and ReLU; 2D dropout [28] with $p=0.1$ precedes the heads. The network has 1.90M parameters (1,898,500 at $C=1$ to 1,901,604 at $C=4$; the modalities differ only in the stem’s input channels, $< 0.2\%$ of parameters).

B. Modality configurations

- D ($C=1$): inpainted depth, normalized per image by subtracting the mean and clipping to $[-1, 1]$;
- RGB ($C=3$): color scaled to $[-0.5, 0.5]$;
- RGB-D ($C=4$): the concatenation of both.

C. Training protocol

All runs use identical code and hyperparameters: Adam [29] at learning rate 10^{-3} , batch size 8, and 30 epochs of 250 iterations each (2,000 augmented samples drawn with replacement per epoch, $\approx 7.5\text{k}$ gradient steps per run), seed 42. Convolution weights use Xavier initialization, following [10]. Images are center-cropped to 350×350 and resized to 224×224 ; training-time augmentation applies a random rotation from $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ and a random zoom in $[0.9, 1.1]$, with grasp rectangles transformed identically. Ground-truth maps are rasterized after augmentation.

TABLE I

Grasp detection accuracy on the Cornell dataset (image-wise split). Ours: mean $\pm$ std over 5-fold cross-validation, best epoch per fold. Published GR-ConvNet values [10] are from a single 80/20 split trained with $\approx 7 \times$ more gradient steps.

Input	Channels	Ours (%)	GR-ConvNet [10] (%)
Depth	1	92.3 ± 0.9	93.2
RGB	3	91.5 ± 2.2	96.6
RGB-D	4	90.8 ± 1.9	97.7

V. Experimental Setup

A. Dataset and protocol

The Cornell Grasping Dataset [3], [8] contains 885 RGB-D images of 240 graspable objects with 5,110 positive grasp rectangles (we discard rectangles containing invalid coordinates); Fig. 2 shows representative samples. We use the image-wise split: the 885 images are shuffled once (seed 42) and partitioned into 5 folds; each configuration trains on 4 folds (708 images) and validates on the held out fold (177 images), yielding $3 \times 5 = 15$ runs. We report, per run, the best validation accuracy over epochs, the standard protocol in this literature [10], [19], and aggregate as mean $\pm$ standard deviation over folds.

B. Implementation and hardware

The pipeline is implemented in PyTorch 2.12 [30] with OpenCV [31] for geometry and rasterization, and runs on a Mac Mini M4 (16 GB unified memory) using the Metal Performance Shaders (MPS) backend; one epoch (250 training iterations plus full rectangle-metric evaluation of the held-out fold) takes ≈ 98 s, and the complete 15-run matrix ≈ 12 h. Depth images are the dataset’s point clouds rasterized to 480×640 and inpainted with the Navier–Stokes method. All code, logs, and figure scripts are released.

VI. Results

A. Grasp detection accuracy

Table I and Fig. 3 summarize the main result. Depth-only input attains $92.3 \pm 0.9\%$ mean accuracy, RGB-only $91.5 \pm 2.2\%$, and fused RGB-D $90.8 \pm 1.9\%$. The spread between the best and worst modality (1.5 percentage points) is smaller than the cross-fold standard deviation of either RGB configuration. Paired fold-wise t -tests confirm that no pairwise difference is significant at $\alpha = 0.05$: D–RGB $+0.8$ pp ($t(4)=0.59$), D–RGB-D $+1.5$ pp ($t(4)=1.51$), RGB–RGB-D $+0.7$ pp ($t(4)=0.64$).

B. Training dynamics

Fig. 4 shows mean training loss and validation accuracy over epochs. The modalities behave differently even though their final accuracies do not: RGB and RGB-D pass 50% accuracy by epoch 5 while depth-only needs 8–9 epochs, but from epoch 15 onward the depth configuration



Fig. 2. Cornell Grasping Dataset samples after the evaluation transform (350×350 center crop resized to 224×224), with ground-truth grasp rectangles overlaid: jaw-opening edges in yellow, plate edges in blue.

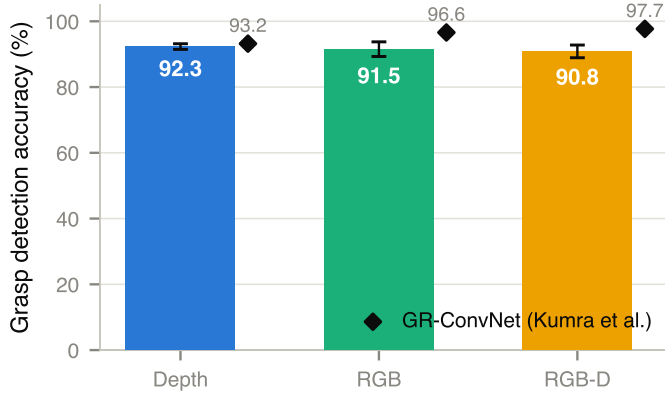


Fig. 3. Mean grasp detection accuracy per input modality with cross-validation standard deviation (error bars). Diamonds mark the published single-split GR-ConvNet results [10].

is the most stable (late-epoch band ± 2 – 3 pp versus ± 8 – 12 pp for RGB), suggesting that color-driven models keep relying on higher-variance texture cues where depth encodes the grasp-relevant geometry more directly.

C. Qualitative behavior

Fig. 5 shows representative validation predictions of an RGB-D model. The quality map concentrates sharply on graspable structure, the width map activates only where quality is high, and the angle map is informative in exactly those regions (elsewhere the atan2 decode of two near-zero maps is noise, which the quality gating makes irrelevant). Failure cases are dominated by thin objects whose predicted rectangle is correct in position but too narrow to clear the 25% Jaccard threshold.

D. Inference latency

Table II reports single-image latency (batch 1, mean of 100 runs after warm-up). On the M4’s GPU all modalities run in ≈ 11 ms (> 85 Hz), comfortably inside the 30–50 Hz closed-loop regime targeted by GG-CNN [4]; even CPU-only inference at ≈ 68 ms supports ≈ 14 Hz visual servoing. The extra input channels change latency by $< 10\%$, so

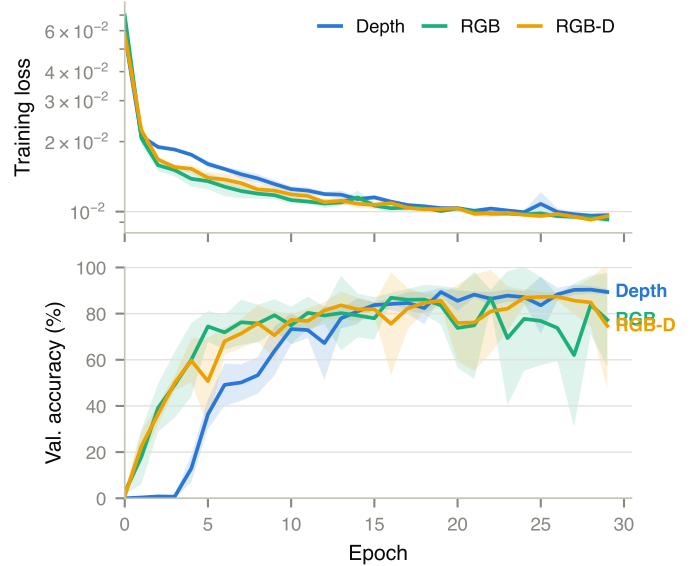


Fig. 4. Training loss (top, log scale) and validation rectangle-metric accuracy (bottom) per input modality: mean over 5 folds, shaded ± 1 std.

TABLE II
Single-image inference latency on Apple M4 (ms, mean $\pm$ std of 100 runs).

Input	Params	CPU	GPU (MPS)
Depth	1.894M	66.7 ± 2.8	10.5 ± 0.1
RGB	1.899M	69.6 ± 2.3	11.2 ± 0.6
RGB-D	1.902M	69.1 ± 2.0	11.5 ± 0.3

modality choice is a sensor-cost decision, not a compute decision.

VII. Discussion and Limitations

A. How much does depth matter?

Under a controlled protocol, the answer is: less than published cross-paper comparisons suggest, at least at this training budget. All three modalities land within 1.5 pp of each other, and the information in a single inpainted depth channel is sufficient to match three color channels or

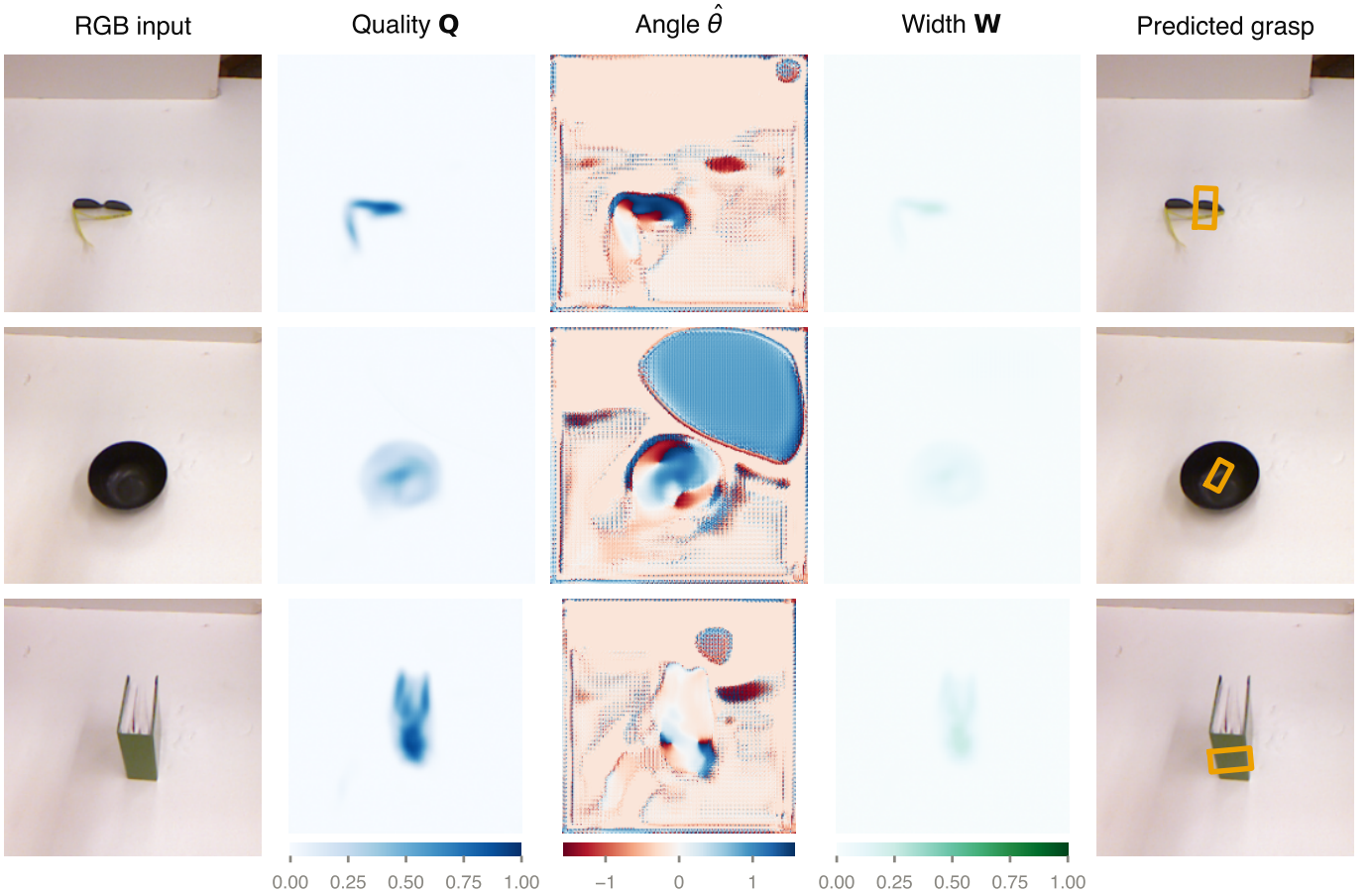


Fig. 5. Qualitative RGB-D predictions on held-out Cornell images (fold 0). Left to right: input, predicted quality $\mathbf{Q}$, decoded angle $\hat{\theta}$, predicted width $\mathbf{W}$, and the extracted best grasp. Angle and width are only meaningful where $\mathbf{Q}$ is high.

the four-channel fusion. This is consistent with the design intuition behind depth-only systems [4], [5]: on a tabletop, grasp geometry is largely a function of shape, which depth encodes explicitly. The practical implication for low-cost robot cells is nuanced: RGB-only performance within ≈ 1 pp of depth-only means a plain camera is a legitimate option when a depth sensor is impractical. Where a depth sensor exists, however, it buys the lowest variance across train/validation splits, which matters when a deployed model is trained once.

B. The reimplement gap

Our depth-only result ($92.3 \pm 0.9\%$) reproduces the published 93.2% [10] within one standard deviation, validating the reimplement. Our RGB and RGB-D results sit 5–7 pp below their published counterparts (96.6%, 97.7%). We attribute this primarily to training budget: our runs use $\approx 7.5\text{k}$ gradient steps versus $\approx 50\text{k}$ in the reference implementation, and Fig. 4 shows color-based configurations still improving and oscillating at epoch 30 while depth has plateaued: depth appears to be the most step-efficient modality, color the most step-hungry. Secondary factors include our coarser 90° -rotation

augmentation and the absence of any pretraining. We report this gap rather than tuning it away because it is itself informative: single high numbers from well-resourced training runs do not automatically transfer to modest reproduction budgets, and modality rankings can invert along the way.

C. Limitations

First, Cornell is small (885 images, 240 objects) and its image-wise split lets near-duplicate views of the same object appear in train and validation folds; absolute numbers are therefore optimistic, as is standard for this benchmark, though the relative modality comparison is unaffected. Second, we evaluate detection accuracy, not physical grasp success; the rectangle metric tolerates grasps a real gripper might miss. Third, our conclusions are specific to $\approx 2\text{M}$ -parameter generative networks at a fixed training budget; larger models or longer schedules may re-open a fusion advantage, as the published RGB-D numbers suggest. Fourth, five folds give limited statistical power ($n=5$ per configuration), so equivalence is our claim only up to detectable effect sizes of a few percentage points.

D. Future work

Natural extensions are: repeating the ablation on Jacquard [21] (where scale removes the small-data caveat) and under the object-wise Cornell split; scaling the training-budget axis explicitly (accuracy versus gradient steps per modality); and closing the loop with physical grasp trials: we plan to deploy the depth-only model on a low-cost 6-axis arm (Arctos-class, wrist RGB-D) to test whether the detection-level equivalence found here survives contact with real gripper tolerances.

VIII. Conclusion

We presented a controlled input-modality ablation of a lightweight generative grasp detection network on the Cornell Grasping Dataset: identical architecture, loss, augmentation, and training budget, with only the input channels varied, evaluated over 5-fold cross-validation. Depth-only input achieved $92.3 \pm 0.9\%$, RGB-only $91.5 \pm 2.2\%$, and fused RGB-D $90.8 \pm 1.9\%$: statistically indistinguishable means, with depth-only training markedly more stable and closely reproducing its published counterpart. At roughly 11 ms per image on consumer Apple silicon, all three configurations support closed-loop grasping. The study suggests that, at modest training budgets, sensor choice for planar grasp detection should be driven by deployment constraints and variance tolerance rather than by expected accuracy gains from fusion, and it provides a fully reproducible baseline for extending the modality question to larger datasets, budgets, and physical trials.

References

- [1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, “Data-driven grasp synthesis—a survey,” *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2014.
- [2] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, “Deep learning approaches to grasp synthesis: A review,” *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023.
- [3] Y. Jiang, S. Moseson, and A. Saxena, “Efficient grasping from RGBD images: Learning using a new rectangle representation,” in 2011 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2011, pp. 3304–3311.
- [4] D. Morrison, J. Leitner, and P. Corke, “Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach,” in *Robotics: Science and Systems (RSS)*, 2018.
- [5] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” in *Robotics: Science and Systems (RSS)*, 2017.
- [6] J. Redmon and A. Angelova, “Real-time grasp detection using convolutional neural networks,” in 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 1316–1322.
- [7] S. Ainetter and F. Fraundorfer, “End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB,” in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 13 452–13 458.
- [8] I. Lenz, H. Lee, and A. Saxena, “Deep learning for detecting robotic grasps,” *The International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, 2015.
- [9] S. Kumra and C. Kanan, “Robotic grasp detection using deep convolutional neural networks,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 769–776.
- [10] S. Kumra, S. Joshi, and F. Sahin, “Antipodal robotic grasping using generative residual convolutional neural network,” in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 9626–9633.
- [11] L. Pinto and A. Gupta, “Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours,” in 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 3406–3413.
- [12] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 421–436, 2018.
- [13] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, “Scalable deep reinforcement learning for vision-based robotic manipulation,” in *Proceedings of the 2nd Conference on Robot Learning (CoRL)*, PMLR 87, 2018, pp. 651–673.
- [14] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, “A survey on learning-based robotic grasping,” *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.
- [15] G. Du, K. Wang, S. Lian, and K. Zhao, “Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.
- [16] F.-J. Chu, R. Xu, and P. A. Vela, “Real-world multiobject, multigrasp detection,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.
- [17] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, “Fully convolutional grasp detection network with oriented anchor box,” in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 7223–7230.
- [18] U. Asif, J. Tang, and S. Harrer, “GraspNet: An efficient convolutional neural network for real-time grasp detection for low-powered devices,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 4875–4882.
- [19] D. Morrison, P. Corke, and J. Leitner, “Learning robust, real-time, reactive robotic grasping,” *The International Journal of Robotics Research*, vol. 39, no. 2–3, pp. 183–201, 2020.
- [20] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
- [21] A. Depierre, E. Dellandréa, and L. Chen, “Jacquard: A large scale dataset for robotic grasp detection,” in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 3511–3516.
- [22] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “GraspNet-1Billion: A large-scale benchmark for general object grasping,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 441–11 450.
- [23] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, “Grasp pose detection in point clouds,” *The International Journal of Robotics Research*, vol. 36, no. 13–14, pp. 1455–1473, 2017.
- [24] A. Mousavian, C. Eppner, and D. Fox, “6-DOF GraspNet: Variational grasp generation for object manipulation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2901–2910.
- [25] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-GraspNet: Efficient 6-dof grasp generation in cluttered scenes,” in 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 13 438–13 444.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [27] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in

Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 448–456.

- [28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 8024–8035.
- [31] G. Bradski, “The OpenCV library,” *Dr. Dobb’s Journal of Software Tools*, vol. 25, no. 11, pp. 120–125, 2000.