Adaptive Multimodal Fusion for Grasping Transparent and Specular Objects

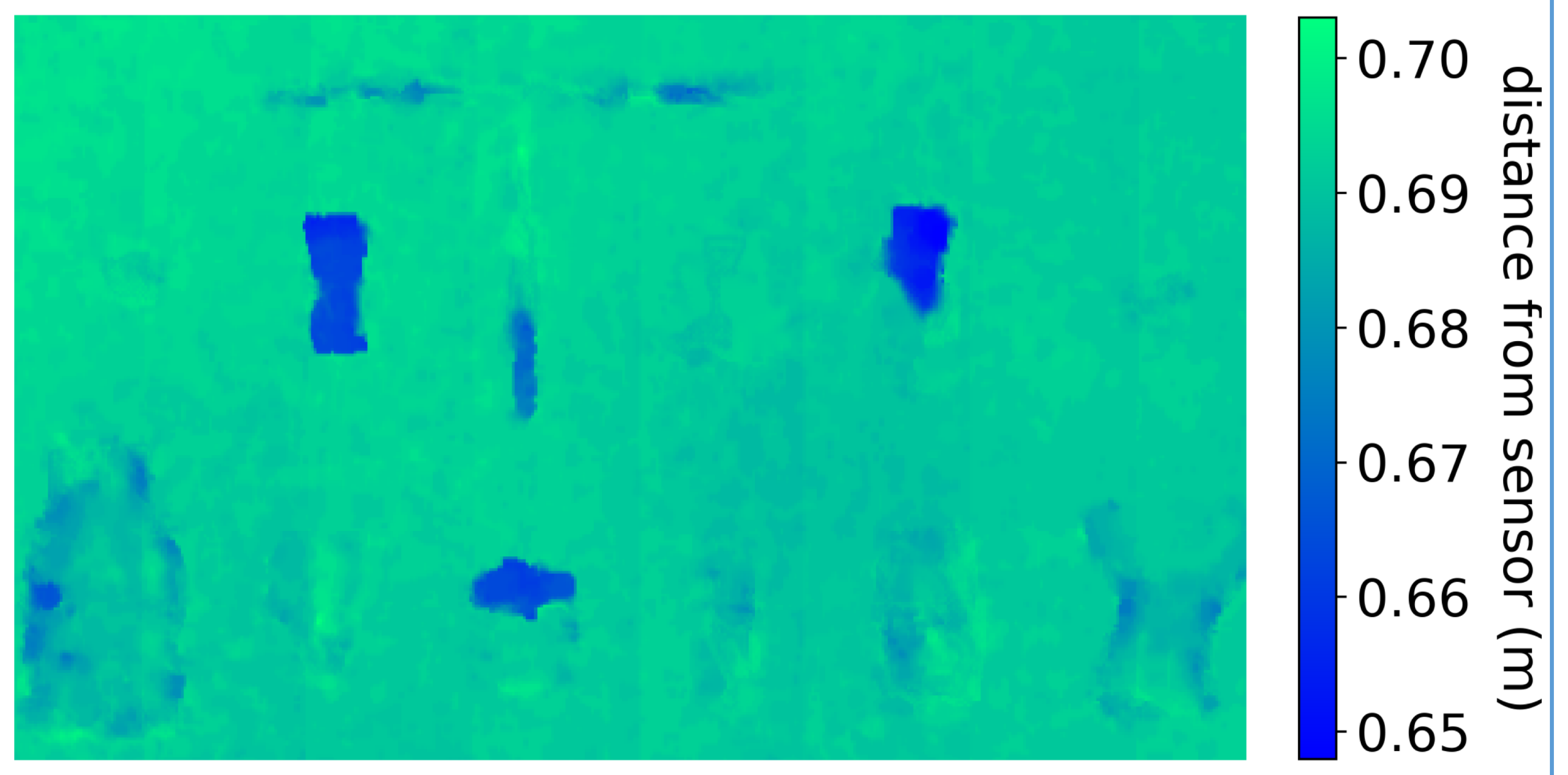
Yimin Tang **Thomas Weng** **David Held**

tangym@shanghaitech.edu.cn tweng@andrew.cmu.edu dheld@andrew.cmu.edu

Introduction

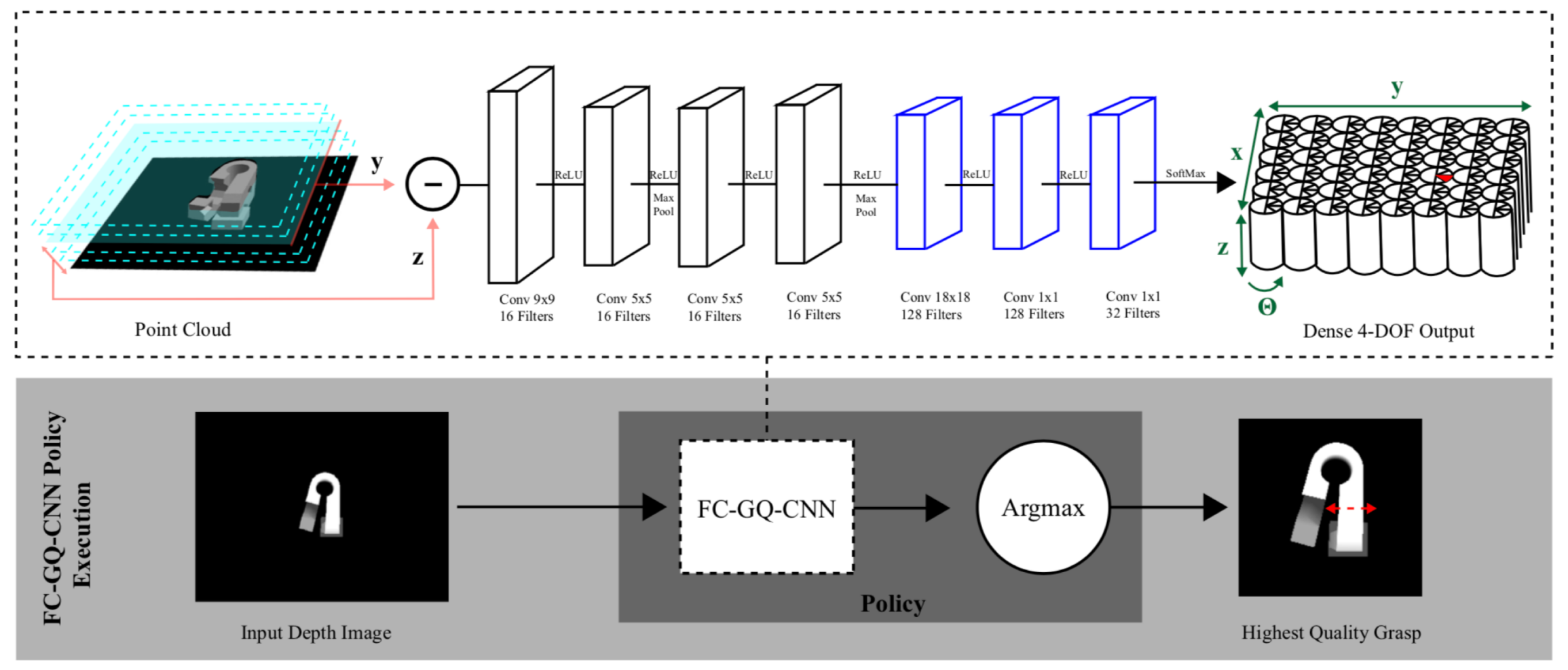
One of the fundamental problems for robotics is object grasping. Many situations and places can use grasping robots to help people complete their jobs and personal issues. However, there are tons of object kinds, object geometries, and material types. We need robust, grasping robots to grasp objects in such a wide range.

RGB image for transparent and reflective objects



Depth image for transparent and reflective objects

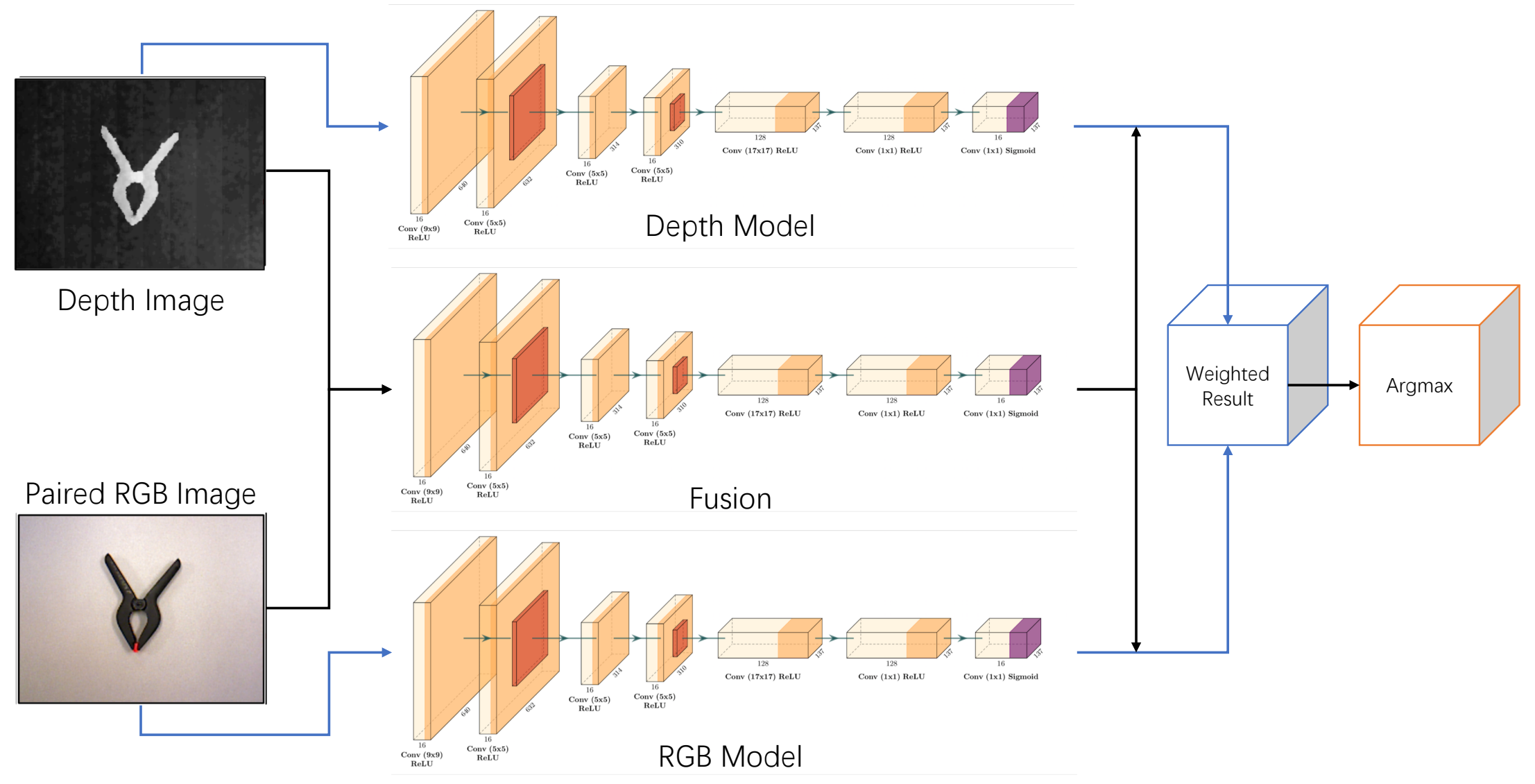
State-of-the-art object grasping methods[4] rely on depth sensing to plan robust grasps, but commercially available depth sensors fail to detect transparent and specular objects. Even for opaque objects if we change light conditions to some particular angle, depth cameras also fail to detect real depth distances.[7] Through depth images, we can training models to predict the success rate for each point and each grasping angel.



State-of-the-art Object Grasping Method Architecture

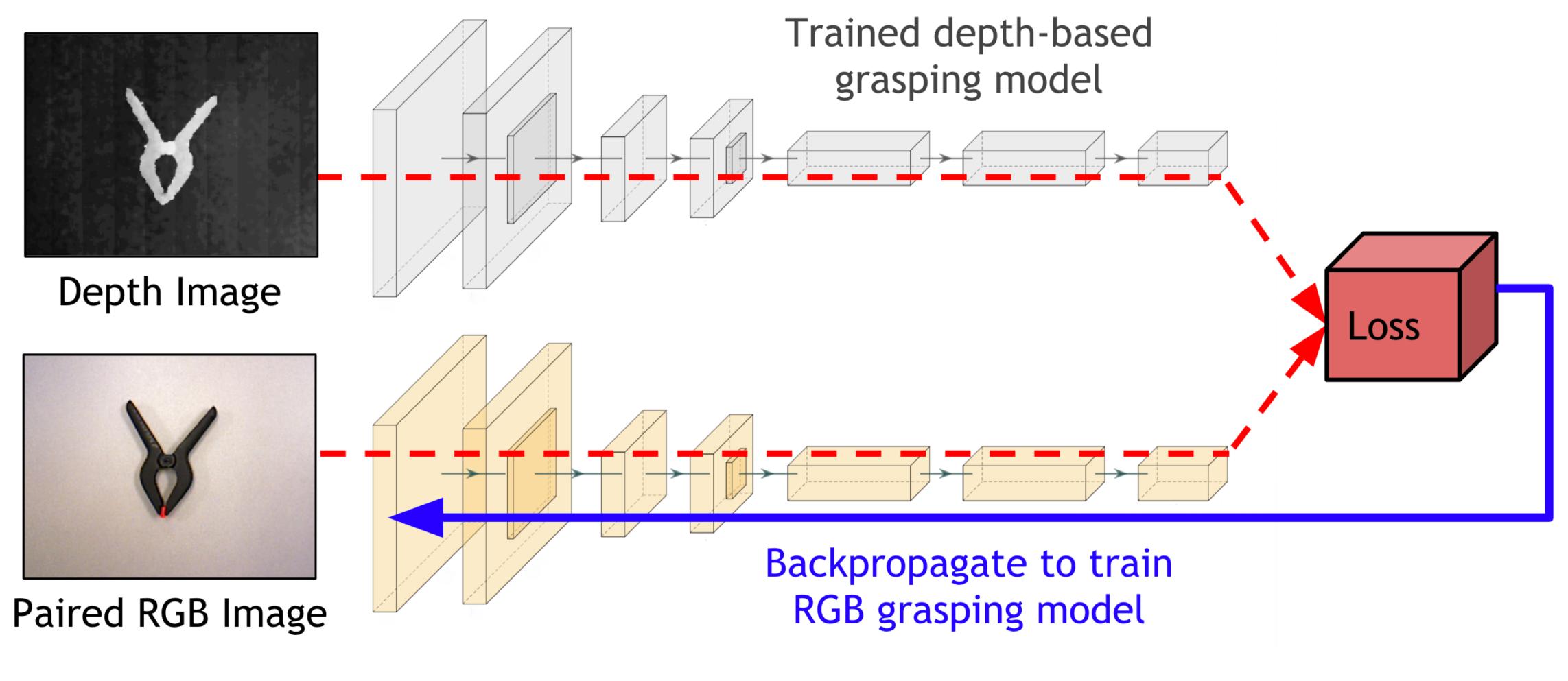
Methods

For this work, we use a very similar representation of the Fully Convolutional Grasp Quality CNNs (FC-GQCNN), which is state of the art depth-based model. And by using this fusion model[6], we combine RGB and depth models.



Our Fusion Model

Thomas et al.[5] using supervision transfer [1], [2], [3] to train a model for a modality (such as RGB). Their dataset D′ only use opaque objects that depth-based grasping methods typically perform well.

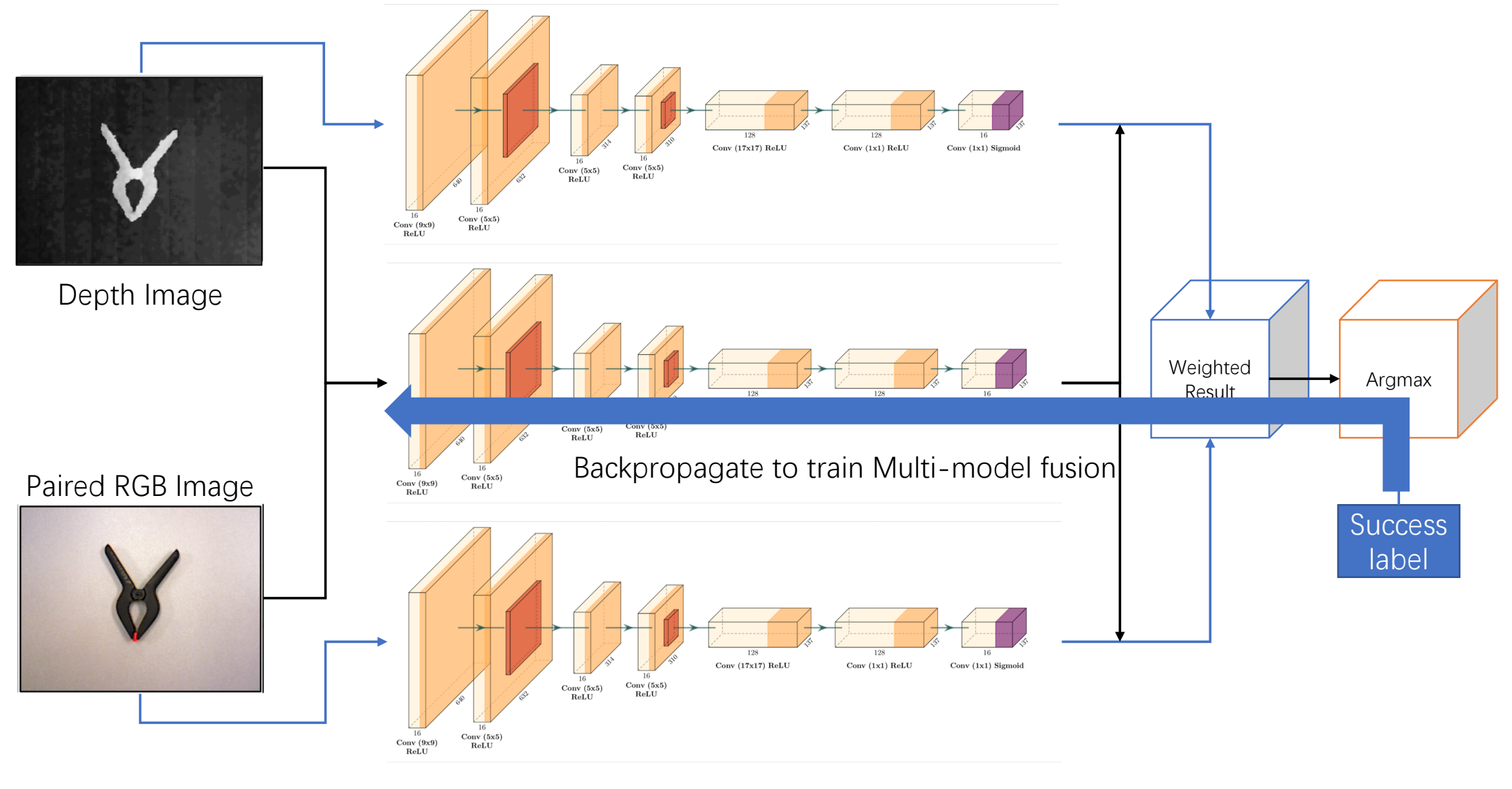

Trained depth-based grasping model
Depth Image
Loss
Paired RGB Image
Backpropagate to train RGB grasping model

Training RGB Model

We train the fusion using a similar representation of MoDE[6]. $G_{w\phi}$ is our fusion model, q_t is the grasping point we choose, t is the final success label, I_d and I_s are input images(Depth and RGB). Loss function:

$$\mathcal{L}(\phi) = \begin{cases} cross\_entropy(G_{w\phi}(q, I_d, I_s), t) & q = q_t \\ 0 & \text{otherwise} \end{cases}$$

$$q_t = argmax_q(G_{w\phi}(q, I_d, I_s))$$



Training Weighted Model

Result

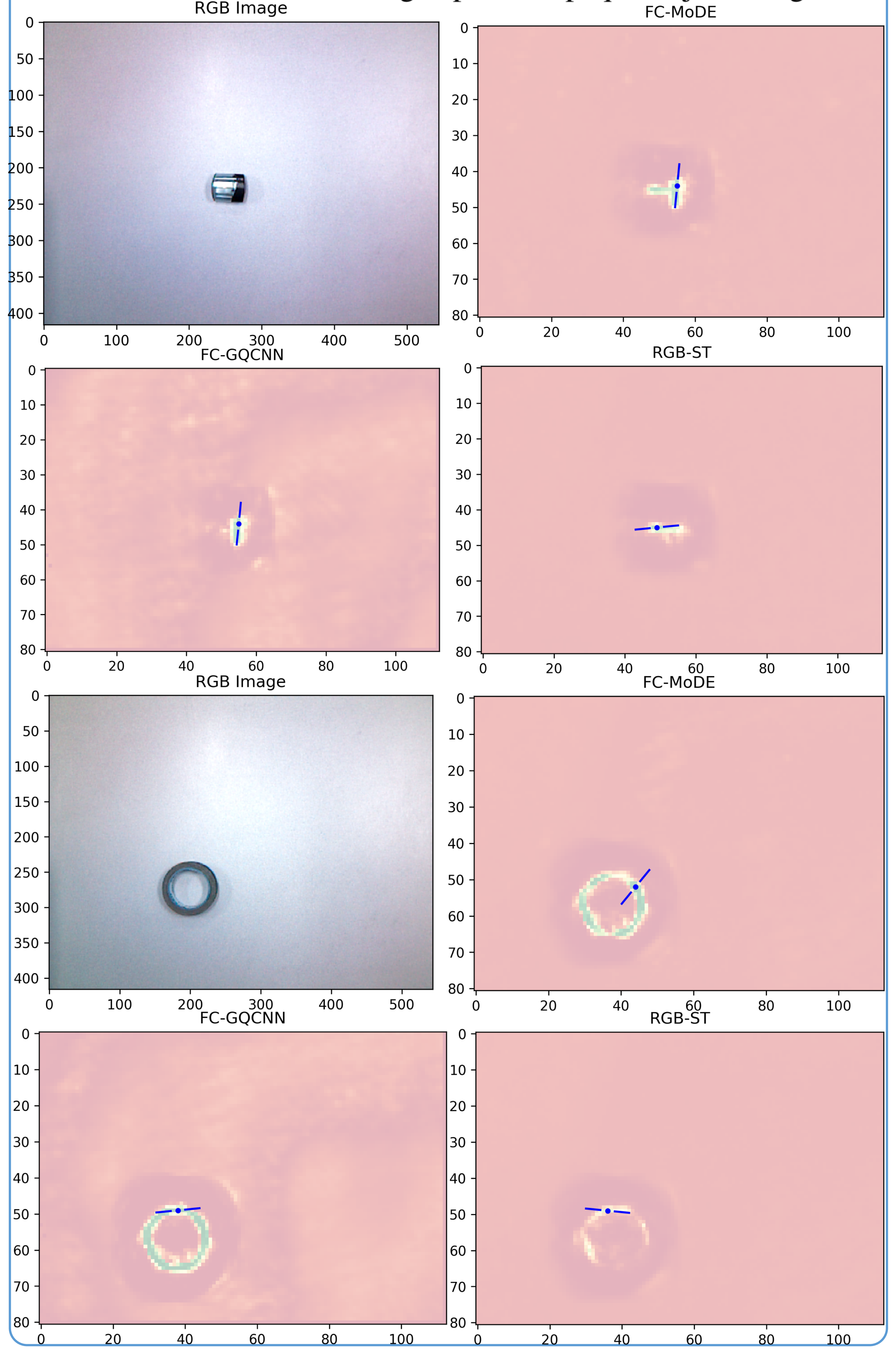
Our new approach is going to combine RGB images and existing depth image network to achieve a higher success rate in grasping transparent and reflective objects. We have achieved higher success rate for transparent and reflective objects than state of the art. (The results of FC-GQCNN, RGB-ST, RGB-G, RGB-C are all from [5])

- FC-GQCNN : State of the art depth-based model
- RGB-ST: Supervision transfer model
- FC-MoDE: Our fusion model
- RGB-G: FC-GQCNN with grayscale image inputs
- RGB-C: FC-GQCNN with RGB image inputs

TABLE I: Performance on individual object grasping

Method	Opaque	Transparent	Specular
FC-GQCNN*	0.733	0.200	0.493
RGB-G*	0.533	0.333	0.413
RGB-C*	0.147	0.240	0.240
RGB-ST***	0.867	0.853	0.640
FC-MoDE**	0.587	0.6	0.507

*Trained on simulated grasps **Trained RGB-D images ***Trained on simulated grasps and opaque object images



Conclusions

We present an adaptive Multimodel fusion for grasping transparent and specular objects. Our model combines depth- based and color-based models under supervising learning, which requires paired depth and RGB images, outputs from two types of models, and does not require any RGBD-based simulation. We show our fusion model has better performances than state of the art model in transparent and specular objects and similar in opaque objects.

While we can use the final success label to train our model, our model can generate itself online when it does real grasp. We are also interested in using this work to more types of models, such as side grasping and multi-finger grasping.

Reference

[1] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2827– 2836, 2016.

[2] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In 2016 IEEE International Conference on Robotics and Automation (ICRA), pages 5032–5039. IEEE, 2016.

[3] Guanbin Li, Yukang Gan, Hejun Wu, Nong Xiao, and Liang Lin. Cross-modal attentional context learning for rgb-d object detection. IEEE Transactions on Image Processing, 28(4):1591–1601, 2018.

[4] Vishal Satish, Jeffrey Mahler, and Ken Goldberg. On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks. IEEE Robotics and Automation Letters, 4(2):1357– 1364, 2019.

[5] Thomas Weng, Amith Pallankize, Yimin Tang, and David Held Oliver Kroemer. Transfer learning for multi-modal grasping on transparent and specular objects. 2019.

[6] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 151–156. IEEE, 2016.

[7] IvoIhrke,KiriakosNKutulakos,HendrikPALensch,MarcusMagnor, and Wolfgang Heidrich. Transparent and specular object reconstruction. In Computer Graphics Forum, volume 29, pages 2400–2426. Wiley Online Library, 2010.

Acknowledgements

First and foremost, I would like to show my deepest gratitude to my supervisor, Thomas Weng and Prof. David Held. I shall extend my thanks to Rachel Burcin and Prof. John M. Dolan for all their kindness and help. I' d like to thank my parents, Fengchang Tang and Yan Chen, for all their help. Last but not least, thank Carnegie Mellon University and ShanghaiTech University.