Multi-modal Transfer Learning for Grasping Transparent and Specular Objects

Thomas Weng[1], Amith Pallankize[2], Yimin Tang[3], Oliver Kroemer[1], and David Held[1]

***Abstract*— State-of-the-art object grasping methods rely on depth sensing to plan robust grasps, but commercially available depth sensors fail to detect transparent and specular objects. To improve grasping performance on such objects, we introduce a method for learning a multi-modal perception model by bootstrapping from an existing uni-modal model. This transfer learning approach requires only a pre-existing uni-modal grasping model and paired multi-modal image data for training, foregoing the need for ground-truth grasp success labels nor real grasp attempts. Our experiments demonstrate that our approach is able to reliably grasp transparent and reflective objects. Video and supplementary material are available at https://sites.google.com/view/transparent-specular-grasping.**

I. Introduction

Robotic grasping is a key prerequisite for a variety of tasks involving robot manipulation. Robust object grasping would enable a wide range of applications in both industrial and natural human environments. The challenge with grasping is that many factors influence the effectiveness of a grasp, such as gripper and object geometries, object mass distribution and friction, and environmental conditions like illumination.

Most state-of-the-art grasping methods rely on depth input from structured light or time-of-flight sensors to determine the best grasp for an object [1], [2], [3]. Under normal operation, such devices emit light patterns onto a scene and use a receiver to construct depth based on changes in the returned pattern. However, such depth sensors fail to detect objects that are transparent, specular, refractive, or have low surface albedo [4], causing depth-based grasp prediction methods to fail. These failures can take the form of both missing depth readings, as is the case with specular objects that deflect structured light patterns, and incorrect depth values, which occur when the emitted light passes through transparent objects (see Fig. 1).

Transparent and specular objects are common in a range of environments, such as in manufacturing facilities, retail spaces, and homes. Under certain lighting conditions and object properties, even seemingly opaque objects can exhibit sensor noise similar to transparency and specularity. The ubiquity of objects with these challenging properties requires us to design methods capable of bridging the sensory gap so that robots can robustly grasp a diverse set of objects.

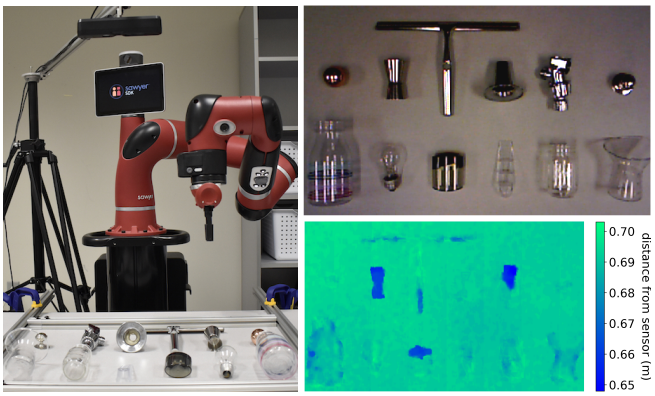


Fig. 1: Transparent and specular objects provide poor depth readings with conventional depth sensors, posing a challenge for depth-based grasping techniques. (left) Robot workspace with fixed overhead sensor for grasping. (top right) Color image of scene from overhead sensor. (bottom right) Depth image of scene showing that most values in depth image are close to the table.

Our contribution in this work is a method for learning to grasp transparent and specular objects that leverages existing depth-based models. Transparent and specular objects are more identifiable in RGB space, where transparencies and specularities produce changes in coloration, rather than the inaccurate or missing values that occur in depth space. Therefore, we make use of both color and depth modalities in our approach. We first train a color-based grasp prediction model from a depth-based one using *supervision transfer* [5], a technique for transferring a learned representation from one modality to another. This transfer technique only requires paired RGB-D images and an existing depth-based grasping method from which to transfer; our method does not require robot grasp attempts nor human annotations.

We conduct real robot grasping experiments on both isolated objects and clutter to show that (1) the RGB-only network produces better grasp candidates for transparent and specular objects, compared to the depth-only network that it was trained from, and (2) the RGB-only network is complementary to the original depth model, such that combining the outputs of both models results in the best overall grasping performance on all three object types. We conduct additional experiments to demonstrate the robustness of our method against slight variations in illuminance, and

This work was supported by the National Science Foundation Smart and Autonomous Systems Program (IIS-1849154), the Sony Corporation, the Office of Naval Research (N00014-18-1-2775), the NSF Graduate Research Fellowship Program (DGE-1745016), the Efort Intelligent Equipment Company, and ShanghaiTech University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the ONR and NSF.

[1]Thomas Weng, Oliver Kroemer, and David Held are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA {tweng, okoremer, dheld}@andrew.cmu.edu

[2]Amith Pallankize is with Microsoft Corporation, Hyderabad, India. ampallan@microsoft.com

[3]Yimin Tang is with ShanghaiTech University, Shanghai, China. tangym@shanghaitech.edu.cn

we discuss failure cases as part of our analysis.

II. Related Work

A. Sensing Transparent and Specular Objects

Sensing transparent and specular objects is a well-studied challenge in the computer vision community. Ihrke *et al.* [4] provide a survey of recent approaches to transparent and specular object reconstruction. Curless *et al.* [6] perform space-time analysis on structured light sensing to achieve better triangulation on transparent objects. Structured light sensing can also be paired with additional equipment like polarization lenses, light fields, or immersion in fluorescent or refractive liquids to detect transparent objects. While structured light sensing is the closest to commercial sensing, the survey also presents methods that improve on multi-view stereo matching to detect transparent and specular objects.

Light field photography for depth reconstruction is another direction for detecting specular and transparent objects [7], [8]. Light field photography has been used in robotics by Oberlin *et al.* [9] applied light field photography to robot manipulation tasks like grasping non-Lambertian objects under running water. However, this method requires capturing a dense set of images in a 3D volume over the scene of interest at both training and test time to construct suitable synthetic images for grasping. In comparison, our proposed method requires a single, static RGB-D sensor, resulting in faster and simpler training and deployment.

Commercial RGB-D sensors (*e.g.,* Intel RealSense, Microsoft Kinect, PrimeSense) use structured-light or time-of-flight techniques to estimate depth. These techniques fail on transparent and specular surfaces, either allowing light emitted by the sensor to pass through or scattering it by reflection. IR stereo and cross-modal stereo techniques have been used to improve depth reconstruction, but the reconstruction quality is still not comparable to that of Lambertian, or diffusely reflective, objects [10], [11], [12]. Lysenkov *et al.* [13], [14] painted over transparent objects to create a dataset of paired transparent and opaque objects, but this approach scales poorly for objects with arbitrary geometries and material properties. Our proposed method is able to use conventional RGB-D sensors without hardware and environmental modifications by combining depth and color information.

B. Grasp Synthesis

Grasp synthesis refers to the problem of finding a stable robotic grasp for a given object and is a longstanding research problem in robotics. Approaches to grasp synthesis can be classified into analytic and empirical methods; see Bohg *et al.* [15] for a survey. Analytic approaches use physics-based contact models to compute force closure on an object, using the shape and estimated pose of the target object [16], [17], [18], but work poorly in the real world due to noisy sensing, simplified assumptions of contact physics, and difficulty in placing contact points accurately.

Empirical approaches, on the other hand, learn to predict the quality of grasp candidates from data on a diverse set of objects, images, and grasp attempts collected through human labeling [19], [20], [21], [22], self-supervision [23], [24], or simulated data [25], [26], [3], [27], [1]. Saxena *et al.* [19] trained a classifier on human-labeled RGB images to predict grasp points, triangulated the points on stereo RGB images, and demonstrated successful grasps on a limited set of household objects, including some transparent and specular objects. However, the predicted grasp points for transparent and specular objects were limited to grasps on points where stereo triangulation was successful. The Cornell Grasping Dataset [20], consisting of 1k RGB-D images of objects and human-labeled grasps parameterized as an oriented bounding box, has been used to train many deep learning-based grasping methods [21], [26], [22]. Self-supervised methods such as those by Pinto and Gupta [23] or Levine *et al.* [24] forego the need for human labels by training a robot to grasp directly from real grasp attempts, but these methods require tens of thousands of attempts to converge.

Recently, approaches trained on data gathered in simulation have demonstrated state-of-the-art performance. The Jacquard dataset by Amaury *et al.* [25] uses a grasp specification similar to the Cornell Grasping Dataset, contains simulated objects and grasp attempts, and has been successfully used for training by Morrison *et al.* 's GG-CNN [26]. Mahler *et al.* [27] developed GQCNN, which was trained on a dataset of simulated grasps generated using analytic model, representing a hybrid empirical and analytic approach.

As we will show, these depth-only grasping approaches fail on transparent and reflective objects. Note that GG-CNN could be modified to incorporate RGB images, which could potentially be used to grasp transparent and specular objects after training on simulated images (such as those in the Jacquard dataset [25]); however, such performance has not been demonstrated; this method has only been demonstrated for depth-based grasping of opaque objects. In this work, we build upon the fully convolutional version of GQCNN (FC-GQCNN) proposed by Satish *et al.* [1], but our method is agnostic to the specific network architecture used. Our method does not require any real-world grasps or labeled data but instead relies on supervision transfer from a pre-trained depth network to obtain a multi-modal grasping method. The pre-trained depth network also may not require real-world grasps or human labels; for example, FC-GQCNN is trained entirely on simulated grasps.

C. Cross-modal Transfer Learning

Supervision transfer has been explored in the past for tasks such as image classification and object detection [28], [5], [29]. These approaches are typically used to transfer image-based networks trained on ImageNet [30] to depth-based or RGB-D based classification or detection networks. To our knowledge, such approaches have not been used previously in the context of multi-modal grasping. We show that such an approach can lead to greatly improved performance for grasping transparent and reflective objects, and can even improve performance on some opaque objects.

III. Approach

Here we describe our approach for supervision transfer, which enables us to transfer a grasping method trained in one modality $\mathcal{M}_d$ to also incorporate an additional modality $\mathcal{M}_s$ without needing any additional real grasp attempts, simulation, nor human-labeled data (other than the data used to train the initial uni-modal grasping method, which in our case is only simulated rendered depth data [1]).

A. Problem Statement

We assume that we initially have a grasping method that takes input from a given modality $\mathcal{M}_d$, such as depth. Specifically, we assume that we have a grasping method that, given a candidate grasp q and an image I_d of modality $\mathcal{M}_d$ (*e.g.,* a depth image), outputs a grasp score $G(q, I_d)$. We wish to transfer this scoring method to a new input modality $\mathcal{M}_s$ (*e.g.,* RGB). Ideally, this new modality $\mathcal{M}_s$ will allow our grasping method to succeed in grasping certain types of objects (e.g. transparent and specular) where the previous modality, $\mathcal{M}_d$, failed. In later sections, we will discuss combining these modalities to create more robust grasping methods.

We assume access to a dataset of image pairs (I_d, I_s), where each pair consists of one image from each modality. We assume that each pair of images was taken at approximately the same time and thus represent images of the same scene under the two modalities $\mathcal{M}_d$ and $\mathcal{M}_s$. Paired images for RGB and depth modalities can be captured using commercially available RGB-D sensors (*e.g.,* Intel RealSense, Microsoft Kinect, PrimeSense).

Note that these paired images can be collected without needing to perform any grasp attempts or human labeling, making the collection of this dataset very efficient. Furthermore, because these paired images are collected in the real world, they contain all of the real-world noise and artifacts that one would encounter in a realistic setting, avoiding the need to create such artifacts in simulation.

B. Supervision Transfer for Multi-modal Perception

In attempting the modality transfer described above, we observe the following: different input modalities (*e.g.,* depth vs RGB) have complementary advantages. In other words, data that is difficult for computing successful grasps in one modality might not be as difficult for another modality, and vice versa. For example, transparent and reflective objects are extremely difficult for depth-based grasping methods, due to the resulting noise or missing data in the depth image. However, our experiments show that RGB-based grasping methods have a much higher success rate for these objects. On the other hand, highly textured objects may present difficulties for RGB grasping methods, but these textures do not manifest in depth-based methods.

Based on this observation, we first filter our dataset D into a new dataset D' for which we expect the grasping method of modality $\mathcal{M}_d$ to perform well. In other words, for images $I_d \in D'$, the grasp score $G(q, I_d)$ should have a high correlation with the success of an executed grasp. In our case, because I_d is a depth image, our filtered dataset D' contains only images of opaque objects, for which depth-based grasping methods typically perform well.

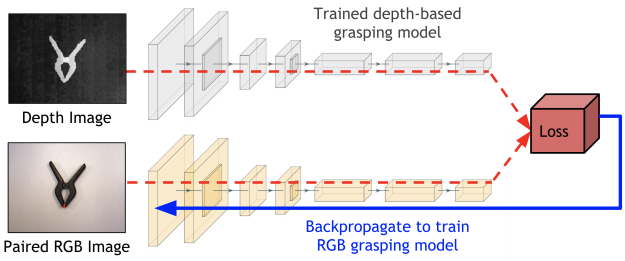


Fig. 2: We train a grasp quality CNN that takes RGB input by supervising the loss of the network on the output of a trained depth model for paired, unlabeled RGB-D image data.

We then train a grasping method for modality $\mathcal{M}_s$ (*e.g.,* RGB) using supervision transfer [28], [5], [29] over dataset D'. For each paired image (I_d, I_s) in dataset D', we compute the grasping score $G(q, I_d)$ for the modality $\mathcal{M}_d$. Because of our filtering, this grasp score is likely to be accurate. We then train a method for computing the grasping score $G_\phi(q, I_s)$ of the second modality $\mathcal{M}_s$ using the grasp score from modality $\mathcal{M}_d$ as the grasp label; thus we define the loss to be

$$\mathcal{L}(\phi) = ||G(q, I_d) - G_\phi(q, I_s)||^2 \quad (1)$$

For paired images of dataset D', we train the grasping method on the new modality $\mathcal{M}_s$ (*e.g.,* RGB) to output the same grasping score as the score output of the previous grasping method on the original modality $\mathcal{M}_d$ (*e.g.,* depth). This procedure is shown in Figure 2.

Because of the complementary nature of the two sensors, this grasping score function will often perform well on data that was originally filtered out of D and not included in D', even though $G_\phi(q, I_s)$ was only trained on data from D'. Specifically, we filter out transparent and reflective objects from D' because depth-based grasping methods perform poorly on these objects. Nonetheless, the image-based grasping method $G_\phi(q, I_s)$ still performs well on images of transparent and reflective objects, because the difference in appearance for these objects in the RGB modality is much smaller than the difference in appearance for these objects in the depth modality. Our experiments confirm this to be the case.

Further, because the modalities are complementary, we show that we can get the best performance by combining the grasping scores from the two modalities. Although there are many potential ways to do this, we evaluate two possibilities. The "early fusion" approach for combining modalities is to transfer from a depth-based grasping network to a RGB-D grasping network ("RGBD-ST", see Fig. 3c). RGBD-ST takes as input both depth and RGB modalities concatenated together. For our second, "late-fusion" approach, we fuse the scores of each modality, averaging the outputs of the depth-based grasping network with a RGB-based grasping network

trained using supervision transfer. We define the multi-modal grasping score as

$$G_\phi(q, I_d, I_s) = \frac{1}{2} \cdot (G(q, I_d) + G_\phi(q, I_s)) \quad (2)$$

This method is referred to below as "RGBD-M" (see Fig. 3d). Both of these approaches share the benefits that they represent multi-modal grasping methods that were trained from a depth-based grasping method only using paired RGB and depth images, without requiring real grasp attempts or human labels.

C. Implementation of Supervision Transfer

Our supervision transfer formulation is agnostic to the specific grasping method or representation we use for grasping in modality $\mathcal{M}_d$. For this work, we use the Fully Convolutional Grasp Quality CNNs (FC-GQCNN) representation as the pre-trained depth model from Satish *et al.* [1], although other depth-based grasping methods could equivalently be used.

FC-GQCNN learns a function $G(q_d, I_d)$ which predicts a grasp success rate for each grasp q_d based on a depth image I_d. In FC-GQCNN, grasps q_d are parameterized as $q_d = (x, y, \theta, z)$, where x and y are horizontal planar coordinates designating the desired grasp point of the gripper, z is the grasp depth relative to the camera, and θ is the clockwise rotation angle of the gripper about the vertical z axis. FC-GQCNN takes as input just a single depth image I_d and outputs a 4-dimensional tensor of grasping scores, producing one score per binned (x,y,z) position as well as binned orientation coordinates θ. FC-GQCNN is designed to be fully convolutional in order to output dense predictions $G(q_d, I_d)$ across the entire depth image. Our methods, shown in Figure 3, use a similarly dense (x, y) output and the same output angular encoding θ.

We wish to use the output of FC-GQCNN to train an image-based grasping method $G(q, I_s)$. Because the image modality does not have access to depth information, for image-based grasping we change the grasping parameterization to just $q = (x, y, \theta)$, without including a parameter for the grasp depth z. With this specification, each grasp starts at an approach height and moves down until it makes contact with either the table or an object before closing the gripper. Due to the difference in grasp representations, we modify our loss slightly, to be:

$$\mathcal{L}(\phi; q, I_d, I_s) = || \max_z G((q, z), I_d) - G_\phi(q, I_s)||^2 \quad (3)$$

where (q, z) is the concatenation of z to a grasp $q = (x, y, \theta)$ to form the new grasp representation (x, y, θ, z). In other words, to compute the target grasp score for some grasp $q = (x, y, \theta)$, we append various depths z to form a depth-based grasp parameterization (x, y, z, θ); for each of these grasp parameterizations we can compute the depth-based grasping score $G((q, z), I_d)$ using our depth-based grasping method (e.g. FC-GQCNN). We then compute the maximum grasp score over the values of z to obtain $\max_z G((q, z), I_d)$.

The network architecture that we use for image-based grasping is very similar to the architecture used in FC-GQCNN for depth-based grasping (see Appendix A). The only modification that we make is that we modify the first layer to accept a 3-channel RGB input rather than a 1-channel depth input. This is accomplished by adding an extra dimension to the first layer convolutional filters. In some of the experiments, we will alternatively use an RGB-D grasping network ("RGBD-ST"), in which case we modify the first layer to accept a 4-channel input, in a similar manner.

IV. Experimental Setup

Following the reproducibility guidelines for grasping research as presented in [12], we describe our experimental setup and protocols below.

A. Physical Components

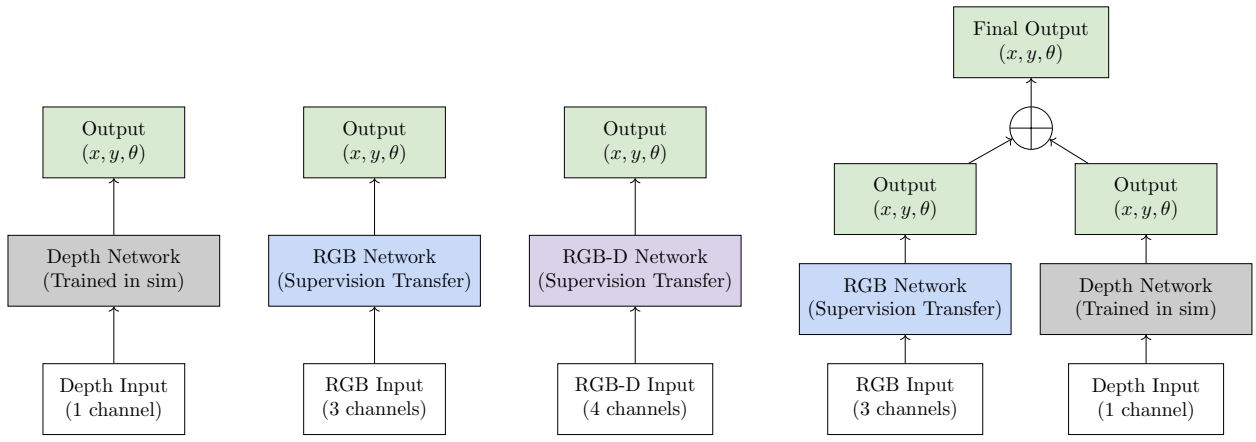
We use an ASUS Xtion Pro Live RGB-D sensor, fixed 0.7 m above and pointing down towards the workspace (see Fig. 4). Robot experiments were performed on a 7 DOF Rethink Robotics Sawyer robot equipped with an electric parallel jaw gripper, though our method can be applied to other robots and end-effectors. The robot's workspace is an approximately 0.65 m $\times$ 0.38 m area that is reachable by the robot with a vertical grasp. Aluminum extrusions enclose the workspace to prevent objects from rolling or sliding out of the space.

All experiments and network training were performed on an Ubuntu 16.04 machine with an NVIDIA GTX 1080 Ti GPU, a 2.1 GHz Intel Xeon CPU, and 32 GB RAM allocated per job. Grasp planning was implemented using off-the-shelf MoveIt! software.

B. Training the Network

We first collected a set of 100 opaque objects from home and office retail stores. Using the ASUS Xtion Pro Live RGB-D sensor fixed above the workspace, we captured 200 paired RGB-D training images and 50 paired validation images of the objects in varying amounts of clutter and with lighting conditions ranging from standard office illuminance (approx. 500 lux) to dimmed illuminance (approx. 175 lux). We resized the images to account for differences between our sensor's intrinsic parameters and those of the pretrained FC-GQCNN model. To increase the amount of training data and improve domain robustness, we applied spatial augmentations (*e.g.,* random rotations and flips) and color-based augmentations (*e.g.,* hue, brightness, and contrast), generating approximately 20k paired training images. This image dataset is available at the URL in the abstract.

The network architecture was implemented in Python using Tensorflow and Keras. The RGB or RGB-D network's weights were randomly initialized, and the model was trained to convergence using an Adam optimizer with cross-entropy loss [31], [32]. We experimented with mean squared error loss, but it performed worse in initial experiments. The loss was supervised from the output of FC-GQCNN, taking the maximum over all values of z as discussed in Sec. III-C. Hyperparameters are provided in Appendix C.



(a) FC-GQCNN [1] (b) RGB-ST (c) RGBD-ST (d) RGBD-M

Fig. 3: Diagrams of the four methods evaluated in this work. We compare against FC-GQCNN [1], which takes a depth image as input and outputs dense grasp scores over image coordinates x, y and rotation θ about the depth axis. RGB-ST and RGBD-ST are both trained using supervision transfer, but differ in the input they accept (3-channel RGB or 4-channel RGB-D input). RGBD-M takes the outputs of the RGB and Depth networks and averages them to produce the final output.

C. Test Objects

We collected objects distinct from the training objects to form three sets of 15 test objects each, one set per category (see Fig. 4). For the opaque object set, we primarily use YCB [33] objects that fit within the 5 cm stroke width of our gripper. We collected our own transparent and specular object sets due to the lack of existing benchmark sets for these categories.

Following typical procedures for grasping evaluations [27], [34], we remove bias related to object pose through the following procedure: objects are shaken in a box and then emptied onto the robot's workspace for each grasp attempt. This procedure is used for both isolated object grasping as well as for grasping in clutter.

V. Experimental Results

We design experiments to answer the following questions:

- To what extent can supervision transfer be used to grasp objects from new modalities (e.g. depth to RGB)?
- To what extent can supervision transfer from depth to RGB be used to learn to grasp transparent and reflective objects?
- Do the depth and image modalities complement each other? That is, will combining both modalities outperform either modality alone?

Note that grasping performance is not directly comparable with previous work like FC-GQCNN [1] as we use a different robot, gripper, and depth sensor.

A. Multi-modal Perception

We evaluate whether multi-modal perception that combines depth and RGB data is better than uni-modal perception using either depth or RGB data alone. We refer to our method for Depth-to-RGB supervision transfer, described in Sections III-B and III-C, as "RGB-ST" (see Fig. 3b).

We evaluate two approaches to multi-modal perception, both of which are described in Sections III-B and III-C. The first "early-fusion" approach uses supervision transfer to directly train an RGB-D grasp prediction network from a depth-based network, called "RGBD-ST" (see Fig. 3c). The second "late-fusion" approach involves taking the mean of the outputs of an RGB-only network and a depth-based network. Specifically, we take the mean of the RGB-ST and FC-GQCNN grasping networks; we call this multi-modal method "RGBD-M" (see Fig. 3d).

The results are shown in Table I. RGBD-ST and RGBD-M both significantly outperform depth-only grasping (FC-GQCNN) on transparent and specular objects, while maintaining comparable performance on opaque objects.

We also see that the multi-modal methods perform similarly to the RGB-based grasping method (RGB-ST) on opaque and transparent objects, but outperform this method on specular objects. These results support the notion that combining both RGB and depth modalities gives better grasping performance than using either modality alone.

TABLE I: Isolated object grasping, averaged over five trials

Method	Opaque	Transparent	Specular
FC-GQCNN*	$\mathbf{0.92 \pm 0.06}$	0.40 ± 0.08	0.48 ± 0.17
RGB-ST†	0.89 ± 0.04	0.79 ± 0.09	0.71 ± 0.04
RGBD-ST†	0.91 ± 0.06	0.77 ± 0.08	$\mathbf{0.83 \pm 0.04}$
RGBD-M†	0.91 ± 0.14	$\mathbf{0.85 \pm 0.06}$	0.81 ± 0.07

*Trained on simulated grasps
†Trained on simulated grasps and opaque object images

B. Grasping in Clutter

We also evaluated our methods for grasping in clutter, as this is important for robots in various cluttered environments like homes and warehouses. The same test objects used in

Fig. 4: Data collection setup and example images from the dataset of all three object types.

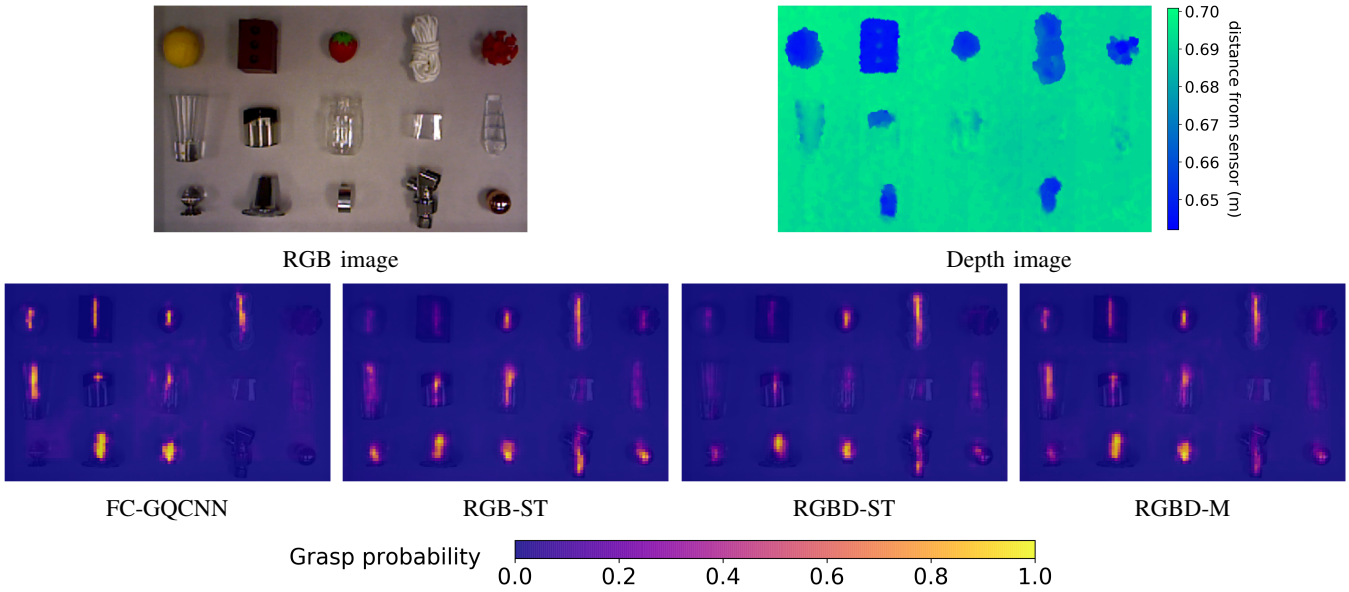


Fig. 5: Probability heatmaps of grasping across methods for the max grasp score of a grasp with fingertips horizontal to the image, centered at the each pixel. Objects from each set are arrayed horizontally such that the top row is opaque objects, the next transparent, and the final one specular.

isolated object grasping were used for clutter experiments. Five trials of grasping in clutter were conducted for each object category. Following the procedure from Viereck *et al.* [34], a trial concluded after all objects were successfully grasped, 3 consecutive failed grasp attempts occurred, or all objects were outside the workspace.

To prevent a network from getting repeatedly stuck on attempting a bad but highly rated grasp, we randomly sample a 0.2m square crop of the input image and select the grasp location within that region with the maximum predicted success probability. All methods including baselines performed similarly or worse without this sampling (see Appendix B). Crops whose grasp probabilities all fall below below a threshold are discarded and resampled to avoid attempting grasps based on noisy sensor readings.

The results are shown in Table II. The results from grasping in clutter corroborate the result of isolated grasping. All methods perform well on opaque objects, although RGBD-M (averaging the output of depth-only grasping and RGB-only grasping networks) performs slightly better than the others. On non-opaque objects (e.g. transparent and specular), FC-GQCNN (e.g. depth-only grasping) performs poorly.

Table II shows that RGB-ST (RGB-only grasping) and RGBD-M (averaging the output of depth-only grasping and RGB-only grasping networks) perform well across all three object categories. We note that, despite averaging across five trials, the results of grasping in clutter have relatively high variance and should be considered accordingly. Overall, our main conclusions are similar to that of isolated object grasping from Section V-A: depth-only grasping performs poorly on transparent and specular objects; with supervision transfer, we can obtain a method that performs much better on grasping transparent and specular objects while maintaining similar performance on opaque objects. This method requires only paired RGB and depth images for training and does not require any real grasp attempts or human

TABLE II: Grasping in clutter, averaged over five trials

Method	Opaque	Transparent	Specular
FC-GQCNN*	0.84 ± 0.06	0.23 ± 0.21	0.35 ± 0.16
RGB-ST†	0.77 ± 0.11	$\mathbf{0.67 \pm 0.10}$	$\mathbf{0.68 \pm 0.12}$
RGBD-ST†	0.86 ± 0.09	0.67 ± 0.27	0.35 ± 0.10
RGBD-M†	$\mathbf{0.97 \pm 0.15}$	0.51 ± 0.32	0.63 ± 0.12

*Trained on simulated grasps
†Trained on simulated grasps and opaque object images

annotations, other than the simulated depth rendering data that was used to train the original FC-GQCNN [1] depth-based grasping method.

C. Lighting Variation Experiments

We note that domain shifts like lighting can be a problem for RGB methods, as mentioned in previous work [12]. To enable our method to be robust to lighting variations, our training images were collected with slight lighting variations, and we applied color-based augmentations like brightness and contrast.

We conducted experiments to evaluate the robustness of the trained networks to lighting variations. We varied the lighting by moving a floor lamp around the robot workspace as shown in Fig. 6a and performed the isolated object grasping experiments for RGBD-M. The additional lighting increased illumination to between 750 and 950 lux. With this variation in lighting, the RGBD-M network performed comparably, achieving grasp success rates of 0.81 ± 0.12 for transparent objects and 0.79 ± 0.09 for specular ones (compare with Table I).

(a) Lighting setup. (b) Extreme lighting.

Fig. 6: (a) Setup for lighting variation experiments. Lighting is controlled using the overhead lights and floor lamp. (b) Failure case in the extreme lighting condition. The method predicts the best grasp to be on the object's shadow.

However, we found that the network performed poorly under more drastic lighting changes, in which we turned off the overhead lights and reduced the height of the floor light, dropping illumination to approx. 175 lux and causing long object shadows to appear. In this case, grasp performance dropped to 0.52 ± 0.18 on transparent objects and 0.60 ± 0.12 for specular ones. In such extreme lighting conditions, we observed the method predicting grasps on shadows for transparent objects (see Fig. 6b). Such drastic lighting would not normally occur in structured applications like bin-picking.

D. Failure Cases

In this section we discuss the most frequent and notable failure cases from our experiments. This section covers failures due to our approach, as well as external factors. Some examples of failure cases discussed in this section can be seen in Fig. 7 and the supplementary video.

Methods that used the depth modality like RGBD-ST and RGBD-M at times selected grasps that were highly rated by the depth network, but did not sufficiently account for transparencies or specularities (Fig. 7, top left). Both the color-based and depth-based networks at times failed to distinguish very transparent objects from the workspace surface, though this was rare and occurred far less frequently than with FC-GQCNN (Fig. 7, top right). Object mass distribution and deformability were not accounted for by our methods (Fig. 7, bottom row).

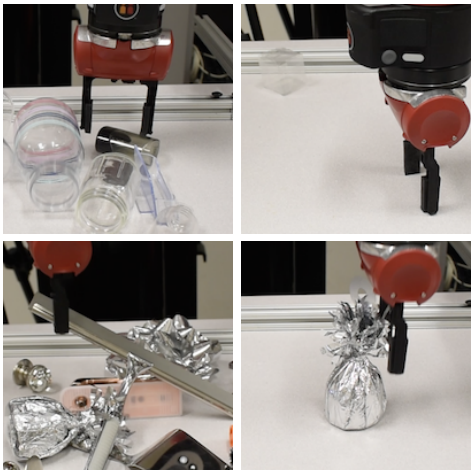


Fig. 7: Examples of failure cases. (top left) Grasp does not account for transparent part of sharpener. (top right) Gripper fails to detect transparent plastic cube and grasps at table. (bottom left) Mass distribution of squeegee causes grasp to fail. (bottom right) Foil on top of balloon weight appears graspable but the gripper passes through.

A failure case external to the methods evaluated involved our gripper hardware. Our parallel electric gripper has a relatively small stroke width, and is unable to execute pinch grasps with a 5cm opening width. This limitation causes grasps on thin parts of objects to fail, because the fingertips do not completely come together. While it is possible to adjust the fingertips to be closer together to enable pinch grasps, the opening width of the gripper would be reduced, which would prevent the gripper from being able to grasp large objects. This issue reduced performance across all methods and would likely be mitigated by other grippers.

Since our paper focused on static grasping, our method fails to grasp objects that start rolling due to perturbation in clutter. Others have investigated ways to address this issue using closed-loop control techniques like visual servoing [2].

VI. Conclusion

We present an approach for improving grasping on transparent and specular objects, for which existing depth-based grasping methods perform poorly. Our method transfers information learned by a depth-based grasping network to RGB or RGB-D networks, enabling multi-modal perception. Our method for supervision transfer requires only real-world paired depth and RGB images, and does not require any human labeling nor real-world grasp attempts. We explore two avenues to multi-modal perception and demonstrate that making use of the RGB modality outperforms depth-only grasping in isolated object grasping as well as grasping in

clutter. The method is extensible to other robots, environments, and end effectors. One potential direction for future work may be to adaptively weight predictions from different modalities instead of averaging them. Another is applying transfer learning techniques to other, less similar modalities like haptics and tactile feedback. Combining different sensor modalities might also be useful in determining the appropriate grasp height for each object.

While we are able to get improved performance without using any real grasping data, we believe that real grasps can be used to further improve the performance of the network. We are also interested in extending this work to other types of grasping, such as 6-DOF, multi-fingered, or suction grasping.

REFERENCES

[1] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1357–1364, 2019.

[2] D. Morrison, J. Leitner, and P. Corke, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[3] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 598–605.

[4] I. Ihrke, K. N. Kutulakos, H. P. Lensch, M. Magnor, and W. Heidrich, "Transparent and specular object reconstruction," in *Computer Graphics Forum*, vol. 29, no. 8. Wiley Online Library, 2010, pp. 2400–2426.

[5] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, "Cross-modal adaptation for rgb-d detection," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5032–5039.

[6] B. Curless and M. Levoy, "Better optical triangulation through space-time analysis," in *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 1995, pp. 987–994.

[7] K. Maeno, H. Nagahara, A. Shimada, and R. Taniguchi, "Light field distortion feature for transparent object recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2786–2793.

[8] G. Wetzstein, R. Raskar, and W. Heidrich, "Hand-held schlieren photography with light field probes," in *2011 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2011, pp. 1–8.

[9] J. Oberlin and S. Tellex, "Time-lapse light field photography for perceiving transparent and reflective objects."

[10] F. Alhwarin, A. Ferrein, and I. Scholl, "Ir stereo kinect: improving depth images by combining structured light with ir stereo," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2014, pp. 409–421.

[11] W. W.-C. Chiu, U. Blanke, and M. Fritz, "Improving the kinect by cross-modal stereo." Citeseer.

[12] J. Mahler, R. Platt, A. Rodriguez, M. Ciocarlie, A. Dollar, R. Detry, M. A. Roa, H. Yanco, A. Norton, J. Falco, *et al.*, "Guest editorial open discussion of robot grasping benchmarks, protocols, and metrics," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1440–1442, 2018.

[13] I. Lysenkov, V. Eruhimov, and G. Bradski, "Recognition and pose estimation of rigid transparent objects with a kinect sensor," *Robotics*, vol. 273, 2013.

[14] I. Lysenkov and V. Rabaud, "Pose estimation of rigid transparent objects in transparent clutter," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 162–169.

[15] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesisa survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[16] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen, "Automatic grasp planning using shape primitives," 2003.

[17] 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.

[18] D. Watkins-Valls, J. Varley, and P. Allen, "Multi-modal geometric learning for grasping and manipulation," *arXiv preprint arXiv:1803.07671*, 2018.

[19] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[20] 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*. IEEE, 2011, pp. 3304–3311.

[21] 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. [Online]. Available: https://doi.org/10.1177/0278364914549607

[22] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 1316–1322.

[23] 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)*. IEEE, 2016, pp. 3406–3413.

[24] 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.

[25] 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)*. IEEE, 2018, pp. 3511–3516.

[26] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, vol. 0, no. 0, p. 0278364919859066, 0. [Online]. Available: https://doi.org/10.1177/0278364919859066

[27] 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," *arXiv preprint arXiv:1703.09312*, 2017.

[28] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[29] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for rgb-d object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1591–1601, 2018.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[31] G. Kahn, A. Villaflor, B. Ding, P. Abbeel, and S. Levine, "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[32] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[33] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[34] U. Viereck, A. t. Pas, K. Saenko, and R. Platt, "Learning a visuomotor controller for real world robotic grasping using simulated depth images," *arXiv preprint arXiv:1706.04652*, 2017.

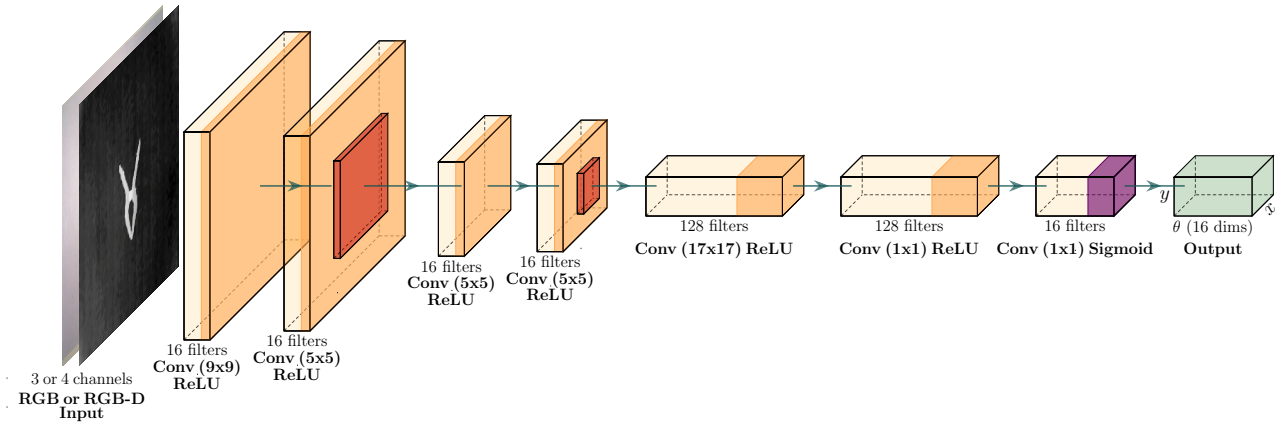


Fig. 8: Architecture diagram for supervision transfer networks, adapted from the FC-GQCNN [1] architecture. The input can be either 3-channel RGB input or 4-channel RGB-D input. The output is a 3D array of grasp quality scores over image coordinates x, y and rotation θ about the depth axis, discretized into 16 bins. The orange color accents correspond to ReLU activations and purple corresponds to sigmoid activation. The red layers are max pooling layers.

Appendix A
Network Architecture

Fig. 8 illustrates the architecture of the networks trained with supervision transfer.

Appendix B
Evaluations without random cropping

Table III provides results for grasping in clutter without random cropping.

TABLE III: Performance on grasping in clutter by method without random cropping, averaged over five trials

Method	Opaque	Transparent	Specular
FC-GQCNN*	0.95 ± 0.05	0.26 ± 0.25	0.35 ± 0.23
RGB-ST†	0.77 ± 0.10	0.77 ± 0.15	0.68 ± 0.15
RGBD-ST†	0.62 ± 0.26	0.67 ± 0.19	0.75 ± 0.08
RGBD-M†	0.75 ± 0.13	0.60 ± 0.18	0.47 ± 0.28

*Trained on simulated grasps
†Trained on simulated grasps and opaque object images

Random cropping refers to sampling a 0.2m square crop from the input image and choosing the grasp with the highest probability from within the crop. Crops which have do not have any objects in them, as determined by whether the max grasp probability within the crop falls below a hand-defined threshold, are discarded and a new crop is sampled. This procedure helps prevent networks from repeatedly choosing highly rated false positive grasps. However, the cropping threshold must be tuned based on the performance of the grasping network. For our experiments, we used a threshold of 0.4.

Appendix C
Hyperparameters

Hyperparameters for networks trained with supervision transfer are:

- Learning rate: 1e-05
- Batch size: 64
- Number of rotation augmentations per image: 32
- Loss: Binary cross-entropy

The FC-GQCNN model we evaluated against was a pre-trained model from `https://berkeleyautomation.github.io/gqcnn/`.