2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Invisible Marker: Automatic Annotation of Segmentation Masks for Object Manipulation

Kuniyuki Takahashi*†, Kenta Yonekura*†

***Abstract*— We propose a method to annotate segmentation masks accurately and automatically using *invisible marker* for object manipulation. *Invisible marker* is invisible under visible (regular) light conditions, but becomes visible under invisible light, such as ultraviolet (UV) light. By painting objects with the *invisible marker*, and by capturing images while alternately switching between regular and UV light at high speed, massive annotated datasets are created quickly and inexpensively. We show a comparison between our proposed method and manual annotations. We demonstrate semantic segmentation for deformable objects including clothes, liquids, and powders under controlled environmental light conditions. In addition, we show demonstrations of liquid pouring tasks under uncontrolled environmental light conditions in complex environments such as inside the office, house, and outdoors. Furthermore, it is possible to capture data while the camera is in motion so it becomes easier to capture large datasets, as shown in our demonstration.** [1] [2]

I. Introduction

Accurate object recognition is one of the important functions for robots in order to manipulate objects. In particular, dealing with deformable objects such as clothes, liquids, and powders is challenging, but it is important for a number of tasks such as laundry tasks and applications in biology and the medical area, which could benefit from robots that do laboratory experiments fully automatically. However, these deformable objects are challenging to recognize accurately.

Deep learning has succeeded in computer vision (CV), natural language processing (NLP) and in the robotics area [1]–[6], but generally requires massive datasets for training to achieve good performance. Even though massive datasets are required, creating datasets consumes enormous costs such as money, time, and human resources. Many objects are still beyond the reach of deep learning recognition because of the difficulty of creating large datasets.

In this study, we focus on automatic annotation of segmentation masks for object manipulation to create large and accurate datasets quickly and inexpensively. We use a marker that we call *invisible marker*, which emits light when light outside the visible spectrum (invisible) is applied, as opposed to visible light, such as ultraviolet (UV) light. This marker does not change the appearance of objects under visible (regular) light (See Fig. 1). We show a comparison between manual annotations and our proposed method, and show its effectiveness for automatic annotation of deformable objects including clothes, liquids, and powders under controlled environmental light conditions (e.g. inside a dark room). Additionally we show demonstrations of pouring liquid under uncontrolled environmental light conditions (referred to as **regular light conditions**) such as in offices, houses, and outdoor conditions even during daylight.

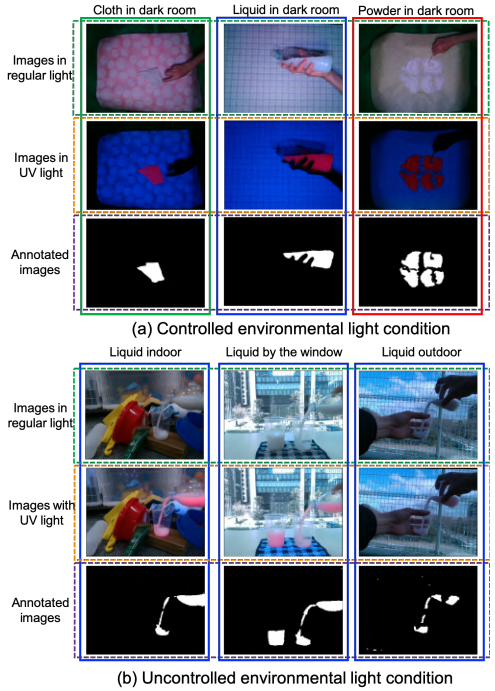

Fig. 1: Dataset creation using *invisible marker* for three kinds of objects: clothes, liquid, and powder. There are two conditions: environmental light is (a) controlled or (b) uncontrolled.

The rest of this paper is organized as follows. Related works is described in II, and our contributions are explained in III, while section IV details our proposed method. Section V outlines our experiment setup and evaluation settings under controlled light conditions, with results presented in VI. Additionally, the same things under regular light conditions are described in sections VII and VIII respectively. Section IX discusses the cost of the proposed method. Finally, conclusions are described as in section X.

II. Related Works & Contributions

Annotation methods have been being developed to reduce the effort of creating datasets. These methods fall in the two categories, manual annotation and automatic annotation.

* The starred authors are contributed equally.

† K. Takahashi and K. Yonekura are associated with Preferred Networks, Inc. takahashi@preferred.jp yoneken@preferred.jp

[1] An accompanying video is available at the following link: https://youtu.be/fnpyDYUvDA4

[2] Dataset is available at the following link: https://github.com/pfnet-research/Invisible_marker_IROS2020

A. Manual Annotation

Popular semantic segmentation datasets, such as Pascal VOC, MS-COCO, and COCO-Stuff, are manually annotated [7]–[9]. Reduction of the workers' manual efforts by annotation tools [10]–[13] and using crowdsourcing [5], [14] enables the creation of large-scale datasets. One of the challenges of manual annotation is that it is prone to human errors and individual judgments in ambiguous cases.

B. Automatic Annotation

To prevent human errors and to create accurate datasets easily, automatic annotation methods have been developed. The studies can be roughly classified into four groups.

1) Approaches that focus on the features of the object itself, such as color tracking [15], object temperature measurement (for example for hot liquids) using thermography [16], and movement through background subtraction [17]: These approaches fail in annotating objects when a feature of an object is shared among multiple objects or the environment.

2) Approaches that focus on features provided in advance such as augmented reality (AR) markers [18], [19]: Even if there are multiple objects, the objects can be distinguished by different markers on each object. However, if a feature like a marker is attached to the object, it doesn't look the same anymore as without the marker.

3) Simulation approaches [20], [21]: This method can artificially create large datasets by switching between various background images and by capturing images of the target object from multiple perspectives. However, the quality of the dataset is usually low for objects that are difficult to simulate, such as deformable objects. The gap between simulation and the real world is also a challenge that needs to be solved, for example through sim-to-real transfer learning.

4) Approaches that focus on in advance provided markers which can be detected by specific devices, or by controlling external factors, for example by using special light. The appearance of the object is not changed greatly under normal conditions. Only the part of the object on which, for example, fluorescent paint has been applied will become visible under UV light. In the medical field, a fluorescent substance that is attached to the material can be observed using a fluorescence microscope [22], [23]. In the computer vision field, a method that involves applying fluorescent paint to objects has been studied [24]. To create optical flow datasets, a computer repeatedly takes a pair of images under both regular light and under UV light, and then moves the scene or camera by a small amount [24].

Our proposed method belongs to the 4th category, which can be used to automatically create a segmentation mask as annotation using *invisible marker* for deep learning.

III. Contributions

Our main contribution is to expand previous work [24] as a general data creation method and to create a large dataset. The details of our contributions are as follows.

1) Application of the method to deformable object manipulation (Section VI-A): In previous research [22]–[24], the target objects and materials have clear contours and are rigid and stationary. In our study, rigid objects can be handled by our proposed method, but deformable objects were the main target. Additionally, we collected datasets of deformable objects during manipulation since the objects are deformed during motions. The challenges that arise during deformable object manipulation and their solutions were described.

2) Comparison with manual annotations by people and our proposed method to evaluate accuracy (Section VI-B): The dataset created with fluorescent paint is treated as ground truth for evaluating other methods, but the accuracy of the method using fluorescent paint has not been evaluated [24]. We compared our proposed method and manual annotations by people using crowdsourcing.

3) Training a deep neural network on a dataset of deformable objects (Section VI-C): The previous study using fluorescent paint [24] was meant for optical flow, thus there was no need for a large dataset. The datasets were too small to apply deep learning, despite the potential to create large datasets. We have collected enough data for training and have verified that the generalization performance of deep learning can cope with deformable objects.

4) Investigation of the appearance changes caused by the fluorescent paint (Section VI-D): In [22]–[24], the appearance change of an object by applying fluorescent paint was out of the scope of their work, because appearance changes do not influence the results in these works. In our study, however, if appearance changes by fluorescent paint are large, the inference performance for semantic segmentation will decrease, particularly for the objects not painted with fluorescent paint. We investigated whether the generalization capabilities of deep learning can still successfully deal with objects to which no fluorescent paint has been applied.

5) Application of several colors of fluorescent paints to multiple objects (Section VI-E): We show that the annotation for instance segmentation can be done by applying different colors of fluorescent paint in to multiple objects in a scene.

6) Proposal of a method to create datasets indoors as well as outdoors during daylight when light conditions cannot be controlled (Section VIII-A): In the previous study [24], datasets could only be collected in a dark room where external light could be controlled. Thus, collecting datasets was only possible in small experiment environments. In this study, we proposed a method for creating datasets indoors as well as outdoors where external light cannot be controlled.

7) Proposal of a method to collect datasets while camera is in motion (Section VIII-B): Although using a camera in a fixed position does not affect the quality of the annotation, it does limit the data collection process. Moving a camera around, however, makes it difficult to collect pairs of images taken subsequently, though it would allow us to easily and quickly create a diverse dataset from different viewpoints. We proposed a method that absorbs the camera movement even if the camera moves when capturing images.

8) Creation of a large dataset (Section V-C & VII-C): Our proposed method can create a dataset easily at high speeds and low costs. 32845 datasets are created in our study.

IV. Invisible Marker

Invisible marker emits light under light outside the visible spectrum, and emits weaker light or none under visible (referred to as regular) light. By using material that is (near) transparent under regular light for the *invisible marker*, the appearance of objects under regular light do not change when *invisible marker* is applied to them; any material can be used as an *invisible marker* as long as it satisfies this condition. Depending on the target object, sprays or liquids can be used as an *invisible marker*. We use fluorescent paint since it can be procured easily. The object painted with fluorescent paint is luminescent under ultraviolet (UV) light; this becomes particularly evident in the dark, but is still visible when UV light is applied even in regular light. Thus, our method can be used under both (1) **controlled** environmental light conditions (referred to as dark conditions) and (2) **uncontrolled** environmental light conditions (referred to as regular light conditions).

1) Objects without fluorescent paint are not visible in the dark even with a UV light turned on, but painted objects are visible. When an image is captured under this UV light condition, only the part painted with the fluorescent paint becomes visible (See second-row in Fig. 1). Only the painted part will be annotated as a class label for segmentation, by applying a threshold value on the RGB values of the image.

2) Under regular light, everything in the scene is still visible but the part to which fluorescent paint has been applied changes in appearance under UV light. Given a pair of images captured in regular light conditions, one with and one without applying UV light, the region where only fluorescent paint was applied to can be obtained by calculating the difference between these images. Then, in the same way as in dark conditions, a segmentation mask can be obtained by applying a threshold to the RGB values.

When moving the camera and taking pictures simultaneously, while also alternately switching between regular and UV light, the camera may move before capturing the same scene under different lighting conditions. Thus, a shift of viewpoints between the captured images with and without UV light will occur. If a segmentation mask is then created using the above method, the annotation mask will no longer match as a paired image captured under regular light because of this shift (The experiment details will be presented in section VIII-B). In particular, when the camera is moved under regular conditions (point (2) from the previous paragraph), this shift will be a problem since a segmentation mask is calculated from the difference in a captured pair of images with and without applying UV light. Therefore, it is necessary to transform the images so that the viewpoints of the two images match. The transformation process is done in three steps: 1) The image features from the images captured in regular and UV light conditions are extracted, 2) the feature point of image captured under regular light corresponding to another feature point of image captured under UV light with the shortest distance is matched, and 3) the image captured under UV light is transformed into a form that fits the image captured under regular light. The feature extraction method uses Oriented FAST and Rotated BRIEF (ORB), because of the calculation cost, matching accuracy [25]. A simple brute-force matcher is used for feature matching. For image transformations, we used homography transformations, which can be used to transform rectangles into trapezoids, unlike affine transformations, by projecting a plane onto another plane through projective transformations.

Fig. 2: Fluorescent paint under regular light and UV light

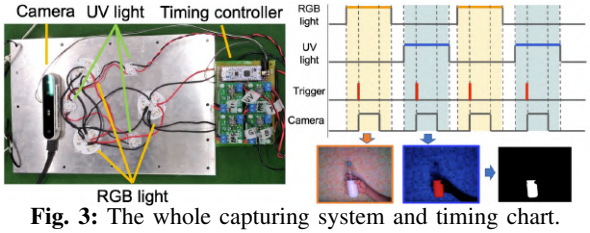

Fig. 3: The whole capturing system and timing chart.

Since only the target object can be extracted by *invisible marker*, our proposed method does not depend on the complexity of the background, or anything not painted with fluorescent paint, and only the fluorescent painted objects are extracted. Moreover, our method can distinguish individual objects from multiple objects by applying different colors to each object since various colors of fluorescent paints can be created by mixing red, green, and blue color (Fig. 2), which is useful for instance segmentation. More details will be discussed in Section VI-E.

V. Experiment Setup For Controlled Light

The purpose of the experiments under controlled environmental light conditions (dark conditions) is to 1) verify the accuracy of the datasets for deformable objects created with the *invisible marker* (Section VI-A), 2) to compare the proposed method with manual annotations by people (Section VI-B), 3) to evaluate the created dataset for deep learning (Section VI-C), 4) to investigate the effect of appearance changes by *invisible markers* (Section VI-D), and 5) to apply several colors of fluorescent paints for multiple objects in one scene and observe the results (Section VI-E).

A. Data Acquisition Device

Datasets for segmentation masks require a set of images under regular light as input data paired with output data that contains a label that tells to which class each pixel of the image belongs. We developed a system that can capture images under the regular light and invisible (UV) light to create such paired images automatically using *invisible marker* (See Fig. 3). Our system is composed of three parts:

- Camera part which captures images of a target object
- Lighting part which controls the lighting output
- Control part which controls the timing of capturing images and changing the light conditions

For the camera part, we use a camera for which the capture timing can be controlled through an external trigger input. When the camera receives an external trigger input from the control part, the camera captures an image of the target object. Then, the image is sent to the control computer. In this experiment, we use a RealSense D415 camera [26].

For lighting, we use an RGB LEDs as regular light, and UV LEDs as "invisible" light. Power LED drivers are used to control them.

For the control part, a NUCLEO-F303K8 board [27] provides trigger signals to the lighting part to control the emission timings and intensities. The NUCLEO-F303K8 board can be controlled from the computer through USB. At the same time, the control part also outputs a trigger signal to the camera part to control the capture timing (See timing chart in Fig. 3). To create datasets of segmentation masks for dynamically changing objects, regular lights and UV lights are alternately switched at high speed.

B. Target objects

Accurate recognition of deformable objects such as clothes, liquids, and powders is important in the industry and daily life to e.g. fold clothes, manipulate medicine, and cook. We prepared a cloth (handkerchief), liquid (water) and powder (baking soda) and mixed them with fluorescent paint.

C. Data Collection

For data collection, images are captured by the data acquisition system described in Section V-A during motions; folding of a cloth, shaking liquid in a plastic bottle, and stirring powder with a spoon. The sampling rate of the camera is 30 Hz. The sampling rate for creating a paired image dataset is 15 Hz because images are taken by alternately switching between regular light and UV light. We prepared six different types of backgrounds (Fig. 4) and perform one motion per background, thus a total of six motions per object. The total number of acquired images for clothes, liquids, and powders are 3920, 3081, and 4199, respectively. The dataset size is sufficient to achieve high accuracy for inference by multiple deep learning models. Details will be described in section VI-C. Datasets of five out of six backgrounds are used for training and the data of one remaining background is used for evaluation through 6-fold cross-validation as all combinations of backgrounds. This means that evaluation is performed on untrained background. Acquired images are resized from 640×480 to 160×120.

D. Deep Learning & Training

In order to show the effectiveness of the created datasets using *invisible marker*, we verify that they can be used to train three typical kinds of deep learning models for semantic segmentation: FCN [1], U-Net [2], and SegNet [3]. These models are trained with datasets composed of images in regular light as input and the annotated data created by the proposed method as output. Chainer, a deep learning library, is used for the implementation [28]. All our network experiments were performed on a machine equipped with

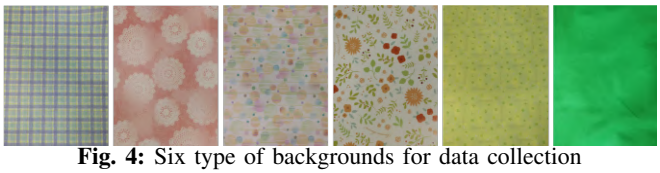

Fig. 4: Six type of backgrounds for data collection

256 GB RAM, an Intel Xeon E5-2667v4 CPU, and eight Tesla P100-PCIE with 12GB. Training time for each material was within 30 minutes by parallel processing with 8 GPUs.

VI. Experiment Results in Controlled Light

A. Annotation result by invisible marker

We first show examples of created datasets for deformable object manipulation (Fig. 1 (a)). In Fig. 1 (a), we can observe that fine unevenness at the edges of the powder is annotated accurately. In addition, only the target object is annotated even though the background includes similar colors to the objects, and other objects such as a hand and a spoon are ignored since they are not painted with fluorescent paint. Results for more complex backgrounds will be described in section VIII. Conventional methods, such as focusing on colors or background subtraction, usually have difficulties to annotate in these situations. Methods that focus on color cannot annotate correctly if the background color is similar to the target object. In the background subtraction method, not only the target object but also the hand, spoon, and plastic bottle will be annotated. Furthermore, these methods have difficulties in handling multiple objects and instance segmentation, as will be described in Section VI-E.

B. Comparison b/w Proposed Method and Manual Method

The purpose of this section is to verify the accuracy of our method. We compare manual annotations to our created annotations created by using the *invisible marker*. In order to strictly compare the proposed method with manual annotations, it is necessary to create a manual annotation dataset and use it to train a network. However, even though the number of images in our dataset is only 11200, this experiment is difficult due to budgeting reasons as the manual annotation cost would be around $ 10000, thus we were unfortunately unable to conduct this experiment.

As an alternative to this costly method, two comparisons from a small sample were performed by using the intersection over union (IoU), which is simply a ratio of the union of two regions and the intersection of the two regions. The two comparisons are as follows:

1) Apply IoU only between manual annotations to measure individual differences, and
2) Apply IoU to the annotations created by the proposed method and the manual annotations to measure how close the proposed method is to manual annotations.

If both IoU values are similar, it means that our method using *invisible marker* is at least as accurate as human annotations. As for the comparisons, we selected three images for each object and each background, that is totally 54 images, and three people per image are assigned for annotation using Amazon Mechanical Turk (AMT).

TABLE I: Comparison of annotations between the manual method and the proposed method.

	cloth	liquid	powder
IoU among manual	94.0% (±3.55%)	93.6% (±2.21%)	84.6% (±10.4%)
IoU b/w manual & *invisible marker*	89.8% (±7.19%)	77.6% (±9.25%)	84.0% (±8.95%)

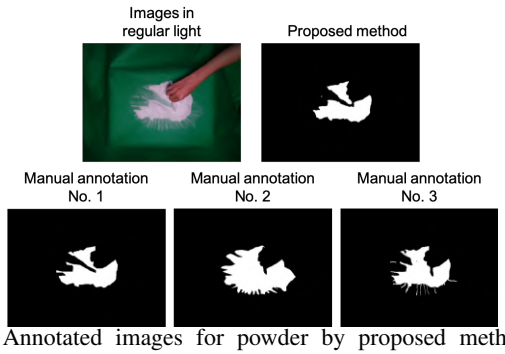

Fig. 5: Annotated images for powder by proposed method and manual by three people.

1) Differences Between Manual Annotations: Table I shows the mean value and standard deviation about the comparisons. In Table I, the IoU among manual annotations shows that individual differences seem to depend on the clarity of boundaries between the target object and other objects. Differences among different annotators are small for clothes and liquids because these boundaries are clear. On the other hand, individual differences are large when there is an unclear boundary between the object and others like small amount of powder around the boundaries. This can also be seen from Fig. 5, which shows images of both manually and automatically annotated powder. It can be seen that the manual annotations are clearly different from each other.

Manual annotator No.1 ignores fine powder area, while manual annotator No.2 tries to cover the entire area of the fine powder, and manual annotator No.3 is a type in between No.1 and No.2. When people annotated ambiguous scenes such as with powder, individual judgments creates variety in the datasets. Our proposed method can control the annotation result by adjusting the amount of fluorescent paint to control the light intensity emitted under UV light, and the threshold for the RGB values for extracting the segmentation mask. In this experiment, the amount of fluorescent paint and the threshold value are adjusted to ignore fine powder like manual annotator No. 1 does, since this area is generally too fine to be manipulated by robots.

2) Gaps b/w the Proposed Method and Manual Method: In Table I, the result of IoU between manual annotations and *invisible marker* shows that differences between manual annotation and *invisible marker* are more significant in liquids than clothes and powders. The first and second rows of Fig. 6 is one of the examples of a large gap between the image under regular light and the annotated images because of the fast movement of the objects and the hand. The movement causes the liquid to shift, and the position of the human hand is also shifted on the cloth. We think that the gaps for the liquid occurred more than for the cloth because the liquid moved more and faster than the cloth due to their inertial properties.

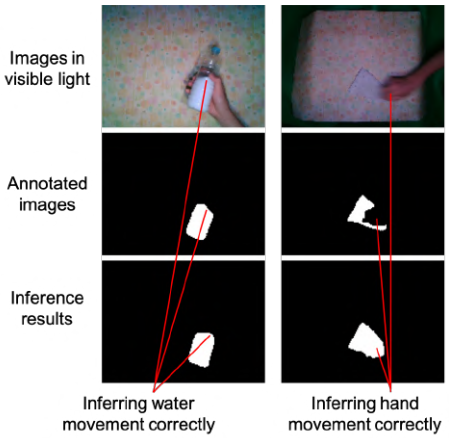

Fig. 6: Images for annotation gaps due to the fast movements and the gaps absorbed by deep learning

TABLE II: Comparison of annotations between the manual and proposed method using stationary liquid

	liquid
IoU among manual	94.6% (±1.95%)
IoU b/w manual & invisible marker	93.6% (±1.41%)

In order to investigate whether the capture timing is the main factor for causing the gaps, we evaluate the IoU in the same way as Table I using stationary liquid on the desk instead of liquid in movement. We selected four images of stationary liquid, and three people per image are assigned for annotation using AMT. The result is shown in Table II. As a result, it can be seen that there is only a small difference between the proposed method and the manual method. This camera hardware challenge can be alleviated if we use a high-speed camera, though there will always be a limitation to the camera sampling rate. In this research, we focus on a software approach to deal with the gaps in the dataset. By training a deep neural network on a large-scale dataset, such gaps are absorbed by the generalization capability of the network. The experiment result will be described in section VI-C.

C. Results of Semantic Segmentation

The purpose of this section is to test whether deep learning can be trained with the invisible marker datasets, and to evaluate if deep learning has enough generalization capabilities to absorb the gap between the captured image under regular light and UV light caused by the fast movement of the objects as described in section VI-B.2. Table III shows the IoU of the inferred segmentation mask on the validation data (which contains only untrained backgrounds), after applying 6-fold cross validation for 6 backgrounds. In U-Net, the accuracy for validation data is over 80% for all objects. Fig. 7 shows the examples of inferred results of clothes, liquids, and powders on untrained background data. As can be seen the accuracies are all around 80% in Table III and from Fig. 7, we can conclude that deep neural networks can be trained correctly with the created datasets.

As described in Section VI-B.2, the fast movement of the object creates a gap between captured images in regular light

TABLE III: IoU of semantic segmentation

IoU	cloth	liquid	powder
FCN	76.7%	84.8%	83.1%
U-Net	81.3%	87.8%	87.7%
SegNet	72.7%	83.4%	84.4%

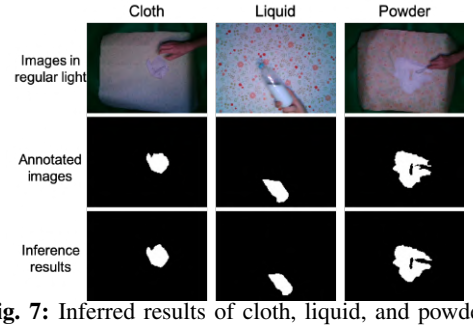

Fig. 7: Inferred results of cloth, liquid, and powder

and UV light (See the first and second rows of Fig. 6). The inferred results in the third row of Fig. 6 shows that the generalization capability of deep learning absorbs these gaps, and thus can still accurately infer the segmentation masks. These images are also inferred with untrained backgrounds. We conclude that our proposed method can create automatic and accurate annotation datasets of deformable object manipulation and that the quality is sufficient to train deep neural networks to correctly infer segmentation masks.

D. Inference for Non-mixed Fluorescent Objects

To investigate the effects of appearance changes due to fluorescent paint, we trained the networks using the datasets created by our method and subsequently validated it on objects which are **not** mixed with fluorescent paint. If the appearance is changed by the fluorescent paint, the network cannot infer the segmentation mask correctly for the object which is not mixed with fluorescent paint because it will simply look different from the training set. Fig. 8 shows the inference result of the liquid. The target object to be inferred is the liquid that would have had the greatest appearance change. The inferred results show that the segmentation mask is still accurate despite the absence of the paint. We can thus conclude that the change in appearance due to the fluorescent paint is small enough (if there is any) to be absorbed within the generalization capability of deep learning.

E. Several Colors of Invisible Marker and Multiple Objects

In the proposed method, the color of *invisible marker* can be changed for each object. Even if there are multiple objects, each object can be painted with different colors. Even in the situation where objects overlap, or same and/or multiple objects, each object can be annotated separately (Fig. 9 (a) shows the results of segmentation for multiple objects, overlapping objects, and instance segmentation of two white colored bottles and a brown colored bottle). In addition, it is also possible to create annotated datasets for only grasp points of an object by applying *invisible marker* to only one part instead of the whole object. Moreover, by giving different colors to each grasp point of the same object, robots can distinguish them from each other and manipulate them

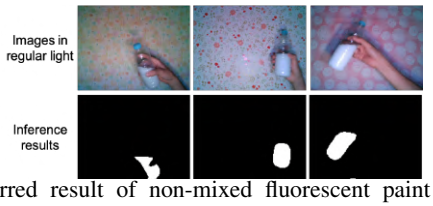

Fig. 8: Inferred result of non-mixed fluorescent paint using the trained networks through the datasets with fluorescent paint

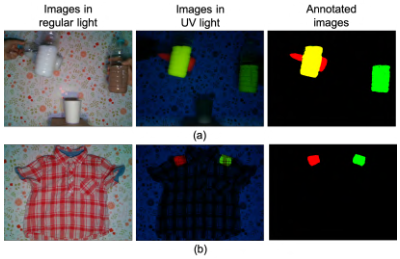

Fig. 9: (a) Semantic segmentation for multiple object, overlapped situation, and instance segmentation (b) Grasping points

according to their purposes (Fig. 9 (b) shows the annotation of two grasping points using two different class labels).

VII. Experiment Setup Under Regular Light

We conducted experiments under regular light conditions (uncontrolled environmental light conditions) similar to what we did under dark conditions (controlled environmental light conditions) described in section V.

A. Data Acquisition Device For Regular Light

The configuration of the data acquisition device is the same except that we only blink a UV light and don't use regular lights anymore since there is enough external environmental light.

B. Target object and Environment Under Regular Light

Among the deformable objects, liquids are widely used indoors and outdoors, such as in medical drug factories and laboratories, and home environments. We conducted liquid pouring tasks in various places e.g. indoors, by the window, and outdoors. The liquid is mixed with fluorescent paint.

C. Data Collection Under Regular Light

For data collection, images are captured during water pouring motions by the data acquisition system described in Section VII-A. The sampling rate of the camera is 30 Hz, thus the sampling rate for creating paired images is 15 Hz because images are taken while switching between UV and regular light. The total number of acquired pairs are 21645. Regular light conditions are difficult since the backgrounds are complicated and varied, so a large dataset was necessary. Details will be described in section VIII-A. Acquired images are resized from 320×240 to 160×120.

VIII. Results under Regular Light Conditions

A. Annotation result under Uncontrolled Light

We show examples of datasets for liquid pouring tasks captured under regular light conditions (uncontrolled environmental light conditions) (Fig. 1(b)). In addition to

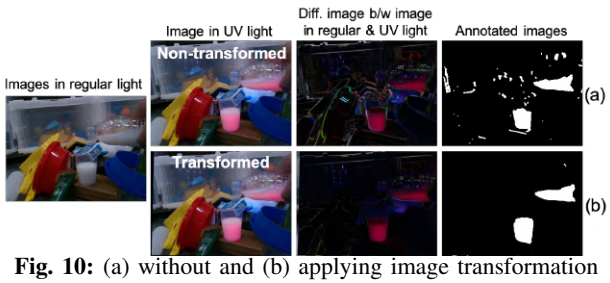

Fig. 10: (a) without and (b) applying image transformation

the examples, experiments are performed at various places indoors, near windows, and outdoors, so please check the attached video file and our published data. It is clear from Fig. 1(b) that the environmental light is not controlled, that is, it cannot be turned off while applying UV light; so not only the fluorescent paint but also everything else is visible. When UV light is applied, it can be seen that only the liquid mixed with the fluorescent paint changes red. It was expected that the fluorescent paint would always react and become red near the windows and outdoors since sunlight contains UV light. However, the UV light in sunlight is not strong enough to influence our experiment. From the difference between the image under regular light and the one under UV light, an annotation for a segmentation mask was achieved by extracting only the part that emitted light because of the fluorescent paint. From the above, it can be seen that the segmentation mask is created correctly even in a situation where the environmental light cannot be controlled and with realistic, complicated backgrounds.

B. Result of Image Transformation For Unfixed Viewpoints

In this section, we show the result of the image transformation to deal with differences of camera viewpoints between the image under regular light and the image under UV light, while moving the camera during capturing images. When moving the camera while taking pictures simultaneously with alternately switching between regular and UV light, a shift of viewpoints between the captured images with and without UV light occurred. Thus, the annotation mask no longer matches as a paired image captured under regular light because of this shift. Therefore, image transformation is necessary for the viewpoints of the two images to match. Fig. 10 shows the comparison before and after applying the image transformation. If the image transformation is not performed, not only the target liquid but also the background edges are annotated due to the viewpoint shift in images captured under regular light and UV light, respectively. On the other hand, when the image transformation is performed, the influence of the shift due to the viewpoint transformation is reduced, and it can be seen that only the target object is annotated correctly. From the above, it is shown that a dataset can be created without fixing the camera viewpoint, by performing this image transformation.

C. Comparison b/w the Proposed and Manual Method

In order to evaluate whether creating datasets under regular light conditions works like under dark conditions, we evaluate the IoU in the same way as we did for Table II, only this time under regular light conditions. We selected 9 images from our collected data in which the liquid is stationary. If there is a gap between the proposed method and the manual method, we can conclude that regular light conditions do not affect our data creation process. Three people per image are assigned for annotation using AMT, the comparison results are shown in Table II. As a result, it can be seen that there is only a small difference between the proposed method and the manual method even under regular light conditions. Thus, we conclude that our method can create dataset accurately even under regular light conditions.

TABLE IV: cf. annotations b/w the manual and proposed method using stationary liquid under regular light cond.

	Indoor	By window	Outdoor	Ave.
IoU among manual	85.2% (±4.82%)	79.3% (±7.65%)	92.1% (±2.08%)	85.5% (±7.40%)
IoU b/w manual & *invisible marker*	83.4% (±2.56%)	77.9% (±6.42%)	89.0% (±3.66%)	83.4% (±6.33%)

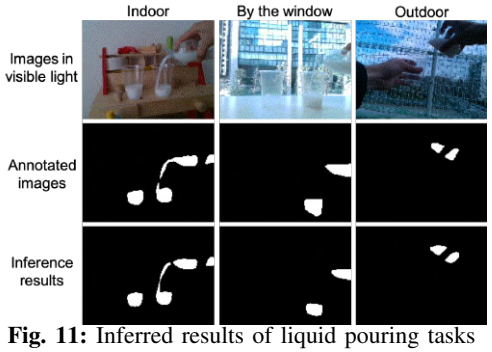

Fig. 11: Inferred results of liquid pouring tasks

D. Results of Semantic Segmentation Under Regular Light

The purpose of this section is to confirm whether deep neural networks can be trained with datasets acquired under regular light conditions. The results of IoU of the inferred segmentation mask on the validation data are 87.1%, 92.5% and, 90.1% for FCN, U-Net, and SegNet, respectively. The networks can infer segmentation mask correctly in complex environments such as indoor, by the window, and outdoor conditions (See Fig. 11. See more examples in the attached video). We can conclude that deep neural networks can be trained correctly with the datasets created by our method.

IX. Discussion for Cost

In this section, we discuss the cost of time and money for manual annotations and the proposed method. Table V shows the total cost for manual annotations and the proposed method. Manual annotations are costly for time and money, whereas coloring objects and the data acquisition system (which is mostly a one-time cost) is cheaper. In the proposed method, the application of the fluorescent paint took only a few minutes per object. Time per image is negligible since painting time can be ignored considering that we only paint once but take a lot of images per object. The cost of the fluorescent paint was $ 18 for all objects for all experiments. For the acquisition system, UV light as an additional device is required for the proposed method, but once we make the data acquisition system, it can be used again. For manual

TABLE V: Cost for manual method and the proposed method

	Manual annotation	Proposed method
Total images	32845	32845
Time in total for objects and image	10 [days] for recording images 22.8 [hours] for annotation[1] 0 [min] for paining	10 [days] for recording images 0 [min] for annotation Less than 5 [min] for painting
Time per image	About 2.5 [min] / image	About 0 [min] / image
Money in total for objects and images	$ 200 for camera $ 75 for LED $ 27590 for annotation[2]	$ 200 for camera $ 75 for LED $ 25 for UV light $ 18 for fluorescent paint
Total cost Money per image	$ 27865 $ 0.85 / image	$ 318 $ 0.0097 / image

[1] Average annotation time by AMT for this study is about 2.5 minutes per image.
[2] Calculated from the minimum recommended semantic segmentation cost of $ 0.84 per image in AMT.

annotations, the total cost increases as the number of images increases and the cost per image does not decrease as the datasets grow in size. Even annotation tools such as LabelMe [10] can only reduce the time and cost per image, but the total costs will still increase as more images are annotated. In the proposed method, cost of time and money per image decrease because colored objects can be reused once objects are painted by the *invisible marker*. Since deep learning requires thousands to tens of thousands of images, it is clear that the proposed method can create datasets quickly and inexpensively compared to manual annotations.

X. Conclusion

In this paper, we proposed a method to create annotations automatically using *invisible marker*, which is visible under UV light and invisible under regular light. By switching between regular light and UV light at high speed, our system can create large datasets of dynamical changing deformable object manipulation in both controlled and uncontrolled environmental light conditions even with unfixed viewpoints. The challenge of annotation gaps due to 1) the shift induced by the capture timing and 2) appearance change by *invisible marker* can be absorbed sufficiently by the generalization capabilities of deep learning. High accuracy of segmentation tasks is shown by multiple deep learning models such as FCN, U-Net, and SegNet trained on datasets created with our method. We conclude that our proposed method can create large datasets accurately, quickly and inexpensively.

These datasets can widen the range of robotic tasks such as folding clothes, cooking, and biomedical applications. For future work, we would like to look into manipulation tasks with a robot such as cooking and laundry folding.

ACKNOWLEDGMENT

The authors would like to thank Shunta Saito for helping to discuss and experiment, and Wilson Ko for proofreading and helping to write this article.

References

[1] J. Long, *et al.*, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[2] O. Ronneberger, *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[3] V. Badrinarayanan, *et al.*, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[4] T. Young, *et al.*, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.
[5] J. Hatori, *et al.*, "Interactively picking real-world objects with unconstrained spoken language instructions," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
[6] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
[7] M. Everingham, *et al.*, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
[8] T.-Y. Lin, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[9] H. Caesar, *et al.*, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218.
[10] B. C. Russell, *et al.*, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
[11] X. Huang, *et al.*, "The apolloscape dataset for autonomous driving," *arXiv: 1803.06184*, 2018.
[12] F. Yu, *et al.*, "BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling," *arXiv e-prints*, p. arXiv:1805.04687, May 2018.
[13] M. Andriluka, *et al.*, "Fluid annotation: a human-machine collaboration interface for full image annotation," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 1957–1966.
[14] A. Chang, *et al.*, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.
[15] C. Liensberger, *et al.*, "Color-based and context-aware skin detection for online video annotation," in *2009 IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2009, pp. 1–6.
[16] C. Schenck and D. Fox, "Detection and tracking of liquids with fully convolutional networks," *arXiv preprint arXiv:1606.06266*, 2016.
[17] S. Brutzer, *et al.*, "Evaluation of background subtraction techniques for video surveillance," in *CVPR 2011*. IEEE, 2011, pp. 1937–1944.
[18] E. Brachmann, *et al.*, "Learning 6d object pose estimation using 3d object coordinates," in *European conference on computer vision*. Springer, 2014, pp. 536–551.
[19] W. Kehl, *et al.*, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1521–1529.
[20] Z.-W. Hong, *et al.*, "Virtual-to-real: Learning to control in visual semantic segmentation," *arXiv preprint arXiv:1802.00285*, 2018.
[21] A. Hämäläinen, *et al.*, "Affordance learning for end-to-end visuomotor robot control," *CoRR*, vol. abs/1903.04053, 2019.
[22] A. Dufour, *et al.*, "Segmenting and tracking fluorescent cells in dynamic 3-d microscopy with coupled active surfaces," *IEEE Transactions on Image Processing*, vol. 14, no. 9, pp. 1396–1410, 2005.
[23] A. Rizk, *et al.*, "Segmentation and quantification of subcellular structures in fluorescence microscopy images using squassh," *Nature protocols*, vol. 9, no. 3, p. 586, 2014.
[24] S. Baker, *et al.*, "A database and evaluation methodology for optical flow," *International journal of computer vision*, vol. 92, no. 1, pp. 1–31, 2011.
[25] E. Rublee, *et al.*, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.
[26] L. Keselman, *et al.*, "Intel (r) realsense (tm) stereoscopic depth cameras," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1267–1276.
[27] "Nucleo-f303k8," https://www.st.com/en/evaluation-tools/nucleo-f303k8.html, 2019.
[28] S. Tokui, *et al.*, "Chainer: a next-generation open source framework for deep learning," in *Workshop on machine learning systems on Neural Information Processing Systems*, 2015.