Deep Gated Multi-modal Learning: In-hand Object Pose Estimation with Tactile and Image

Tomoki Anzai¹, Kuniyuki Takahashi*²

Abstract—In robot manipulation tasks, especially in-hand manipulation, estimation of the position and orientation of an object is an essential skill to manipulate objects freely. However, since in-hand manipulation tends to cause occlusion by the hand itself, image information only is not sufficient. For the challenge, combining tactile sensors is one of the approaches. The advantage of using multiple sensors (modals) is that the other modals can compensate for occlusion, noise, and sensor malfunctions. Even though the decision making of each modal reliability corresponding to the situations is important, the manual design of the model is difficult to deal with various situations. Therefore, in this study, we propose deep gated multi-modal learning using end-to-end deep learning in which the network self-determines the reliability of each modal. As experiments, an RGB camera and a GelSight tactile sensor were attached to the gripper of the Sawyer robot, and the poses were estimated during grasping. A total of 15 objects were used in the experiments. In the proposed model, the reliability of the modal was determined according to the noise and failure of each modal, and it was confirmed that the pose was estimated even for unknown objects.

I. INTRODUCTION

Robots are expected to be working not only in factories, but also in home environments. Personal robots are assumed to grasp objects in various environments. Sometimes obstacles in the environment disturb to realize the target grasping pose in reaching motion directly. In such a situation, the robots need to grasp the object once and then place it on the desk to grasp it again in order to achieve the target pose, or to perform a change-over motion for the object being grasped, in-hand manipulation. However, placing the object takes some time compared to in-hand manipulation. Furthermore, placing tables do not always exist in environments, where the method can not be used. For robot manipulation tasks, especially in-hand manipulation [1], [2], object pose estimation grasped by a robotic hand is essential to manipulate objects freely and accurately.

Although, many researchers have developed object pose estimation based on image information only, images alone are not sufficient for in-hand manipulation, where occlusion is likely to occur by the hand itself. One of the approach for the challenge is to combine multiple sensors (modals).

Hence, we propose a method that we call deep gated multi-modal learning (DGML) which uses end-to-end deep learning in which the network can predict and determine the reliability of each modal by DGML itself. By virtue of end-to-end deep learning, this method is capable of generalizing for unknown objects when estimating their pose in hand.

The rest of this paper is organized as follows. Related works and the contribution of this paper are described in Section II while Section III explains our proposed method. Section IV outlines our experiment setup and evaluation settings with results presented in Section V. Finally, future work and conclusions are described in Section VI.

II. RELATED WORKS & CONTRIBUTIONS

A. Object Pose Estimation with Depth and Image

Object pose estimation is a well-studied problem in computer vision and important for robotic tasks. Particularly many researchers have been developing methods using depth data (point cloud) or RGB-D data [3]–[5]. Classical approaches with depth data are mainly based on point cloud matching method such as interactive closest point (ICP) [6]. These methods can achieve high accuracy. However, since this method requires 3D models of objects, unknown objects...
can not be applied.

Recently, deep learning has become an active research area and especially computer vision field has achieved a lot of success. Then, object pose estimation with image through combination of deep learning with model-based approaches has been studied [7]–[9]. The convergence error by initial position deviation of the 3D model with the target object for ICP can be suppressed by deep-learning-based methods. Since this method also requires a 3D model, adapting to unknown objects remains challenging. In state of arts in pose estimation, methods that do not require 3D models have been studied by using deep learning [10]–[12]. Although this makes it possible to adapt to unknown objects without a 3D model, objects with occlusion remain as challenges.

Note that OpenAI implemented the task of aligning the direction of the cube using a multi-fingered hand [13]. Although this can be realized on the basis of an image including occlusion, it remains a challenge about precise pose estimation because required accuracy of pose estimation is only to align the six specified faces of the cube.

B. Tactile-based Object Pose Estimation

As common to pose estimation researches described in Section II-A, most of them mainly use images and depth information only. These methods are difficult to apply to in-hand manipulation because of the occlusion by the hand in the image or depth. Since tactile sensors can observe the contact state without causing a hidden state by the object or the hand, tactile sensors are attracting attention [14]–[17]. There are some researches in which the tactile and image information are combined [18]–[20], and the effectiveness of the fusion of tactile and visual sensing was demonstrated for object grasping. Some of the research performed object pose estimation with model based approach using 3D model [21]. To the best of our knowledge, there is still little research on object pose estimation using tactile sensors that does not require a 3D model for unknown objects.

C. Multi-modal Learning

A variety of methods have been employed in existing works on multi-modal learning. The majority of these methods falls in either of the three categories about how to pay attention to each modal.

1) Equally attention to all modals: Each modal information is input to the each network for extracting features, and the obtained features are simply combined to estimate information such as grasping point and motion. For example, the combination of image and tactile described in Section II-B force and image [22], language and image [23], and, image and sound [24].

2) Attention in each modal: Only the important parts of each modal will be used. Part to be attention in each modal is acquired by a network itself. For example, in the case of images, the pixel of interest in the image is used [25].

3) Attention to the modal: Only the important modals are used while other modals are ignored. The network makes decision which modal should be used from modals.

For example, choose important modals from language and images according to the situation [26].

In this study, we develop a new approach that the network determines the reliability of each modal, and it changes the amount used of each modal according to the reliability of each modal.

D. Development of Tactile Sensors

To achieve the skillful manipulation by robots, many tactile sensors have been developed and some of them have been applied to robotic hands [27]. The majority of these sensors falls in either of the following two categories. 1) Multi-touch enabled sensors with sensing capabilities limited to one axis force information vertical to the surface of the sensor, which resembles the pressure sense [28]–[31] or 2) three-axis sensing enabled sensors for only a single cell [32]. Two of the few exceptions are the uSkin [15] and the GelSight [17], [33] which can obtain share force as well as pressure force with multi-touch. The commercialized uSkin sensor [15] utilizes embedded magnets in a silicon rubber to measure the deformation of silicon during contact by monitoring changes of the magnetic fields. Using this method, it is able to measure both normal as well as shear forces per sensor unit for 16 contact points in this prototype [14]. Instead of magnet, GelSight [16], [19] is an optical-based tactile sensor, which uses a camera to capture and measure the deformation of its attached elastomer during contact with a surface.

E. Contributions

The target of our method is to estimate object pose in hand. The main contributions of this article are following:

• Combine image and tactile information to cope with occlusion situation.
• Use end-to-end learning without assuming a 3D model to estimate the pose of unseen objects.
• Propose the new approach in which the network itself decide the reliability and contribution of each modal.
• Investigate details of noise, malfunction, and occlusion behavior in sensor information unique to the robot field.

III. DEEP GATED MULTI-MODAL LEARNING

We propose an object pose estimation method with image and tactile sensing based on end-to-end deep learning. In order to consider the reliability of each modal, we propose a method using gate. Used in long short-term memory (LSTM), information from each neuron can be adjusted by opening and closing the gate. In this study, the amount of information in each modal is adjusted by changing the gate continuously from 0 to 1 instead of 0 and 1. We show the concept of proposed network model in Fig. 2. DGML is composed of four components for in-hand pose estimation as follows:

• Extraction feature parts from image and tactile by convolutional neural network (CNN)
• Reliability part of each modal by gate
• Reflection of reliability to each modal
• Estimation of the object pose using time-series input
information that reflects reliability by LSTM and full connected (FC)

An image and a tactile data are the input, as well as, a relative pose of an object is the output. Since our proposed method address without 3D models given in advance, we estimate the relative pose the basis of which is the pose at the moment the object is grasped, not the absolute pose. In other words, how much the object moves within the hand after the robot start to grasp the object is estimated.

Extraction feature parts are calculated with training data \( D = \{(input_1, output_1), ..., (input_T, output_T)\} \) with \( T \) step as follow:

\[
\begin{align*}
    x_{img} &= f_{\theta_1}(I_{img}) = s_{img}(W_1 I_{img} + b_1) \\
    x_{tac} &= f_{\theta_2}(T_{ac}) = s_{tac}(W_2 T_{ac} + b_2) \\
\end{align*}
\]  

Then, reliability value \( \alpha \) in image is given by:

\[
\alpha = \text{Reliability}(x_{img}, x_{tac}) = \text{sigmoid}(f_{\theta_3}(x_{img}) + f_{\theta_4}(x_{tac}))
\]  

The reliability values \( \alpha \) and \( \beta \) of each module in image and tactile has condition:

\[
\alpha + \beta = 1
\]  

Gate determines relative modal reliability from all modal information. Therefore, the modal reliability is calculated by integrating the input from each modal. By calculating so that the sum of each modal gate is 1, relative reliability is expressed. The active function of \( \text{Reliability} \) is sigmoid. \( \alpha \) and \( \beta \) values are between 0 and 1. The lower the modal reliability, the closer the value is to 0 and the smaller the contribution to the output. On the other hand, the higher the reliability of the modal, the closer the value is to 1, and the greater the output contribution. In other words, if the value is 0, the modal is completely ignored.

Then, reliability is reflected to each modal as:

\[
\begin{align*}
    h_{img} &= \alpha \cdot x_{img} \\
    h_{tac} &= \beta \cdot x_{tac}
\end{align*}
\]  

Finally, the object pose as output is calculated:

\[
\begin{align*}
    y' &= f_{\theta_5}(\text{concat}[h_{img}, h_{tac}]) \\
    &= \text{FC}(LSTM(\text{concat}[h_{img}, h_{tac}]))
\end{align*}
\]  

To minimize the cost function \( L \) as follows

\[
    \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, y'_i)
\]  

where \( s \) is the activation functions, \( \theta = (W, b) \) are the parameters to be trained, \( f_{\theta} \) includes the activation function and \( \theta \), and \( L \) is Huber loss function, respectively. \( y \in \mathbb{R}^d \) is the expected output, and \( y' \in \mathbb{R}^d \) is the inferred output from input \( x \in \mathbb{R}^d \).

There is no teacher signal for the reliability values \( \alpha \) and \( \beta \), and these are calculated by the network itself to minimize the output error by \( \theta \). Then, the reliability values are multiplied by each modal in \( \theta \). For this reason, when training the networks with connecting multiple sensors, modals that do not contribute to train data are ignored. Therefore, it is not necessary to design which sensor to use in advance manually.

### IV. EXPERIMENTAL SETUP

The purpose of experiments is to verify the DGML in occlusion situation, followed by details investigation for noise, and malfunction in sensor information.

We note that our values of the hyper-parameters provided in this section are tuned by random search.

#### A. Hardware Setup

1) Tactile sensor: The GelSight tactile sensor we duplicate from article [16], [19] is an optical-based tactile sensor, which captures 640 \( \times \) 480 pixel image by a camera. The sensor is capable of measuring applied pressing force and shear force in the x, y, and z axes from captured image. We use captured raw image as tactile sensor. Applying an excess amount of force results in tearing the silicon layer from the sensor’s container. To prevent this problem, we applied baby powder on its silicon surface to reduce friction between the surface and the grasped object.
Fig. 3: Setup used in our experiments. Custom printed end-effector with both a tactile skin sensor and a web camera. The Sawyer robot let the gripper move to the minus x-axis, y-axis direction and rotation of yaw.

Fig. 4: Trained materials (red) and unknown materials (blue)

2) Gripper: We developed a gripper to grasp objects shown in Fig. 3. This gripper is a parallel gripper which has two fingers driven by a servo motor (Dynamixel XM430-W350-R). GelSight tactile sensor is attached to one fingertip, and the other fingertip has a sponge. A camera (BUFFALO BSW200MBK) is mounted on the center of the gripper.

3) Sawyer: To perform our experiments, we use a Sawyer 7-DOF robotic arm with the gripper as end-effector (See Fig.3). The Sawyer, GelSight sensor, gripper, and camera are connected to a PC running Ubuntu 16.04 with ROS Kinetic.

B. Objects

For the target objects, we have prepared 15 objects with various size and shape (See Fig. 4). 11 of these objects are used for training, while the remaining 4 were used to evaluate our trained network as unknown materials.

C. Data Collection

Fig. 3 shows one of the initial position the robot starts to manipulate the object for data collection. A table is placed in front of the robot and the object is fixed on the table with a double-sided tape. The robot grasps the object, and the image and tactile data are recorded while the robot slides the object in its hand. We estimate the three DoF object pose. Given the coordinate system of the hand which is defined as shown in Fig. 3, the estimated values are \( x, y, \) and \( \theta \).

While we estimate the object pose relative to the pose when the robot start to grasp it, change of the object pose can be calculated by change of the hand pose since the object is fixed to the table. We define homogeneous transformation matrix \( H_t \) as the pose of the gripper at time \( t \) described in base link coordinate system. This can be calculated easily by forward kinematics. The change of the object pose from time \( t \) to \( t + 1 \) can be represented as \( H_{t+1}^{-1} H_t \).

We prepare the movement patterns including both translational and rotational motions, and collect data for each object. In order not to depend on the features of background movement as the robot moves, green clothes cover the background and desk. The maximum movement in translation was about 30 mm as well as in rotation was about 40 degree. For small objects such as tapes, cups, scale, and wrench, only rotational movement was performed. In each object, the numbers of motion for translation, rotation, and combination of both of them are 10, 10, and 12, respectively. In trained object in Fig. 4, datasets of 6 out of 10 in translation, 6 out of 10 in rotation, and 8 of 12 in combination of both are used for training, and the remaining data are used for evaluation.

Images, tactile, and object poses were acquired at 30 Hz, and the dataset used for training was resampled to 15 Hz. The captured images were converted to grayscale because the object pose is independent of the color of the objects. One motion of the training data set has a length of about 150 steps, and each step composes of \( 80 \times 120 \) pixels for image, \( 160 \times 120 \) pixels for tactile, and 3 for object pose.

D. Network Design

The architecture of our network model is composed of two CNNs with gate, and LSTM and FC to perform DGML as shown in Fig. 2 as described in Section III. We use Chainer [34]–[36] as deep learning library, for implementation. More details on the network parameters are shown in Table I. All our network experiments were conducted on a machine equipped with 256 GB RAM, an Intel Xeon E5-2667v4 CPU, and eight Tesla P100-PCIE with 12GB resulting in about 24 to 48 hours of training time.

V. Results

A. Learning Curve of DGML

As a comparison with DGML, we prepared several networks that use only images, only tactile, and both images and tactile with simple connection without using gate. Fig. 5 shows their learning curves and gate values of DGML. As for the values of \( \alpha \) and \( \beta \), the learning of images progresses at first, but it can be seen that the reliability of tactile information gradually increases. As one of the characteristics of DGML, computational epoch to the convergence of DGML is fastest in the models since training progresses from easy-to-train modals.
TABLE I: Network Design

| Layer      | In | Out | Filter size | Activation function |
|------------|----|-----|-------------|---------------------|
| **Image**  |    |     |             |                     |
| 1st conv.  | 1  | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (4,4)       | -                   |
| 2nd conv.  | 32 | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (2,2)       | -                   |
| 3rd conv.  | 32 | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (2,2)       | -                   |
| **Tactile**|    |     |             |                     |
| 1st conv.  | 1  | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (4,4)       | -                   |
| 2nd conv.  | 32 | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (2,2)       | -                   |
| 3rd conv.  | 32 | 32  | (3,3)       | ReLu                |
| Average pooling | 32 | 32  | (2,2)       | -                   |
| **Reliability** | |    |             |                     |
| 1st img (FC) | 1120 | 1 | - | - |
| 1st tact (FC) | 2240 | 1 | - | - |
| 2nd Gate | 2 | 1 | - | sigmoid |
| **LSTM**   |    |     |             |                     |
| LSTM       | 3360 | 170 | - | sigmoid & tanh |
| **FC**     |    |     |             |                     |
| 1st FC     | 170 | 3  | - | - |

1. In and out are the number of channel for image and tactile, whereas these are the number of neurons for gate, LSTM, and FC. Stride and padding for n-th convolution in image and tactile are (1, 1).

B. Gate Values of Objects

We visualize the reliability of image modal by creating the histogram with all the inferred values of $\alpha$ through untrained motions (See Fig. 6). The average value of $\alpha$ smaller than 0.5 indicates that the network relies more on the tactile. In this experiment, since the occlusion by the robot gripper itself and the object itself always occurs, the reliability of the image becomes low.

The images of the objects in the histogram are displayed on $\alpha$ that are often appeared (See Fig. 6). The ratio between the each object size in the image and the actual object size is the same. Since objects such as tapes, wrench, and cups are prone to tactile changes when they are grasped, the network relies on more tactile information for these objects. Whereas, image information is reliable for objects with no surface irregularities that are difficult to observe features by the tactile sensor. Even though the object has no features on the surface, the reliability of the image is low when a large object makes occlusion in the image. For example, the coffee plastic bottle and orange triangle object are small, so movement can be observed from the image. In contrast, large objects such as the mustard and green plastic bottles make occlusions, which reduces the reliability of image information, therefore the network uses more tactile information. Therefore, we can say that the proposed method is effective to determine the reliability of each module according to the object size and shape.

C. Inference Result of Object Pose

Table II shows the object pose inference error of in-hand manipulation. The evaluation conditions include without noise and the absence of either image or tactile input. Under such conditions, we compared four models: an image only, a model using only tactile, a model using image and tactile (no gate), and the proposed model. From the result without noise, tactile information should be included in the models for accurate in-hand pose estimation. In this condition, there is no difference in performance between the models including the tactile sensor. The networks can predict correctly not only for known objects but also for unknown objects.

From the result with the absence of one of the modals in Table II, the performance of the proposed DGML is the best. This is because if one of the models is absence, the gate can make decision that the reliability of the modal should be reduced, then, the modal is almost ignored (See Table III). Even though the training dataset does not include data that the absence one of the modals, the gate of DGML can deal with these situations. Even though the training dataset does not include data in which one of the modals is absent, the gate of DGML can deal with these situations.
**TABLE II:** Inference error of in-hand object pose estimation

| Cond. | Model | Known obj. | | | | | Unknown obj. | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | trans. [mm] | Rot. [deg] | trans. [mm] | Rot. [deg] | | | | | | | | |
| w/o noise | Image | 2.10 | 1.45 | 3.25 | 5.47 | | | | | | | | |
| | Tactile | 5.99 x 10^{-1} | 1.43 | 1.09 | 2.04 | | | | | | | | |
| w/o gate | DGML | 7.48 x 10^{-1} | 8.38 x 10^{-1} | 1.01 | 1.70 | | | | | | | | |
| w/o image | w/o gate | 3.00 | 8.13 | 3.01 | 9.00 | | | | | | | | |
| | DGML | 1.04 | 2.35 | 1.39 | 3.33 | | | | | | | | |
| w/o tactile | w/o gate | 5.17 | 1.35 x 10^{1} | 5.02 | 1.44 x 10^{1} | | | | | | | | |
| | DGML | 3.49 | 1.19 x 10^{1} | 3.87 | 1.39 x 10^{1} | | | | | | | | |

**TABLE III:** The average value of $\alpha$ and $\beta$

| Cond. | Known obj. | Unknown obj. | | | | | |
|---|---|---|---|---|---|---|---|
| w/o noise | 0.261 | 0.739 | 0.272 | 0.727 | | | |
| w/o image | 0.069 | 0.931 | 0.072 | 0.928 | | | |
| w/o tactile | 0.994 | 0.006 | 0.994 | 0.006 | | | |

**D. Gate Values in Different Noise**

In this section, we discuss the change of reliability when a noise is applied to an input of a modal. Fig. 7 shows the reliability value $\alpha$ when the network infers from the dataset with added noises to the tactile input. The noises are generated by the normal distribution with various variance. From Fig. 7, it can be seen that the reliability of the image increases as the noise of tactile increases. The gate can correctly recognize the noise of the input signal in the tactile. The strength of using DGML is that the DGML helps to understand the network through the values of the gate because the gate represents which and how much modal should be used. Furthermore, if the sensor is broken, the reliability value is close to 0, so DGML can judge the sensor failure.

**VI. CONCLUSION**

In this paper, we proposed a method to estimate in-hand object pose using tactile and image with modals’ reliability, called deep gated multi-modal learning (DGML). The proposed method can estimate not only known object pose but also unknown object pose without 3D model. Moreover, the modals’ reliability can be changed by the network itself depending on the situations such as sensor failure and different size of noise level. Visualization of reliability helps to understand the network behavior such as which modal and how much to use. Furthermore, by using the proposed method, computation efficiency for training has been improved.

For future work, we will develop a system of in-hand object manipulation to integrate the object pose estimator.

**REFERENCES**

[1] Hanna Yousef, Mehdi Boukallel, and Kaspar Althoefer. Tactile sensing for dexterous in-hand manipulation in robotics—a review. Sensors and Actuators A: Physical, 167(2):171 – 187, 2011. Solid-State Sensors, Actuators and Microsystems Workshop.

[2] Chavan-Dafle Nikhil, Holladay Rachel, and Rodriguez Alberto. In-hand manipulation via motion cones. Robotics: Science and Systems (RSS), 2018.

[3] C. Choi, Y. Taguchi, O. Tuzel, M. Liu, and S. Ramalingam. Voting-based pose estimation for robotic assembly using a 3d sensor. In 2012 IEEE International Conference on Robotics and Automation, pages 1724–1731, May 2012.

[4] C. Choi and H. I. Christensen. 3d pose estimation of daily objects using an rgb-d camera. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3342–3349, Oct 2012.

[5] A. Aldoma, Z. Marton, F. Tombari, W. Wohlkinger, C. Potheast,
Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. In Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.

B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. IEEE Robotics Automation Magazine, 19(3):80–91, Sept 2012.

[6] J. Bimbo, S. Rodriguez-Jimenez, H. Liu, X. Song, N. Burrus, L. D. Senerivatine, M. Abderrahim, and K. Althoefer. Object pose estimation and tracking by fusing visual and tactile information. In 2012 IEEE International Conference on Multimedia Sensor Fusion and Integration for Intelligent Systems (MFI), pages 65–70, Sept 2012.

[7] Roberto Calandra, S. Rodríguez-Jimenez, H. Liu, X. Song, N. Burrus, L. D. Senerivatine, M. Abderrahim, and K. Althoefer. Object pose estimation and tracking by fusing visual and tactile information. In 2012 IEEE International Conference on Multimedia Sensor Fusion and Integration for Intelligent Systems (MFI), pages 65–70, Sept 2012.

[8] Y. Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseNet: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

[9] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cesu Lu, Li Fei-Fei, and Silvio Savarese. Densefuson: 6d object pose estimation by iterative dense fusion. arXiv preprint arXiv:1901.04780, abs/1901.04780, 2019.

[10] M. Schwarz, H. Schulz, and S. Behnke. RGB-d object recognition and pose estimation based on pre-trained convolutional neural network features. In 2015 IEEE International Conference on Robotics and Automation (ICRA), pages 1329–1335, May 2015.

[11] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlenBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihre, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 19–34, 2018.

[12] Philipp Mittendorfer and Gordon Cheng. Humanoid Multimodal Tactile-sensing Modules. IEEE Transactions on Robotics, 27(5):401–410, 2011.

[13] Akiba Takuya, Fukuda Keisuke, and Suzuki Shuji. ChainerMN: A next-generation open source framework for deep learning. In Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.

[14] Akiba Takuya, Fukuda Keisuke, and Suzuki Shuji. ChainerMN: Scalable Distributed Deep Learning Framework. In Proceedings of Workshop on ML Systems in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[15] Yusuke Niitani, Toru Ogawa, Shunta Saito, and Masaki Saito. Chainrerec: a library for deep learning in computer vision. In ACM Multimedia, 2017.