## Supplementary Materials of

# Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes

*Shaowei Cui, Rui Wang, Junhang Wei, Jingyi Hu, Shuo Wang*
**IEEE Robotics and Automation Letters (IROS Option)**

## 1. Dataset D0

Since the visual and tactile data provided by dataset **D0** are both RGB images, we set both visual and tactile encoding functions as the CNN architectures for the proposed model. In this paper, the first four layers of ResNet18 [1] are used as the feature extract network for both the two CNNs, and output two $7 \times 7 \times 512$ feature maps.

The detailed parameters of the VTFSA model on dataset **D0** is shown in Table 1. Subsequently, we constructed some baselines using the direct-fusion (DF) method, the model parameters with different inputs (**I0**, **I1**, **I2**, and **I3**) are shown in Table 2. Note that the DF model with input **I0** is the original model of [2].

Table 1: Detailed network parameters of the VTFSA model on Dataset **D0**.

| **Functions** | **Operations (I0, I1, I2, I3)** | **Output Shape** |
|---|---|---|
| $E_v$ | Resnet18 (The first 4 layers) | $7 \times 7 \times 512$ |
| $E_t$ | Resnet18 (The first 4 layers) | $7 \times 7 \times 512$ |
| $F_{v,t,p}$ | $\oplus$ | $49 \times 49 \times 1024$ |
| $\mathbb{F}_{v,t}$ | VTFSA module | $49 \times 49 \times 1024$ |
| $\hat{\mathbb{F}}_{v,t}$ | AdaptiveAvgPool2d ((1, 1)) | $1 \times 1 \times 1024$ |
| $\mathbb{F}_c$ | FC(1024, 128), FC(128, 2) | $1 \times 1 \times 2$ |

## 2. Dataset D1

The detailed parameters of the DF model and VTFSA model on **D1** dataset are shown in Table. 3.

Table 2: Detailed network parameters of DF model on Dataset **D0**.

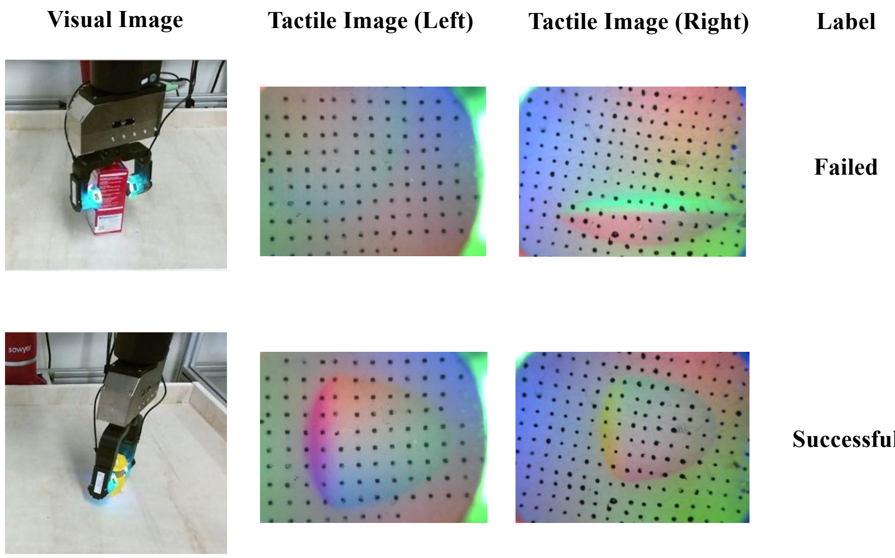
| **Layers** | **Operations (I0)** | **Output Shape** |
|---|---|---|
| Visual CNN | Resnet18 (avg-pool) | 2048 ($V_{pre}$, $V_{dur}$) |
| Left Tactile CNN | Resnet18 (avg-pool) | 2048 ($TL_{diff}$, $TL_{dur}$) |
| Right Tactile CNN | Resnet18 (avg-pool) | 2048 ($TR_{diff}$, $TR_{dur}$) |
| Concatenation | $V_{pre} \oplus$ ... $\oplus TR_{dur}$ | $2048 \times 6$ |
| $\text{FC}_1$ | Linear ($2048 \times 6$, 128) | 128 |
| $\text{FC}_2$ | Linear (128, 2) | 2 |
| **Layers** | **Operations (I1, I2, I3)** | **Output Shape** |
| Visual CNN | Resnet18 (avg-pool) | 2048 ($V$) |
| Tactile CNN | Resnet18 (avg-pool) | 2048 ($T$) |
| Concatenation | $V \oplus T$ | $2048 \times 2$ |
| $\text{FC}_1$ | Linear ($2048 \times 2$, 128) | 128 |
| $\text{FC}_2$ | Linear (128, 2) | 2 |

Table 3: Detailed network parameters of models on Dataset **D1**.

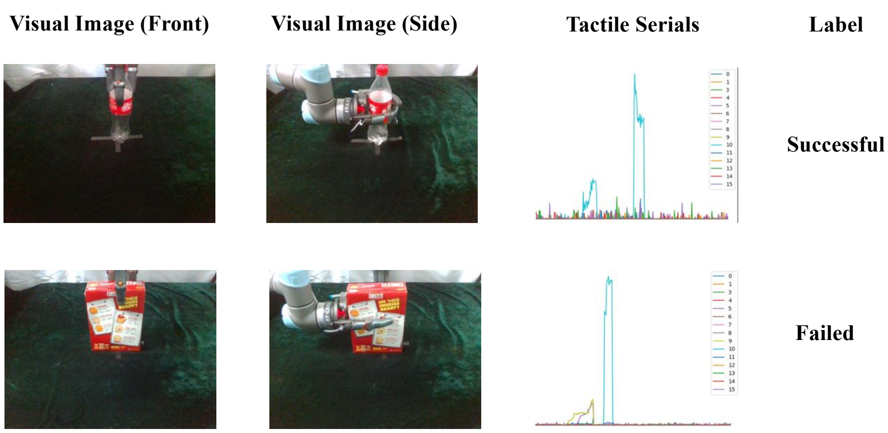
| **Functions** | **Operations (VTFSA)** | **Output Shape** |
|---|---|---|
| $E_v$ | Resnet18 (The first 4 layers) | $7 \times 7 \times 512$ |
| $E_t$ | LSTM (layers 1, hidden 256) | $1 \times 1 \times 64$ |
| $F_{v,t,p}$ | $\oplus$ | $7 \times 7 \times 576$ |
| $\mathbb{F}_{v,t}$ | VTFSA module | $7 \times 7 \times 576$ |
| $\hat{\mathbb{F}}_{v,t}$ | AdaptiveAvgPool2d ((1, 1)) | $1 \times 1 \times 576$ |
| $\mathbb{F}_c$ | FC(576, 72), FC(72, 2) | $1 \times 1 \times 2$ |
| **Layers** | **Operations (DF)** | **Output Shape** |
| Visual CNN | Resnet18 (avg-pool) | 2048 ($V$) |
| Tactile LSTM | LSTM (layers 1, hidden 256) | 64 ($T$) |
| Concatenation | $V \oplus T$ | 2112 |
| $\text{FC}_1$ | Linear (2112,128) | 128 |
| $\text{FC}_2$ | Linear (128, 2) | 2 |

## 3. Some examples of datasets

Some examples of the two grasping datasets are shown in Fig. 2. Since they are public datasets developed by other scholars, we do not add them to our manuscript, but add them in our supplementary materials.




(a) **D0**




(b) **D1**

Figure 2: Some examples of the two grasping datasets. Please note that these images are from dataset **D0** [3] and **D1** [4].

## References


[1] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, “Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, NV, USA, Jun. 2016, pp. 770-778.

[2] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, “The feeling of success: Does touch sensing help predict grasp outcomes?," *arXiv preprint arXiv:1710.05512*, 2017.

[3] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, “The feeling of success: Does touch sensing help predict grasp outcomes?," *arXiv preprint arXiv:1710.05512*, 2017.

[4] C. Yang, P. Du, F. Sun, B. Fang, and J. Zhou, “Predict Robot Grasp Outcomes based on Multi-Modal Information," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, Dec. 2018, pp. 1563–1568.