

Towards Comprehensive Multimodal Perception: Introducing the Touch-Language-Vision Dataset

Ning Cheng¹, You Li¹, Jing Gao¹, Bin Fang², Jinan Xu¹, and Wenjuan Han^{1,3,4,✉}

¹ Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China

² School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

³ China Railway Design Corporation, Tianjin, China

⁴ National Engineering Research Center for Digital Construction and Evaluation of Urban Rail Transit, Tianjin, China
wjhan@bjtu.edu.cn

Abstract. Tactility provides crucial support and enhancement for the perception and interaction capabilities of both humans and robots. Nevertheless, the multimodal research related to touch primarily focuses on visual and tactile modalities, with limited exploration in the domain of language. Beyond vocabulary, sentence-level descriptions contain richer semantics. Based on this, we construct a touch-language-vision dataset named TLV (Touch-Language-Vision) by human-machine cascade collaboration, featuring sentence-level descriptions for multi-mode alignment. The new dataset is used to fine-tune our proposed lightweight training framework, STL_V-Align (Synergistic Touch-Language-Vision Alignment), achieving effective semantic alignment with minimal parameter adjustments (1%). Project Page: <https://xiaoen0.github.io/touch.page/>.

Keywords: Tactile-related multimodal perception, Tactile dataset, Modal Alignment.

1 Introduction

Tactile perception occupies a distinctive and pivotal role within the human sensory system, constituting a fundamental basis for our cognitive comprehension of the environment, coexisting harmoniously with other sensory modalities, such as vision and audition. Tactility allows us to perceive the texture, temperature, and hardness of objects etc., and enables us to explore environments and perform intricate tasks, such as grasping and manipulating. The significance of touch is evident not only in humans [1, 2] but also in robotic applications [3, 4], where the acquisition and processing of tactile information are crucial for enhancing the perceptual capabilities and interaction efficiency of these applications.

Despite the undeniable significance of touch, tactile-related multimodal research predominantly focuses on the visual and tactile [5-7], with limited exploration in the

domain of language. While there are some works related to language, they remain primarily at the lexical level, serving as labels for classification purposes [8-10]. This arises from the heightened challenges associated with annotating lengthier texts, including intricate narratives and elevated expenses.

Continuous innovation in image-to-text models [11, 12] enables the generation of fluent text from prompts and images, thereby offering opportunities for tactile annotation with longer texts. In this work, we introduce a tactile-related multimodal dataset, named TLV (Touch-Language-Vision), through human-machine cascade collaborative annotation. TLV incorporates three modalities: touch, language, and vision, with pairwise correspondence between any two modalities, aiming to strengthen alignment between touch and language. Compared to a set of vocabularies (i.e., lexical-level descriptions), the descriptions in TLV are at the sentence level, capable of conveying more rich and more complete semantic information.

To assess TLV's efficacy, we employ it as the training dataset and present a lightweight unsupervised training method, STLV-Align (Synergistic Touch-Language-Vision Alignment). This method maps all modalities to a shared embedding space, enabling effective semantic alignment. To improve training efficiency, we employ Low-Rank Adaptation (LoRA) [13] for fine-tuning and only 1% of the parameters are adjusted. Subsequently, we evaluate the performance of STLV-Align on various tactile classification tasks using a cross-domain dataset. Experimental results demonstrate the potential of the TLV dataset. This paper presents the following contributions:

- Introducing TLV, a new touch-language-vision dataset with sentence-level descriptions annotated by human-machine cascade collaboration, Addressing the challenge of tactile annotation for longer texts.
- Proposing STLV-Align, a lightweight joint pretraining framework characterized by independence from labeled data, the utilization of a smaller dataset, the adjustment of model parameters, and acceptable performance.
- Validating the effectiveness of our dataset and method and providing direction for further optimization on tactile-related tasks.

2 Related Work

2.1 Tactile Perception

Extracting and leveraging tactile information, encompassing surface texture, elasticity, and temperature, holds substantial promise for advancements in both robotics and AI research [14-16]. Current tactile sensors primarily rely on vision, employing a camera and illumination system to record deformations in a curved elastomeric gel. This structure has given rise to diverse perception systems, including GelSight [8, 17-23], DIGIT [7, 24, 25], and GelSlim [26, 27]. These systems aim to comprehensively record high-resolution, detailed tactile information. Among them, GelSight stands out as one of the most widely used tactile perception systems, offering elaborate capture of depth, shear,

and surface orientation. This work leverages tactile observations primarily from GelSight.

2.2 Tactile Datasets

A significant challenge in learning from the tactile modality lies in the substantial human effort and time required to construct high-quality datasets. Despite this hurdle, the research community’s continuous efforts have yielded several publicly available datasets: Objectfolder 2.0 [28] (featuring 1,000 implicitly represented objects generated through simulation), SSVTP [7] (containing 4.5K spatially aligned image-tactile pairs acquired using DIGIT), the Feeling of Success [20] (employing a two-finger gripper and GelSight sensor), Touch and Go [8] (a high-quality, in-the-wild dataset encompassing diverse categories and quantitative visuo-tactile pairs) and VisGel [17] (a dataset comprising over 12K touch instances and 3 million vision-touch frames). However, these datasets are in the absence of rich textual descriptions, hindering their potential for realizing higher-level cross-modal alignment. This work addresses this limitation by incorporating detailed and qualitative captions, fostering a more comprehensive and advanced cross-modal understanding.

2.3 Multimodal Alignment

Effectively aligning semantics from diverse modalities is fundamental and crucial in multimodal research, yet it had been challenging to construct a high-dimensional joint embedding space incorporating features of different modalities. CLIP [29] achieved remarkable performance and generalization ability through self-supervised contrastive pretraining on a massive dataset of 400 million image-text pairs scraped from the internet. Subsequent works such as ALIGN [30], Flamingo [31], OpenCLIP [32] have further advanced the field towards more robust and accurate alignment. Beyond vision and language modalities, significant research efforts aim to bridge the gap between even more diverse modalities, including 3D points [33, 34] and audio [35, 36]. ImageBind [37] significantly extended the joint embedding space to encompass six distinct modalities through image-centered contrastive learning, further promoting comprehensive cross-modal understanding. Along this line, LanguageBind [38] proposed a language-centered alignment strategy to fully leverage the rich semantic information within the text, achieving significant performance improvements. This work builds upon these advancements by further expanding cross-modal alignment to concurrently include touch alongside other modalities.

3 TLV Dataset

The TLV dataset aims to associate tactile and visual perceptions with sentence-level descriptions for multimodal alignment. As shown in Fig. 1, the construction process of TLV consists of three stages: touch and vision collection (Sec. 3.1), touch localization (Sec. 3.2), and tactile labeling (Sec. 3.3).

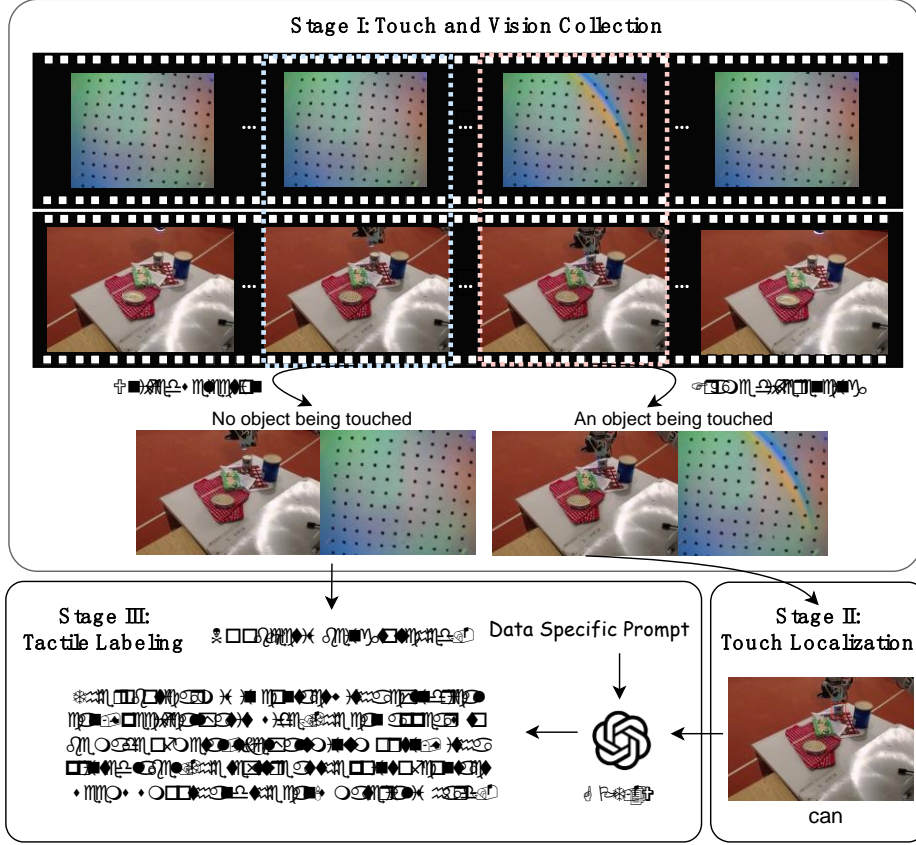


Fig. 1. Construction process of the TLV dataset.

3.1 Stage I: Touch and Vision Collection

We collect paired tactile and visual observations from VisGel, a large vision-touch dataset collected by a video camera, and a tactile sensor called GelSight. VisGel captured synchronized videos of the scenes where the robotic arm touched the objects and recorded timestamps to synchronize visual and tactile images. Among the synchronized videos captured, 10,000 videos were used to construct the training dataset. We utilize the synchronized visual and tactile images from these 10,000 synchronized videos for touch and vision collection.

From the visual videos, we observe that the first frame depicts the starting state of the robotic arm when it is away from objects. As time progresses, the arm gradually approaches an object until it makes contact, remains in contact for a period, and then slowly withdraws. Based on the above observations, for each pair of synchronized videos, we select two sets of synchronized visual and tactile frames: one set depicting an object being touched, and the other set showing no object being touched. To obtain frames where an object is being touched, we use the first frame as the background and

apply frame differencing [39]. The frame with the maximum difference from the background is selected as the frame where the object is being touched. Through observation, we uniformly select the 40th frame as the frame where no object is being touched.

3.2 Stage II: Touch Localization

The two modalities of touch and vision can be regarded as different views containing the same semantics. From this standpoint, we recruit participants to label the object being touched in visual images from Stage I. For visual images where no object is touched, we do not consider them. This serves as preliminary touch localization, preparing for the next step of tactile labeling using GPT-4V [11]. touch localization comprises two parts: highlighting the touched object with a red box in visual images and providing a name for the enclosed object. The labeling of object name is open-ended, and we do not provide a predefined set of candidate object names. In the process of object labeling, we found that due to issues during the collection of the original dataset (i.e., VisGel), certain data could not be annotated. For example, instances where the touched object is occluded or the entire video does not involve interaction with any object. We have filtered out such data.

3.3 Stage III: Tactile Labeling

From the perspective of containing identical semantics in both touch and vision, we utilize GPT-4V for the annotation of texts. For each visual image with the highlighted box from Stage II, we employ a thoughtfully designed, data-specific prompt. This prompt instructs GPT-4V to generate detailed descriptions, taking into account factors such as the name of the touched object, the specific location of the contact, the material composition at the point of contact, and the texture characteristics and softness/hardness of the touched area. For visual images where no object is touched, we refrain from using GPT-4V for annotation and instead provide a uniform description: *No object is being touched.*

3.4 Dataset Statistics

We have annotated text-based descriptions for 20,000 pairs of synchronized tactile and visual observations collected from VisGel, including 10,000 pairs with an object being touched and 10,000 pairs without an object being touched. For the cases where an object is touched, we filtered out data that cannot be annotated as mentioned in Stage II, resulting in the annotation of 9,843 instances. For the cases without an object being touched, we annotated all 10,000 instances. Thus, we ultimately obtained a total of 19,834 annotated data entries. To our knowledge, this is the first touch-language-vision dataset with sentence-level descriptions.

4 Method

We propose STLV-Align (Synergistic Touch-Language-Vision Alignment), an unsupervised and lightweight joint training method designed to leverage the TLV dataset we constructed. The visual observations in TLV can be considered as auxiliary information, assisting in learning the alignment between touch and language and enhancing the zero-shot classification ability of touch. The method primarily consists of three components: multi-modal encoder, LoRA fine-tuning, and joint training, as illustrated in Fig. 2.

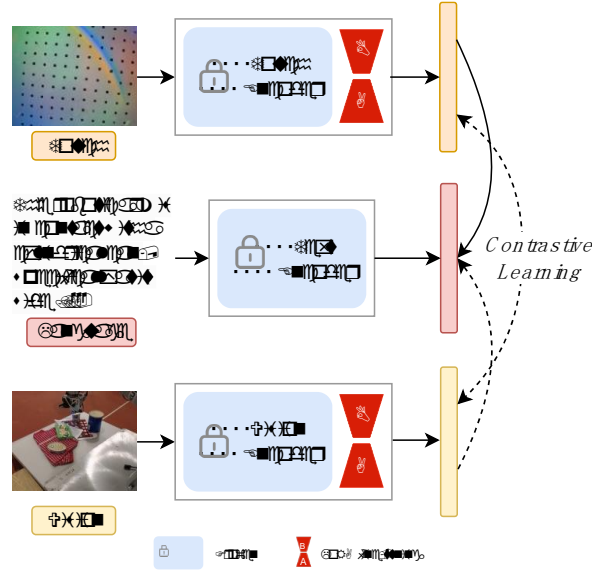


Fig. 2. Overview of our lightweight joint training method.

4.1 Multi-modal Encoders

STLV-Align involves three modalities: touch, language, and vision. We treat the touch modality as RGB images for processing. Therefore, for both touch and vision modalities, we use the Vision Transformer (i.e., ViT) [40] for encoding. The touch and vision encoders are instantiated as OpenCLIP vision encoders. For the text encoder, we instantiate it as an OpenCLIP text encoder.

4.2 LoRA Fine-tuning

Differing from the prior approach [41], we do not utilize large-scale datasets for pre-training. Instead, we employ LoRA for lightweight fine-tuning on the TLV dataset. For a modality-agnostic encoder $f(\cdot)$ with a weight matrix $W_0 \in R^{d \times k}$, we maintain the

weight matrix W_0 frozen while learning a new weight matrix BA . The forward pass can be formalized as follows:

$$f(x) = W_0x + BAx \quad (1)$$

where $B \in R^{d \times r}$, $A \in R^{r \times k}$, with r being the minimum of d and k .

4.3 Joint Training

Joint learning aims to align touch and language better. While learning the alignment between touch and language, we also acquire knowledge about the alignment between vision and language, as well as the alignment between touch and vision. The text encoder from OpenCLIP has demonstrated good generalization in text, so during the joint learning process, we freeze the text encoder and only update the touch encoder and vision encoder. The update to the vision encoder is made to assist the update to the touch encoder. To ensure alignment across different modalities, we perform contrastive learning principles [29] for joint learning.

$$L_{T,L} = -\frac{1}{K} \sum_{i=1}^K \log \frac{\exp(x_i^T y_i / \tau)}{\sum_{j=1}^K \exp(x_i^T y_j / \tau)} \quad (2)$$

$$L_{V,L} = -\frac{1}{K} \sum_{i=1}^K \log \frac{\exp(z_i^T y_i / \tau)}{\sum_{j=1}^K \exp(z_i^T y_j / \tau)} \quad (3)$$

$$L_{T,V} = -\frac{1}{K} \sum_{i=1}^K \log \frac{\exp(x_i^T z_i / \tau)}{\sum_{j=1}^K \exp(x_i^T z_j / \tau)} \quad (4)$$

where x, y, z represent the observations of tactile, language, and visual modalities, respectively, and τ and K are the scalar temperature and batch size. In practice, we use a symmetric joint loss $(L_{T,L} + L_{L,T}) + \alpha(L_{V,L} + L_{L,V}) + \beta(L_{T,V} + L_{V,T})$.

5 Experiments

5.1 Setup

We evaluate our model and dataset on various tactile classification tasks, including material, hard/soft, and rough/smooth classification, using the Touch and Go dataset [8]. This means that zero-shot evaluation is conducted on a cross-domain dataset. STLV-Align is extended based on OpenCLIP-large and fine-tuned on our TLV dataset in an unsupervised and lightweight manner. Because the visual modality is considered as auxiliary information, both α and β are set to 0.1 in the symmetric joint loss. We use accuracy as the metric.

5.2 Results and Analysis

We contrast our model with ViT-LENS-2 [41], a state-of-the-art multi-model that excels in zero-shot performance on tactile tasks. The comparative results of different models can be found in Table 1. While the accuracy of \model may be not optimal, exhibits an 8.3% improvement in material classification compared to our foundation, OpenCLIP. Especially significant is the marked improvement in both hard/soft and rough/smooth classifications, with \model's performance advancing by more than 30%. Nevertheless, VIT-LENS-2 (I) demonstrated an improvement ranging from 7% to 9% compared to their foundation, ImageBind. Despite VIT-LENS-2 (I+T) displaying a notable 41.6% boost in material classification, there was a 6% decline in rough/smooth classification. This reflects the effectiveness of the TLV dataset and the efficiency of STLV-Align in utilizing data. Certainly, we recognize certain performance limitations of STLV-Align, thus prompting us to analyze distinctions in training paradigm, #training data, parameter tuning ratio, and cross-domain evaluation between STLV-Align and ViT-LENS-2, as illustrated in Table 2. It can be observed that \model is characterized by independence from labeled data, the utilization of a smaller dataset, a light-weight training approach, and evaluation across various data domains. This could make it more attractive for specific application scenarios.

Table 1. Accuracy of different models on various tactile classifications. Those with performance improvement compared with their respective foundation within 30% are marked as green and above 30% are marked as red. V-L-2: VIT-LENS-2; (I): Anchored by images; (I+T): Anchored by images and texts.

Model	Size	Touch and Go		
		Material	Hard/Soft	Rough/Smooth
ImageBind	Base	24.2	65.7	69.8
V-L-2 (I)	Base	29.9 (+5.7%)	72.4 (+6.7%)	77.9 (+8.1%)
V-L-2 (I)	Large	31.2 (+7.0%)	74.3 (+8.6%)	78.2 (+8.4%)
V-L-2 (I+T)	Large	65.8 (+41.6%)	74.7 (+9.0%)	63.8 (-6.0%)
OpenCLIP	Large	17.7	32.2	42.7
STLV-Align	Large	26.0 (+8.3%)	65.1 (+32.9%)	75.6 (+31.9%)

Table 2. Comparison of Ours and VIT-LENS-2 in training paradigm (TP), #training data (#TD), parameter tuning ratio (PTR), cross-domain evaluation (CDE).

Model	TP	#TD	PTR	CDE
VIT-LENS-2	Supervised	91,982	100%	✗
STLV-Align	Unsupervised	19,843	1%	✓

5.3 Ablation Study

We conduct the ablation study in Table 3 to illustrate the impact of vision information. Simultaneously aligning both touch and text with visual information enhances tactile classification, yielding a positive overall effect. Conversely, aligning either touch or text with visual information has a detrimental effect.

Table 3. Impact of vision information at different levels. -TV: Do not align touch with vision; -VL: Do not align language with vision; -(TV&VL): Do not involve visual information.

Model	Touch and Go		
	Material	Hard/Soft	Rough/Smooth
STLV-Align	26.0	65.1	74.6
-TV	27.8	52.8	52.7
-VL	26.5	55.3	49.1
-(TV&VL)	32.5	56.5	56.6

6 Conclusion

In this work, we construct the first touch-language-vision dataset, TLV, featuring sentence-level descriptions for multimodal alignment. To demonstrate the effectiveness of the TLV dataset, we extended OpenCLIP and proposed STLV-Align, an unsupervised lightweight training approach. Preliminary experiments validate that the TLV dataset facilitates better alignment between touch and language. The proposed method may apply to specific scenarios, but there is room for improvement in terms of performance, and further enhancements are needed. Additionally, we intend to extend the application of TLV to more tasks to fully exploit its potential.

Acknowledgments. This work is supported by the Talent Fund of Beijing Jiaotong University (2023XKRC006) and the Pattern Recognition Center, WeChat AI, Tencent Inc.

References

1. Abend, O., Reichart, R., Rappoport, A.: Unsupervised argument identification for semantic role labeling. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 28-36. (2009)
2. Johansson, R.S., Flanagan, J.R.: Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience* 10, 345-359 (2009)
3. Qi, H., Yi, B., Suresh, S., Lambeta, M., Ma, Y., Calandra, R., Malik, J.: General in-hand object rotation with vision and touch. In: Conference on Robot Learning, pp. 2549-2564. PMLR, (2023)
4. Hansen, J., Hogan, F., Rivkin, D., Meger, D., Jenkin, M., Dudek, G.: Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 8298-8304. IEEE, (2022)
5. Dave, V., Lygerakis, F., Rückert, E.: Multimodal Visual-Tactile Representation Learning through Self-Supervised Contrastive Pre-Training. In: Proceedings/IEEE International Conference on Robotics and Automation. Institute of Electrical and Electronics Engineers, (2024)
6. Yang, F., Zhang, J., Owens, A.: Generating visual scenes from touch. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22070-22080. (2023)

7. Kerr, J., Huang, H., Wilcox, A., Hoque, R., Ichnowski, J., Calandra, R., Goldberg, K.: Self-supervised visuo-tactile pretraining to locate and follow garment features. arXiv preprint arXiv:2209.13042 (2022)
8. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and Go: Learning from Human-Collected Vision and Touch. *Advances in Neural Information Processing Systems* 35, 8081-8103 (2022)
9. Gao, R., Taunyazov, T., Lin, Z., Wu, Y.: Supervised autoencoder joint learning on heterogeneous tactile sensory data: Improving material classification performance. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10907-10913. IEEE, (2020)
10. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5580-5588. (2017)
11. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
12. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
13. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: *International Conference on Learning Representations*. (2021)
14. Cui, S., Wang, R., Wei, J., Hu, J., Wang, S.: Self-attention based visual-tactile fusion learning for predicting grasp outcomes. *IEEE Robotics and Automation Letters* 5, 5827-5834 (2020)
15. Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J.B., Rodriguez, A.: See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics* 4, eaav3123 (2019)
16. Lin, J., Calandra, R., Levine, S.: Learning to identify object instances by touch: Tactile recognition via multimodal matching. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 3644-3650. IEEE, (2019)
17. Li, Y., Zhu, J.-Y., Tedrake, R., Torralba, A.: Connecting touch and vision via cross-modal prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10609-10618. (2019)
18. Johnson, M.K., Cole, F., Raj, A., Adelson, E.H.: Microgeometry capture using an elastomeric sensor. *ACM Transactions on Graphics (TOG)* 30, 1-8 (2011)
19. Gomes, D.F., Paoletti, P., Luo, S.: Generation of gelsight tactile images for sim2real learning. *IEEE Robotics and Automation Letters* 6, 4177-4184 (2021)
20. Calandra, R., Owens, A., Upadhyaya, M., Yuan, W., Lin, J., Adelson, E.H., Levine, S.: The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes? In: *Conference on Robot Learning*, pp. 314-323. PMLR, (2017)
21. Yuan, W., Zhu, C., Owens, A., Srinivasan, M.A., Adelson, E.H.: Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 951-958. IEEE, (2017)
22. Si, Z., Yuan, W.: Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters* 7, 2361-2368 (2022)
23. Dong, S., Yuan, W., Adelson, E.H.: Improved gelsight tactile sensor for measuring geometry and slip. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 137-144. IEEE, (2017)

24. Suresh, S., Si, Z., Anderson, S., Kaess, M., Mukadam, M.: Midastouch: Monte-carlo inference over distributions across sliding touch. In: Conference on Robot Learning, pp. 319-331. PMLR, (2023)
25. Lambeta, M., Chou, P.-W., Tian, S., Yang, B., Maloon, B., Most, V.R., Stroud, D., Santos, R., Byagowi, A., Kammerer, G.: Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters* 5, 3838-3845 (2020)
26. Donlon, E., Dong, S., Liu, M., Li, J., Adelson, E., Rodriguez, A.: Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1927-1934. IEEE, (2018)
27. Taylor, I.H., Dong, S., Rodriguez, A.: Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 10781-10787. IEEE, (2022)
28. Gao, R., Si, Z., Chang, Y.-Y., Clarke, S., Bohg, J., Fei-Fei, L., Yuan, W., Wu, J.: Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10598-10608. (2022)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763. PMLR, (2021)
30. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, pp. 4904-4916. PMLR, (2021)
31. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35, 23716-23736 (2022)
32. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2818-2829. (2023)
33. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems* 36, (2024)
34. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1179-1189. (2023)
35. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 976-980. IEEE, (2022)
36. Chen, J., Zhang, R., Lian, D., Yang, J., Zeng, Z., Shi, J.: iquery: Instruments as queries for audio-visual sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14675-14686. (2023)
37. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180-15190. (2023)
38. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z.: LanguageBind: Extending Video-Language Pretraining to N-modality by Language-

- based Semantic Alignment. In: The Twelfth International Conference on Learning Representations. (2023)
39. Zaki, W.M.D.W., Hussain, A., Hedayati, M.: Moving object detection using keypoints reference model. *EURASIP J. Image Video Process.* 2011, 13 (2011)
 40. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations. (2020)
 41. Lei, W., Ge, Y., Yi, K., Zhang, J., Gao, D., Sun, D., Ge, Y., Shan, Y., Shou, M.Z.: ViT-Lens-2: Gateway to Omni-modal Intelligence. *arXiv preprint arXiv:2311.16081* (2023)