URBAN ROBOTICS LAB

# ICCV 2019

## ( International Conference on Computer Vision )

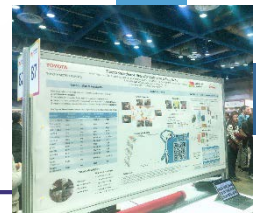
Lab Seminar (2019.11.12)

발표자 : 이승희

KAIST

URBAN ROBOTICS LAB

# Toyota Smarthome: Real-World Activities of Daily Living



TOYOTA

TOYOTA MOTOR EUROPE

# TOYOTA SMARTHOME: REAL WORLD ACTIVITIES OF DAILY LIVING

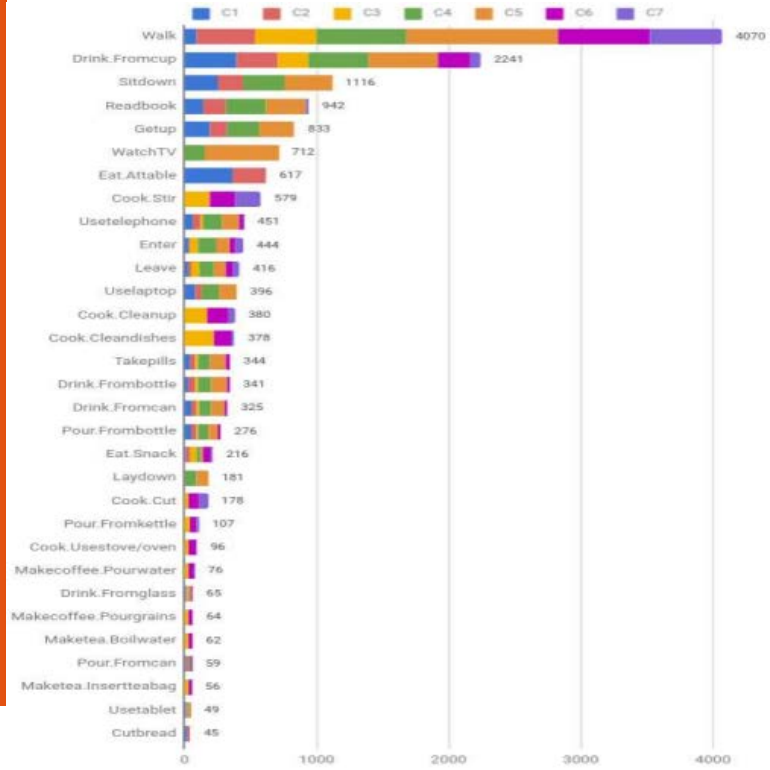



Figure 2. Number of video clips per activity in Smarthome and the relative distribution across the different camera views. C1 to C7 represent 7 camera views. All the activity classes have multiple camera views, ranging from 2 to 7.

Table 1. Comparative study highlighting the challenges in real-world setting datasets

| Dataset | Context | Duration variation | Cross-view challenge | Composite activities | View Type | Spontaneous acting | Camera framing | Fine-grained activities | Type |
|---|---|---|---|---|---|---|---|---|---|
| ACTEV/VIRAT [7] | free | Medium | Yes | No | Monitoring | Medium | Low | No | Surveillance |
| SVW [32] | biased | Low | No | No | Shooting | High | High | No | Sport |
| HMDB [21] | biased | Low | No | No | Shooting | Medium | High | No | Youtube |
| Kinetics [4] | biased | Low | No | No | Shooting | Medium | High | No | Youtube |
| AVA [15] | biased | Low | No | No | Shooting | Medium | High | No | Movies |
| EPIC-KITCHENS [6] | free | High | No | Yes | Egocentric | Medium | High | Yes | Kitchen |
| Something-Something [14] | free | Low | No | No | Shooting | Low | High | Yes | Object interaction |
| MPII Cooking2 [31] | free | High | Yes | Yes | Monitoring | Medium | Medium | Yes | Cooking |
| DAHLIA [42] | free | High | Yes | No | Monitoring | Medium | Medium | No | Kitchen |
| NUCLA [46] | free | Low | Yes | No | Shooting | Low | High | No | Object interaction |
| NTU RGB+D [33] | free | Low | Yes | No | Monitoring | Low | High | No | ADL |
| Charades [35] | free | Low | Yes | Yes | Shooting | Low | High | Yes | ADL |
| **Smarthome** | free | High | Yes | Yes | Monitoring | High | Low | Yes | ADL |

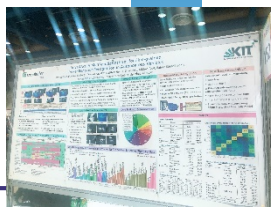
https://project.inria.fr/toyotasmarthome/


Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., & Francesca, G. Toyota Smarthome: Real-World Activities of Daily Living.


KAIST

URL URBAN ROBOTICS LAB

# Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles



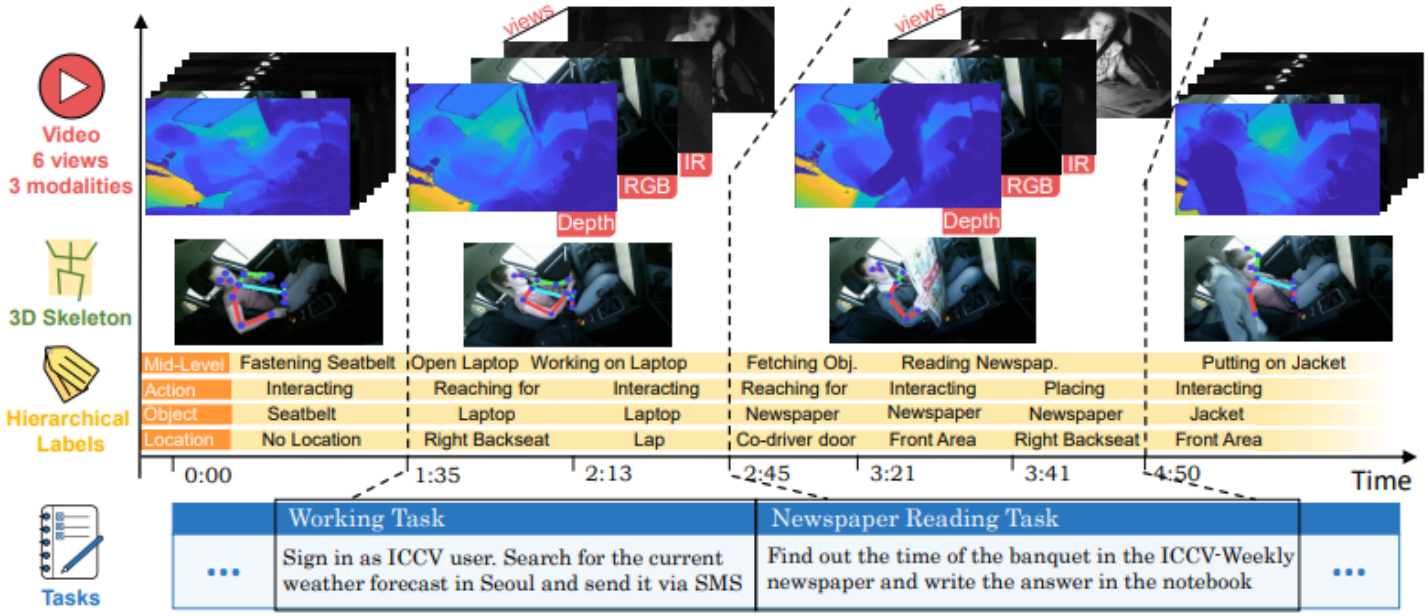



Figure 1: Overview of the Drive&Act dataset for driver behavior recognition. The dataset includes 3D skeletons in addition to frame-wise hierarchical labels of 9.6 Million frames captured by 6 different views and 3 modalities (RGB, IR and depth).

- www.driveandact.com
- 83 manually annotated hierarchical activity labels:
  - Level 1: Long running tasks (12)
  - Level 2: Semantic actions (34)
  - Level 3: Object Interaction tripplets [action|object|location] (6|17|14)

- 12h video data in 29 long sequences (9.6 M frames)
- Calibrated multi view camera system with 5 views
- Multi modal videos: NIR, Depth and Color data
- Markerless motion capture: 3D Body Pose and Head Pose
- Model of the static interior of the car

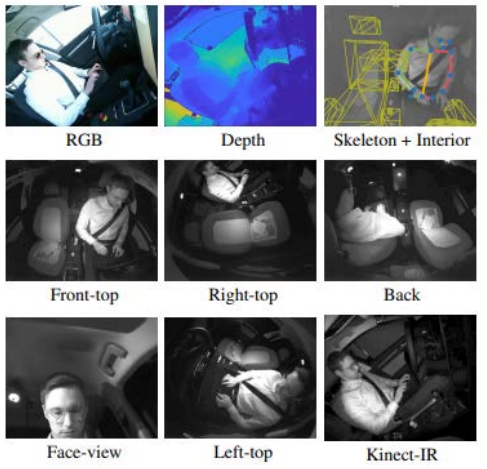



Figure 2: Example images of the *working on laptop* activity

Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., & Stiefelhagen, R. Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles.

KAIST

# Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles

| | SoA conven. AR | Multi-mod. AR | Driver Activity Recognition Datsets | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Kinetics [7] | NTU [43] | HEH [36] | Ohn et al. [35] | Brain4Cars [19] | D.P.-Night [50] | D.P.-Real [50] | AUC-D.D. [2] | **Drive&Act** |
| Year | 2017 | 2016 | 2014 | 2014 | 2015 | 2016 | 2016 | 2017/18 | 2019 |
| Publicly available | ✓ | ✓ | ✓ | – | ✓ | – | – | ✓ | ✓ |
| Manual driving | – | – | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Autonomous driving | – | – | – | – | – | – | – | – | ✓ |
| RGB/Grayscale | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Depth | – | ✓ | ✓ | N/A[b] | – | – | – | – | ✓ |
| NIR | – | ✓ | – | – | – | ✓ | – | – | ✓ |
| Skeleton | – | ✓ | – | – | – | – | – | – | ✓ |
| Video | ✓ | ✓ | ✓ | N/A[b] | ✓ | ✓ | ✓ | N/A[b] | ✓ |
| N° images | >76M | 4M | N/A[b] | 11K | 2M | 29K | 18K | 17K | > 9.6M |
| N° synch. views | 1 | 3 | 1 | 2 | 2 | 1 | 1 | 1 | 6 |
| Resolution | N/A[c] | 1920×1080[a] | 680×480 | N/A[b] | 1920×1088 | 640×480 | 640×480 | 1920×1080 | 1280×1024[d] |
| N° subjects | N/A[b] | 40 | 8 | 4 | 10 | 20 | 5 | 31 | 15 |
| Female / male | N/A[b] | N/A[b] | 1 / 7 | 1 / 3 | N/A[b] | 10 / 10 | N/A[b] | 9 / 22 | 4 / 11 |
| N° Classes | 400 | 60 | 19 | 3 | 5 | 4 | 4 | 10 | 83 |
| Multi-level annot. | – | – | – | – | – | – | – | – | ✓ |
| N° Levels | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| Continuous labels | – | – | – | N/A[b] | – | ✓ | ✓ | N/A[b] | ✓ |
| Object annot. | ✓ | – | – | – | – | – | – | – | ✓ |

[a] RGB resolution, IR/Depth resolution is 512×424
[b] information not provided by the authors
[c] variable resolution
[d] NIR-camera resolution

Table 1: Comparison of driving and non-driving related datasets for action recognition. In this table, we depict the characteristics of the recording modalities, the content of the dataset and the properties of the provided reference labels.

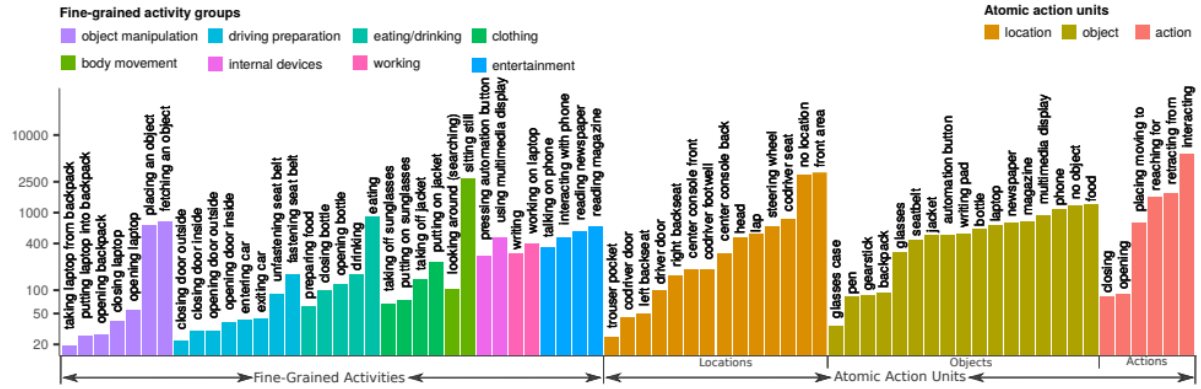



Figure 3: Sample frequency of fine-grained activities (left) and atomic actions (right) by class (logarithmic scale). A sample corresponds to a 3s snippet with the assigned label. Colors denote the activity group (e.g. food-related activities).

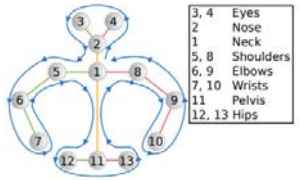



Figure 4: Annotated body joints and joint sequence used for the spatial stream


Figure 2: Interior of the simulator depicting the modified dashboard and camera positions.

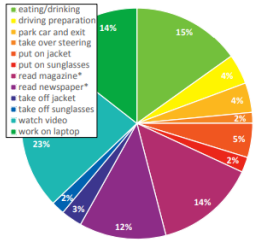



Figure 4: Distribution of the scenarios/tasks in our dataset. *these tasks consist of both finding information about a previously asked question by reading a newspaper/magazine and of writing the answer into a notebook.

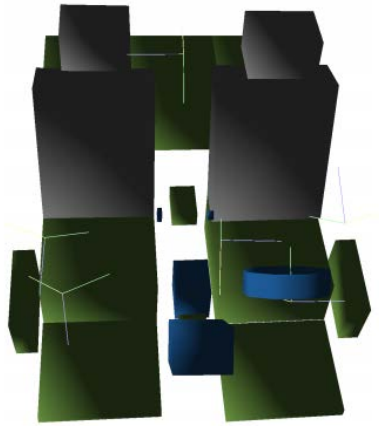
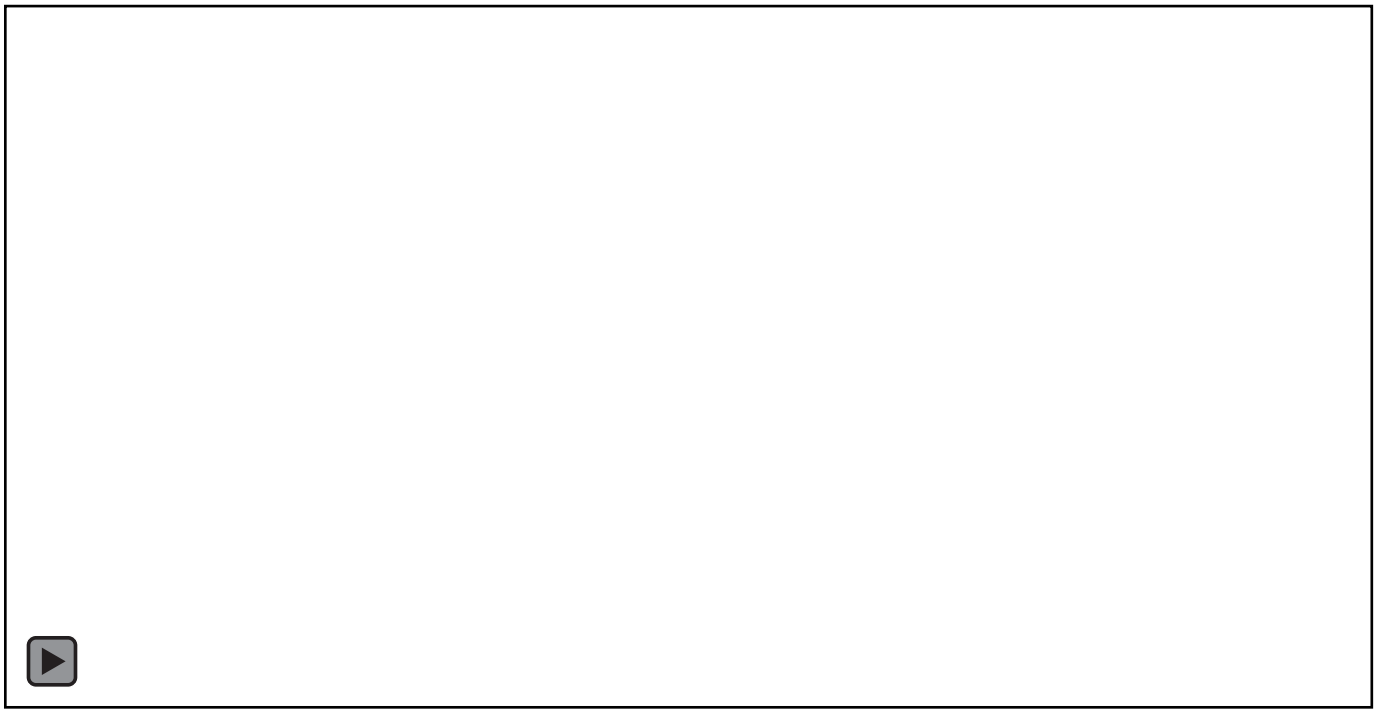
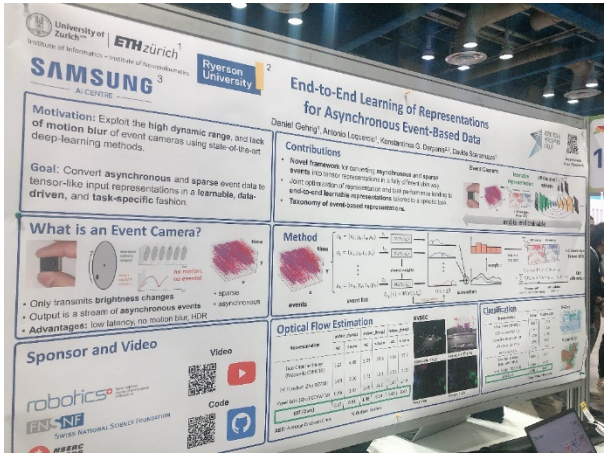


Figure 3: The interior model of the car. Green denotes storage areas, blue denotes car controls and gray remaining regions. Camera positions are depicted as coordinate systems.

Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., & Stiefelhagen, R. Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles.

# End-to-End Learning of Representations for Asynchronous Event-Based Data





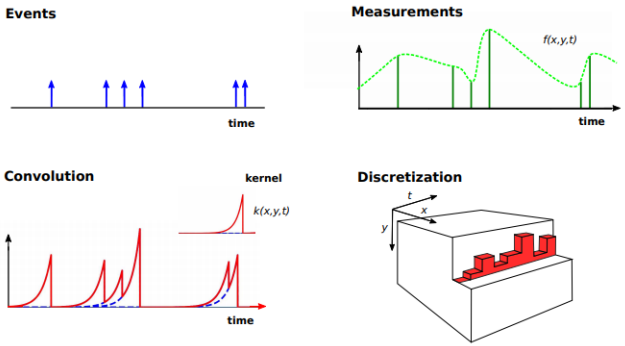




Figure 2. An overview of our proposed framework. Each event is associated with a measurement (green) which is convolved with a (possibly learnt) kernel. This convolved signal is then sampled on a regular grid. Finally, various representations can be instantiated by performing projections over the temporal axis or over polarities.

| Representation | Dimensions | Description | Characteristics |
|---|---|---|---|
| Event frame [53] | $H \times W$ | Image of event polarities | Discards temporal and polarity |
| Event count image [36, 69] | $2 \times H \times W$ | Image of event counts | Discards time stamps |
| Surface of Active Events (SAE) [7, 69] | $2 \times H \times W$ | Image of most recent time stamp | Discards earlier time stamps |
| Voxel grid [70] | $B \times H \times W$ | Voxel grid summing event polarities | Discards event polarity |
| Histogram of Time Surfaces (HATS) [59] | $2 \times H \times W$ | Histogram of average time surfaces | Discards temporal information |
| Event Spike Tensor (EST, our work) | $2 \times B \times H \times W$ | Sample event point-set into a grid | Discards the least amount of in |

Table 1. Comparison of grid-based event representations used in prior work on event-based deep learning. $H$ and $W$ denc height and width dimensions, respectively, and $B$ the number of temporal bins.

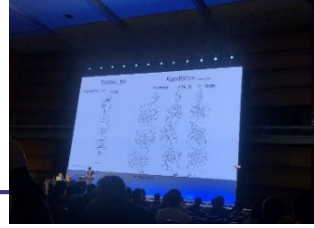
https://github.com/uzh-rpg/rpg_event_representation_learning


Gehrig, D., Loquercio, A., Derpanis, K. G., & Scaramuzza, D. (2019). End-to-End Learning of Representations for Asynchronous Event-Based Data. *arXiv preprint arXiv:1904.08245*.


KAIST

URL URBAN ROBOTICS LAB

# Exploring Randomly Wired Neural Networks for Image Recognition



- What we call deep learning today descends from the connectionist approach to cognitive science.
- However, like the wiring patterns in ResNet, DenseNet, the NAS network generator is hand designed and the space of allowed wiring patterns is constrained in a small subset of all possible graphs.
- **<u>What happens if we loosen this constraint and design novel network generators?</u>**
  - Design a Network Generator not an Individual Network! (main topic)
  - Relation to Neuroscience
    - Turing analogized the unorganized machines to an infant human's cortex.
    - "At birth, the construction of the most important networks is largely random."

Xie, S., Kirillov, A., Girshick, R., and He, K. (2019). Exploring randomly wired neural networks for image recognition. *arXiv preprint arXiv:1904.01569*.

# Exploring Randomly Wired Neural Networks for Image Recognition

- By using three Random graph model (ER, BA, WS), randomly wired neural networks are generated.

| network | top-1 acc. | top-5 acc. | FLOPs (M) | params (M) |
|---|---|---|---|---|
| MobileNet [15] | 70.6 | 89.5 | 569 | 4.2 |
| MobileNet v2 [40] | 74.7 | - | 585 | 6.9 |
| ShuffleNet [54] | 73.7 | 91.5 | 524 | 5.4 |
| ShuffleNet v2 [30] | **74.9** | 92.2 | 591 | 7.4 |
| NASNet-A [56] | 74.0 | 91.6 | 564 | 5.3 |
| NASNet-B [56] | 72.8 | 91.3 | 488 | 5.3 |
| NASNet-C [56] | 72.5 | 91.0 | 558 | 4.9 |
| Amoeba-A [34] | 74.5 | 92.0 | 555 | 5.1 |
| Amoeba-B [34] | 74.0 | 91.5 | 555 | 5.3 |
| Amoeba-C [34] | **75.7** | **92.4** | 570 | 6.4 |
| PNAS [26] | 74.2 | 91.9 | 588 | 5.1 |
| DARTS [27] | 73.1 | 91.0 | 595 | 4.9 |
| **RandWire-WS** | **74.7**$_{\pm 0.25}$ | **92.2**$_{\pm 0.15}$ | 583$_{\pm 6.2}$ | 5.6$_{\pm 0.1}$ |

Table 2. **ImageNet: small computation regime** (*i.e.*, <600M FLOPs). RandWire results are the mean accuracy (±std) of 5 random network instances, with WS(4, 0.75). Here we train for 250 epochs similar to [56, 34, 26, 27], for fair comparisons.

| network | top-1 acc. | top-5 acc. | FLOPs (B) | params (M) |
|---|---|---|---|---|
| ResNet-50 [11] | 77.1 | 93.5 | 4.1 | 25.6 |
| ResNeXt-50 [52] | 78.4 | 94.0 | 4.2 | 25.0 |
| **RandWire-WS**, $C$=109 | **79.0**$_{\pm 0.17}$ | **94.4**$_{\pm 0.11}$ | 4.0$_{\pm 0.09}$ | 31.9$_{\pm 0.66}$ |
| ResNet-101 [11] | 78.8 | 94.4 | 7.8 | 44.6 |
| ResNeXt-101 [52] | 79.5 | 94.6 | 8.0 | 44.2 |
| **RandWire-WS**, $C$=154 | **80.1**$_{\pm 0.19}$ | **94.8**$_{\pm 0.18}$ | 7.9$_{\pm 0.18}$ | 61.5$_{\pm 1.32}$ |

Table 3. **ImageNet: regular computation regime** with FLOPs comparable to ResNet-50 (top) and to ResNet-101 (bottom). ResNeXt is the 32×4 version [52]. RandWire is WS(4, 0.75).

| network | test size | epochs | top-1 acc. | top-5 acc. | FLOPs (B) | params (M) |
|---|---|---|---|---|---|---|
| NASNet-A [56] | $331^2$ | >250 | 82.7 | 96.2 | 23.8 | 88.9 |
| Amoeba-B [34] | $331^2$ | >250 | 82.3 | 96.1 | 22.3 | 84.0 |
| Amoeba-A [34] | $331^2$ | >250 | 82.8 | 96.1 | 23.1 | 86.7 |
| PNASNet-5 [26] | $331^2$ | >250 | 82.9 | 96.2 | 25.0 | 86.1 |
| **RandWire-WS** | $320^2$ | 100 | 81.6$_{\pm 0.13}$ | 95.6$_{\pm 0.07}$ | 16.0$_{\pm 0.36}$ | 61.5$_{\pm 1.32}$ |

Table 4. **ImageNet: large computation regime**. Our networks are the same as in Table 3 ($C$=154), but we evaluate on 320×320 images instead of 224×224. Ours are only trained for 100 epochs.

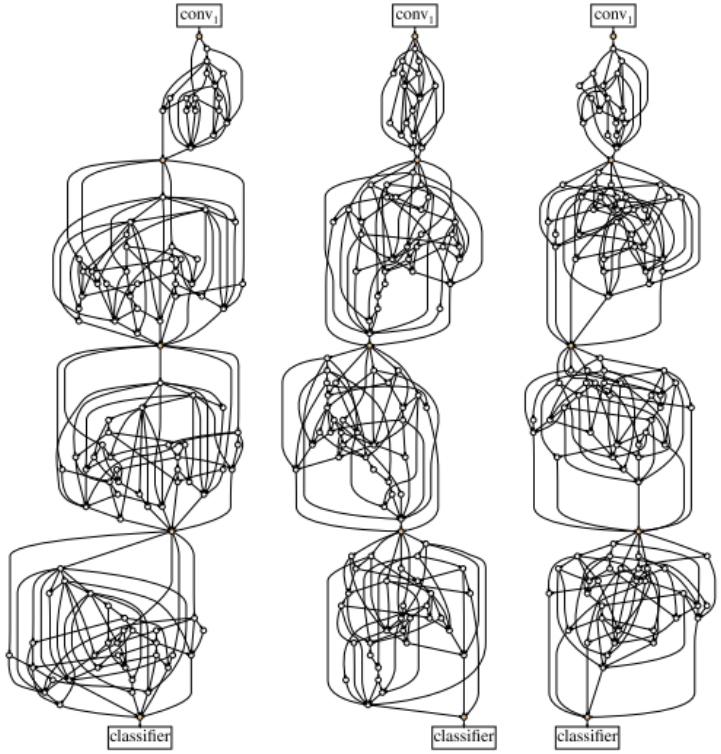



Figure 1. **Randomly wired neural networks** generated by the classical Watts-Strogatz (WS) [50] model: these three instances of random networks achieve (left-to-right) 79.1%, 79.1%, 79.0% classification accuracy on ImageNet under a similar computational budget to ResNet-50, which has 77.1% accuracy.

- Exploring new generator designs may yield new, powerful networks designs.

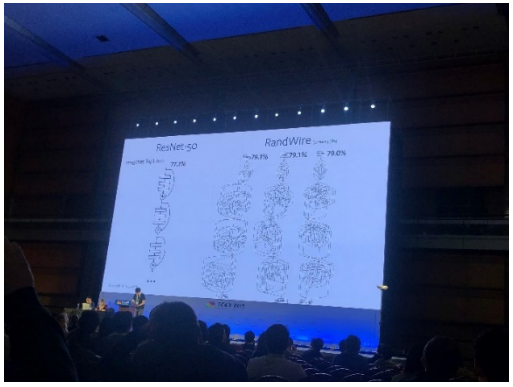

Xie, S., Kirillov, A., Girshick, R., and He, K. (2019). Exploring randomly wired neural networks for image recognition. *arXiv preprint arXiv:1904.01569*.

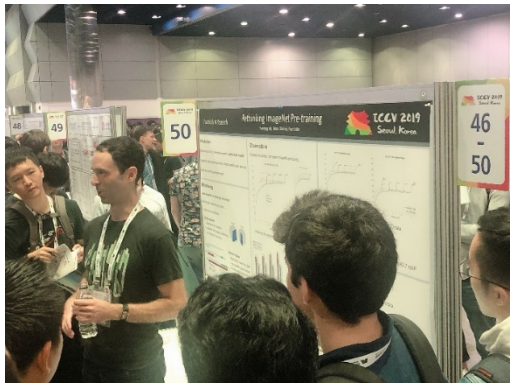
# Appendix

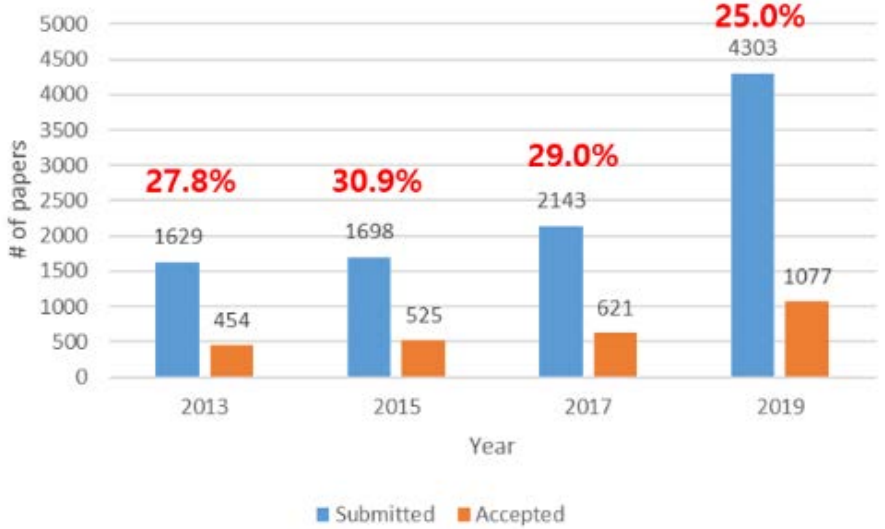
- 7,500 attendees (more than 3,000 attendees from Korea)

https://github.com/hoya012/ICCV-2019-Paper-Statistics






Thank you ☺