TBDLGR: Transformer-Based Dactile Language Gesture Recognition

Станислава Иваненко
Роман Горбунов
Максим Шугаев
Анжелина Абдулаева
Кирилл Зайцев

23 декабря 2025 г.

[GitHub]

Дактиль (33+1 класса)

Bukva: Russian Sign Language Alphabet, 2024
[Paper]

Архитектура

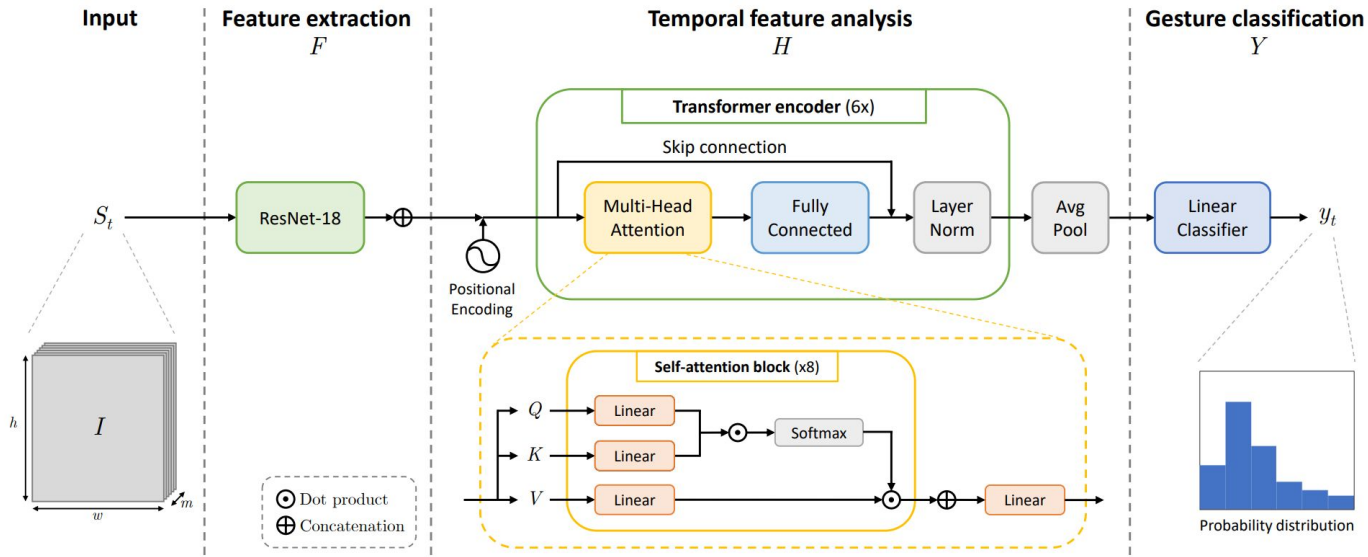
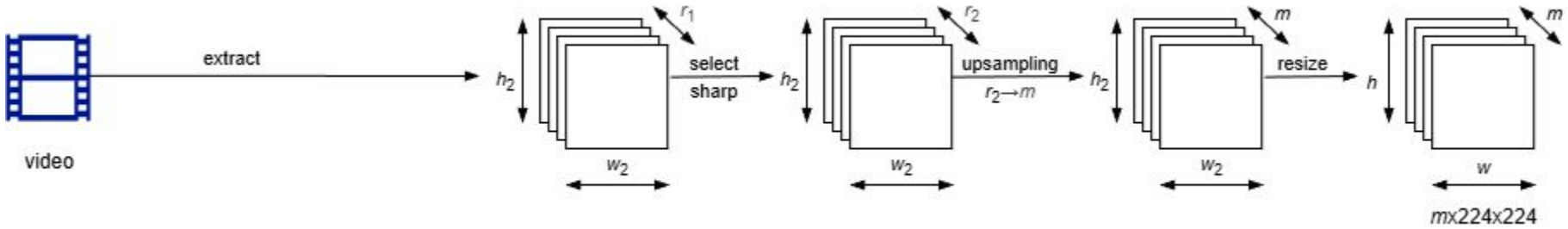


Figure 1. Overview of the proposed method. The temporal feature analysis, computed after the feature extraction performed by the ResNet-18 model, is highlighted showing the architecture of the transformer encoder and the self-attention block.

A Transformer-Based Network for Dynamic Hand Gesture Recognition, 2020
[Paper] [Project Page] [GitHub]

Framer


extract
video
h_2
w_2
r_1
select
sharp
r_2
upsampling
$r_2 \rightarrow m$
m
resize
h
w
mx224x224

Датасет Bukva

~4000

HD видеозаписей демонстрации жестов разными людьми

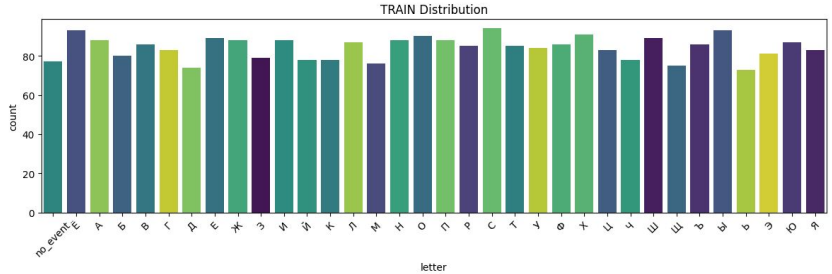
>100

видеозаписей на каждый жест

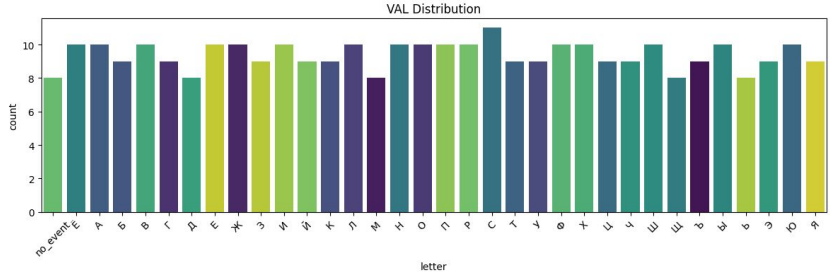
Bukva: Russian Sign Language Alphabet, 2024
[Paper]

Датасет Bukva

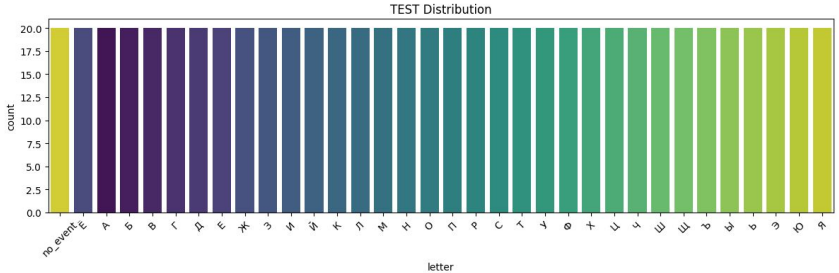
Train: 2863



Val: 319



Test: 680

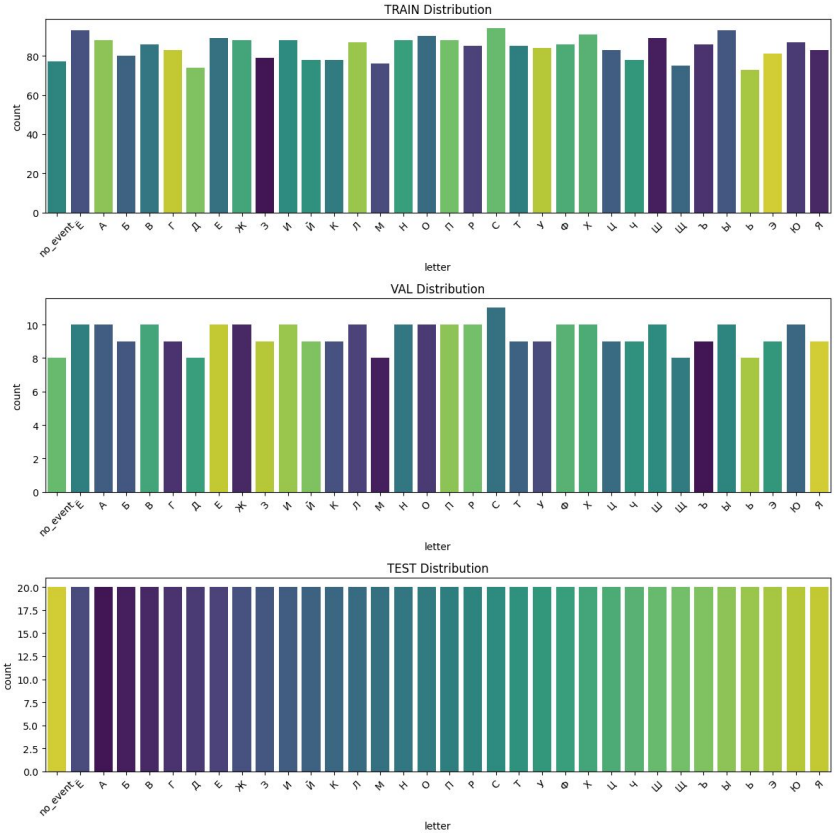


Датасет Bukva

Train: 2863

Val: 319

Test: 680



- NVIDIA GeForce RTX 3060 (12 GB)
- AdamW:
  - weight_decay=1e-4
  - base_lr=1e-4
- 100 эпох
- политика “best” (84-я эпоха)
- batch_size=8

75.29%

mean accuracy

Accuracy на похожих датасетах

Method	Modality	Accuracy
color	Spat. st. CNN [40]	54.6%
	iDT-HOG [45]	59.1%
	Res3ATN [12]	62.7%
	C3D [42]	69.3%
	R3D-CNN [33]	74.1%
	Ours	**76.5%**
	I3D [8]†	78.4%
depth	SNV [47]	70.7%
	C3D [42]	78.8%
	R3D-CNN [33]	80.3%
	I3D [8]†	82.3%
	Ours	**83.0%**
infrared	R3D-CNN [33]	63.5%
	Ours	**64.7%**
flow	iDT-HOF [45]	61.8%
	Temp. st. CNN [40]	68.0%
	Ours	72.0%
	iDT-MBH [45]	76.8%
	R3D-CNN [33]	**77.8%**
	I3D [8]†	83.4%
normals	**Ours**	**82.4%**
color	*Human [33]*	*88.4%*

Table 1. Unimodal results on NVGestures [33]. Previous results are taken from the respective papers and from [33, 1]. † indicates models pre-trained on Kinetics [23], in addition to ImageNet [11].

#	Modality	Accuracy
1	infrared (ir)	64.7%
	color	76.5%
	normals	82.4%
	depth	83.0%
2	color + ir	79.0%
	depth + ir	81.7%
	normals + ir	82.8%
	color + depth	84.6%
	color + normals	84.6%
	depth + normals	**87.3%**
3	color + ir + depth	85.3%
	color + ir + normals	85.3%
	color + depth + normals	86.1%
	depth + normals + ir	**87.1%**
4	**color + depth + normals + ir**	**87.6%**

Table 2. Multimodal results on NVGestures [33] using several combinations of modalities. # refers to the number of used modalities.

NVGestures - 25 классов

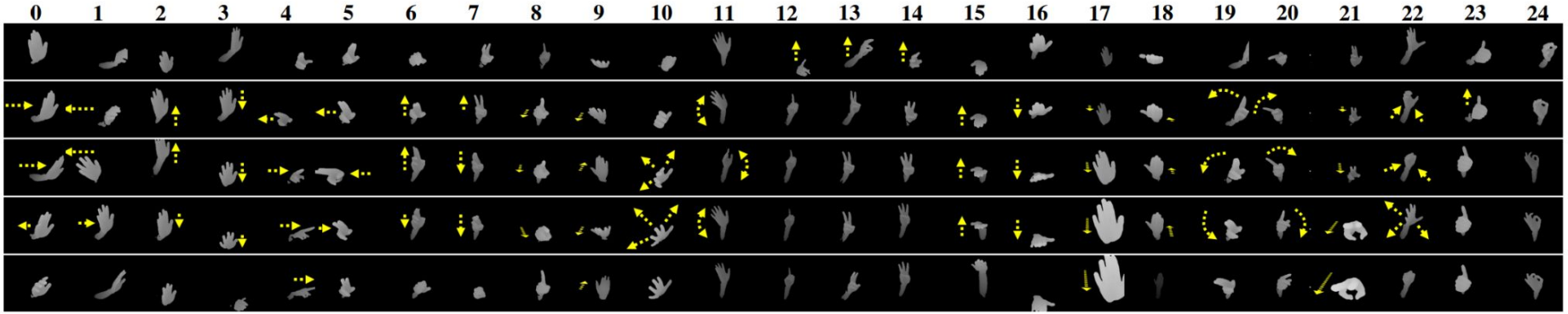


Figure 3: Twenty-five dynamic hand gesture classes. Some gestures were adopted from existing commercial systems [1] or popular datasets [23, 27]. Each column shows a different gesture class (0 – 24). The top and bottom rows show the starting and ending depth frames, respectively, of the nucleus phase for each class. (Note that we did not crop the start and end frames in the actual training and evaluation data.) Yellow arrows indicate the motion of each hand gesture. (A more detailed description of each gesture is available in the supplementary video.)

Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks, 2016
[Paper] [Project Page]

Accuracy на похожих датасетах

Method	Modality	Accuracy
color	Spat. st. CNN [40]	54.6%
	iDT-HOG [45]	59.1%
	Res3ATN [12]	62.7%
	C3D [42]	69.3%
	R3D-CNN [33]	74.1%
	Ours	**76.5%**
	I3D [8]†	78.4%
depth	SNV [47]	70.7%
	C3D [42]	78.8%
	R3D-CNN [33]	80.3%
	I3D [8]†	82.3%
	Ours	**83.0%**
infrared	R3D-CNN [33]	63.5%
	Ours	**64.7%**
flow	iDT-HOF [45]	61.8%
	Temp. st. CNN [40]	68.0%
	Ours	72.0%
	iDT-MBH [45]	76.8%
	R3D-CNN [33]	**77.8%**
	I3D [8]†	83.4%
normals	**Ours**	**82.4%**
color	*Human [33]*	*88.4%*

Table 1. Unimodal results on NVGestures [33]. Previous results are taken from the respective papers and from [33, 1]. † indicates models pre-trained on Kinetics [23], in addition to ImageNet [11].

#	Modality	Accuracy
1	infrared (ir)	64.7%
	color	76.5%
	normals	82.4%
	depth	83.0%
2	color + ir	79.0%
	depth + ir	81.7%
	normals + ir	82.8%
	color + depth	84.6%
	color + normals	84.6%
	depth + normals	**87.3%**
3	color + ir + depth	85.3%
	color + ir + normals	85.3%
	color + depth + normals	86.1%
	depth + normals + ir	**87.1%**
4	**color + depth + normals + ir**	**87.6%**

Table 2. Multimodal results on NVGestures [33] using several combinations of modalities. # refers to the number of used modalities.

#	Modality	Accuracy
1	color	90.6%
	depth	92.4%
	ir	95.1%
	normals	**95.8%**
2	color + depth	94.1%
	depth + ir	95.1%
	color + ir	95.5%
	depth + normals	96.2%
	color + normals	96.5%
	ir + normals	**97.2%**
3	color + depth + ir	95.1%
	color + depth + normals	95.8%
	color + ir + normals	96.9%
	depth + ir + normals	**97.2%**
4	**color + depth + ir + normals**	**96.2%**

Table 4. Unimodal and multimodal results obtained on Briareo. # refers to the number of used modalities.

Briareo - 12 классов

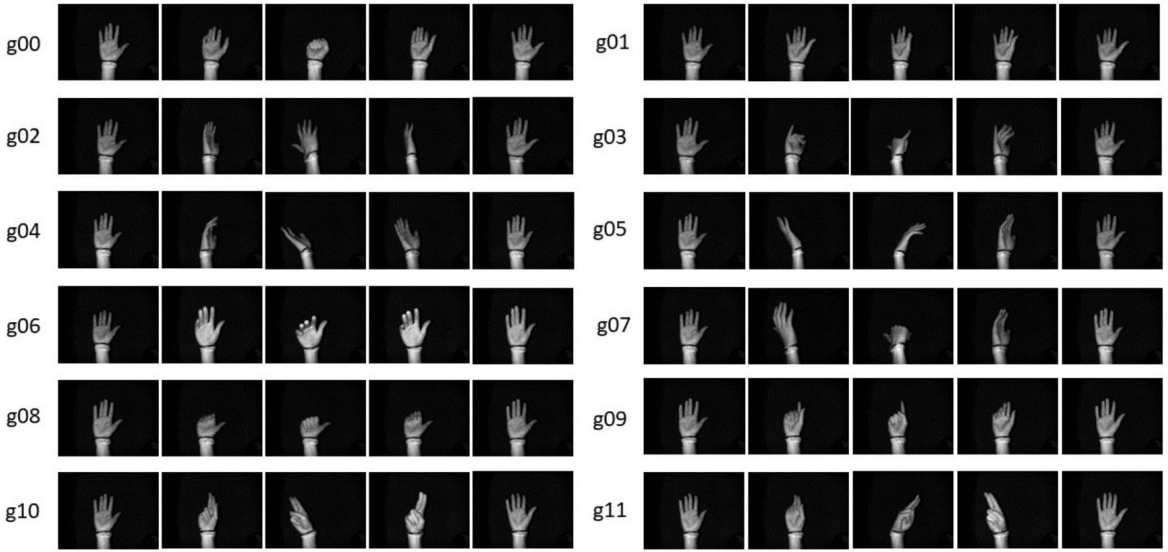
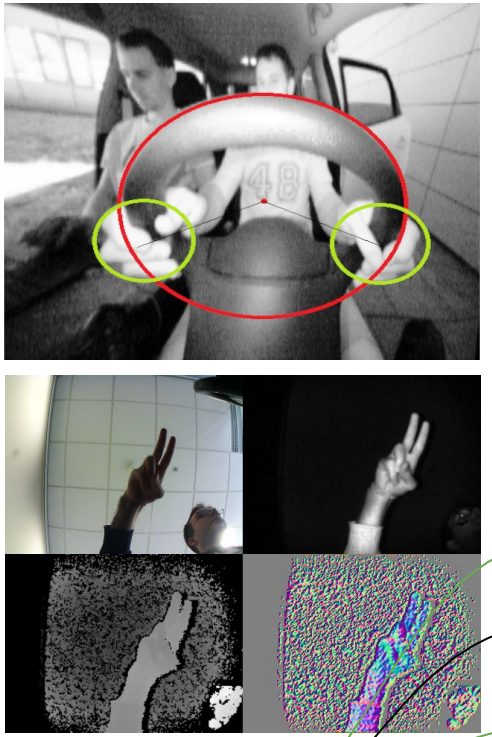
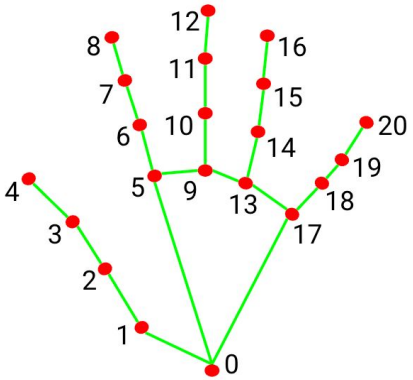


Fig. 1. Gesture classes included in the *Briareo* dataset. As shown, only *dynamic* gestures are present in the dataset. For further details, see Section 3.2.

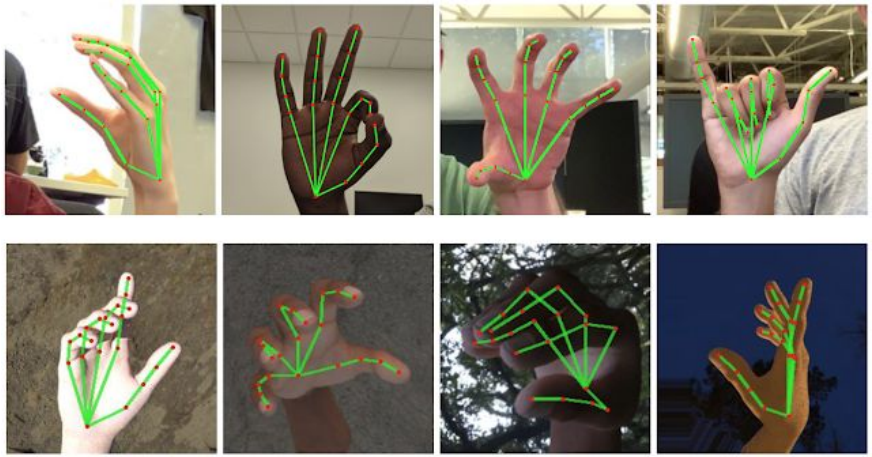


Hand Gestures for the Human-Car Interaction: the Briareo dataset, 2019
[Paper] [Project Page]

MediaPipe Hands



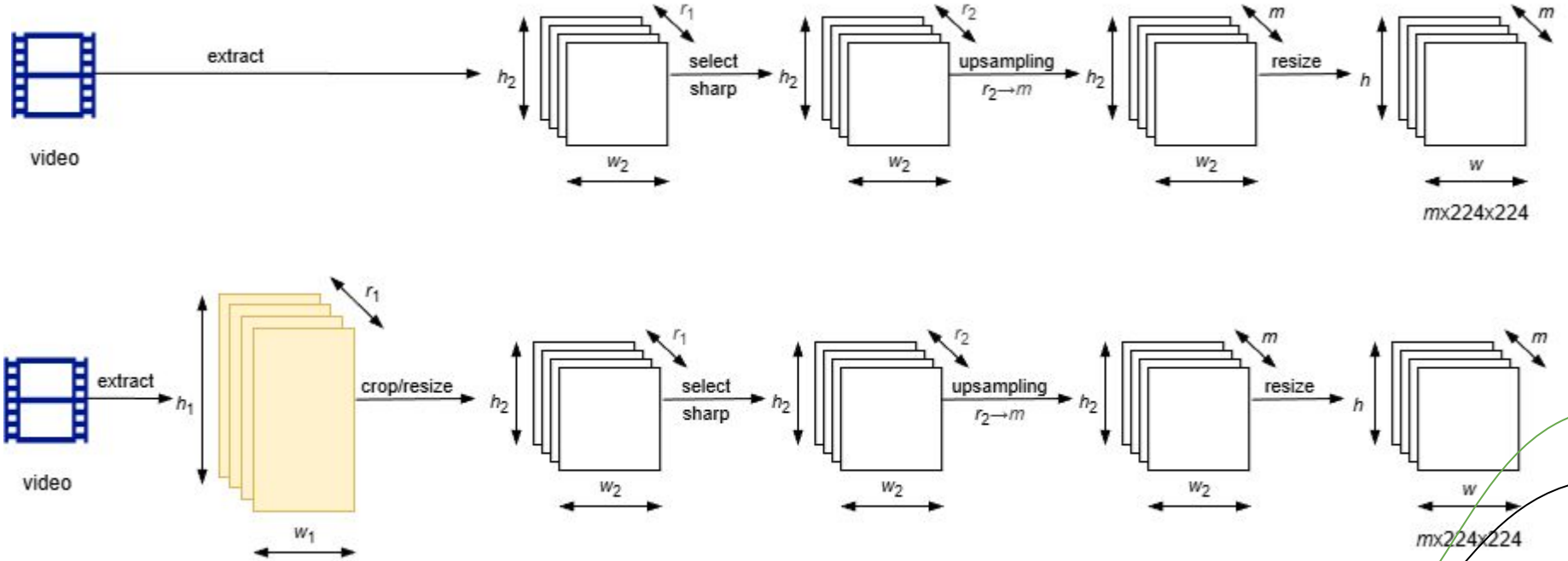
0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP



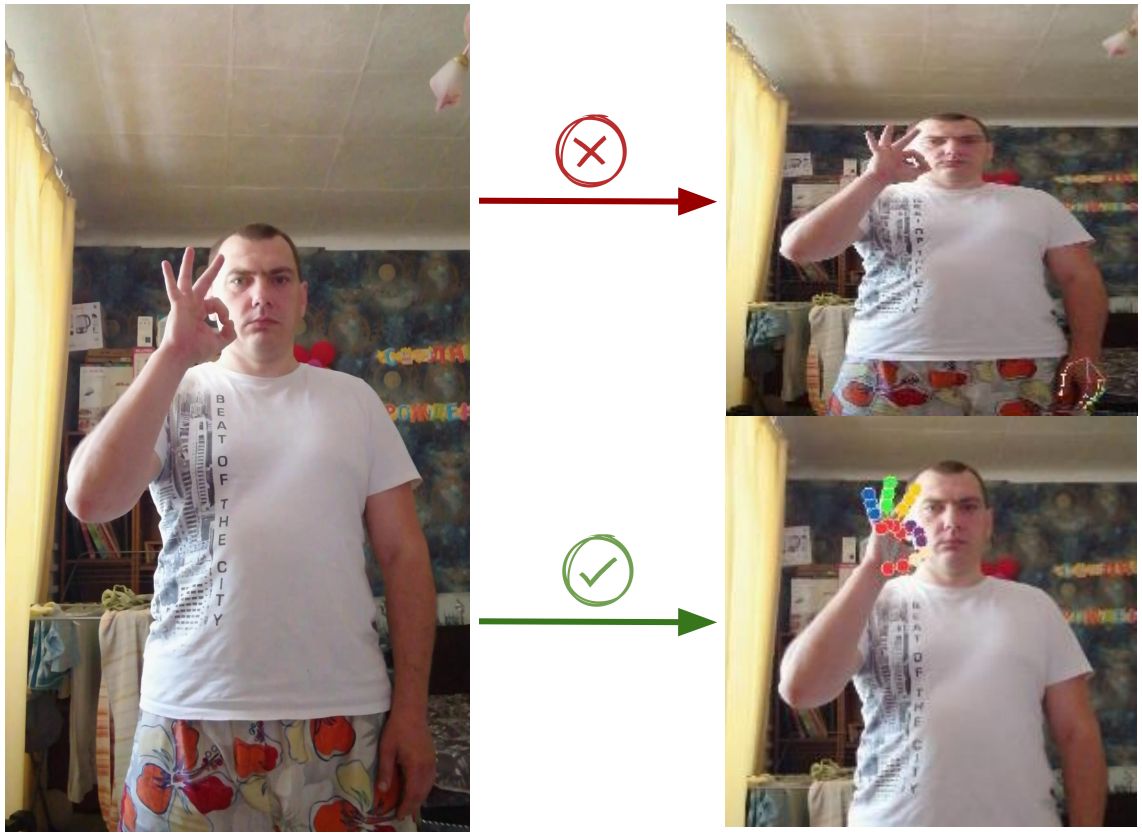
[Project Page]

MediaPipe Hands

Новый framer

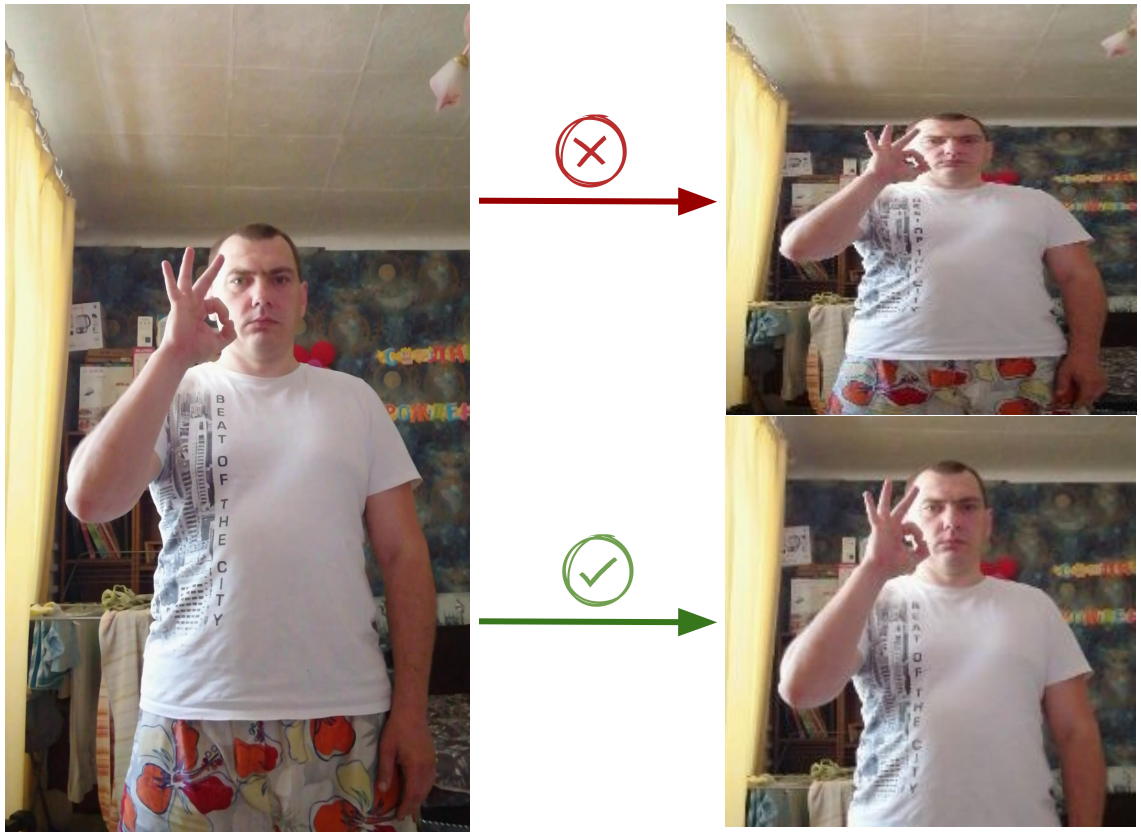


MediaPipe Hands (новый framer)



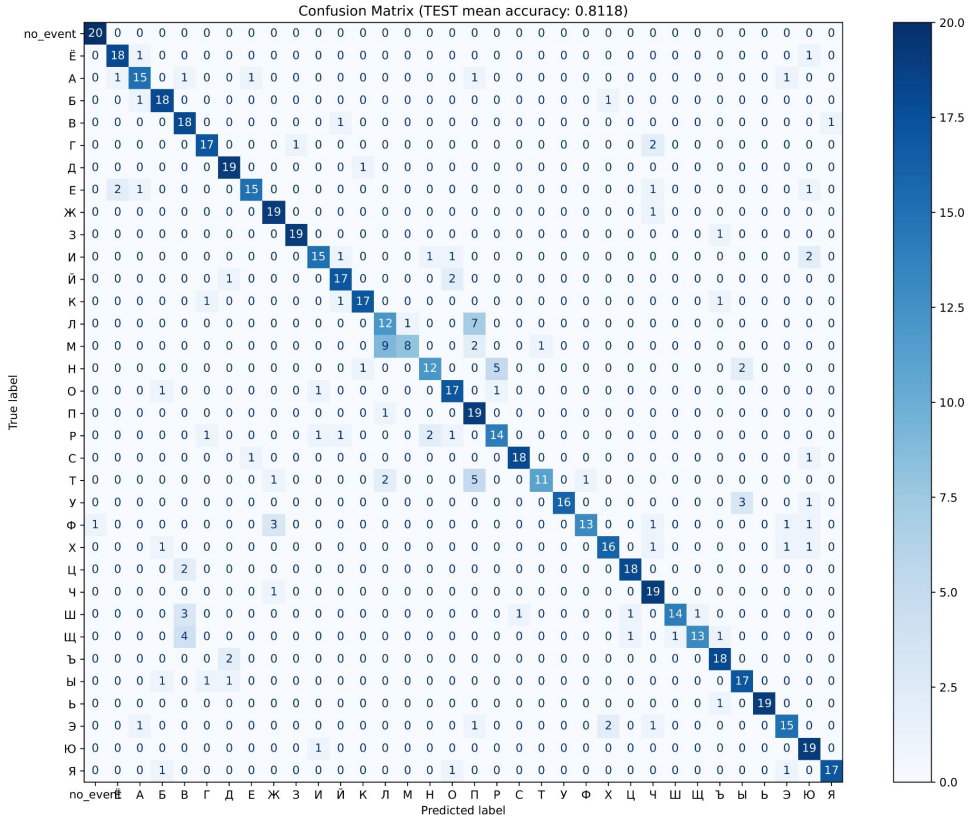
78.24% ← 75.29%

Новый framer без MediaPipe

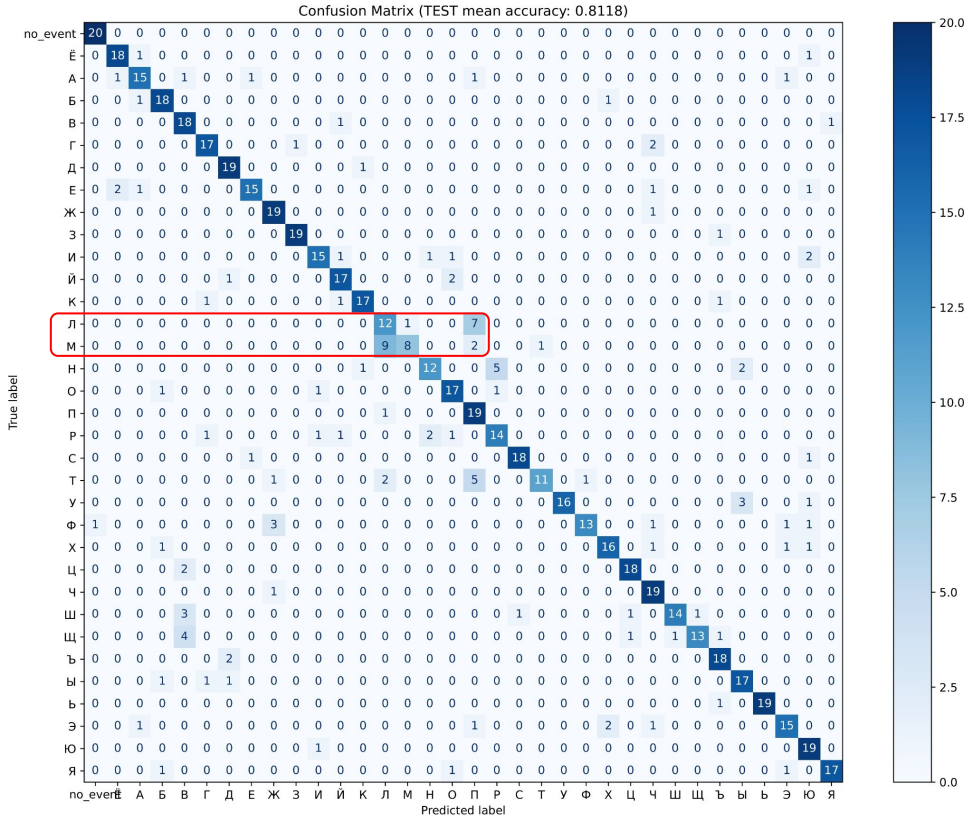


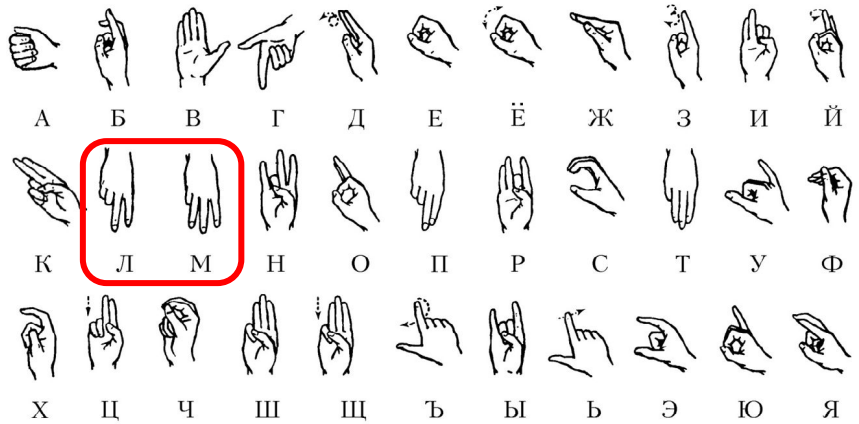
81.18% ← 78.24% ← 75.29%

На тесте: mean accuracy **81.18%**

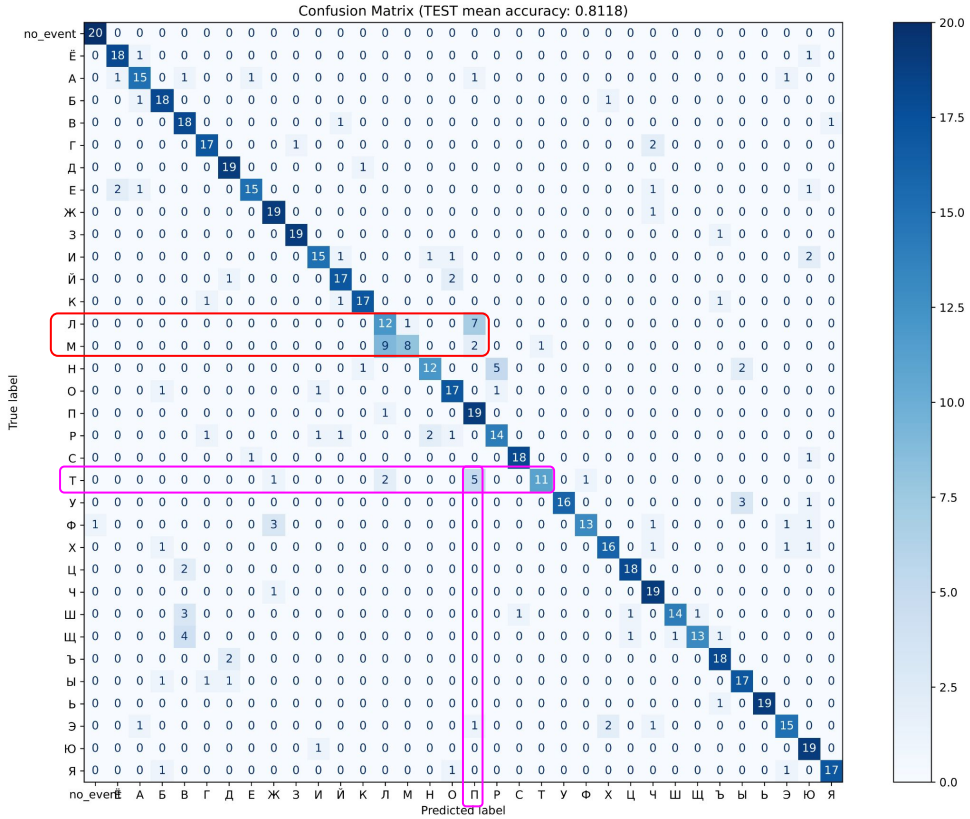


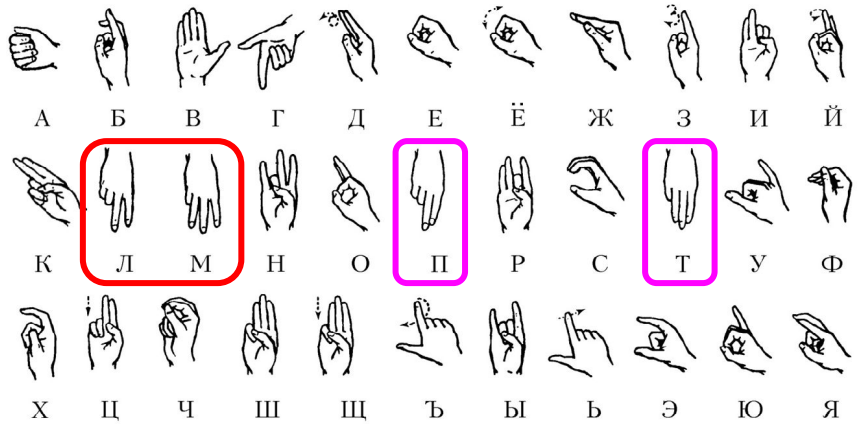
На тесте: mean accuracy **81.18%**





На тесте: mean accuracy **81.18%**





На тесте: mean accuracy **81.18%**