Vilniaus universitetas

VEIKLOS ATASKAITA

Kelių asmenų kūno padėties sekimas realiuoju laiku taikant mašininio mokymosi metodus

Doktorantūros pradžios/pabaigos metai: 2022-2026
Studijų metai: 2023/2024 antras pusmetis

Doktorantas: Algimantas Skuodis
Vadovas: Olga Kurasova, prof. Dr

Vilniaus universitetas

Studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	2	2
II (2023/2024)	**2**	**2**
III (2024/2025)		
IV (2025/2026)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2022/2023)										
II (2023/2024)	**1**	**1**						**1**	**1**	**Publi kuota**
III (2024/2025)	1				1					
IV (2025/2026)	1				1					
Iš viso:	2		1		2			1		

Vilniaus universitetas

Egzaminai 2023/2024 (II pusmetis)		
Planas	**Įvykdyta**	**Būklė**
Gilieji neuroniniai tinklai (2024 m. birželio mėn.)	Gilieji neuroniniai tinklai (2024 m. birželio mėn.)	Išlaikytas

Dalyvavimas konferencijose 2023/2024 (II pusmetis)		
Planas	**Įvykdyta**	**Konferencijos tipas**
Dalyvavimas The 16th International Baltic Conference on Digital Business and Intelligent Systems (2024 birželio 20 – liepos 3)	Algimantas Skuodis and Olga Kurasova, Evaluation of Deep Learning-Based Models for Recognition of Skydiving Formations, The 16th International Baltic Conference on Digital Business and Intelligent Systems, 2024 birželio 20 – liepos 3	Tarptautinė

Vilniaus universitetas

Publikacijos 2023/2024 (II pusmetis)			
Planas	**Įvykdyta**	**Būklė**	**Publikacijos tipas**
Communications in Computer and Information Science, vol 2157. Springer	Skuodis, A., Kurasova, O. (2024). Evaluation of Deep Learning-Based Models for Recognition of Skydiving Formations. In: Lupeikienė, A., Ralyté, J., Dzemyda, G. (eds) Digital Business and Intelligent Systems. DB&IS 2024. Communications in Computer and Information Science, vol 2157. Springer, Cham. https://doi.org/10.1007/978-3-031-63543-4_14	Publikuota: 2024-05-23	Be cituojamumo rodiklio

Vilniaus
universitetas

Tyrimų objektas

Mašininio mokymosi metodai kelių asmenų kūno padėties sekimui realiuoju laiku.

Vilniaus universitetas

Tikslas

Sukurti mašininio mokymosi metodą kelių asmenų kūno padėties sekimui realiuoju laiku, skirtą kūno padėties sekimui ir taškų skaičiavimui parašiutų sporto laisvojo kritimo derinių disciplinoje.

Uždaviniai

- Ištirti modernius mašininio mokymosi metodus, naudojamus kelių asmenų kūno padėties sekimui realiuoju laiku.
- Sukurti duomenų rinkinį, skirtą mašininio mokymosi metodų, skirtų kelių asmenų kūno padėties sekimui realiuoju laiku parašiutų sporto laisvojo kritimo derinių disciplinoje, vertinimui.
- Pasiūlyti mašininio mokymosi metodą, tinkamą kelių asmenų kūno padėties sekimui realiuoju laiku parašiutų sporto laisvojo kritimo derinių disciplinoje.
- Atlikti eksperimentinius tyrimus ir įvertinti pasiūlytą metodą naudojant sukurtą duomenų rinkinį.
- Atlikti eksperimentinius tyrimus ir įvertinti pasiūlytą metodą naudojant kitus atvirus duomenų rinkinius, dažniausiai naudojamus kelių asmenų kūno padėties identifikavimo metodų vertinimui.

Laisvojo kritimo deriniai

- 4 kilometrai
- 10 turų / šuolių
- 35 sekundės laisvojo kritimo
- 5-6 burtų būdu ištrauktos figūros
- Viso 16 figūrų ir 22 blokai (2 figūros)
- t.y., 60 skirtingų formacijų

Round 10 - 451 "Great Britain" 4-Way Female

Problema

- Nėra galimybės matyti preliminarų rezultatą gyvai

Kuriamas duomenų rinkinys

- 12 varžybų (2019-2022)
- 60 formacijų (figūrų)
- 26 436 kadrai

Competition	Rounds	Teams
Airspace2022	10	36
Eloy2022	10	25
Tanay2021	6	13
Eloy2019	10	20
FlyspotOpen2021	10	6
FlyspotOpen2022	10	11
DIPC2021	10	6
USPANationals2022	10	12
SwissNationals2022	8	3
USPANationals2022Advanced	10	15
1STDITC2021	10	7
AbhuDhabiFirstOpen2022	6	4

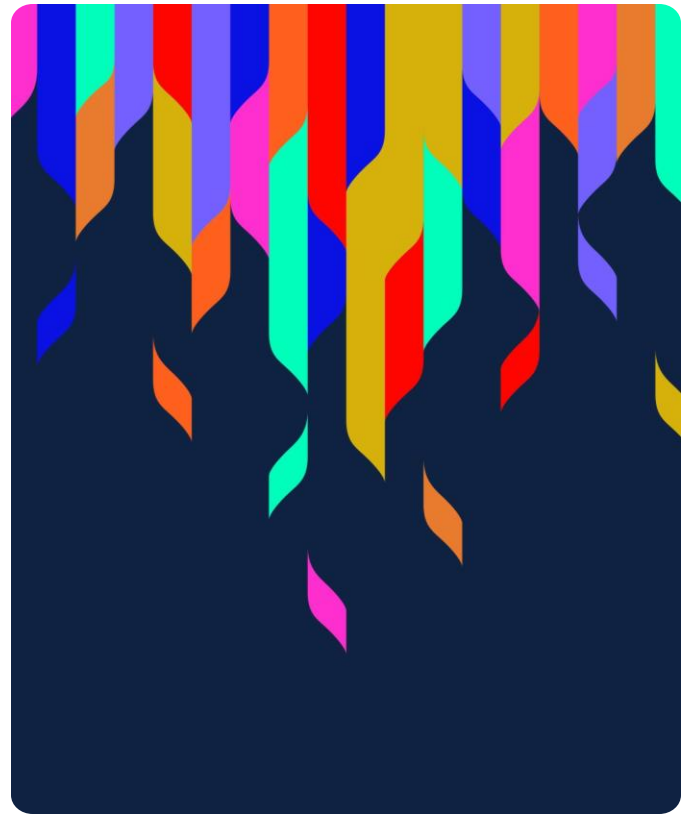
Experimentinis duomenų rinkinys

- Pirmos šešios figūros (A,B,C,D,E,F),
- 2946 vaizdai,
- 6 klasės,
- Suvienodintas dydis,
- Pašalintos šiukšlės,
- Padalinta 80/20.

Pasirinkti modeliai

- ResNet,
- EfficientViT,
- FastViT,
- ConvMixer.

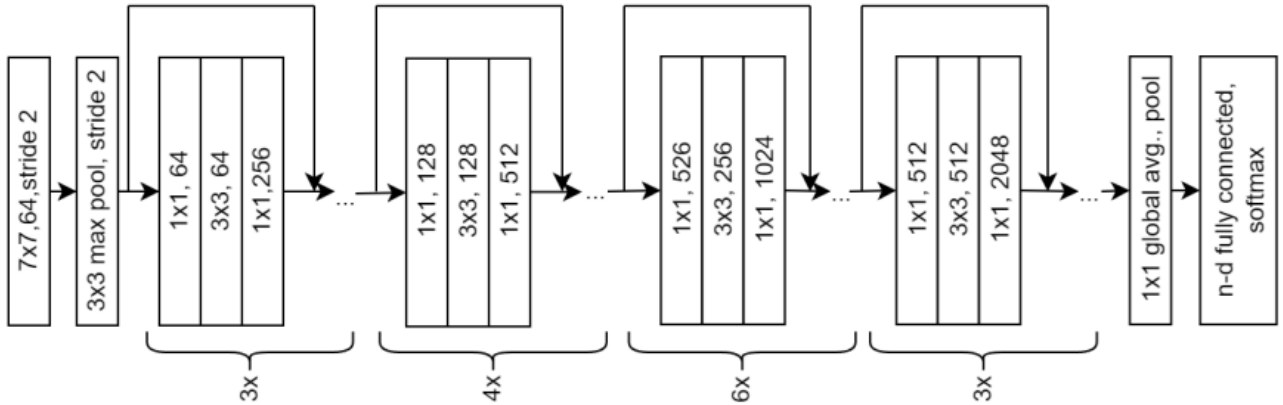
Kodėl šie?

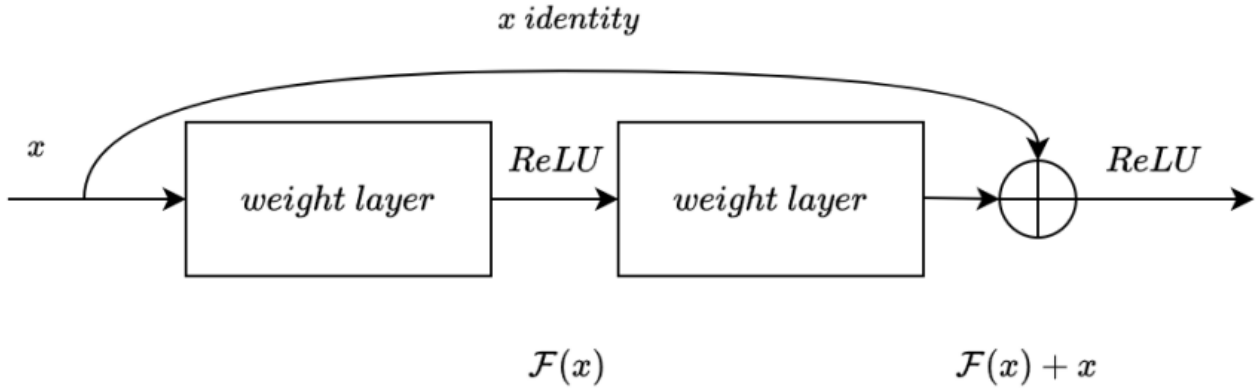


ResNet

https://arxiv.org/pdf/1512.03385.pdf

- ResNet-50,
- Liekanų neuroniniai trinklai,
- Liekanų blokai,
- 50 sluoksnių su svoriais,
- Daugelio sudėtingesnių modelių fundamentas.





Visual Transformers (ViT)

An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale

Paskutiniu metu gerus rezultatus rodo transformerių tinklų architektūra paremti asmenų kūno padėties atpažinimo ir sekimo metodai,

- Vieni jų naudoja konvoliucinius tinklus kaip pagrindą ir vėliau naudoja transformerius kūno dalių esminių taškų susiejimui,
- Antri, bando naudoti tiktai transformerius kūno padėčių atpažinimui,
- ViT - labai daug variantų (pvz https://github.com/lucidrains/vit-pytorch),
- Ilgas mokymo laikas,
- Didelis parametrų kiekis,
- Reikalauja didelio duomenų kiekio mokymams,

Todėl pasirinkti **EfficientViT** ir **FastViT**:

- Abiejų tikslas buvo sumažinti parametrų kiekį ir klasifikavimo laiką,
- t.y., pritaikyti mobiliems įrenginiams ir realaus laiko apdorojimui.

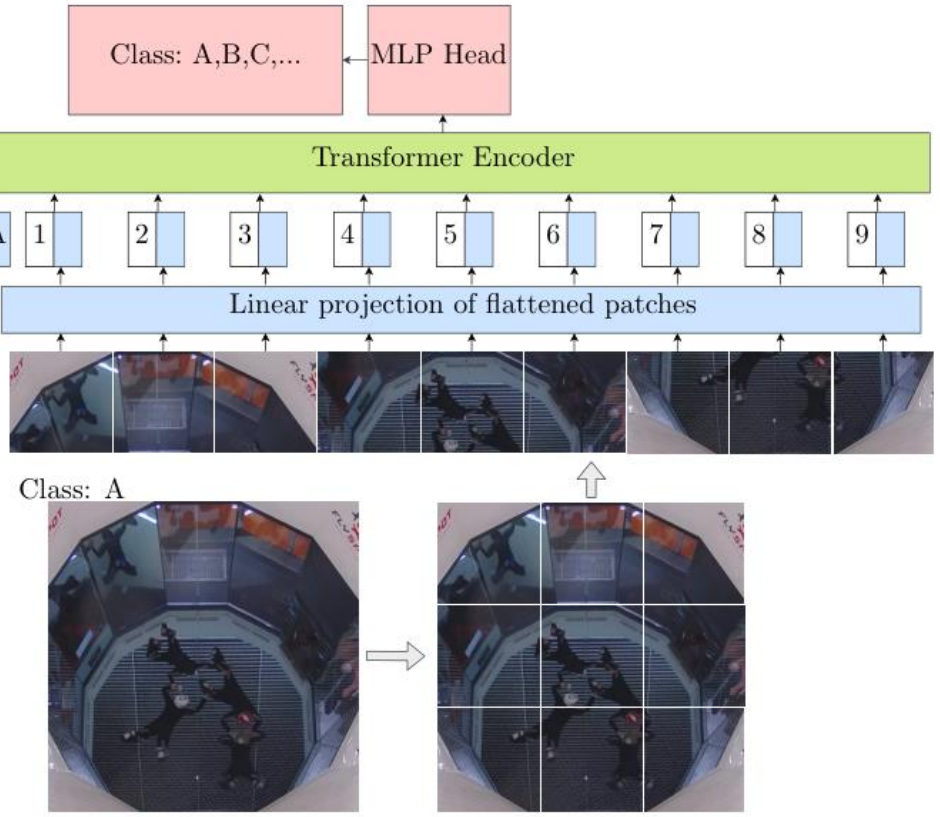


Figure 3: Architecture of Vision Transformers (ViT).

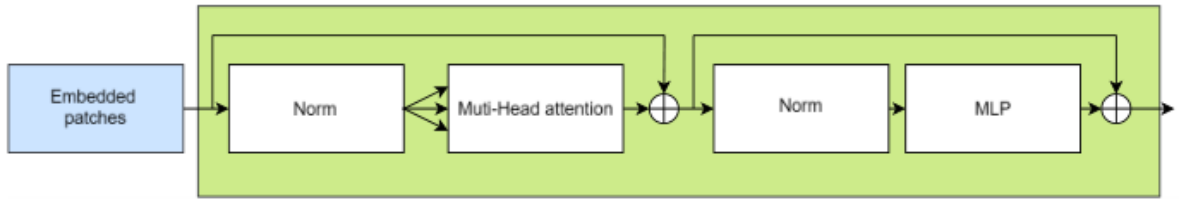
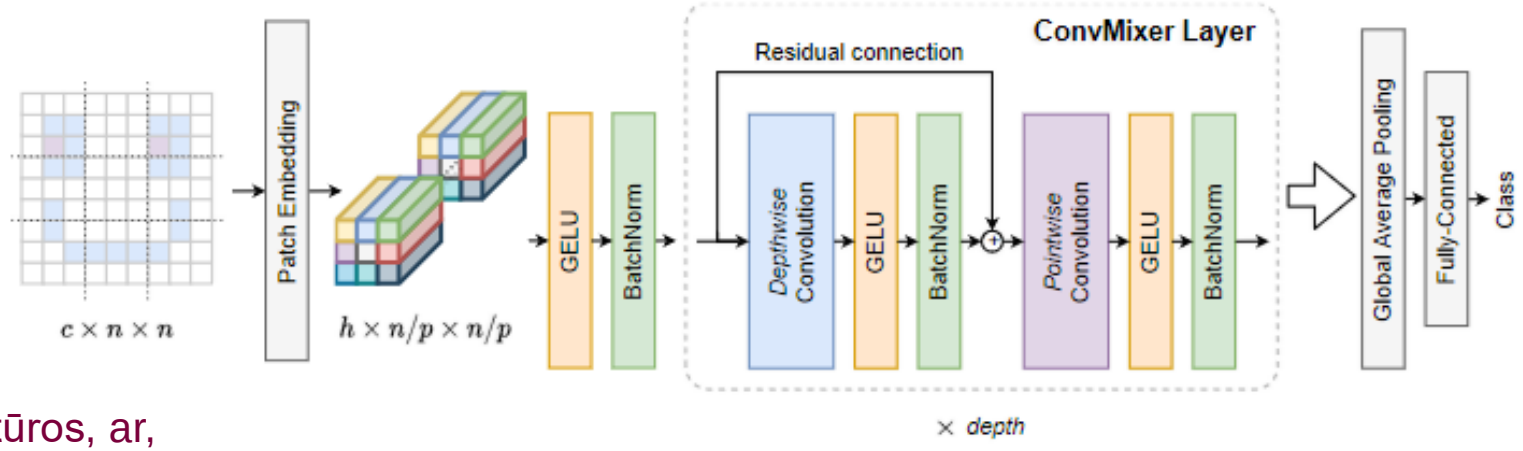


Figure 4: Transformer Encoder.

ConvMixer

Patches Are All You Need?



Ar ViT geresni dėl Transformerių architektūros, ar, iš dalies, dėl skiaučių kaip įvesties panaudojimo ?

- Nebėra Transformerių blokų,
- Tačiau pradinė įvestis kaip ViT - skiautėmis,
- Toliau d kiekis ConvMixer blokų,
- Labai paprasta architektūra,
- Parametrų mažiau nei ViT,
- Rezultatai geresni nei ResNet-152 ar kai kurių ViT modelių*

(* pagal autorių eksperimentų rezultatus)

Aplinka

https://mif.vu.lt/itwiki/hpc

- VU HPC GPU telkinys,
- Kiekvienai užduočiai naudojami 2 GPU.

Augmentavimas

- Atsitiktinis (RandAugment),
- Stiprumas 9,
- Galima stiprumo sklaida 0.5,
- Augmentacijos - 2 iš:
- 'AutoContrast','Equalize','Invert','Rotate','Posterize','Solarize','SolarizeAdd','Color','Contrast','Brightness','Sharpness','ShearX','ShearY','TranslateXRel','TranslateYRel'.

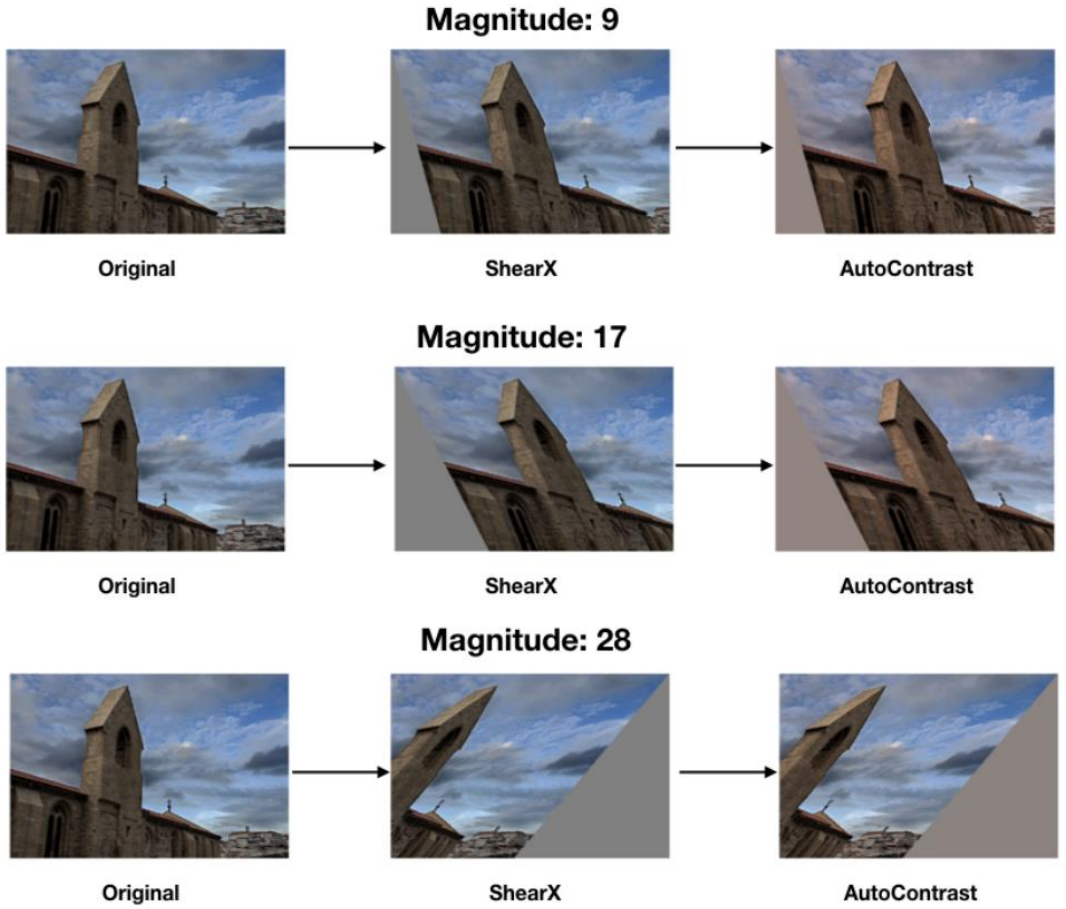


Figure 1. **Example images augmented by RandAugment.** In these examples N=2 and three magnitudes are shown corresponding to the optimal distortion magnitudes for ResNet-50, EfficientNet-B5 and EfficientNet-B7, respectively. As the distortion magnitude increases, the strength of the augmentation increases.

Šaltinis: https://arxiv.org/pdf/1909.13719v2.pdf

Rezultatai

- Perkėlimo mokymo atveju paimti modeliai prieš tai apmokyti ImageNet-1k,
- ConvMixer_h_d kur h skiaučių gylis ir d ConvMixer blokų kiekis,
- FastVit-MA36, MA36 žymi variantą su didesne įterpinių dimensija,
- EfficientViT-M0, M0 architektūros variantas (žiūrėti autorių darbą).

Model	ResNet50	ConvMixer_1536_20	EfficientViT_m0	ConvMixer_1024_20
Pretrained	FALSE	FALSE	TRUE	TRUE
Parameters	23,520,326	50,098,182	2,157,594	23,364,614
Accuracy of				
class A	0.4390	0.8415	0.9390	0.9634
class B	0.1129	0.4355	0.8548	0.9839
class C	0.6471	0.2857	0.8908	0.9832
class D	0.4286	0.7959	0.7653	0.9082
class E	0.5600	0.7800	0.9400	0.9700
class F	0.1527	0.7939	0.8550	0.9924
F1 score (weighted)	0.3778	0.6415	0.8731	0.9677

Model	ResNet50	FastViT_ma36	ConvMixer_1536_20	ConvMixer_768_32
Pretrained	TRUE	TRUE	TRUE	TRUE
Parameters	23,520,326	42,858,306	50,098,182	20,345,862
Accuracy of				
class A	0.9634	**0.9756**	0.9634	**0.9756**
class B	0.9677	0.9839	**1.0000**	**1.0000**
class C	0.9748	0.9832	**1.0000**	0.9916
class D	0.9184	0.9286	0.9286	**0.9592**
class E	0.9700	0.9600	**0.9900**	**0.9900**
class F	**1.0000**	0.9924	0.9924	1.0000
F1 score (weighted)	0.9678	0.9712	0.9797	**0.9865**

Kūno dalių ir žmonių atpažinimo modelių bandymai

- Pasirinkti modeliai:
  - YOLOv8n-pose
  - YOLOv8n
  - YOLOv8x-pose-640
  - DETR-ResNet-50

https://youtu.be/fPjFh2v--Mk

Rezultatai

Table 4: Proportion of frames with the lowest confidence levels of 0.50, 0.75, and 0.90 for detecting 4 and 3 human bodies

File	Environment	Total Frames	Confidence Threshhold	Frames ($\geq$ 3 bodies)	Proportion ($\geq$ 3 bodies)	Frames ($\geq$ 4 bodies)	Proportion ($\geq$ 4 bodies)
1.mp4	Indoor	290	0.5	252	0.87	204	0.70
			0.75	218	0.75	150	0.52
			0.9	151	0.52	85	0.29
2.mp4	Outdoor	450	0.5	449	1.00	447	0.99
			0.75	448	1.00	441	0.98
			0.9	441	0.98	401	0.89
3.mp4	Indoor	450	0.5	442	0.98	414	0.92
			0.75	431	0.96	369	0.82
			0.9	392	0.87	256	0.57
4.mp4	Outdoor	443	0.5	397	0.90	334	0.75
			0.75	322	0.73	198	0.45
			0.9	191	0.43	70	0.16

Rezultatai

Table 5: Average proportion of frames across all selected video files where three (or more) and four (or more) bodies were detected with corresponding confidence thresholds

Confidence Threshhold	Average proportion of frames ($\geq$ 3 bodies)	Average proportion of frames ($\geq$ 4 bodies)
0.5	0.93	0.84
0.75	0.86	0.69
0.9	0.7	0.48

Išvados

- Galima pasiekti gana aukštą klasifikavimo rezultatą naudojant standartinius vaizdų klasifikavimui skirtus modelius (6 klasių atveju),
- Perkėlimo mokymas stipriai pagerina rezultatą (pvz ResNet nuo 0.4 iki 0.9),
- ConvMixer pasiekė geresnį rezultatą su mažesniu parametrų kiekiu nei ViT ir ResNet modeliai,
- Pasirinktuose vaizdo įrašuose standartiniai YOLOv8n ir YOLOv8n-pose modeliai sunkiai atpažįsta žmones ir jų kūno dalis.
- Pasirinktuose vaizdo įrašuose standartinis transformerių architektūra paremtas DETR-ResNet-50 modelis atpažįsta trijų ar daugiau žmonių kūnus su klasifikavimo įverčiu >=0.9 vidutiniškai 70% vaizdo kadrų.
- Pasirinktuose vaizdo įrašuose standartinis transformerių architektūra paremtas DETR-ResNet-50 modelis atpažįsta keturių ar daugiau žmonių kūnus su klasifikavimo įverčiu >=0.9 vidutiniškai 48% vaizdo kadrų.

Tolimesni darbai

- Eksperimentai su didesniu klasių kiekiu (viso yra 60),
- Nauji modeliai, įvertinti ConvMixer modelio panaudojimo galimybę vaizdų segmentavimui ir spartesniam formacijų identifikavimui,
- Pradėti ruošti publikaciją į žurnalą su citavimo rodikliu.

Vilniaus universitetas

Šaltiniai

1. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. AnImage is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs]. URL: http://arxiv.org/abs/2010.11929
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs]. URL: http://arxiv.org/abs/1512.03385
3. Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention, May 2023. arXiv:2305.07027 [cs]. URL: http://arxiv.org/abs/2305.07027
4. Asher Trockman and J. Zico Kolter. Patches Are All You Need?, January 2022. arXiv:2201.09792 [cs]. URL: http://arxiv.org/abs/2201.09792
5. Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization, August 2023. arXiv:2303.14189 [cs]. URL: http://arxiv.org/abs/2303.14189
6. Phil Wang. lucidrains/vit-pytorch, February 2024. original-date: 2020-10-03T22:47:24Z. URL: https://github.com/lucidrains/vit-pytorch

Vilniaus
universitetas

Klausimai?