Segmentation-guided Privacy-preservation in Visual Surveillance Monitoring

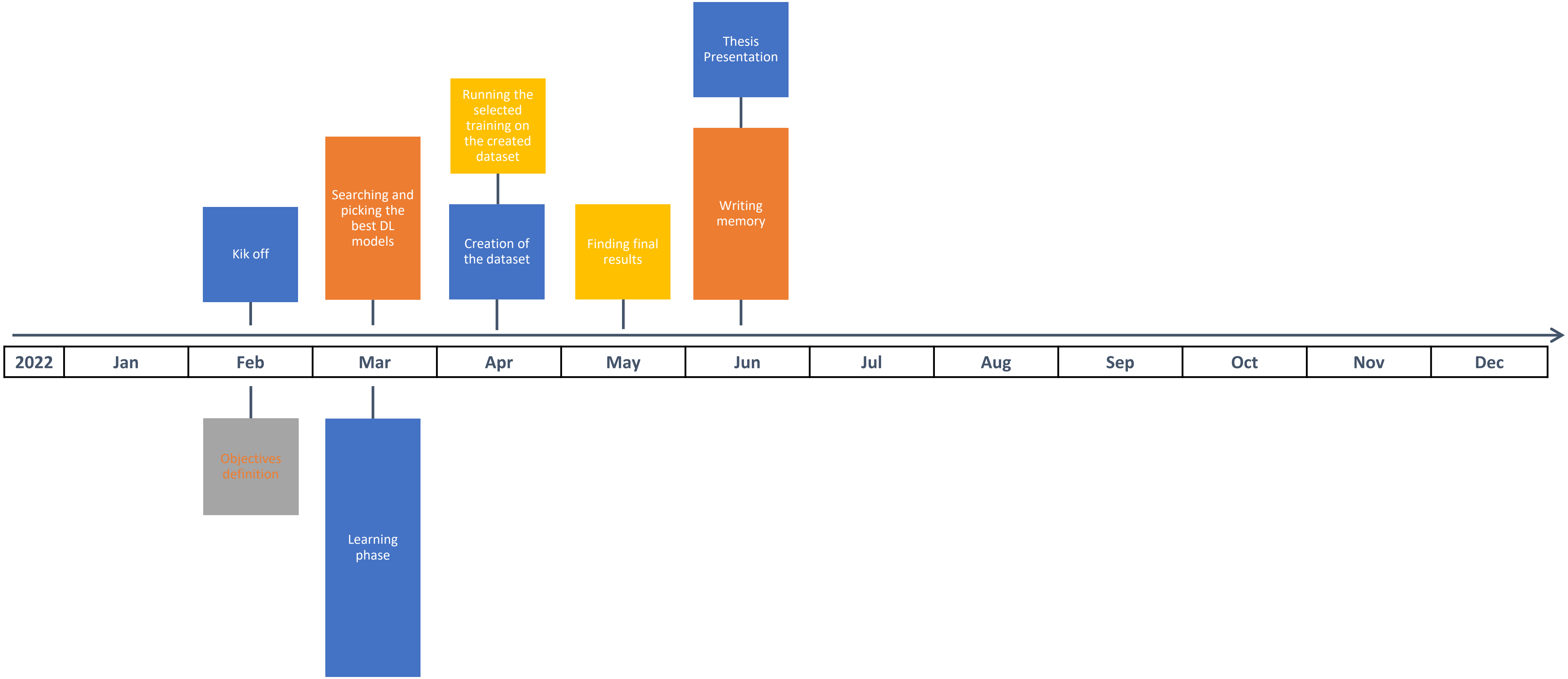
Author: Daniel Geryous Fares

Directors:

Dr. Sergio Escalera Guerrero
Zenjie Li
Kamal Nasrollahi

CONTENT

Timeline



Thesis Presentation
Running the selected training on the created dataset
Searching and picking the best DL models
Kik off
Creation of the dataset
Finding final results
Writing memory
2022
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Objectives definition
Learning phase

Introduction & Motivation

Introduction & Motivation

Problem

1. In today's world, video surveillance has become a necessity for safety and security.
2. CCTV cameras are installed in private and public spaces.
3. Privacy is compromised for monitored individuals.

Introduction & Motivation

Solution

1. Real-time visualization of privacy protected video sequences.
2. Make sure monitored persons' personal information is protected.
3. Maintain the ability to monitor and identify potential risk behaviors.

Privacy preservation in video sequences

Privacy preservation in video sequences

GDPR & AI

GDPR: Protection of individuals' fundamental rights and freedom as well as giving them control over their collected data and how it is being processed.

Region of Interest: Any data that can be used to identify a person in a video sequence.

AI within GDPR context:

- Privacy by design system.
- Irreversible results safeguarding personal data.
- Minimization of accessible/visible personal data.
- Minimization of quasi-identifiers that cannot be fully protected (gait, way of speaking. Language).
- Ensuring integrity, confidentiality, privacy.

Existing methods.

- Blanking and masking.
- Obfuscation and scrambling.
- Pixelation and Blurring.
- Mosaic.
- Cartooning.
- Warping.
- Morphing.
- Visual Abstraction.
- Tokenization.
- False Colors.
- Hashing.

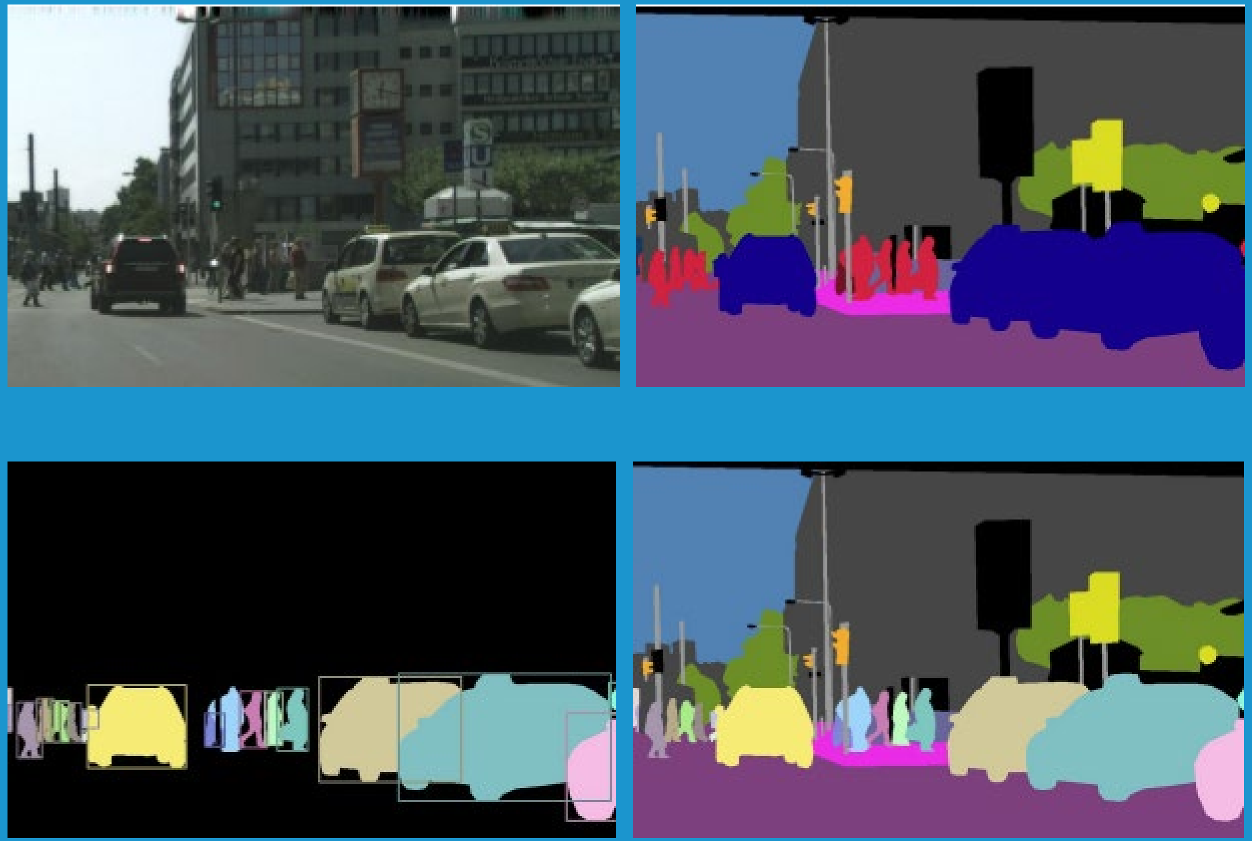
Original Frame (a) Blur (b) Pixelate (c) Mosaic (d) Cartooning (e) Masking (f) Warp (g) Morph (h) Abstraction with 3D Avatar (i-1) Face Scrambling (i-2) Frame Scrambling (j) False Colours

Semantic Segmentation

Image segmentation

Semantic segmentation

- Types:
  - Semantic segmentation
  - Instance segmentation
  - Panoptic segmentation
- Evaluation metrics:
  - Pixel accuracy
  - Intersection over Union (IoU)
  - Mean (IoU)
  - Dice coefficient



Dataset

Dataset

Surveillance-like dataset

Merging 3 datasets:

- Identifying images that has human class.
- Isolate these images.
- Binarize these images (1 for human and 0 for any other class).
- Analysis and evaluation.
- Data cleansing.
- Merge and split datasets.
- Revaluation of the result dataset.

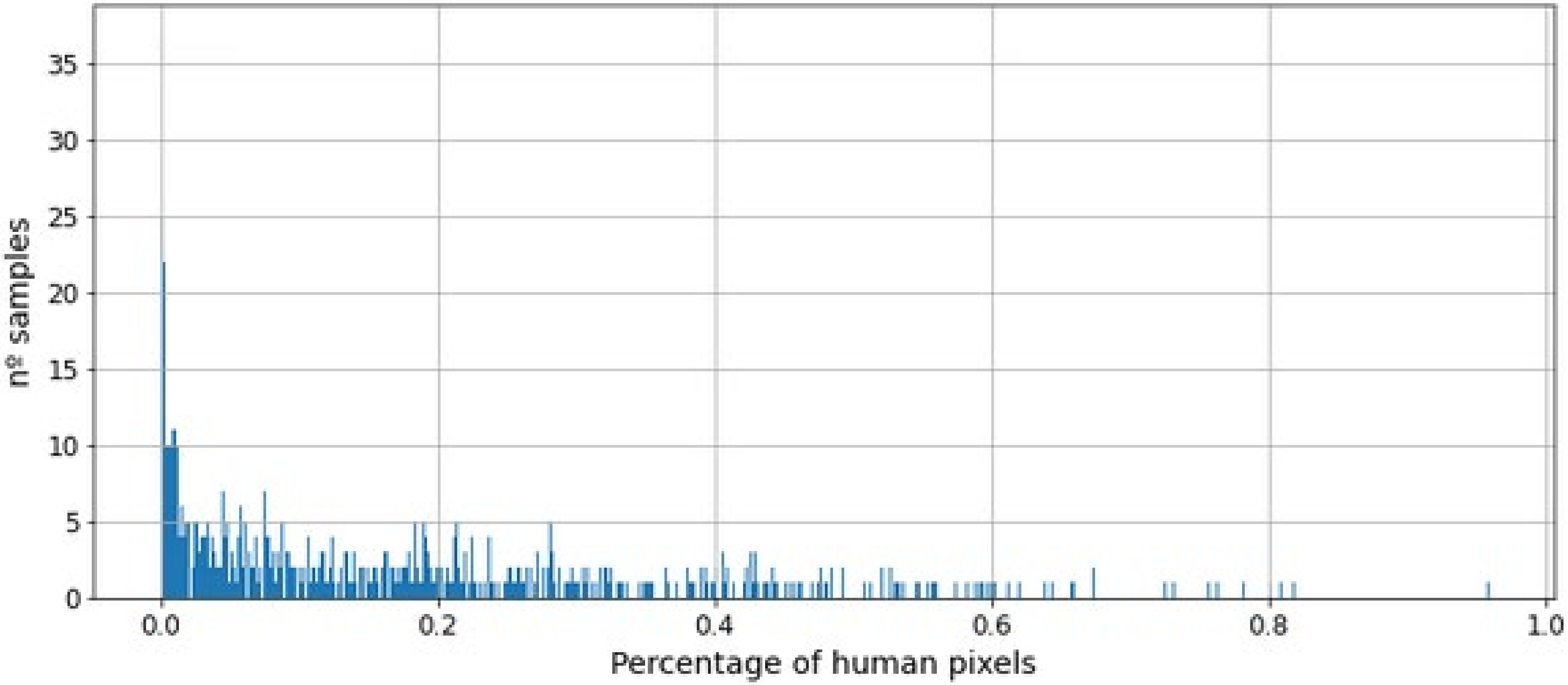
Dataset

Surveillance-like dataset

Original Dataset	shape	Total Size	Train Size	Test Size	Human Train Size	Human Test Size	+ No-human images	Human Train Size	Human Test Size	Human Pixels % (training)	Human Pixels % (validation)
VOC Pascal	512x512	11540	5717	**5823**	3270	817	20%	4038	980	**0.153**	**0.236**
cocoStuff164	512x512	**123287**	**118287**	5000	**63965**	**2681**	**20%**	**76758**	**3217**	0.13	0.13
ADE20K	512x512	22210	20210	2000	5069	510	10%	5569	610	0.052	0.051
TOTAL		**157,037**	**144,214**	**12,823**	**72,255**	**4,030**	**19.5% approx.**	**86,365**	**4,807**	**0.126**	**0.135**

Dataset

Surveillance-like dataset

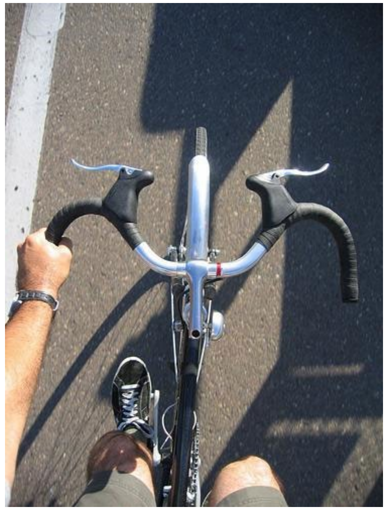


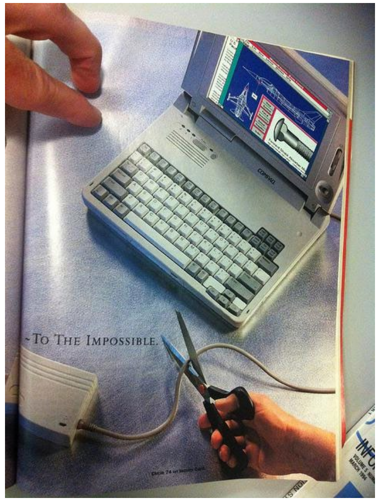
Dataset

Cleaning process

Dataset

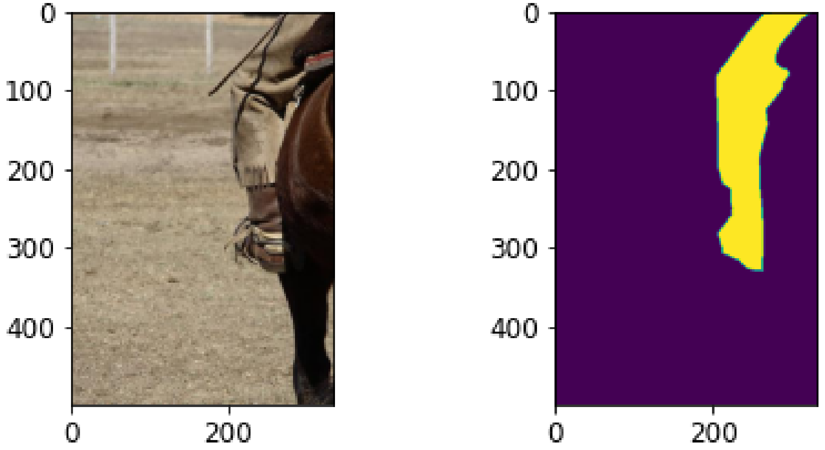
Cleaning process

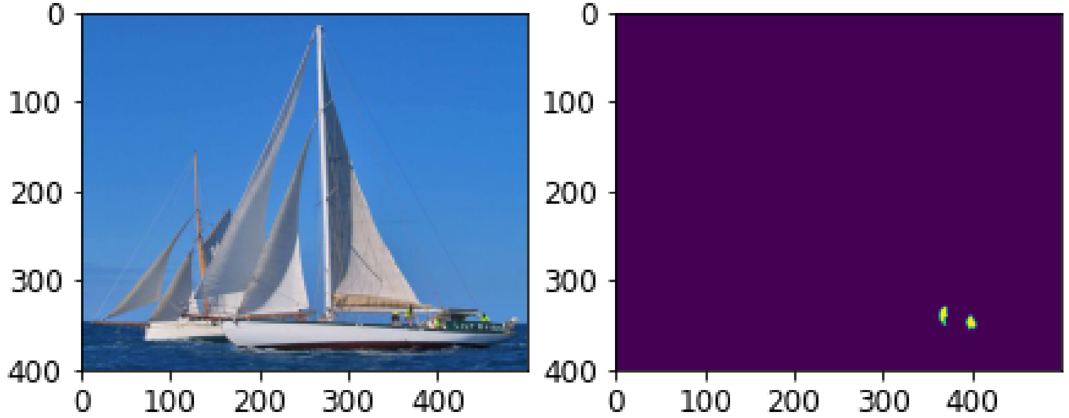



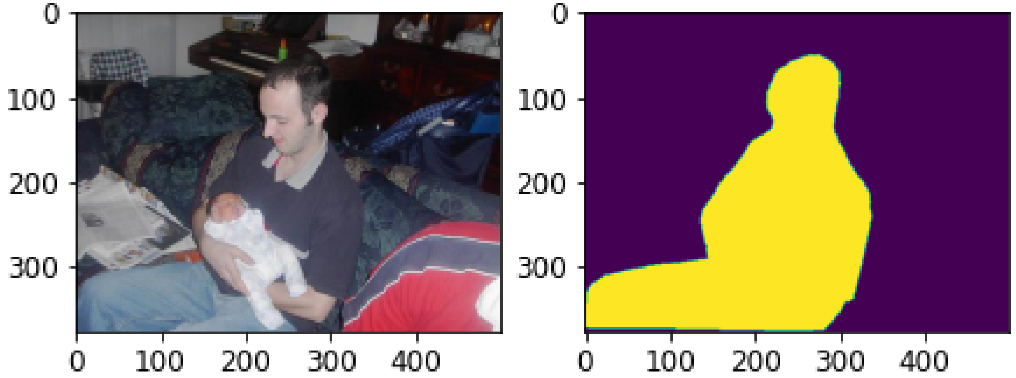


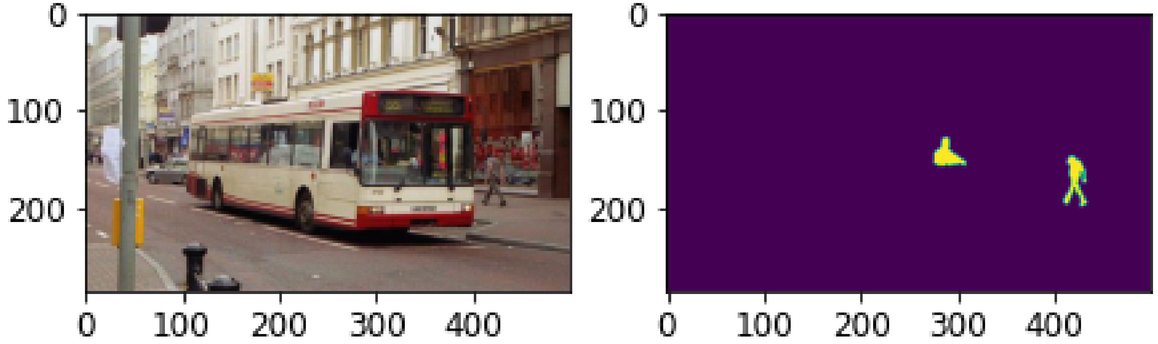
Dataset

Surveillance-like dataset









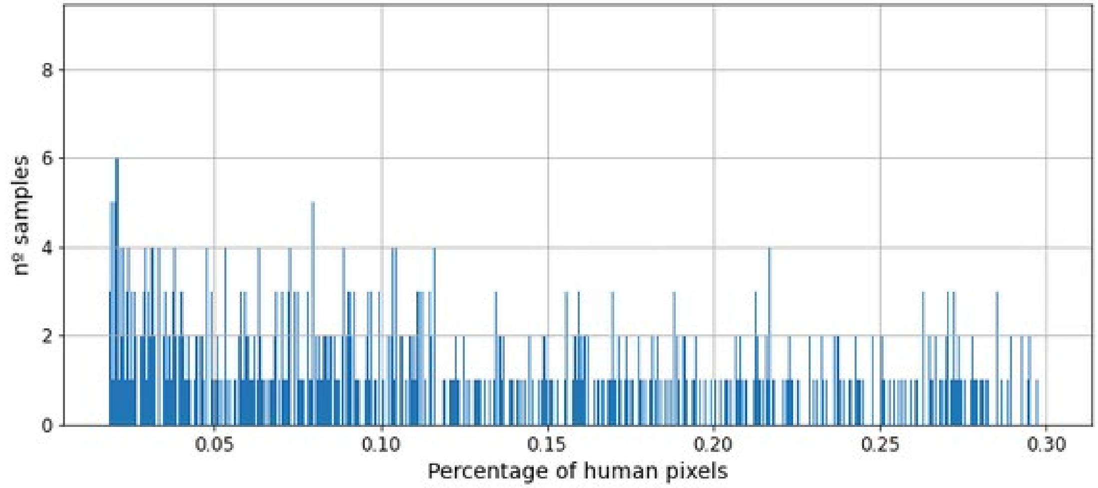
Dataset

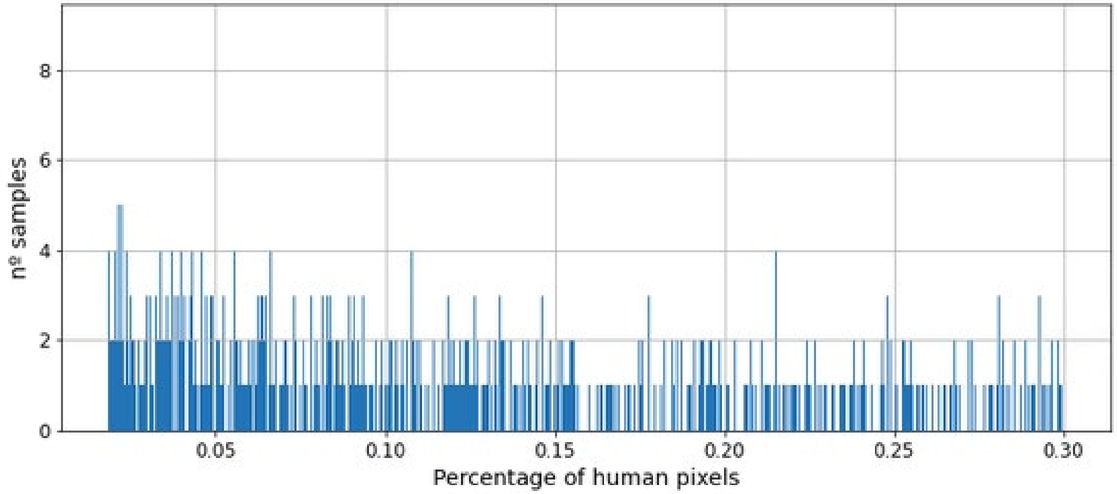
Cleansing process

Dataset	Human Train Size	Human Test Size	Human Pixels % (training)	Human Pixels % (validation)
Old	86,365	4,807	0.126	0.135
New	43,675	4,855	0.11	0.11

New Dataset	Total+noHuman20%	Removed
Validation	4,413 + 442	1,676
Training	39,704+3,971	28,050

Dataset
Cleansing process





Dataset

Split and merge

Dataset	Human Train Size	Human Validation Size	Human Test Size	Human Pixels % (training)	Human Pixels % (validation)	Human Pixels % (test)
new	43,675	4,855		0.11	0.11	-
Result	41,492	2,183	4,855	0.11	0.11	0.11

New Dataset	Total+noHuman20%	Removed
Validation	1977 + 206	-
Training	37727 + 3,765	2,183
Test	4,413 + 442	-

Methods

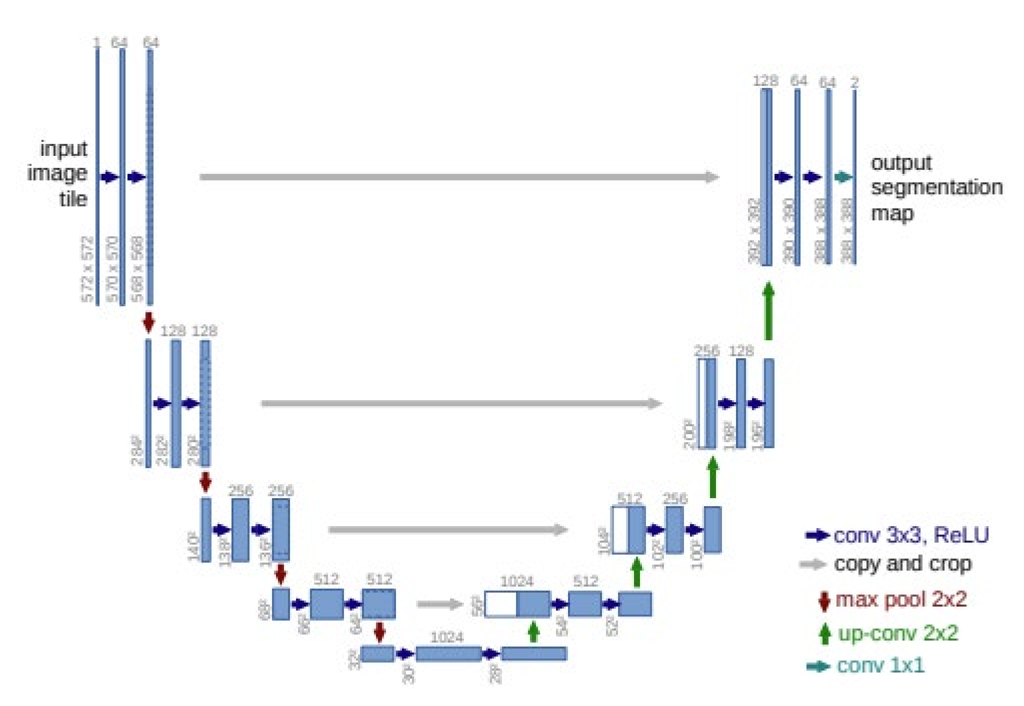
First steps to chose our model was by running through the results provided by mmsegmentation.

- This is one of the tables that were created in the early stages of this research:
  - Yields a better understanding and a clearer picture on the most extensively used datasets in semantic segmentation task.
  - Visualize how different and modern state-of-the-art approaches are behaving on them.
- Speed benchmark:
  - 8 NVIDIA Tesla V100 (32G) GPUs
  - Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz

CityScapes	resolution	Lr schd (iter)	FPS	mIoU
ISANet	512x1024	80000	2.35	80.32
FCN		40000	2.66	75.45
PSPNet		40000	2.68	78.34
UNet		160000	3.05	69.1
DeepLabV3 (FP16)		80000	3.86	80.48
SegFormer (B1)		160000	4.3	78.56
DeepLabV3+ (FP16)		80000	7.87	80.46
Semantic FPN		80000	10.29	75.8
STDC1 (No Pretrain)		80000	23.06	71.82
STDC2 (No Pretrain)		80000	23.71	73.15
ICNet	832x832	80000	27.12	68.14
BiSeNetV2	1024x1024	160000	31.77	73.21
BiSeNetV1 (No Pretrain)	1024x1024	160000	31.77	74.44

Models
UNet

- Built upon Fully Convolutional Network (FCN).
- Consists of two paths:
  - Contracting path
  - Expansive path
  - Plus, two 3x3 conv + ReLU and one final layer of 1x1 convolution.

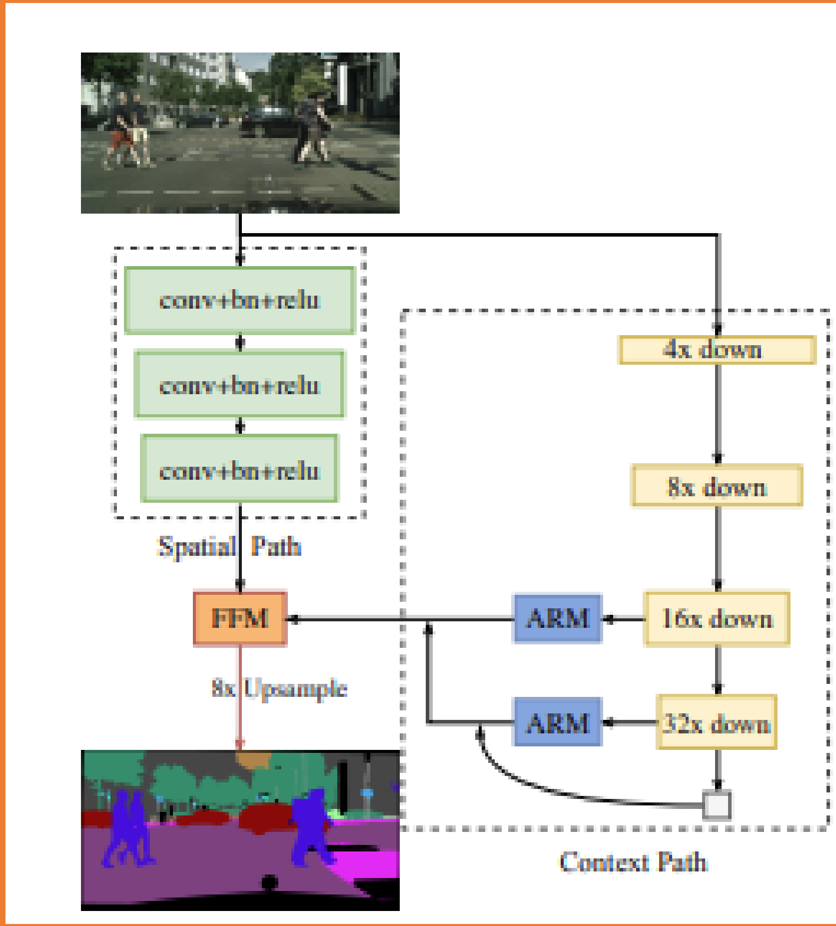


Models

BiSeNet

Bilateral Segmentation Network

- BiSeNet has two parts:
  - Spatial Path (SP) to tackle the loss of spatial information problem.
  - Context Path (CP) to tackle the shrinkage of the receptive field.



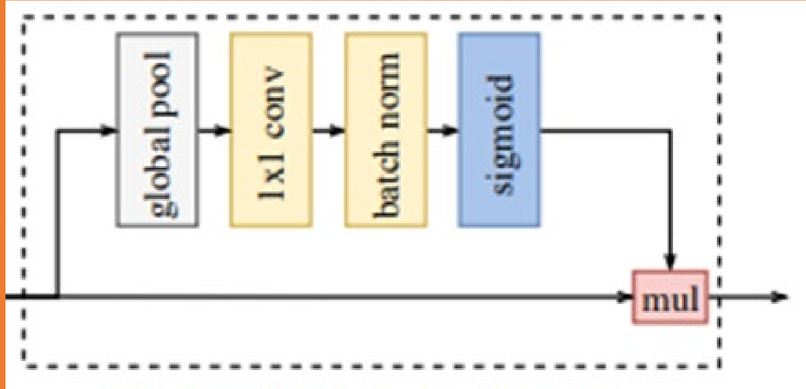
The length of block indicates the spatial size, while the thickness represents the number of channels.

Models

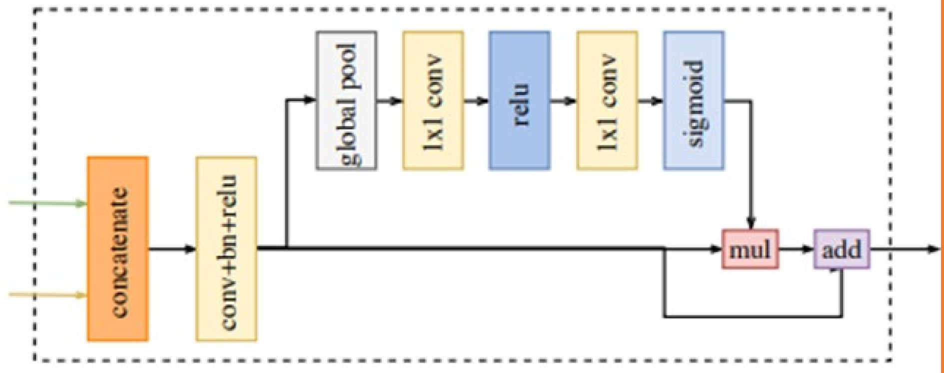
BiSeNet

Bilateral Segmentation Network

- For better accuracy without loss of speed, they introduce FFM and ARM:
  - Feature Fusion Module (FFM) to fuse the two paths.
  - Attention Refinement Module (ARM) for the refinement of final prediction.



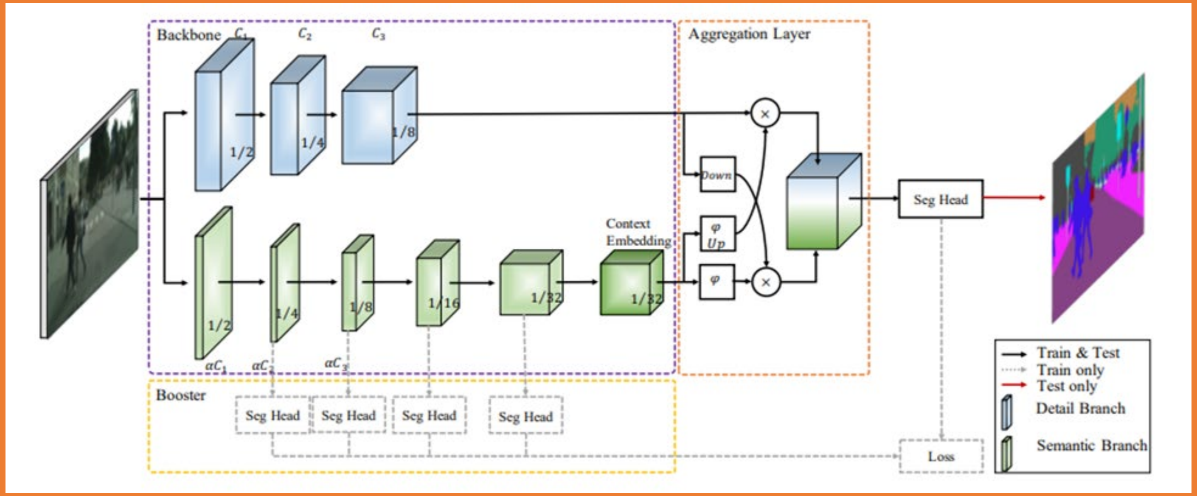
(b) Attention Refinment Module

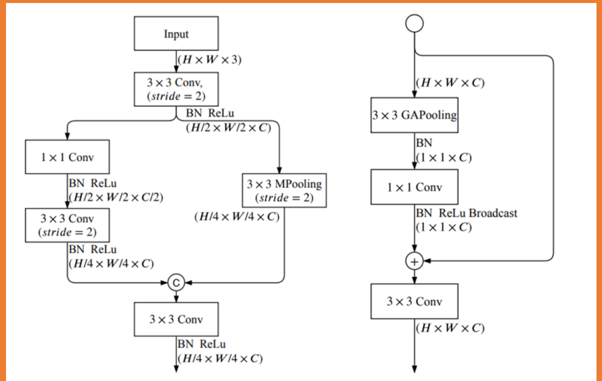


(c) Feature Fusion Module

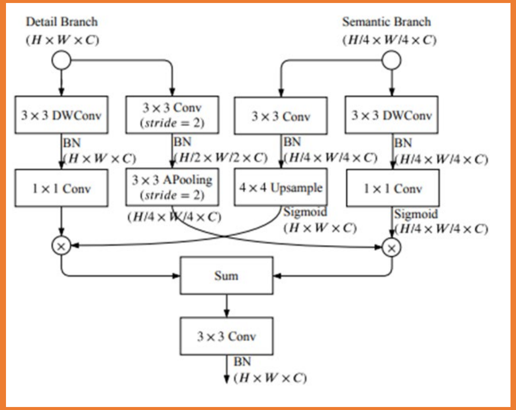
Models
BiSeNet V2
Bilateral Segmentation Network

- Two-pathway architecture:
  - **Detail Branch** to capture the spatial details with wide channels and shallow layers.
  - **Semantic Branch** to extract the categorical semantics with narrow channels and deep layers.
- **Guided Aggregation layer**: merges both types of feature representation.
- **Booster** (auxiliar prediction heads) to improve segmentation without increasing computation cost.





(a) Stem Block (b) Context Embedding Block

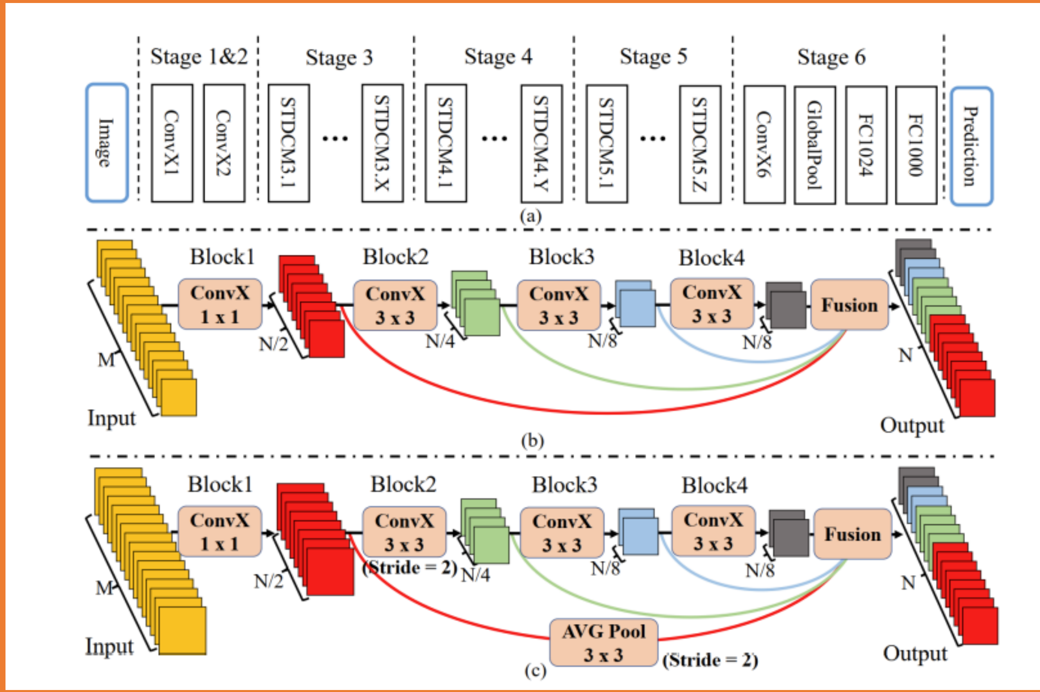


Models

STDC

Short-Term Dense Concatenate

- Dense Concatenate Module to extract deep feature with scalable receptive field and multiscale information.
- Replaces the extra path of **BiSeNet V1** with **Detail Aggregation Module**
- Guide the low-level layers for the learning of spatial details by generating detail ground truth.
- More precise preservation of spatial details in low-level layers without the extra computation cost during the inference time.



STDC module	Block1	Block2	Block3	Block4	Fusion
RF(S = 1)	1×1	3×3	5×5	7×7	$1 \times 1, 3 \times 3$ $5 \times 5, 7 \times 7$
RF(S = 2)	1×1	3×3	7×7	11×11	3×3 $7 \times 7, 11 \times 11$

Models

STDC

Short-Term Dense Concatenate

- **Stage 1 and 2:**
  - They are regarded as low-level layers. For sake of efficiency there is only one convolutional bloc in each stage.
- **Stages 3,4,5:**
  - They are used to produce the feature maps with down sample of 1/8, 1/16, 1/32, respectively. Adopted from the context path from **BiSeNet** to encode the context information using **pretrained network** as backbone of the encoder.
  - **Global average pooling** to provide global context information with large receptive field.
  - **U-shape structure** to up-sample the features stem from the global feature and combine each of them with the counter part from the last two stages 4 and 5 during the encoding phase.
  - **Feature Refinement Module** is used to fuse the output from stage 3 in the encoder and the counterpart from the decoder.
- **Stage 6:**
  - The stage 6 outputs the prediction maps by one convolution layer and one global average pooling and two fully connected convolutional layers.

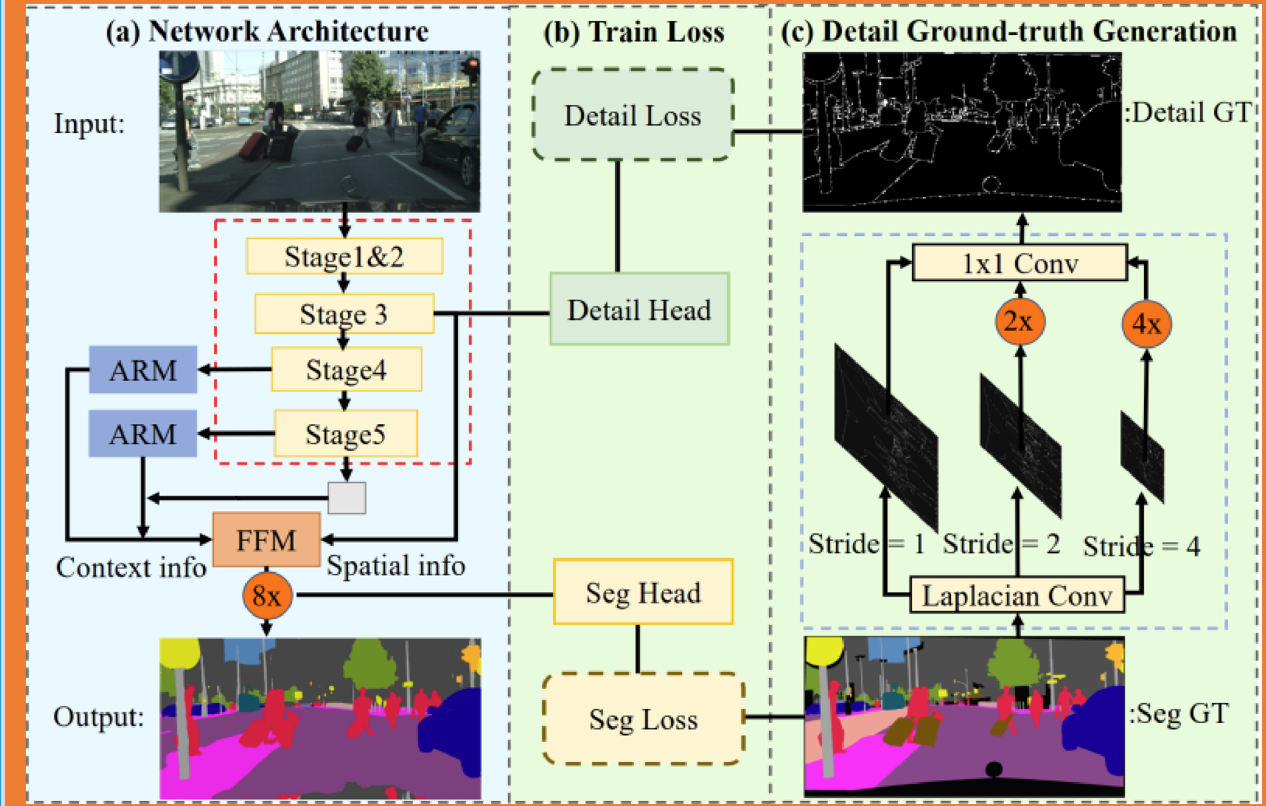
Stages	Output size	KSize	S	STDC1		STDC2	
				R	C	R	C
Image	224×224				3		3
ConvX1	112×112	3×3	2	1	32	1	32
ConvX2	56×56	3×3	2	1	64	1	64
Stage3	28×28 28×28		2 1	1 1	256	1 3	256
Stage4	14×14 14×14		2 1	1 1	512	1 4	512
Stage5	7×7 7×7		2 1	1 1	1024	1 2	1024
ConvX6	7×7	1×1	1	1	1024	1	1024
GlobalPool	1×1	7×7					
FC1					1024		1024
FC2					1000		1000
FLOPs				813M		1446M	
Params				8.44M		12.47M	

Models

STDC

Short-Term Dense Concatenate

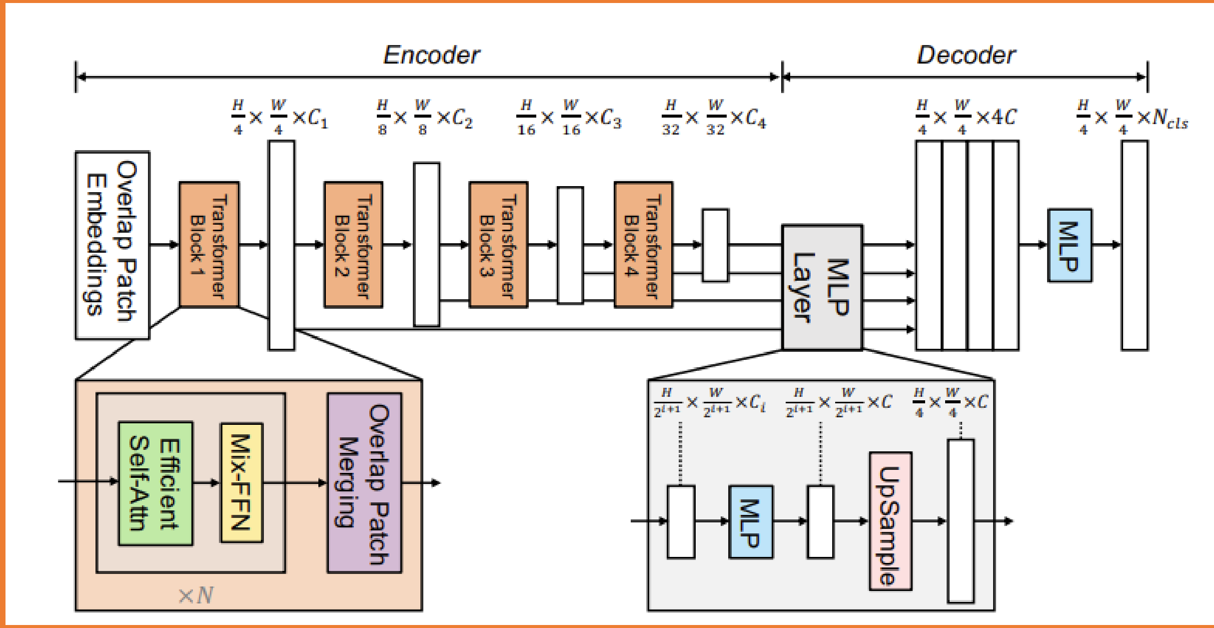
- **Detail Guidance of low-level Feature:**
  - **Detail Ground-truth Generation.**
  - **Detail head** inserted in stage 3 Guide the low-level layers to learn features of spatial details.
- **Detail Ground-truth Generation:**
  - Generates the **detail feature map** from the semantic segmentation ground-truth.
  - This is carried out by 2-D conv Laplacian kernel and trainable 1x1 conv.
  - The Laplacian operator is used to produce soft, thin detail feature maps with different strides to obtain multi-scale details information.
  - Up-samples the details feature maps to the original size and fuse it with a trainable 1 x 1 conv for dynamic re-weighting.
  - Adopts a threshold 0.1 to convert the predicted details to the final binary detail ground-truth with boundary and corner information.
- **Detail Head:**
  - Produces the detail map to guide the shallow layer to encode spatial information.
  - Includes a 3 x 3 conv-BN-ReLU operator followed with a 1 x 1 convolution to get the output detail map.



Models

SegFormer

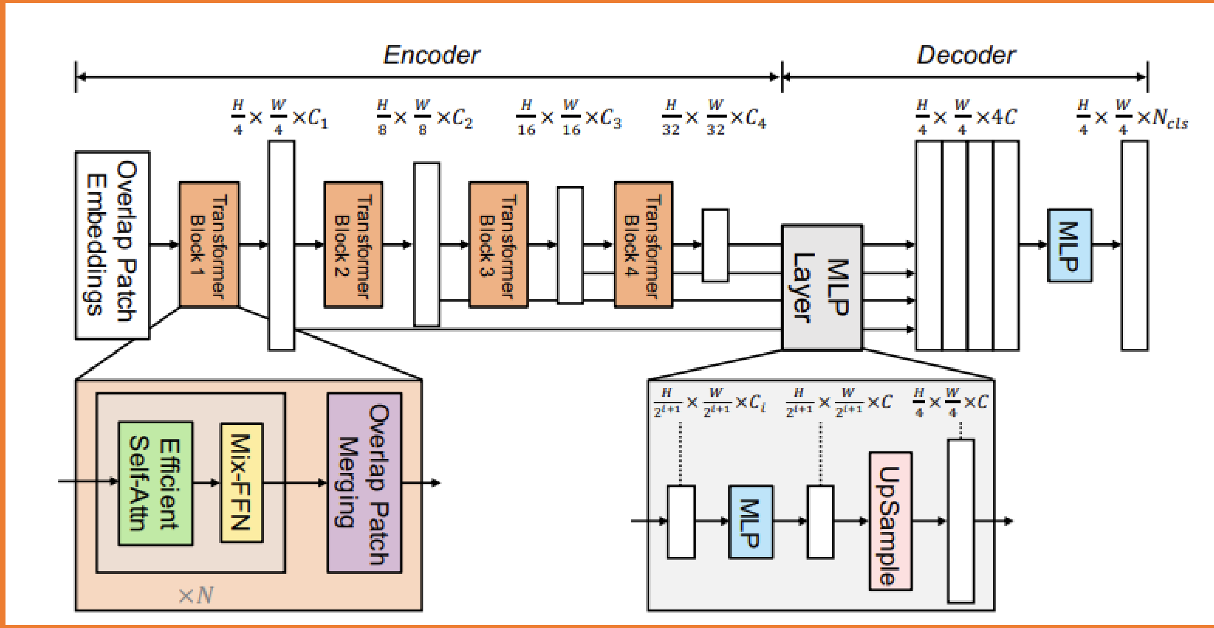
- Encoder-decoder model:
  - The encoder par is a hierarchical Transformer module to extract coarse and fine features.
  - The decoder part is a Lightweight MLP model.
- This architecture divides the image into patches as input to the hierarchical Transformer encoder.
- The extracted multi-level features are passed to the MLP decoder to predict the segmentation mask.



Models

SegFormer

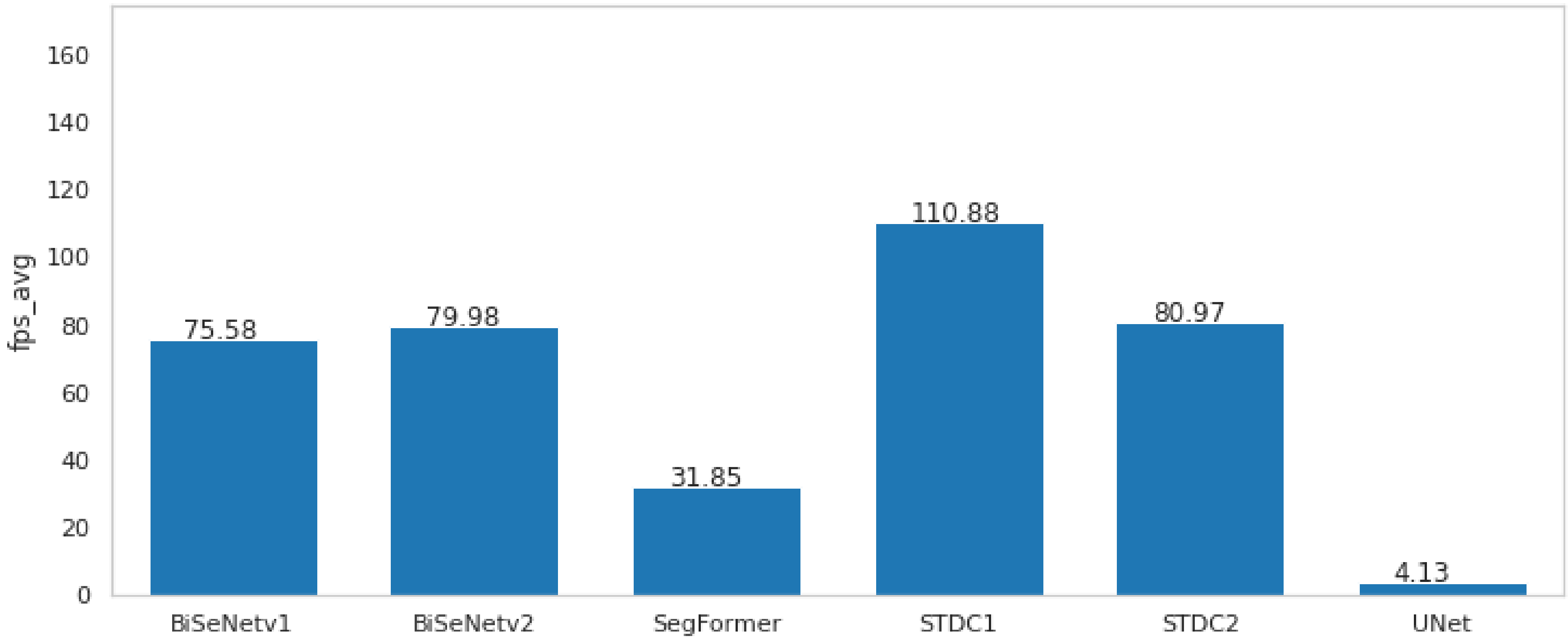
- Overlap Patch Merging
  - Shrinks the hierarchical features from Ci to Ci+1 with stride = 2.
- Mix-FFN
  - By using a 3x3 conv in feed forward network FFN allow to use different test resolution then the training one. And it sufficient to provide positional information for Transformer.
- Efficient Self-Attention module
  - Instead of using the original multi-head self-attention process it uses the sequence reduction process to reduce the length of the sequence thus, reduces it cost from O(N^2) to O($\frac{N^2}{R}$) where R is reduction ratio



Experiments & results

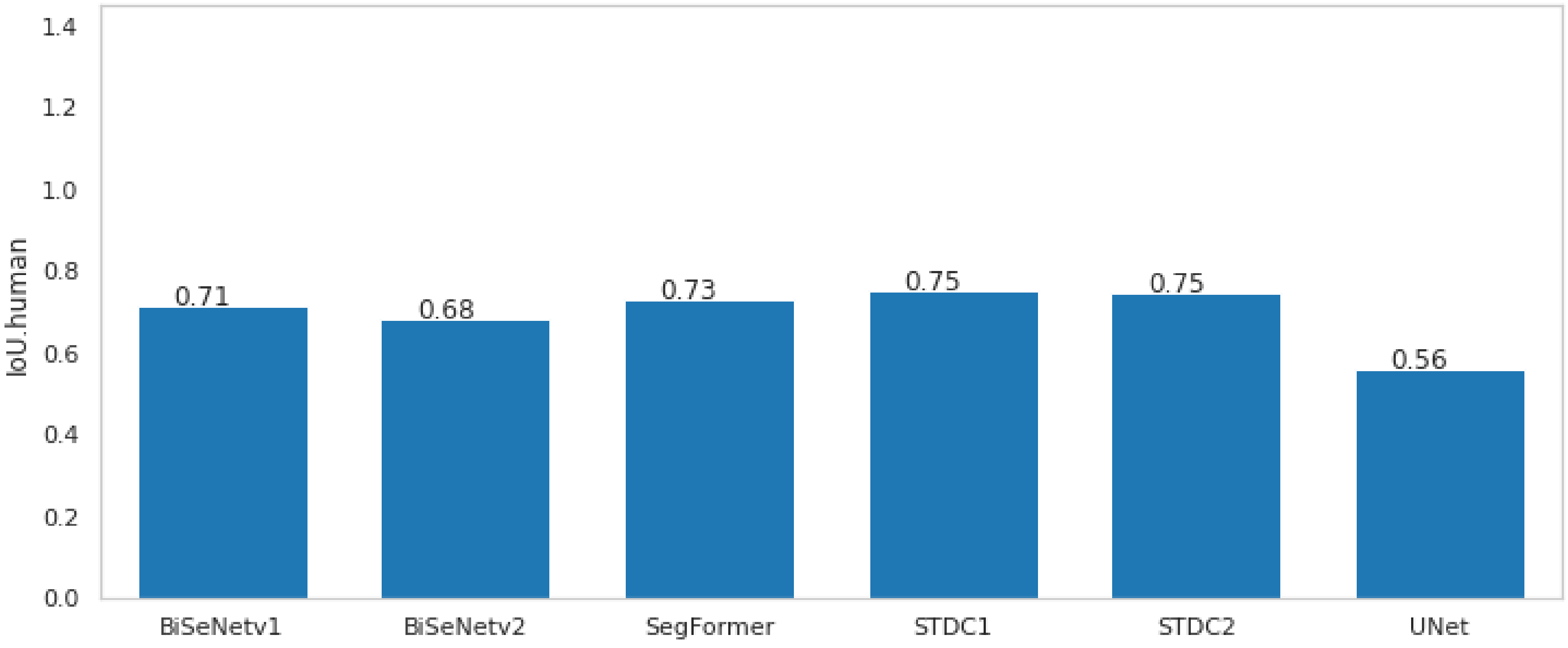
Results

Speed experiments



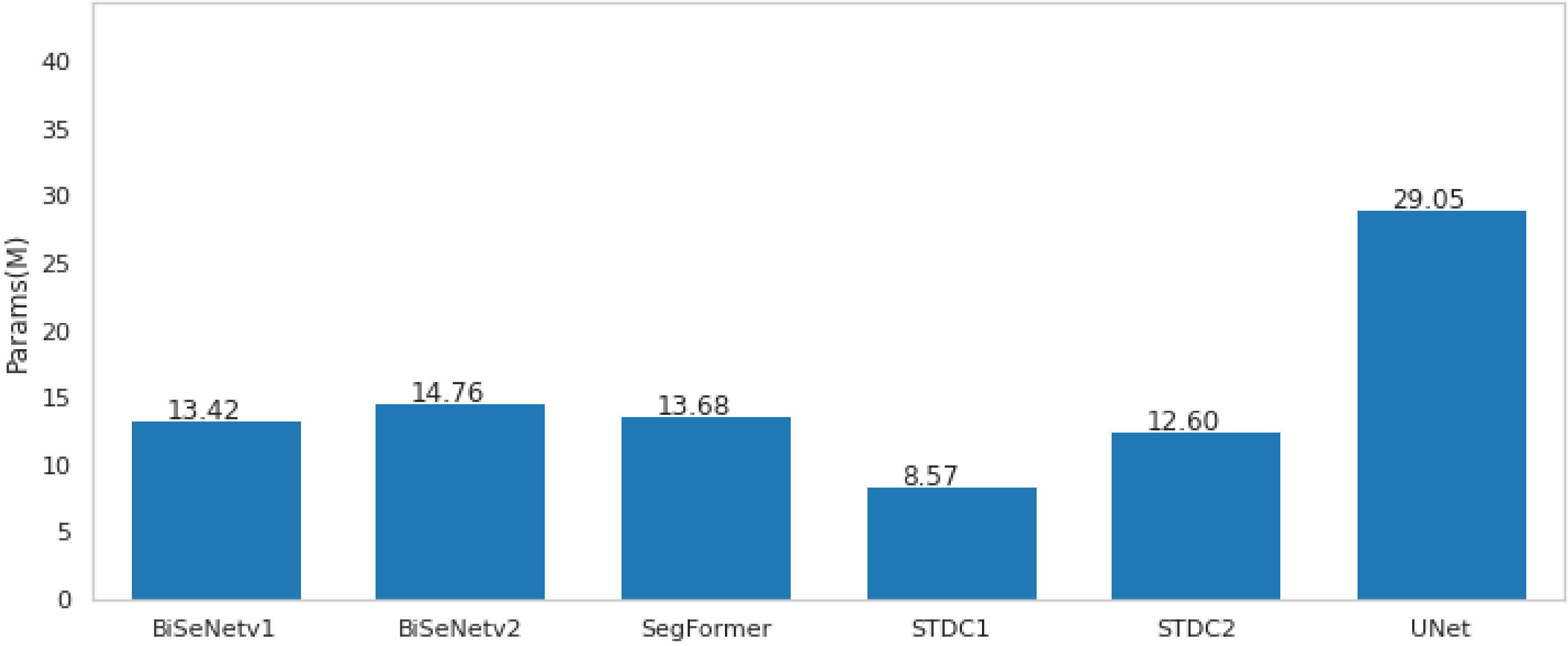
Results

Accuracy experiments



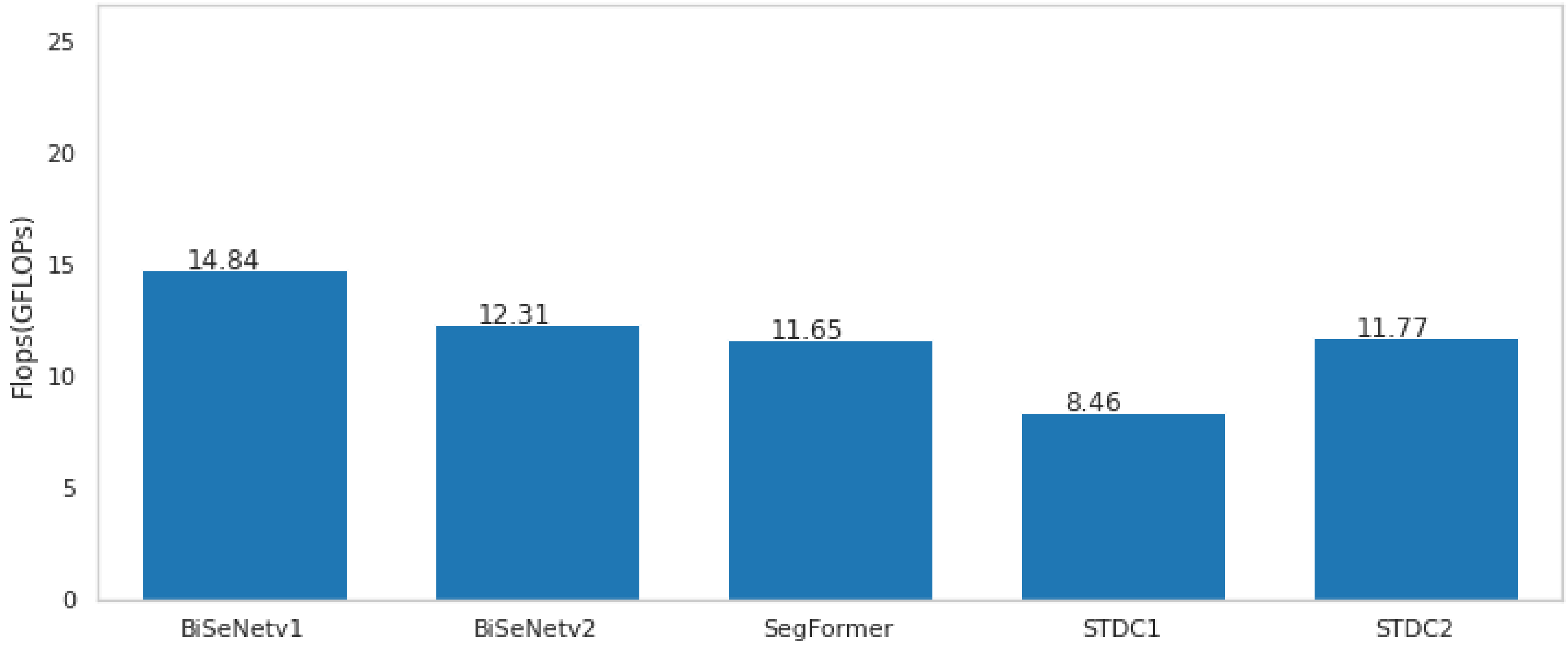
Results

Complexity experiments



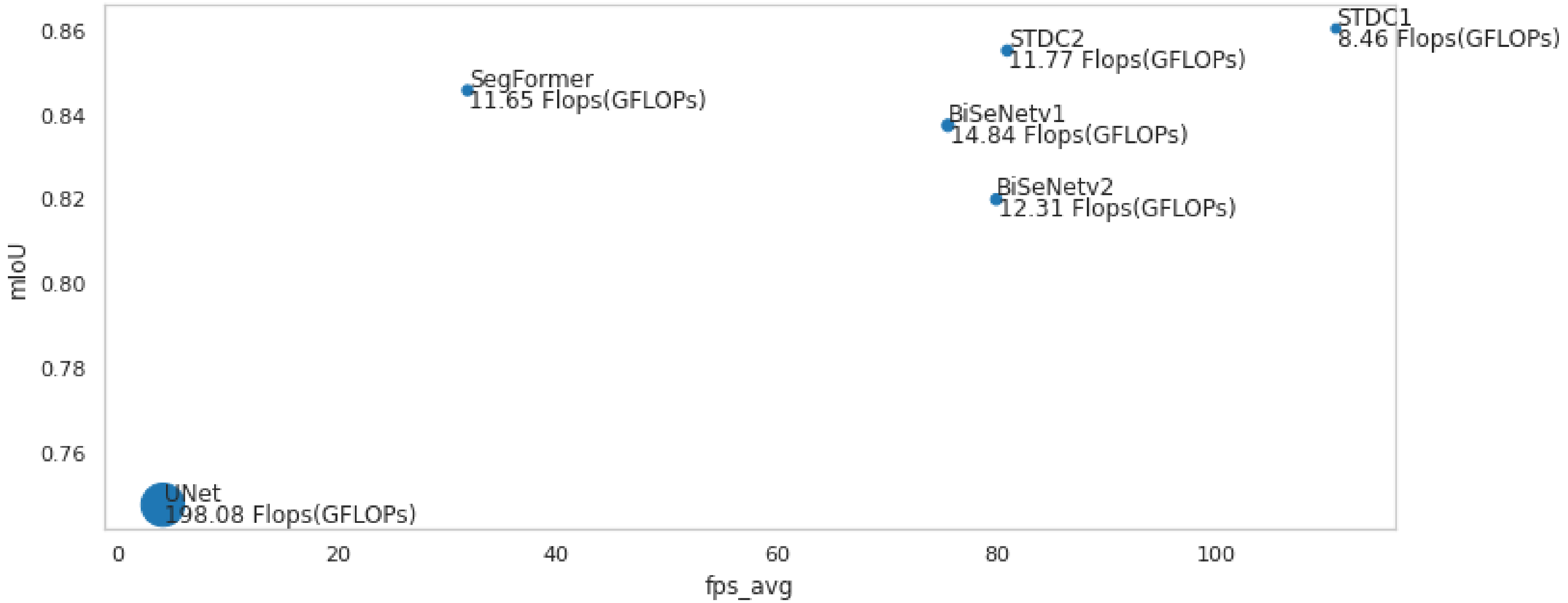
Results

Complexity experiments



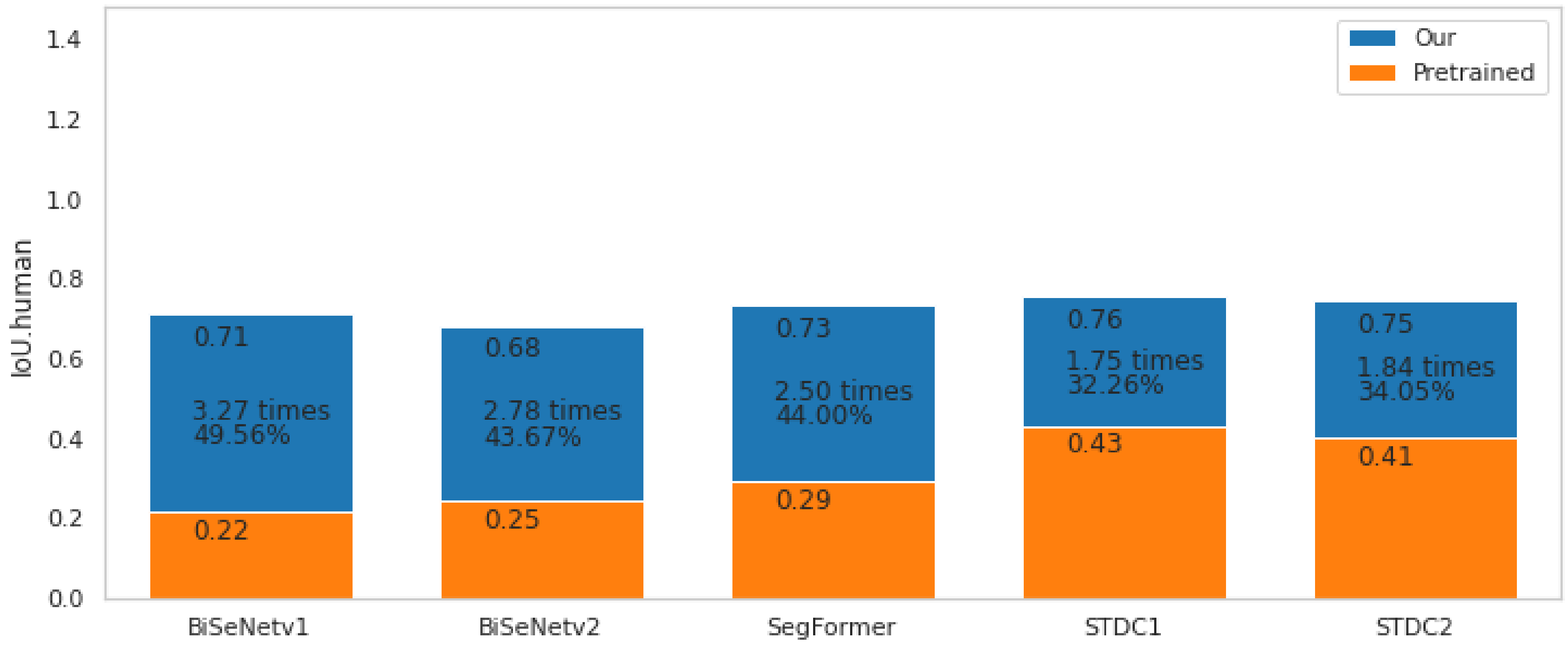
Results

Overall



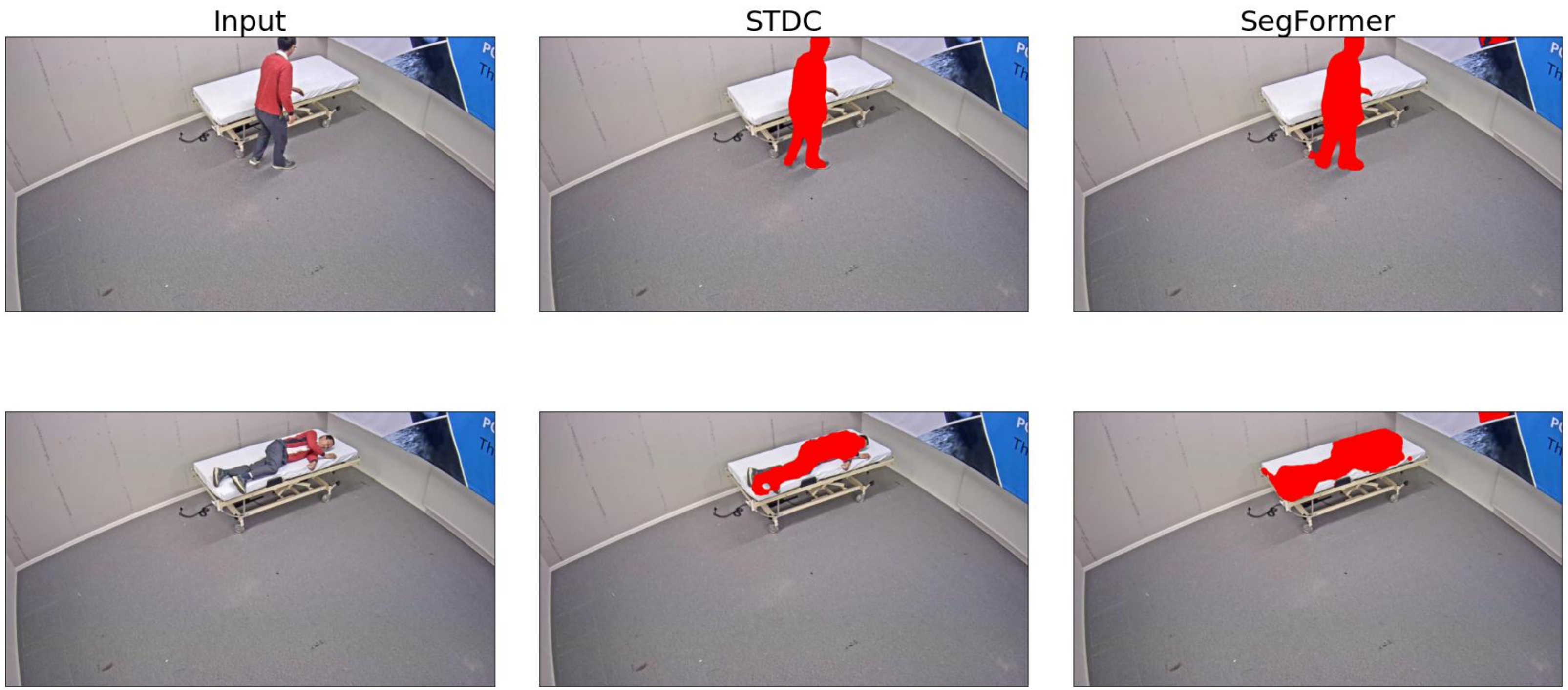
Results

Comparison



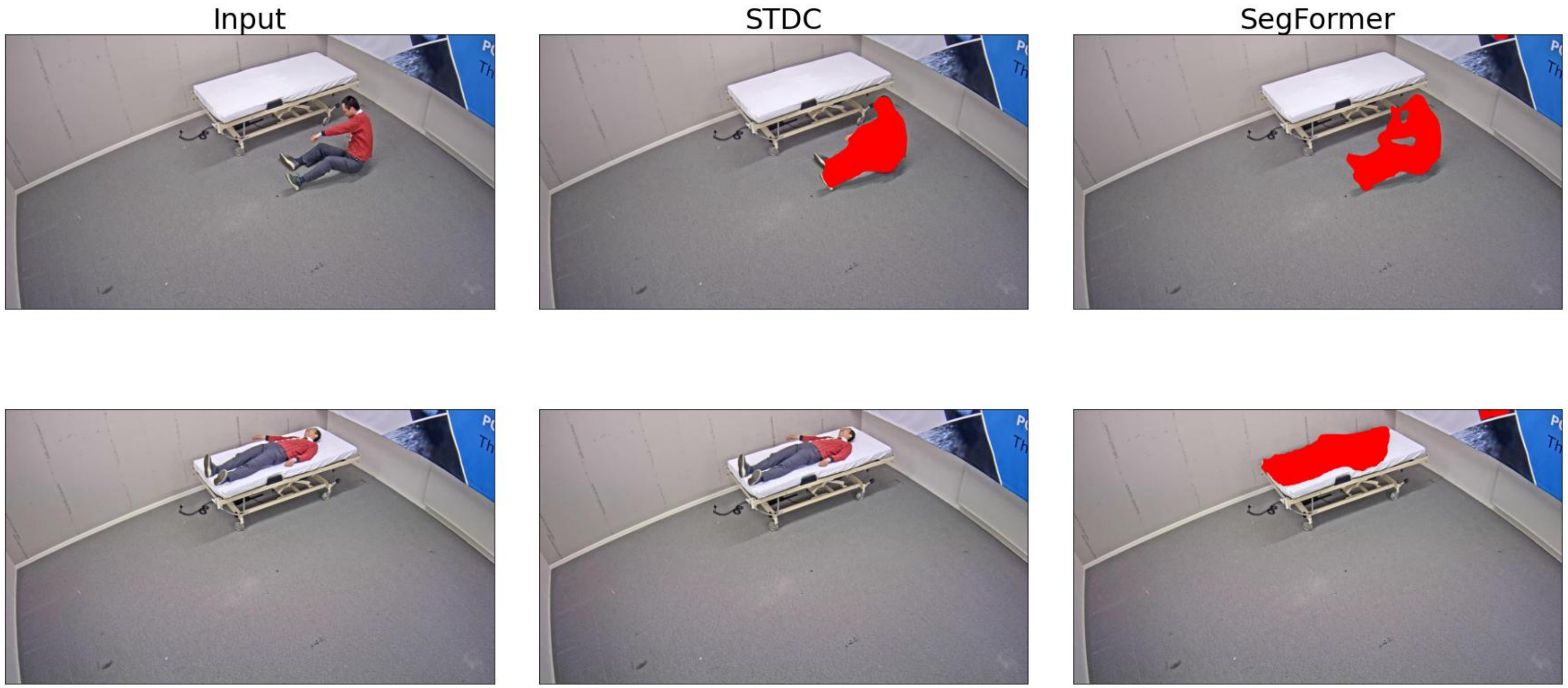
Privacy preservation

Privacy preservation
Result



Privacy preservation

Result



Privacy preservation

Demo

Future work

Future work

Induced ideas

1. We are working on more inclusive dataset.
2. Understanding the definition of surveillance-like images.
3. In health-care context such as segmenting humans in laying or sitting positions.

Conclusions

Thank you