UrbanPipe Challenge on Fine-grained Video Anomaly Recognition
Tech Report

Jiawei Dong, Bo Zhang, Zongjie Yu, Chen Hu, Shuo Wang

Shanghai Paidao Intelligent Technology Co., Ltd.

1. Data Description

Dataset examples

Classes	1 normal class 16 defect classes
Classification Type	Multi-Label
Video numbers	9609
Instance Labels Num Range	1 – 5 labels
Average Labels Num	1.4 labels
Total Duration	55H
Average Duration	20.7s
Duration Range	0.7s – 385.2s

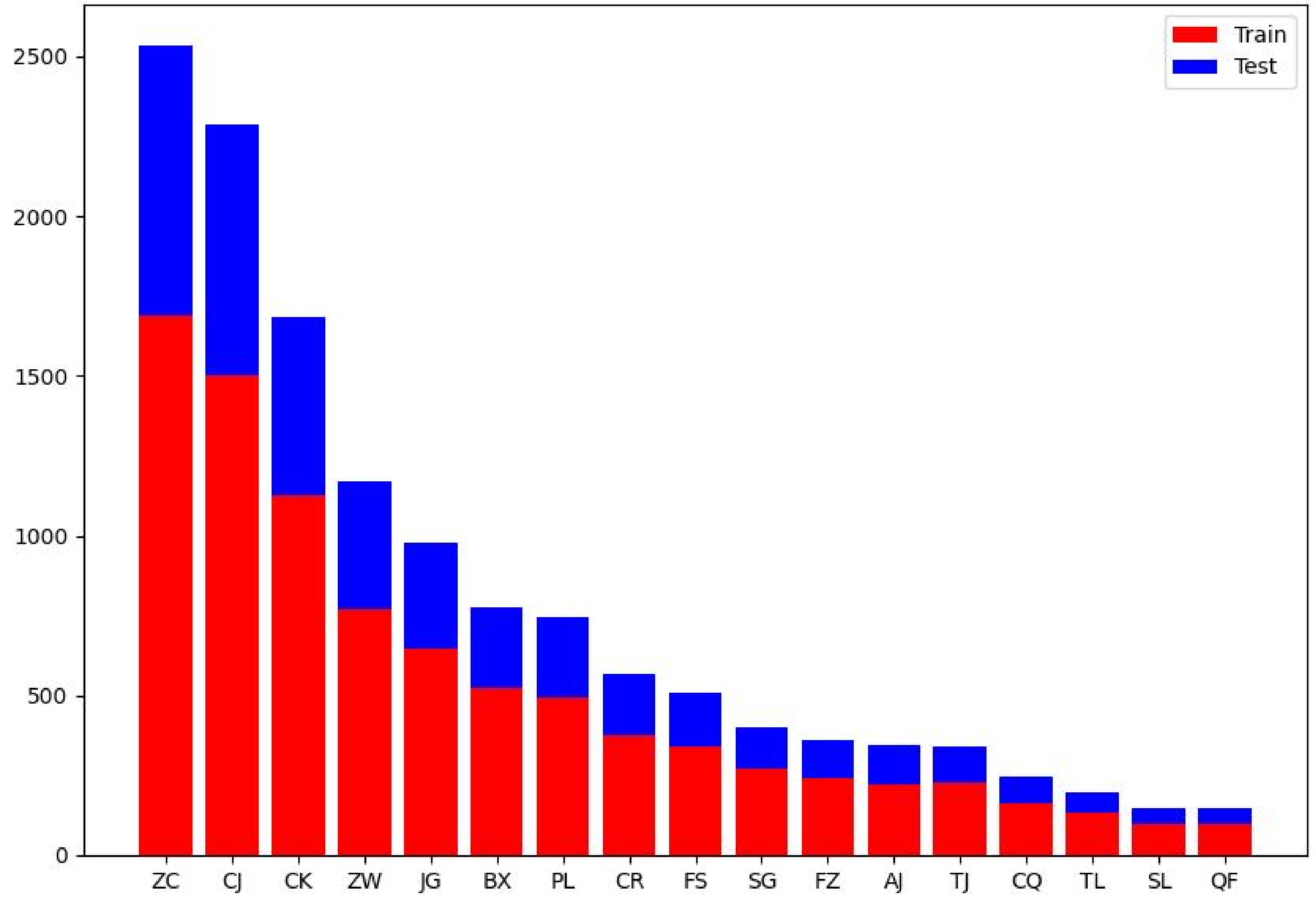
Dataset statistics

Pipe situation is complex, multiple defects often appear at the same time, so each video is annotated by multiple labels. To obtain accurate annotations of defect instances, professional engineers are asked to check all the videos multiple rounds with cross validation. Given a QV video, our goal is to predict multiple labels of pipe defects in this video.

Deeper Action

1. Data Description

Data Splits and Distribution



Dataset distribution

The 9.6k videos are divided into train set and test set according to the ratio of 2:1. As shown in Figure, the data exhibits the natural **long-tailed distribution**.

Video Anomaly Detection Benchmark Comparison

Dataset	Multi-Labeled	Class	Video	Duration	Anomaly Type
UCSD Ped1 [1]	x	2	70	5 mins	Human Action
UCSD Ped2 [1]	x	2	28	5 mins	Human Action
Subway Entrance [2]	x	2	1	1.5 hours	Human Action
Subway Exit [2]	x	2	1	1.5 hours	Human Action
Avenue [3]	x	2	37	30 mins	Human Action
UMN [4]	x	2	5	5 mins	Human Action
RealWorld [5]	x	13	1,900	128 hours	Human Action
UrbanPipe	✓	**17**	**9,609**	**55 hours**	**Pipe Defect**

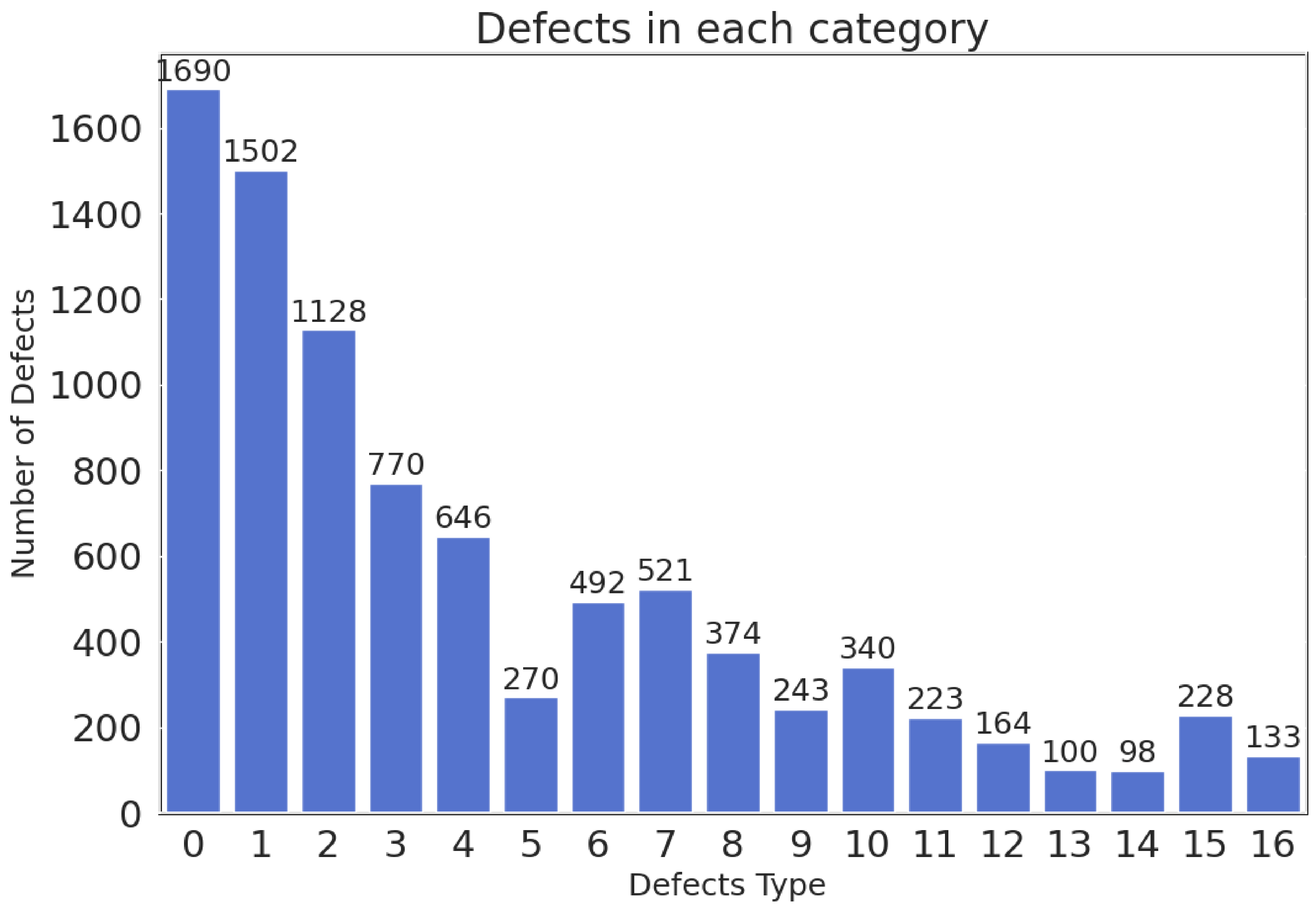
Dataset comparison

1. UrbanPipe is large scale.
2. UrbanPipe contains multiple anomaly categories, and these categories are fine-grained.
3. The previous datasets mainly works on human. Alternatively, the **domain shift is large** for urban pipe inspection.

2. Data Preprocess

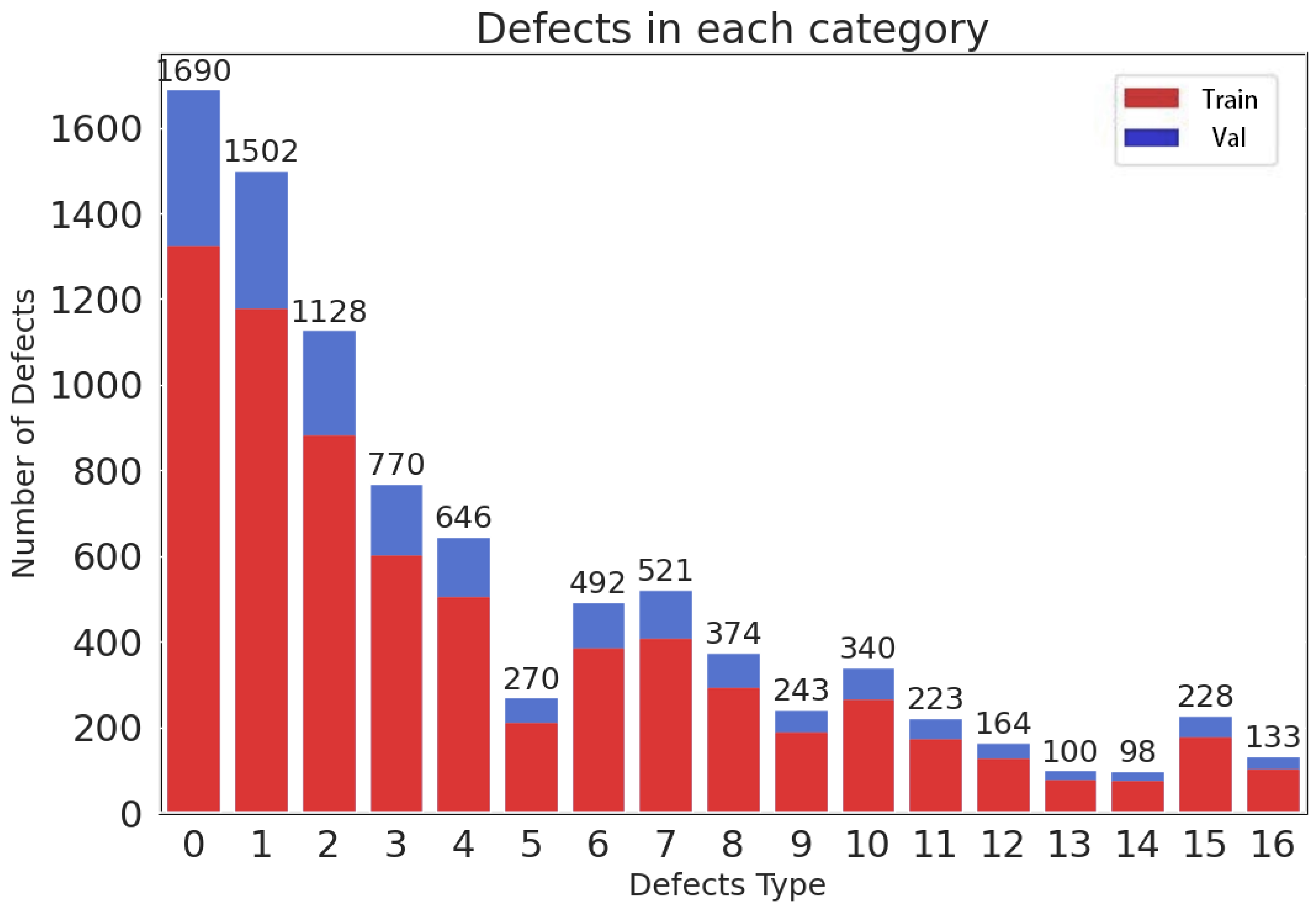
5 Folders Split —— Iterative stratification for multi-label data

We use **IterativeStratification** from **skmultilearn.**
The idea behind this stratification method is to assign label combinations to folds based on how much a given combination is desired by a given fold, as more and more assignments are made, some folds are filled and positive evidence is directed into other folds, in the end negative evidence is distributed based on a folds desirability of size.



Categories Distribution in 6202 Training Videos

Iterative Stratification



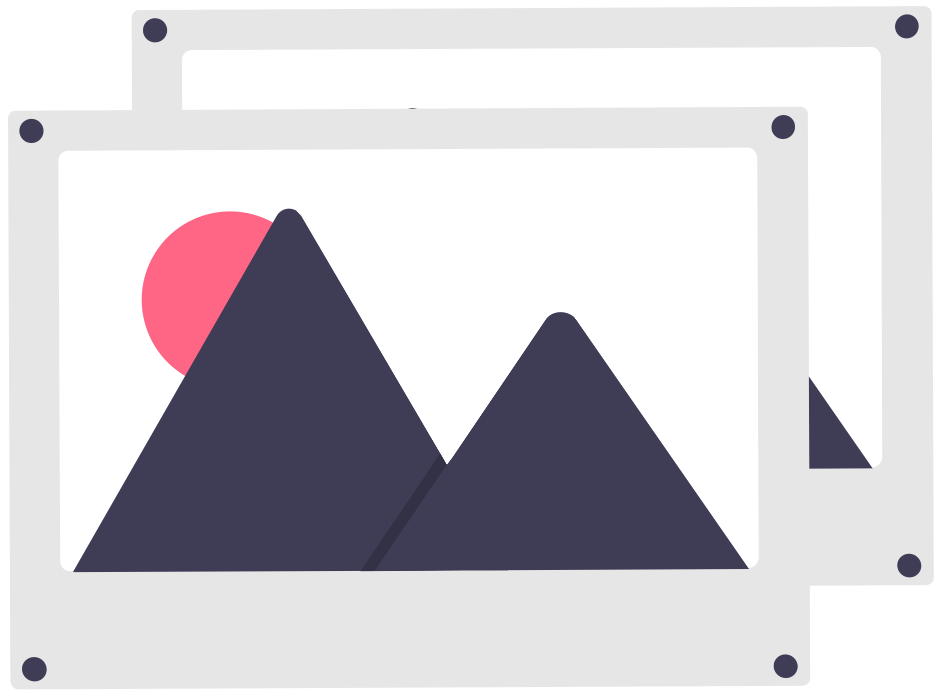
Folder 0 Categories Distribution

3. Method

Task Definition

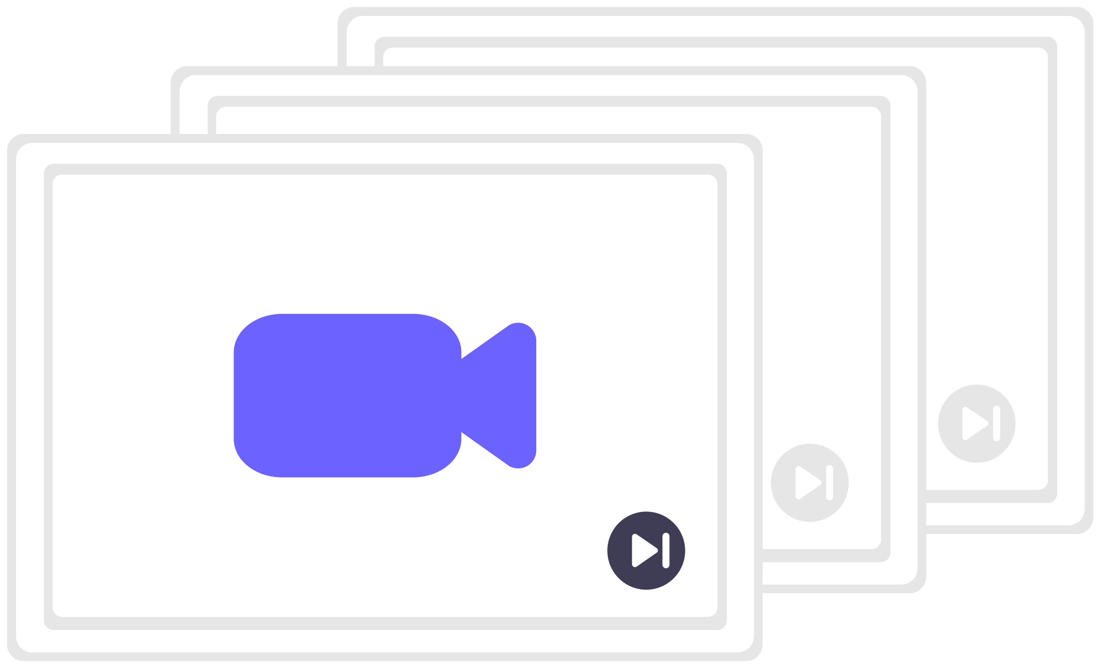
01 Frame-Based Task

Multi-label video classification using frame-based predictions based on an image classification network.



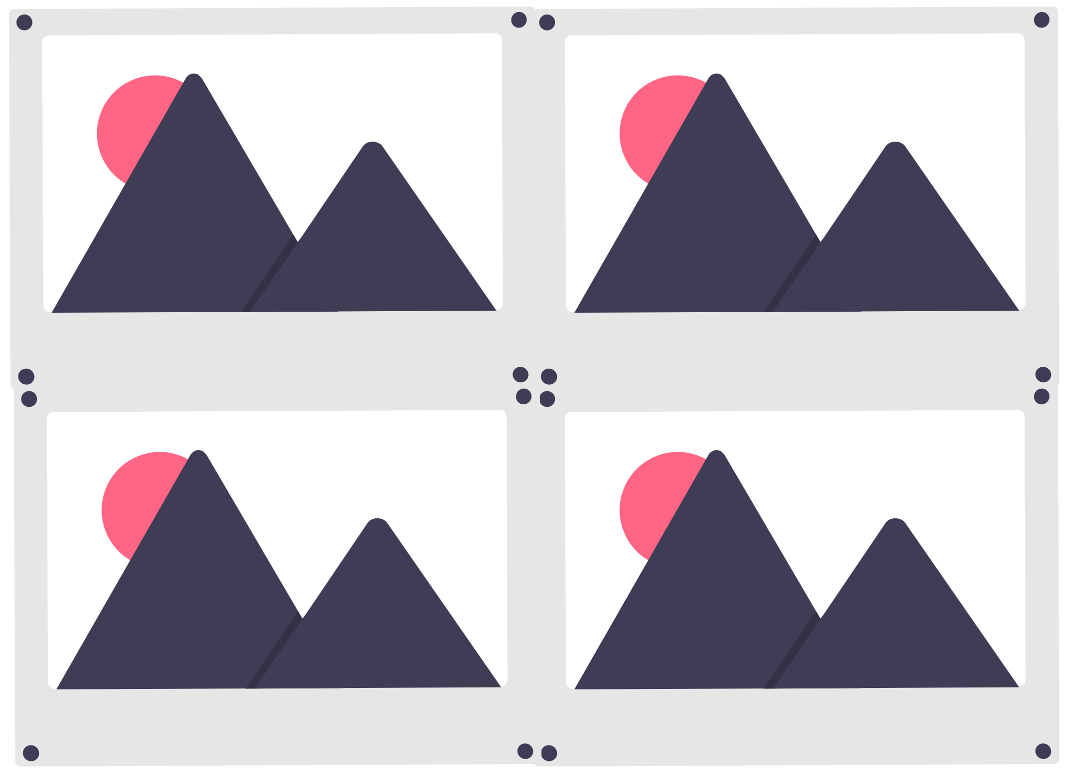
02 Video-Based Task

Multi-label video classification directly based on video classification network.



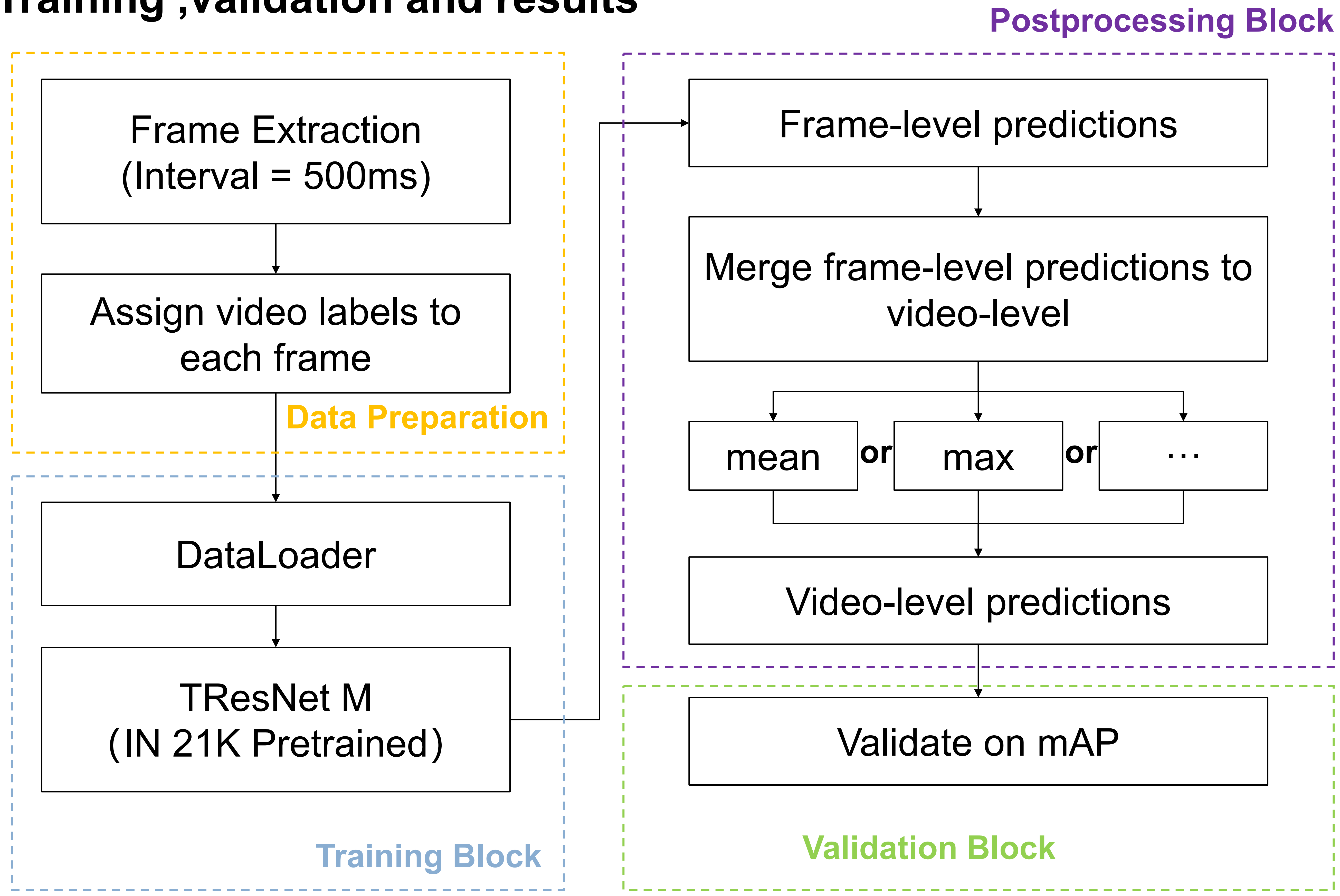
03 Super-image Task

Multi-label Video classification based on super-image method, using image classification network.



• Frame-Based Method

Training ,validation and results



Training and Validation Flowchart

- First, assign the video labels to the frame images.
- Second, TResNet image classification network is used for training.
- Third, collecting frame-level predictions, for single video, average(or maximum and median) the predictions of all frames , and output it as the predictions of this video.

Using this simple method, we achieved a validation score of 55.2%, which is a good start.

Model	Params	Lr Schd	Pred Merge Method	Val mAP (%)
TResNet M	41M	30ep	Mean	55.20
			Max	48.43

Results on Leaderboard

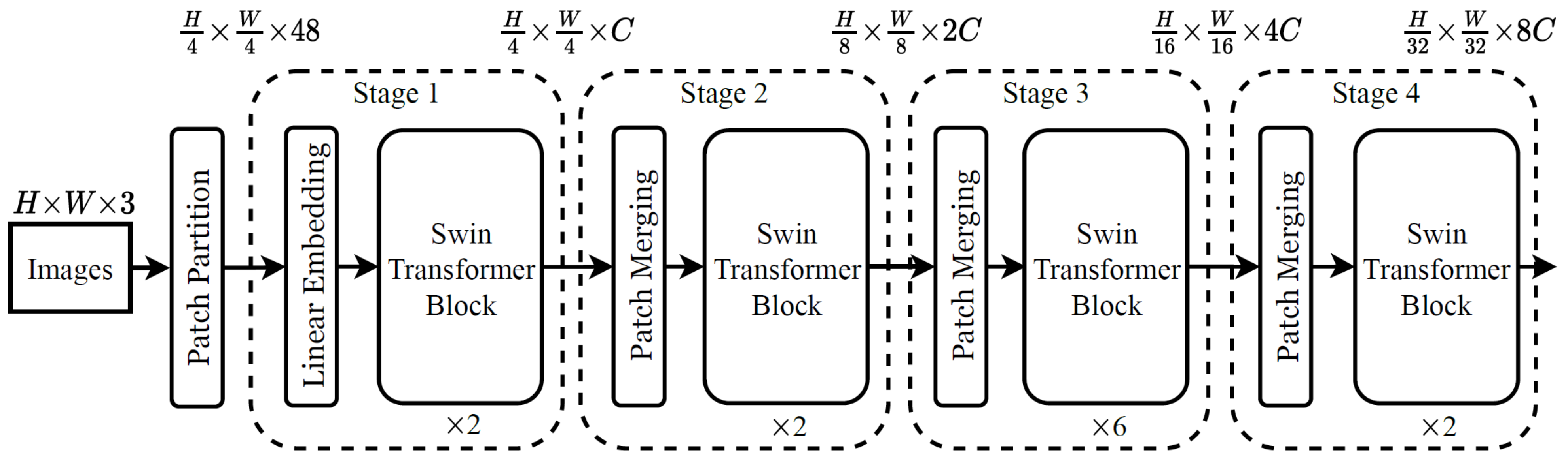
• Video-Based Method

Video Swin Transformer Paper: Video Swin Transformer

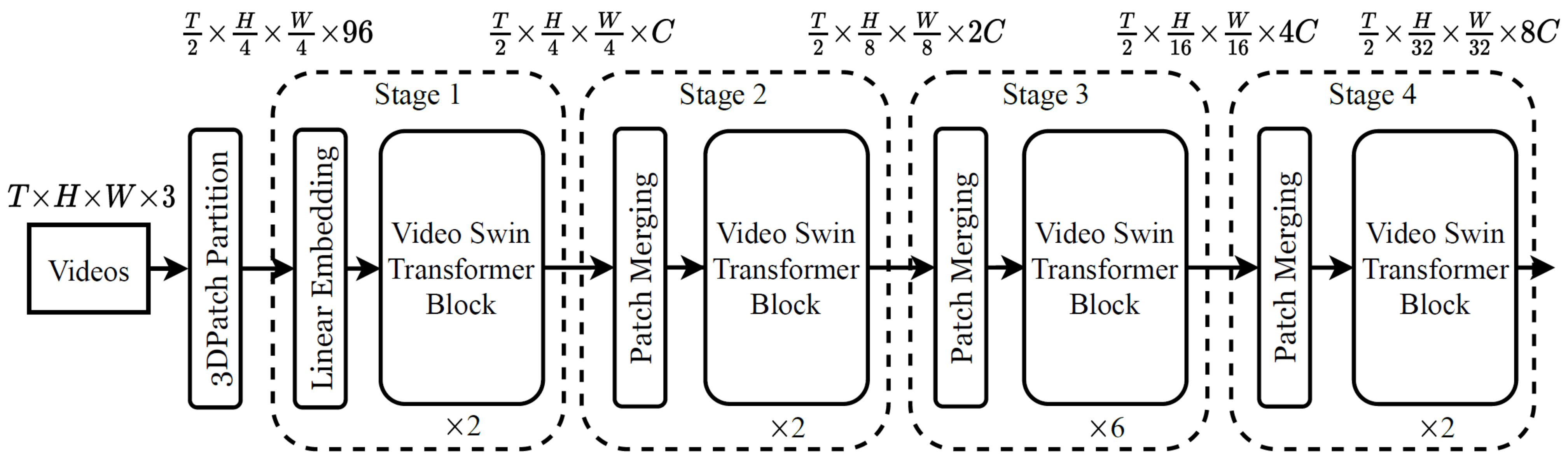
A pure-transformer architecture for video recognition that is based on spatiotemporal locality inductive bias. This model is adapted from the Swin Transformer for image recognition, and thus it could leverage the power of the strong pre-trained image models. The proposed approach achieves state-of-the-art performance on three widely-used benchmarks, Kinetics-400, Kinetics-600 and Something-Something v2.

Method	Pretrain	Top-1	Top-5	Views	FLOPs	Param
R(2+1)D [37]	-	72.0	90.0	10 × 1	75	61.8
I3D [6]	ImageNet-1K	72.1	90.3	-	108	25.0
NL I3D-101 [40]	ImageNet-1K	77.7	93.3	10 × 3	359	61.8
ip-CSN-152 [36]	-	77.8	92.8	10 × 3	109	32.8
CorrNet-101 [39]	-	79.2	-	10 × 3	224	-
SlowFast R101+NL [13]	-	79.8	93.9	10 × 3	234	59.9
X3D-XXL [12]	-	80.4	94.6	10 × 3	144	20.3
MViT-B, 32×3 [10]	-	80.2	94.4	1 × 5	170	36.6
MViT-B, 64×3 [10]	-	81.2	95.1	3 × 3	455	36.6
TimeSformer-L [3]	ImageNet-21K	80.7	94.7	1 × 3	2380	121.4
ViT-B-VTN [29]	ImageNet-21K	78.6	93.7	1 × 1	4218	11.04
ViViT-L/16x2 [1]	ImageNet-21K	80.6	94.7	4 × 3	1446	310.8
ViViT-L/16x2 320 [1]	ImageNet-21K	81.3	94.7	4 × 3	3992	310.8
ip-CSN-152 [36]	IG-65M	82.5	95.3	10 × 3	109	32.8
ViViT-L/16x2 [1]	JFT-300M	82.8	95.5	4 × 3	1446	310.8
ViViT-L/16x2 320 [1]	JFT-300M	83.5	95.5	4 × 3	3992	310.8
ViViT-H/16x2 [1]	JFT-300M	84.8	95.8	4 × 3	8316	647.5
Swin-T	ImageNet-1K	78.8	93.6	4 × 3	88	28.2
Swin-S	ImageNet-1K	80.6	94.5	4 × 3	166	49.8
Swin-B	ImageNet-1K	80.6	94.6	4 × 3	282	88.1
Swin-B	ImageNet-21K	82.7	95.5	4 × 3	282	88.1
Swin-L	ImageNet-21K	83.1	95.9	4 × 3	604	197.0
Swin-L (384↑)	ImageNet-21K	84.6	96.5	4 × 3	2107	200.0
Swin-L (384↑)	ImageNet-21K	**84.9**	**96.7**	10 × 5	2107	200.0

Comparison to state-of-the-art on Kinetics-400



The architecture of a Swin Transformer (Swin-T)



The architecture of Video Swin Transformer (tiny version, referred to as Swin-T)

Video Swin Transformer: https://arxiv.org/abs/2106.13230

• Video-Based Method

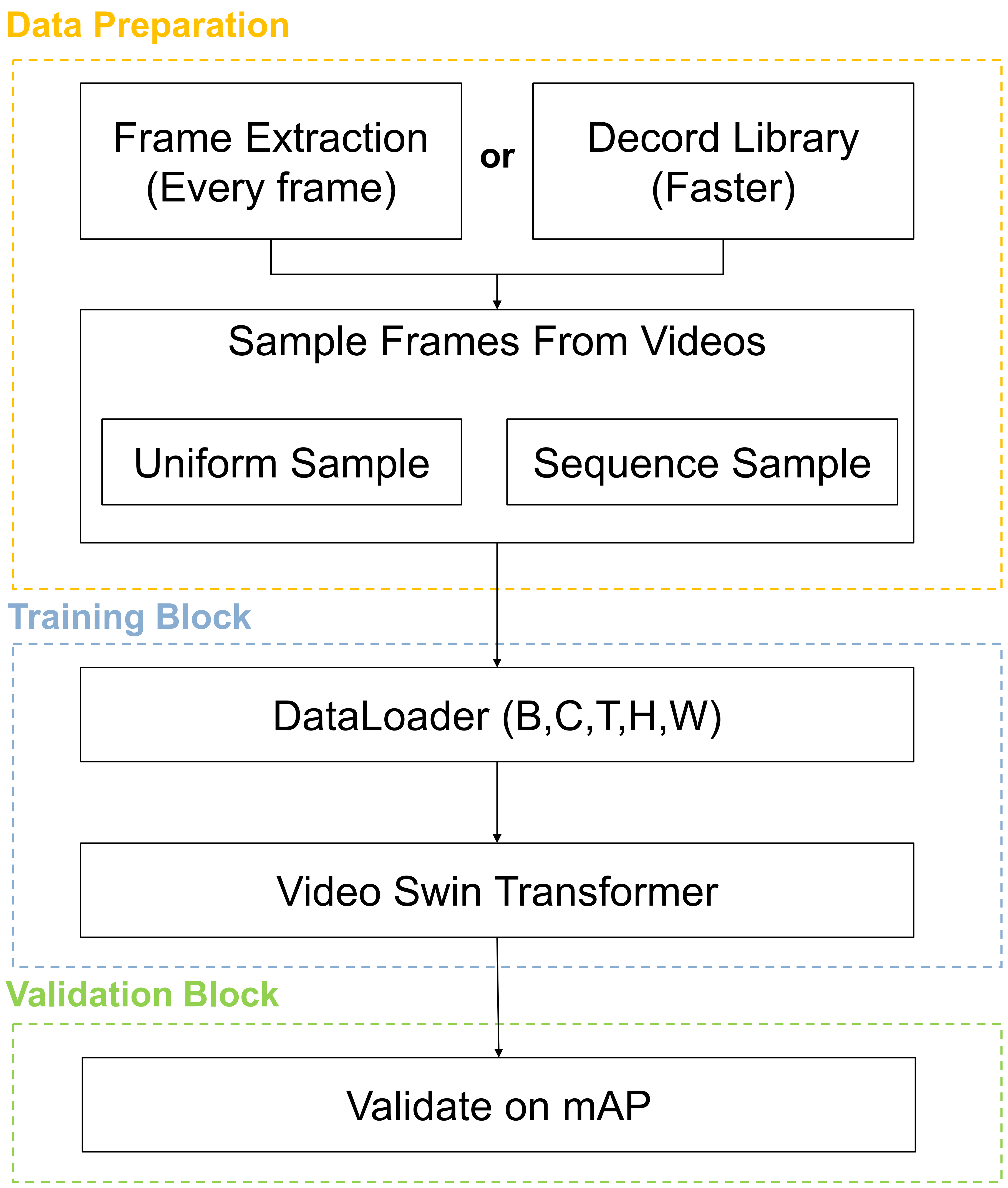
Training, validation and results

Model	Backbone	Params	Lr Schd	Pretrain	Val mAP (%)
Video Swin Transformer	Swin-B	88M	30ep	Kinectic 600	64.512
				Kinectic 400	64.798
				Something-Something V2	64.714

Results on Leaderboard

Using video classification network based on the Video Swin Transformer, and using different backbone for training, the mAP score on the Leaderboard reached 64.79%. Compared with the method based on single-frame prediction, video-based method boost score by nearly 10%.

- **Temporal features may not be critical**. Even if the order of all frames is disrupted, the trained model even had a slight improvement compared to regular trained model. Therefore, we infer that what is relatively important in this task is the ability to extract spatio features.
- **The backbone for spatio feature extracting lacks flexibility**. There are fewer pre-training weights to choose from, which makes it difficult to improve the model capacity by ensembling multiple structure networks.

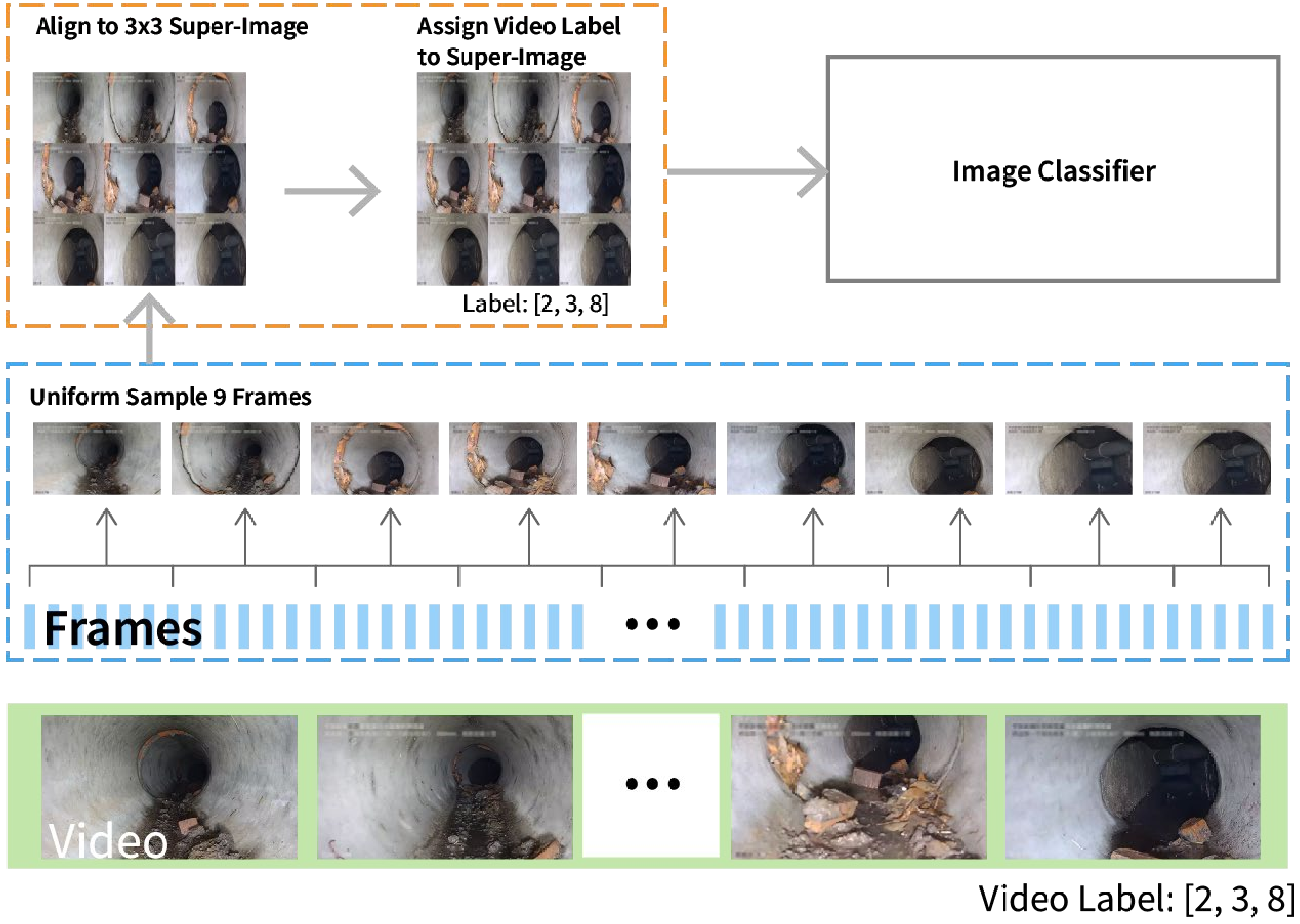


Training and Validation Flowchart

• Super-Image Method

Following our previous observation, we abandoned the temporal part of the action classification network and attempted to turn the problem into a pure image classification problem. In this way, not only can the model be trained more efficiently, but in terms of model capacity improvement, more different structured models and more pre-trained models on different datasets can be chosen, which is highly flexible.

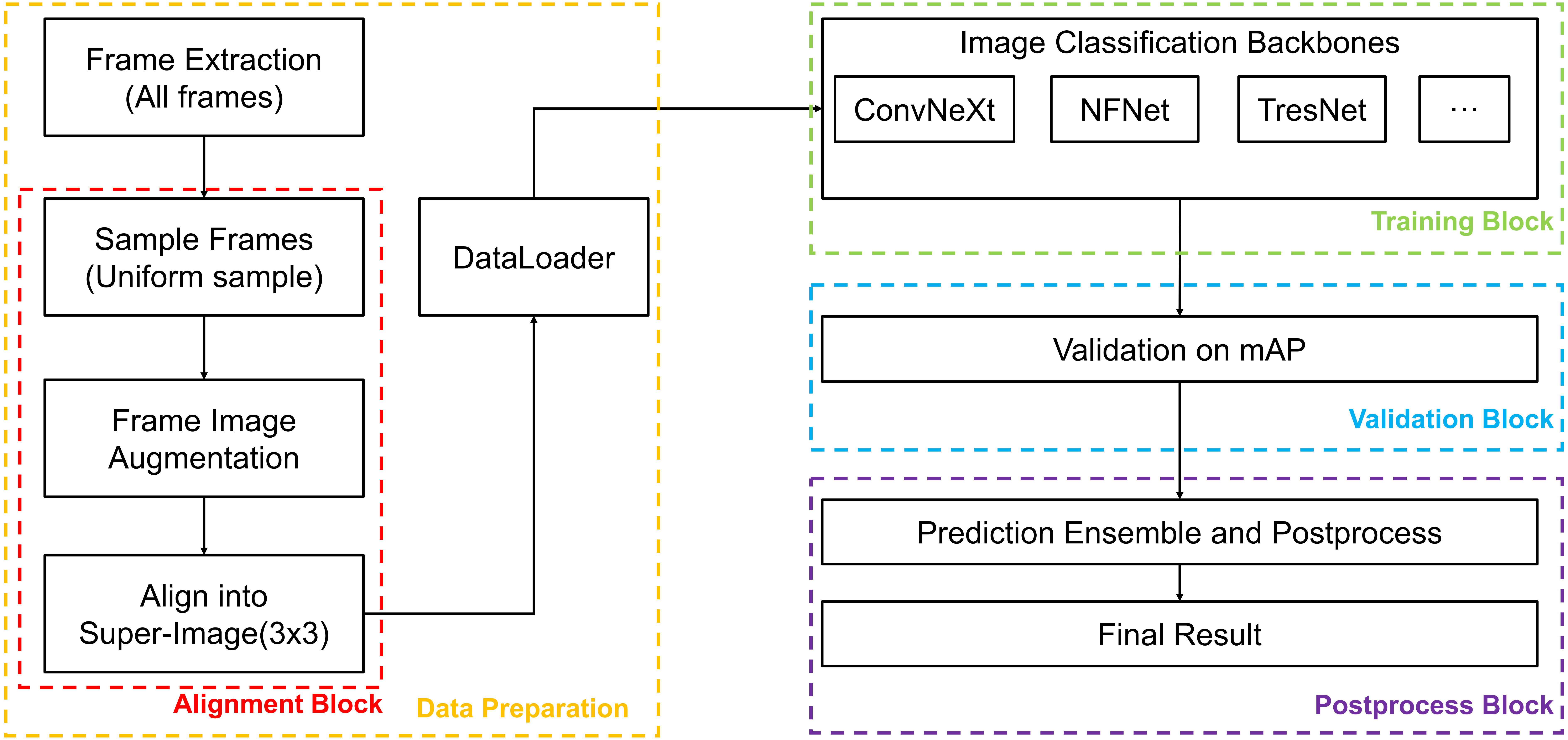
Inspired by the mosaic data augmentation, we wondered whether it is possible to convert the video into a grid image composed of frames through a similar processing method, here we define this grid image as super-image.



• Super-Image Method

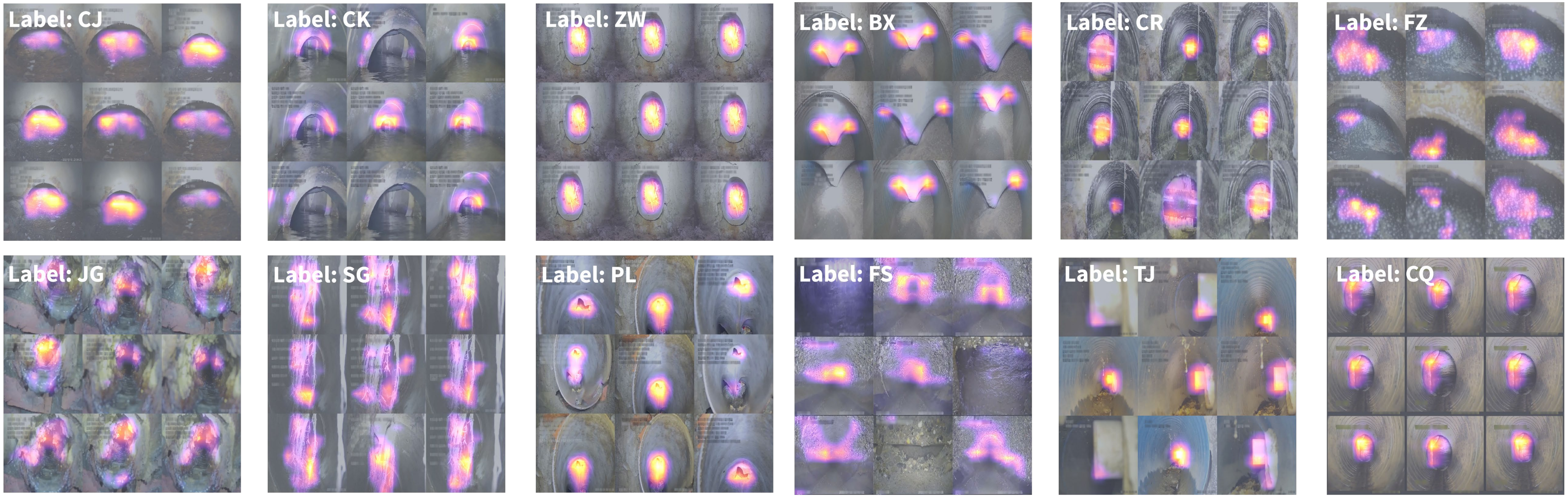
Training ,validation and postprocessing

Training and Validation Flowchart



• Super-Image Method

Visualizations



We use ConvNeXt-Base as the base network to extract the feature map generated by layer4 of the network for visualization. It can be seen from the visualization results that the network's response to the 16 types of defects is close to the real situation.

4. Experimental Result

Ablation Study - Model

Results on Leader Board

Model	Pretrain	Params	Input Size	Super-Img Grid	Data Aug	Optim	Lr Schd	Mean Val mAP(%)
Tresnet XL + MLDecoder	IN21K (Input Size 640)	78M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	67.19
ConvNeXt Base	IN22Kft1K (Input Size 384)	88M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	69.89
NFNET F3	ImageNet 1K (Input Size 416)	254M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	71.41
NFNET F6	ImageNet 1K (Input Size 576)	438M	1152 (384*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	70.45
ECA ResNet 269d	ImageNet 1K (Input Size 352)	102M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	70.69
Swin Transformer Large	IN22Kft1K (Input Size 384)	196M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	71.11
EfficientNet L2	ImageNet 1K (Input Size 800)	480M	1334 (448*3)	3x3	Horizonal Flip + Tiles Shuffle	AdamW	OneCycle 30e	70.95

We also tried many other backbones, such as ConvNeXt Large, Coat, EfficientNet v2, MaxVIT and so on, they were not included in the table due to poor model performance. And also, they will not be added to the ensemble.

ImageNet Dataset Difference

4. Experimental Result

Ablation Study - Ensemble

Models for ensemble

Model	Val mAP(%)	Ensemble Weight	Ensembled LB mAP(%) (post-processed)
Tresnet XL + MLDecoder	67.194	0.1	72.689
ConvNeXt Base	69.891	0.1	
NFNET F3	71.405	0.15	
NFNET F6	70.453	0	
ECA ResNet 269d	70.689	0.15	
Swin Transformer Large	71.106	0.2	
EfficientNet L2	70.853	0.2	
Video Swin Transformer	68.251	0.1	

General ensemble methods

Ensemble Method
Mean
Max
Median
Class-based
Mix folder
One best folder for each model

Finally, we use the previously trained model for ensemble. Here we use the weighted average ensemble method, as it has been the most stable and interpretable method. The post-processed predictions achieves the highest score on Leaderboard of 72.689.

The post-processing here refers to, for each prediction, if prob of ‘ZC’ above 0.9, set prob of ‘ZC’ to 1, set other prob of classes to 0.

Deeper Action

4. Experimental Result

Ablation Study - Other

Boosting Experiments

<table>
<tr><th>Level</th><th>Type</th><th>Description</th><th>Boosted(%)</th></tr>
<tr><td rowspan="4">Data</td><td>Size</td><td>Large input size (448)</td><td>+1</td></tr>
<tr><td rowspan="2">Augment</td><td>Horizonal flip</td><td>+0.6</td></tr>
<tr><td>Tiles shuffle</td><td>+1</td></tr>
<tr><td>Sample</td><td>Uniform sample</td><td>+2.2</td></tr>
<tr><td rowspan="7">Model</td><td rowspan="3">Learning Strategy</td><td>Long warmup epoch</td><td>+0.9</td></tr>
<tr><td>Big learning rate</td><td>+1.8</td></tr>
<tr><td>Onecycle scheduler</td><td>+0.5</td></tr>
<tr><td rowspan="3">Batch Strategy</td><td>Accumulate gradients</td><td rowspan="3">+2</td></tr>
<tr><td>Mixed precision</td></tr>
<tr><td>Gradient checkpoint</td></tr>
<tr><td>Other</td><td>Ema models</td><td>+5</td></tr>
<tr><td>Ensemble</td><td colspan="2">5 folders ensemble, mix folder ensemble</td><td>+1.8</td></tr>
<tr><td>Postprocess</td><td colspan="2">For each prediction, if prob of ‘ZC’ above 0.9, set prob of ‘ZC’ to 1, set other prob of classes to 0.</td><td>+0.12</td></tr>
</table>

Not working Experiments

<table>
<tr><th>Level</th><th>Type</th><th>Description</th><th>Boosted(%)</th></tr>
<tr><td rowspan="5">Data</td><td rowspan="3">Augment</td><td>Randaug</td><td>-1</td></tr>
<tr><td>Autoaug</td><td>-0.6</td></tr>
<tr><td>Rotate, vertical flip, color jitter</td><td>-1.6</td></tr>
<tr><td rowspan="2">Sample</td><td>Sequence sample</td><td>-2.2</td></tr>
<tr><td>Larger super-image grid(4x4, 5x5)</td><td>-1</td></tr>
<tr><td rowspan="4">Model</td><td rowspan="4">Weakly Supervised Model</td><td>SimCLR + TransMIL</td><td>-19.4 (local)</td></tr>
<tr><td>SimCLR + MLDecoder</td><td>-19.7 (local)</td></tr>
<tr><td>MAE + TransMIL</td><td>-22 (local)</td></tr>
<tr><td>MAE + MLDecoder</td><td>-25 (local)</td></tr>
<tr><td rowspan="3">TTA</td><td colspan="2">Horizonal flip, Vertical flip</td><td>-3</td></tr>
<tr><td colspan="2">Resample video</td><td>-0.3</td></tr>
<tr><td colspan="2">Grid shuffle</td><td>-0.5</td></tr>
<tr><td>Ensemble</td><td colspan="2">Ensemble by max mAP of each class</td><td>-1.6</td></tr>
<tr><td>Postprocess</td><td colspan="2">Set threshold for each class</td><td>-1.3</td></tr>
</table>

5. Conclusion

- Frame-based methods inevitably assign wrong labels to frames, causing the model to learn data with large deviation.
- Method based on video classification are relatively general, but the lack of flexibility makes it difficult to use more backbones to increase model capacity. The method is also less efficient in training due to learning more complex temporal information, and temporal information are also proved to be less important in this task.
- It is also possible to transform this task into a weakly supervised multi-instance learning task, but pre-training of feature extractors such as MAE and SimCLR is a critical step, and they are also time-consuming. If the feature extractor can be pretrained well on the dataset of similar domain, the score can definitely be improved a lot.
- The super-image-based method is relatively effective in this task. The network only needs to learn the spatio information in the super-image, and can replace the multi-structure backbone and multi-domain pretrained weights at any time, which is of great significance in improving the model capacity. And the mapping between labels and groundtruth will be more accurate as the super-image size increases, but obviously its size is limited by hardware.

Thank you for listening!