**REPORT**

# I. Summary

- Publisher’s information:

+ Computer Vision Foundation

+

- Topic: video object segmentation (segment the most prominent object in a video sequence)

+ Problem: complex backgrounds and multiple foreground objects

- Algorithm:

+ unsupervised

+ a guided slot attention network

- Contribution:

+ achieves state-of-the-art performance on two popular datasets: DAVIS-16, FBMS. we use three datasets for network train- ing: DUTS [28], DAVIS-16 [20], and YouTube-VOS [31], and two datasets for network testing: DAVIS-16 [20], FBMS

+ propose a guided slot attention network to reinforce spatial structural information and obtain better foreground–background separation. Details: The foreground and background slots, which are initialized with query guidance, are iteratively refined based on interactions with template information. Furthermore, to improve slot–template interaction and effectively fuse global and local features in the target and reference frames, K-nearest neighbors filtering and a feature aggregation transformer are introduced.

- Git link: https://github.com/ Hydragon516/GSANet

- Illustration

A group of people standing in front of a map

Description automatically generated

A screenshot of a document

Description automatically generated

# II. Important components

## 2.1. Experiment dataset

- Training and testing dataset:

+ Three datasets for network training: DUTS [28], DAVIS-16 [20], and YouTube-VOS [31]. Two datasets for network testing: DAVIS-16 [20], FBMS

+ The most widely used dataset is DAVIS 2016, which includes 30 training videos and 30 validation videos. FBMS [18] is also commonly used datasets to validate the performance of VOS models.

+ We utilize **DUTS [28], to pretrain the model** and prevent overfitting (DUST does not does not contain optical flow maps, only the RGB encoders and decoders of the RGB stream are pretrained ). Secondly, the pretrained parameters of the RGB stream are applied equally to the optical flow stream. Lastly, the **entire model is fine-tuned with the training set of the DAVIS-16 [20] and YouTube-VOS [31] dataset**

+ The optical flow map required for training is generated using RAFT [25], a pre-trained optical flow estimation model.

+ Other method comparsion

A collage of images of a red bike

Description automatically generated

## 2.2. Measurement metric

A math equations and formulas

Description automatically generated with medium confidence

## 2.3. Implementation Detail

- we randomly sample the reference frames during the training phase and uniformly sample them dur- ing the testing phase

- RGB images optical flow maps are uniformly resized to 384 × 384 pixels for both training and inference.

- experiments are conducted on a two NVIDIA RTX 3090 GPUs and are implemented using the PyTorch deep-learning framework.

- Configuration:

- the clustering count M of the local feature extractor to 64

- the number of reference frames NR to 3

- The number of KNN-filtered samples N in GSA to 16.

- The Adam optimizer [8] is used for network training and fine-tuning with hyperparameters β1 = 0.9, β2 = 0.999, and ε = 10−8.

- The learning rate decreases from 10−4 to 10−5 using a cosine annealing scheduler

- The total number of epochs is set to 200, with a batch size of 12.

# III. Architecture

- proposed model uses one target frame image and NR reference frame images as inputs

A diagram of a machine

Description automatically generated