# 1st Place Solution for VSPW Challenge at ICCV2021: Exploiting Spatial-Temporal Semantic Consistency for Video Scene Parsing

Xingjian He[1,2], Weining Wang[1,2], Zhiyong Xu[1,3], Hao Wang[1,2], Jie Jiang[1,2], Jing Liu[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences

[3] School of Automation and Electrical Engineering, University of Science and Technology Beijing

{xingjian.he,weining.wang,jie.jiang,jliu}@nlpr.ia.ac.cn, g20198748@xs.ustb.edu.cn wanghao2019@ia.ac.cn

## Abstract

*Compared with image scene parsing, video scene parsing introduces temporal information, which can effectively improve the consistency and accuracy of prediction. In this paper, we propose a Spatial-Temporal Semantic Consistency method to capture class-exclusive context information. Specifically, we design a spatial-temporal consistency loss to constrain the semantic consistency in spatial and temporal dimensions. In addition, we adopt an auto-labelling strategy to enrich the training dataset. We obtain the scores of 59.84% and 58.85% mIoU on development (test part 1) and testing set of VSPW, respectively. And our method wins the 1st place on VSPW challenge at ICCV2021.*

## 1. Introduction

Video scene parsing aims to assign pixel-wise semantic labels to each video frame. Directly using image scene parsing methods on video frames, without exploiting the temporal information of video, may lead to the problem of discontinuous prediction results. Therefore, it is valuable to explore how to use video temporal information to improve the continuity of prediction.

By analyzing the prediction results of adjacent frames of multiple videos obtained by the image scene parsing, we find that when the object remains unchanged but the surrounding scene changes, the prediction results of the object may be inconsistent. As shown in Figure 1, in two adjacent frames, the “printer” remains unchanged, while the surroundings of the “printer” change. As a result, the predictions of the “printer” in the two frames are inconsistent. This is because these methods predict by learning the class-inclusive context, which represents information of pixels in all classes in the surroundings of the object. When the surroundings of the object changes, the prediction results of the object may also change, leading to discontinuous video prediction results. Therefore, we consider that the model should learn the class-exclusive context information for prediction, in which the class-exclusive context represents the information of pixels in the same category around the pixel. In this way, when the surrounding environment changes, the pixel prediction results between adjacent frames will remain relatively consistent. The difference between the class-inclusive context and class-exclusive context is shown in Figure 1 (bottom).

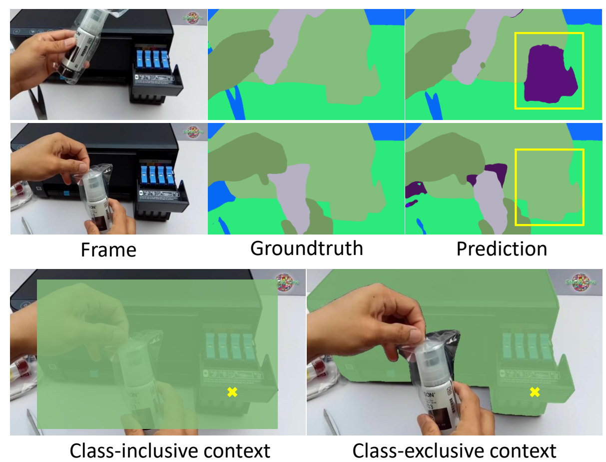


Figure 1. The first two rows show the discontinuity of predicted results on two adjacent frames. The third row shows the class-inclusive context which means information of pixels in all classes in the surroundings of the object and the class-exclusive context which means the information of pixels in the same category around the pixel.

In this paper, we propose a Spatial-Temporal Semantic

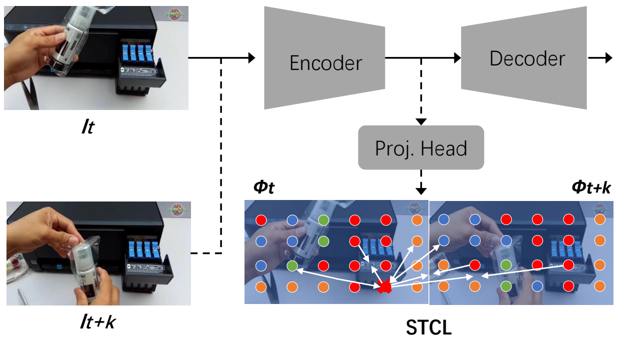

Figure 2. The overview pipeline of our network. STCL denotes spatial-temporal consistency loss.

Consistency method to explore the class-exclusive context information. Specifically, we take adjacent frames as the input of the network, and propose a spatial-temporal consistency loss (STCL) to constrain the intra-class consistency in spatial and temporal dimensions. Thus, the network can pay more attention to the feature of the object itself and reduce the dependence on other categories among the surrounding environment. In other words, the model could capture class-exclusive context information and avoid noisy class-inclusive context information, thus the predicted results are more stable and robust. In addition, we put forward an auto labelling method to help the network improve the representation learning with limited labeled data.

We carry out extensive ablation experiments on Video Scene Parsing in the Wild (VSPW) dataset [12] to verify the effectiveness of our proposed method. We finally obtain the scores of 59.84% and 58.85% mIoU on test part 1 and testing set of VSPW, respectively. And our solution is ranked 1st place on VSPW challenge at ICCV2021.

# 2. Method

In this section, we first describe the overview pipeline of our network. And then, we introduce our Spatial-Temporal Semantic Consistency method. Next, we describe the autolabelling strategy of unlabelled images. Finally, we describe the training and inference process of the network.

## 2.1. Overview

As illustrated in Figure 2, given a video sequence, we take the current frame as the query frame, and a few adjacent frames of the current frame as the reference frames. The query and reference frames are fed into the encoder to extract the pixel-level feature. On the one hand, the encoded features pass through the decoder to recover the resolution for pixel-level prediction, which is supervised by the commonly used cross-entropy loss. On the other hand, the encoded features are transformed by the projection head, and the transformed features are supervised by the proposed spatial-temporal consistency loss.

**Encoder** We apply the CNN-based network (e.g. ResNet [7], ResNeSt [15]) and Transformer-based network (e.g. Swin Transformer [11]) as encoder to extract image features.

**Decoder** Following UperNet [14], we use Feature Pyramid Network [9] to build high-level semantic feature maps at all scales and adopt multi-level feature aggregation for obtaining decoded features.

## 2.2. Spatial-Temporal Semantic Consistency

Most of the recent methods [6, 5] construct class-inclusive context representation for each pixel to assist the semantic prediction. However, the class-inclusive context makes the network weaken the features learning of the object itself, resulting in the network heavily relying on class-inclusive context to achieve pixel prediction. For video sequences, the surroundings of the current pixel is fixed in spatial dimension, while may change in temporal dimension. When the surroundings changes and the object remains unchanged between the adjacent frames, the prediction of the object may be inconsistent. In this work, we propose a spatial-temporal consistency loss to capture class-exclusive context, establishing semantic consistency in spatial and temporal dimensions. The loss encourages the same category of features to be consistent, regardless of whether the surrounding environment is the same or not.

As shown in Figure 2, for input frames $I_t, I_{t+k}$ from the time $t$ and $t+k$, the $I_t, I_{t+k}$ first pass through the encoder to obtain the feature maps $F_t$ and $F_{t+k}$, respectively. Next, similar to [1], they are fed into a projection head to get transformed features $\phi_t$ and $\phi_{t+k}$, where the projection head consists of (FC)→(BN)→(ReLU). For a pixel $i$ with groundtruth semantic label $\hat{c}$, the positive samples are other pixels belonging to the same class $\hat{c}$ in $\phi_t$ and $\phi_{t+k}$, while the negatives are the pixels belonging to the other classes $C \backslash \hat{c}$. The $C$ denotes all classes in dataset. Formally, our spatial-temporal semantic consistency loss is defined as:

$$L(i) = \frac{1}{|P_i|} \sum_{i^+ \in P_i} -\log \frac{\exp(i \cdot i^+ / \tau)}{\exp(i \cdot i^+ / \tau) + \sum\limits_{i^- \in N_i} \exp(i \cdot i^- / \tau)} \tag{1}$$

$$L_{stcl} = \frac{1}{|\Omega|} \sum_{i \in \Omega} L(i) \tag{2}$$

where $\Omega$ denotes the all pixels in $\phi_t$, $P_i$ and $N_i$ denote positive and negative samples for pixel $i$, respectively. $\tau$ denotes temperature coefficient.

By minimizing the above loss function, when the surrounding environment changes, the pixel prediction results between adjacent frames will remain relatively consistent. Thus, the prediction results are more stable and accurate.

### 2.3. Auto Labelling on VSPW

For video data, there are a lot of unlabeled data. Inspired by semi-supervised learning [2, 8], which aims to exploit unlabeled data to further improve the representation learning when given limited labeled data, we adopt an auto-labelling strategy for VSPW to enlarge the training dataset.

In particular, we first build a teacher network to provide a target probability for each of $N$ classes for every pixel in each image. And we adopt a hard labelling strategy, whereby for a given pixel, we select the top class prediction of the teacher network. In order to make auto label have higher confidence, we threshold the label based on teacher network output probability. Those pixels whose maximum prediction probability exceed the threshold will be set as true label, otherwise the pixels will be marked as "ignore" class. In our experiments, the threshold is set to 0.5.

### 2.4. Training & Inference

**Training** We jointly learn the semantic segmentation and the class-inclusive context information in an end-to-end mode. The total loss is defined as:

$$L = \lambda_1 L_{seg} + \lambda_2 L_{stcl} \tag{3}$$

where $L_{seg}$ denotes cross-entropy loss, $\lambda_1$ and $\lambda_2$ are hyper-parameters used to balance the two losses. In practice, $\lambda_1 = 1$ and $\lambda_2 = 0.2$.

**Inference** In the inference phase, we process each frame of the input video independently. Therefore, no additional computational overhead is introduced in the inference stage.

## 3. Experiments

In this section, we first introduce the dataset and implementation details. After that we provide the ablation studies to validate the effectiveness of our proposed method. Finally, we report the results on the challenge test server.

### 3.1. Dataset and Evaluation Metrics

The Video Scene Parsing in the Wild (VSPW) dataset [12] covers a wide range of real-world scenarios and categories. Over 96% of the captured videos are with high spatial resolutions from 720P to 4K. It also provides the dataset with the resolution of 480P. It densely annotates 3,536 videos, including 251,633 frames from 124 categories. The training set, validation set and testing set of VSPW contain 2,806/343/387 videos with 198,244/24,502/28,887 frames, respectively.

In this paper, we adopt mean Interaction over Union (mIoU) and Weight IoU (WIoU) as evaluation metrics for scene parsing. In addition, to evaluate the stability across frames in a video, following [12], we report Video Consistency (VC) metric to evaluate the category consistency among long-range adjacent frames. In particular, we report VC8 and VC16, which evaluate the consistency among 8 frames and 16 frames, respectively.

| Backbone | Decoder | mIoU | VC8 | VC16 |
|---|---|---|---|---|
| ResNet-50 | UperNet | 0.3430 | 0.8050 | 0.7450 |
| ResNest-101 | UperNet | 0.3758 | 0.8430 | 0.7920 |
| Cswin-L | UperNet | 0.5530 | 0.8860 | 0.8467 |
| Swin-L | UperNet | **0.5627** | **0.8864** | **0.8539** |

Table 1. Experiments of different backbones on VSPW val set.

| Method | mIoU | VC8 | VC16 |
|---|---|---|---|
| Swin-L | 0.5627 | 0.8864 | 0.8539 |
| Swin-L + ADE | 0.5670 | 0.8792 | 0.8451 |
| Swin-L + COCO | 0.5704 | 0.8906 | 0.8578 |
| Swin-L + COCO-pre | **0.5803** | **0.8916** | **0.8611** |

Table 2. Experiments of extra data augmentation methods on VSPW val set.

| Method | dim | $\tau$ | mIoU | VC8 | VC16 |
|---|---|---|---|---|---|
| Swin-L | - | - | 0.5803 | 0.8916 | 0.8611 |
| Swin-L + STCL | 512 | 0.07 | **0.5930** | **0.9007** | **0.8687** |
| Swin-L + STCL | 512 | 0.05 | 0.5876 | 0.8989 | 0.8649 |
| Swin-L + STCL | 256 | 0.07 | 0.5872 | 0.8986 | 0.8651 |

Table 3. Experiments of spatial-temporal consistency loss on VSPW val set.

| Method | val set | Test part 1 | | |
|---|---|---|---|---|
| | **mIoU** | **mIoU** | **VC8** | **VC16** |
| Swin-L | 0.5930 | - | - | - |
| Teacher | **0.6085** | 0.5638 | 0.9120 | 0.8769 |
| Swin-L + Pesudo | 0.6006 | **0.5743** | **0.9253** | **0.8989** |

Table 4. Experiments of auto labelling methods on VSPW val set and development (test part 1).

### 3.2. Implementation details

The Pytorch [13] framework is employed to implement our network. In our experiments, we use CNN-based networks and Transformer-based networks as encoder. In training phase, for the CNN-based networks, the backbone is pretrained on ImageNet-1K [3] and SGD optimizer is employed with an initial learning rate of 0.01, a weight decay of 0.0001. For the Transformer-based networks, the backbone is pretrained on ImageNet-22K and AdamW optimizer is adopted with an initial learning rate of $7 \times 10^{-5}$, and a linear warmup of 1,500 iterations. All models in our experiments are trained on 8 GPUs with 2 images per GPU for 20K iterations. For data augmentation, random horizontal flipping, random cropping (cropsize $768 \times 768$) and random resizing within ratio range [0.5, 2.0] are applied. During the testing phase, the sliding window method is used.

| Team | mIoU | WIoU | VC8 | VC16 |
|---|---|---|---|---|
| CASIA_IVA | **0.5984** | **0.7307** | **0.9451** | **0.9222** |
| BetterThing | 0.5806 | 0.7205 | 0.9325 | 0.9053 |
| CharlesBLWX | 0.5738 | 0.7210 | 0.9046 | 0.8669 |
| ustcvim | 0.5497 | 0.7188 | 0.9017 | 0.8674 |
| Arlen | 0.5477 | 0.6991 | 0.9127 | 0.8825 |

Table 5. Comparisons with other methods on the VSPW-challenge-2021 (development) test part 1.

| Team | mIoU | WIoU | VC8 | VC16 |
|---|---|---|---|---|
| CASIA_IVA | **0.5885** | **0.7178** | **0.9477** | **0.9259** |
| CharlesBLWX | 0.5744 | 0.7205 | 0.9129 | 0.8770 |
| BetterThing | 0.5735 | 0.7210 | 0.9328 | 0.8621 |
| Arlen | 0.5562 | 0.6963 | 0.8997 | 0.8621 |
| ustcvim | 0.5549 | 0.7038 | 0.9094 | 0.8746 |

Table 6. Comparisons with other methods on the VSPW-challenge-2021 final test set.

### 3.3. Ablation Studies

**Backbone** We explore the CNN-based and Transformer-based networks as our backbone. The CNN-based networks employ convolutional operations to extract semantic information, which are often affected by limited receptive fields in recognition. The Transformer-based networks employ the self-attention mechanism to model long-range dependencies and obtain competitive results. With large dataset for pre-training, the Transformer-based models show a strong ability of feature representation. The experimental results of different backbones are shown in Table 1. From the table we can find that, the Transformer-based networks (Cswin-L [4], Swin-L [11]) outperform the CNN-based networks (ResNet [7], ResNeSt [15]) for a large margin, which demonstrates that the Transformer-based networks are strong to extract image features. In the following experiments, we employ Swin-L as our backbone network.

**Extra Data Augmentation** To enrich the dataset, we also utilize extra data from other datasets for training and pre-training. ADE20K dataset [16] is a dataset for image scene parsing task that spans diverse annotations of scenes, objects, parts of objects, and in some cases even parts of parts. It contains nearly 20K densely annotated images and 150 classes, which can provide a lot of additional training data. COCO dataset [10] yields approximately 118K images in total and we use the images with panoptic labels for training in experiments. As shown in Table 2, when we add ADE20K dataset and COCO dataset in the training phase, the model obtains improvements with these extra data. Further, when employing COCO dataset in the pre-training phase, the model achieves greater improvements (from 0.5627 to 0.5803), which demonstrates that extra data is more suitable for pre-training.

**Spatial-Temporal Consistency Loss** We propose a spatial-temporal consistency loss to improve the video consistency and accuracy of prediction results. The experimental results are shown in Table 3. We analysis the model performance under different feature dimensions and temperature coefficients. The results show that with dimension set to 512 and $\tau$ set to 0.07, the model improves the video consistency (VC8) of the basic model from 0.8916 to 0.9007, demonstrating the effectiveness of our method.

**Auto labelling** To further make full use of the unlabelled data, we apply an auto-labelling strategy to generate pseudo labels on unlabelled data. We combine the pseudo data and training set to obtain a new dataset. Then, we train our model on the new dataset to improve the performance. As shown in Table 4, we first build a teacher network, which ensembles Swin-L+ADE20K (in Table 2) and Swin-L+STCL (in Table 3) models. The teacher model achieves 0.6085 on val set. And we submit the prediction results on testing set to competition server, obtaining 0.5670, 0.8792 and 0.8451 in terms of mIoU, VC8 and VC16, respectively. We use the strong teacher model to generate pseudo labels and train the model with the obtained pseudo labels. We achieve 0.5704, 0.8906 and 0.8578 in terms of mIoU, VC8 and VC16, respectively.

### 3.4. Comparisons

We use different image ratios of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75] for multi-scale testing, and apply horizontal flipping for each scale. Moreover, we ensemble three different models for boosting the performance. Table 5 and Table 6 show that our method surpasses others by a large margin. Specifically, in the development (test part 1), our solution achieves 0.5984 mIoU, 0.7307 WIoU, 0.9451 VC8, 0.9222 VC16, which is a large gap to the second method on four evaluation metrics. In the final testing set, our model maintains the 1st place on the four metrics.

## 4. Conclusion

In this work, we propose a Spatial-Temporal Semantic Consistency method for video scene parsing. We propose a spatial-temporal consistency loss to encourage the model to learn class-exclusive context information, improving intra-class consistent. In addition, we employ an auto-labelling strategy to make full use of unlabelled data to enrich the training dataset. With the proposed method and other practical tricks, e.g. stronger backbone, extra data augmentation, and ensembling method, our solution outperforms the others significantly and wins the 1st place on VSPW challenge at ICCV2021.

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106400), National Natural Science Foundation of China (61922086, 61872366), and Beijing Natural Science Foundation (4192059, JQ20022).

## References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[2] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[4] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.

[5] Jun Fu, Jing Liu, Jie Jiang, Yong Li, Yongjun Bao, and Hanqing Lu. Scene segmentation with dual relation-aware attention network. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[12] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[14] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 418–434, 2018.

[15] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.