STC-GAN: Spatio-Temporally Coupled Generative Adversarial Networks for Predictive Scene Parsing

Mengshi Qi, *Member, IEEE*, Yunhong Wang, *Fellow, IEEE*,
Annan Li, *Member, IEEE*, and Jiebo Luo, *Fellow, IEEE*

***Abstract*—Predictive scene parsing is a task of assigning pixel-level semantic labels to a future frame of a video. It has many applications in vision-based artificial intelligent systems, *e.g.,* autonomous driving and robot navigation. Although previous work has shown its promising performance in semantic segmentation of images and videos, it is still quite challenging to anticipate future scene parsing with limited annotated training data. In this paper, we propose a novel model called *STC-GAN*, *Spatio-Temporally Coupled Generative Adversarial Networks* for predictive scene parsing, which employ both convolutional neural networks and convolutional long short-term memory (LSTM) in the encoder-decoder architecture. By virtue of STC-GAN, both spatial layout and semantic context can be captured by the spatial encoder effectively, while motion dynamics are extracted by the temporal encoder accurately. Furthermore, a coupled architecture is presented for establishing joint adversarial training where the weights are shared and features are transformed in an adaptive fashion between the future frame generation model and predictive scene parsing model. Consequently, the proposed STC-GAN is able to learn valuable features from unlabeled video data. We evaluate our proposed STC-GAN on two public datasets, *i.e., Cityscapes* and *CamVid*. Experimental results demonstrate that our method outperforms the state-of-the-art.**

***Index Terms*—Predictive Scene Parsing, Generative Adversarial Networks, Coupled Architecture, Spatio-Temporal Features.**

I. Introduction

THE goal of predictive scene parsing [1]–[3] is to assign a semantic label to every pixel in an unobserved video frame, which is critical for real-time decision making and scene understanding in a wide range of applications. As shown in Fig. 1, predictive scene parsing can benefit self-driving, traffic surveillance and automotive assistance.

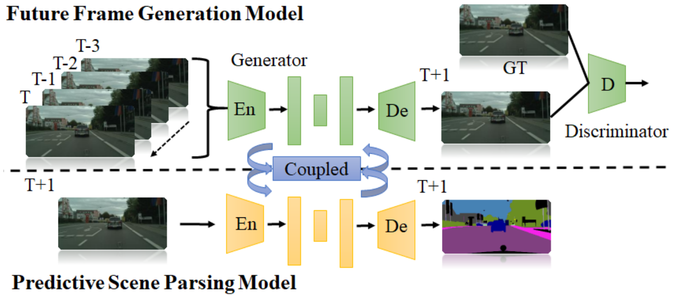

Fig. 1. Illustration of the proposed STC-GAN. STC-GAN consists of two models: the upper part is *a future frame generation model* that generates the next frame given a sequence of observed frames in a video; while the bottom part is *a predictive scene parsing model* that produces a semantic parsing map by inputing the generated future frame. In addition, a coupled architecture is adopted in STC-GAN, which employs a weight-sharing constraint and a feature adaptation transform between two models. *En* and *De* denote encoder and decoder, respectively. *GT* refers to the ground-truth.

Most of the extensive efforts in computer vision invariably focus on images that have been observed, *e.g.,* object detection, image and video semantic segmentation. Conventional state-of-the-art approaches, *e.g.,* SegNet [4], U-net [5], DeepLab [6], PSPNet [7], are all based on fully convolutional networks (FCN) [8], which already achieve a remarkable success in semantic segmentation. However, they only tackle the task of semantic segmentation for existing images or videos in a frame-by-frame manner, but cannot predict parsing results for future images or frames. Moreover, the estimated labeling results across frames by applying current video segmentation methods often suffer from jittering, noise and inconsistency, due to the overlook of the significant temporal context information. In addition, these existing deep learning based methods require a large amount of training data with pixel-level annotation, which are too expensive to acquire.

Predicting or anticipating the future image or frame is a valuable but yet challenging problem in computer vision. How to model the inherent uncertainty and possible events that may happen in the future is the main difficulty in predictive scene parsing. Only limited work tried to address this issue, such as autoregressive convolution neural network [3], [9] and predictive learning model [1], [2]. So far, these methods capture appearance features but fail to incorporate important semantic contextual cues and motion dynamics across frames, leading to noisy and inconsistent labeling results. Capturing semantic context and motion information of a dynamic scene should provide further improvements.

In this paper, to address the above-mentioned issues, we propose a novel deep framework, called *Spatio-Temporally*

Manuscript received August 2, 2019; revised February 13, 2020; accepted March 17, 2020. Date of current version April 6, 2020. This work was supported by the National Key Research and Development Plan of China under Grant 2016YFB1001002. The work of Mengshi Qi was supported by Beihang University when he was a Ph.D. Student. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ioannis Kompatsiaris. *(Corresponding author: Yunhong Wang.)*

Mengshi Qi was with the Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China. He is now with the Computer Vision Laboratory, École polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: mengshi.qi@epfl.ch).

Yunhong Wang and Annan Li are with the Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: yhwang@buaa.edu.cn; liannan@buaa.edu.cn).

Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: jiebo.luo@gmail.com).

Digital Object Identifier 10.1109/TIP.2020.2983567

Coupled Generative Adversarial Networks (STC-GAN) for predictive scene parsing. Specifically, our proposed STC-GAN includes two models: a future frame generation model and a predictive scene parsing model. Given the preceding frames of a video, STC-GAN learns powerful temporal representations by extracting rich dynamic features and high-level spatial contexts via a spatio-temporal encoder in the future frame generation model. Furthermore, we introduce a coupled structure to share weights and transform features in an adaptive fashion between the future frame generation model and the predictive scene parsing model during the adversarial training process. By virtue of such a coupled architecture, our STC-GAN is capable of transferring valuable representations from unlabeled video data to predictive scene parsing. It is worth noting that our main task is *predictive* scene parsing in a video. In other words, the key challenge is *not* segmentation or parsing itself but an effective mechanism to adapt the parsing model to an unseen new frame using prior knowledge from previous frames. What we try to develop is *not* a general solution of scene parsing, but a better method for such a specific scenario. Consequently, experimental results are conducted to demonstrate that our model can learn rich dynamic features and produce more accurate and temporally consistent parsing results without extra supervision. Our main contributions are summarized as follows:

- We propose a novel Spatio-Temporally Coupled Generative Adversarial Networks (STC-GAN) for predictive scene parsing, which captures contextual appearance information and dynamic temporal representation from prior frames to generate future scene parsing results.
- We introduce a coupled architecture into our STC-GAN, and employ a weight-sharing strategy and a feature adaptation transform in the adversarial training to capture the joint representation between future frame generation and the corresponding scene parsing with limited training data. To the best of our knowledge, we are the first to propose such a coupled architecture for predictive scene parsing.
- Extensive experiments on two public benchmarks, *i.e., Cityscapes* and *CamVid*, validate the performance of the proposed method over the state-of-the-art.

This manuscript is organized as the following. In Section II, we provide a brief overview of the literature related to semantic segmentation in images and videos, predictive scene parsing, and generative adversarial networks. In Section III, we elaborate details of the proposed STC-GAN architecture. In Section IV, we tabulate the performance of the proposed approach, and end in Section V with a conclusion of this work.

II. Related Work

In this section, we briefly review three related aspects, *i.e.,* semantic segmentation, predictive scene parsing, and generative adversarial networks.

A. Semantic Segmentation

With the emergence of deep learning methods, semantic segmentation has seen large improvements in recent years. Long *et al.* [8] present an end-to-end Fully Convolutional Network (FCN) for pixel-to-pixel semantic segmentation. SegNet [10] is an FCN-based encoder-decoder network for road and indoor scene understanding. Ronneberger *et al.* [5] propose a similar encoder-decoder architecture named U-net composed of a contracting path to capture contextual features and a symmetric expanding path for precise localization. Recently, DeepLab [6] combines fully-connected Conditional Random Field (CRF) and deep convolutional neural networks, and employs atrous spatial pyramid pooling to encode objects at multiple scales. PSPNet [7] utilizes a multi-scale pyramid module to exploit global contextual information from pyramid layers for scene parsing. Furthermore, a growing number of methods attempt to solve video semantic segmentation [11]–[20] with deep learning methods. In contrast, our model is proposed to handle the unobserved future scene parsing task.

B. Predictive Scene Parsing

Limited work strive for this new topic. Jin *et al.* [1], [2] employ predictive feature learning for video scene parsing and optic flow anticipation, where a predictive learning network is integrated to produce structure-preserving parsing results. Luc *et al.* [3] develop an autoregressive convolutional neural network [9] that can iteratively generate multiple frames for semantic segmentation of future frames. Rochan *et al.* [21] utilize a convolutional LSTM as encoder to capture the representation of observed frames for future parsing map prediction. Chen and Han [22] introduce a multi-timescale context encoding approach for scene parsing prediction, which can simultaneously extract both short-term and long-term temporal relations from the preceding frames, and model semantic interdependencies with an attention mechanism. Zhou *et al.* [23] design a depth embedded recurrent predictive parsing network to address the same challenge, by leveraging binocular stereo images to mine 3D structure information and a LSTM to capture temporal consistence between observed frames. In this work, we present a novel STC-GAN that can learn spatial and temporal information jointly, and employ them to generate better future frames and produce the corresponding parsing results.

C. GAN

Generative Adversarial Networks (GAN) are first introduced to generate images from random noise [24], and have been widely used in many fields including image synthesis [25]–[27], image editing [28], semantic inpainting [29], future prediction learning [9], [30], [31] and representation learning [32], [33]. The main idea of GAN is to employ an adversarial loss to force the generated images to be indistinguishable from real images. Inspired by CoGAN [34], our proposed STC-GAN model shares the knowledge and representations between a future frame generation model and a predictive scene parsing model with a coupled architecture, by employing a weight-sharing constraint and feature adaptation transform in adversarial training.

III. Proposed Approach

A. Overview

The framework of our STC-GAN is illustrated in Fig. 1. Given an input video x, we define $x_t \in \mathbb{R}^{w \times h \times c}$ as the

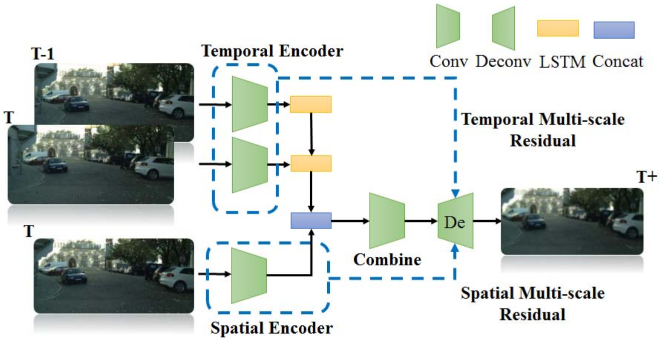

Fig. 2. Overall architecture of our proposed Spatio-Temporal Encoder-Decoder. We feed a sequence of frame differences into the temporal encoder and input the last observed frame to the spatial encoder. We then concatenate and combine the output from them to form the joint spatio-temporal features, and forward them to the predictive decoder (denoted as *De*) with multi-scale residual connections (denoted as blue dashed arrows) to generate the next frame.

t-th frame, where w, h, c denote the width, height and the number of channel, respectively. Our goal of predictive scene parsing is to generate the future frame $\hat{x}_{t+1}$ and the corresponding scene parsing result $S_{\hat{x}_{t+1}}$, through observing previous consecutive frames (*e.g.,* three prior frames denoted as $x_{t-2:t}$). STC-GAN consists of two models: *i.e.,* a future frame generation model and a predictive scene parsing model, while a coupled architecture is incorporated into the adversarial training to share the weights between them. An encoder-decoder based CNN and convolutional LSTM [35] are introduced in the future frame generation, thus building a model composed of a spatial encoder, a temporal encoder and a predictive decoder. Our predictive scene parsing model also employs an encoder-decoder architecture similar to SegNet [4], where the encoder is initially trained with the shared weights, and the decoder utilizes a deconvolutional segmentation model with feature adaptation transform. We will describe the details in the following section.

B. Spatio-Temporal Encoder

As shown in Fig. 2, the Spatial-Temporal Encoder includes two encoders: one is the spatial encoder for extracting appearance and layout information from the last observed frame, and the other is the temporal encoder for capturing dynamic representation considering the difference between the previous frames.

1) Spatial Encoder: The spatial encoder is designed to capture the spatial content representation from a single video frame, *i.e.,* the appearance features of object and background, the structural layout of the scene, etc. We employ a Convolutional Neural Nework (CNN) as the spatial encoder to extract the deep features from the last observed frame x_t:

$$I_t = F^{\text{spat}}(x_t), \tag{1}$$

where $I_t \in \mathbb{R}^{w' \times h' \times c'}$ is the encoding spatial feature tensor, F^{spat} denotes the convolution operation on the last observed frame, w', h', c' denote the width, height and the number of channel of the output, respectively.

2) Temporal Encoder: The temporal encoder is utilized to capture the motion dynamics of the scene. Notably, we adopt a convolutional LSTM [35] layer, and distinguish the difference between two adjacent frames (*i.e.,* x_t and x_{t-1}) recurrently via element-wise subtraction. The output will be the hidden representation H_t that encodes the motion dynamics of the scene, which is formulated as

$$[H_t, C_t] = F^{\text{temp}}(x_t - x_{t-1}, H_{t-1}, C_{t-1}), \tag{2}$$

where $H_t \in \mathbb{R}^{w' \times h' \times c'}$ denotes the encoding temporal feature tensor, $C_t \in \mathbb{R}^{w' \times h' \times c'}$ is the memory cell in LSTM for retaining the observed dynamic representation across time, and F^{temp} is the convolutional LSTM operation.

C. Residual Connection

Inspired by ResNet [36], [37], we employ residual connections between the spatial-temporal encoder and the predictive decoder for preserving more spatio-temporal representation after the pooling operation. Consecutive convolution layers and a linear layer for rectification are also adopted. Moreover, we utilize pyramid pooling to capture multi-scale contextual features for encoding, and each scale feature can have one residual connection. The residual feature at layer l is formulated as:

$$r_t^l = F^{\text{res}}([I_t^l, H_t^l])^l, \tag{3}$$

where r_t^l is the residual output feature at layer l, $[I_t^l, H_t^l]$ denotes the concatenation of the spatial and temporal representation extracted from the spatio-temporal encoder at layer l, and F^{res} means the residual operation at layer l.

D. Predictive Decoder

The predictive decoder is utilized to generate the next future frame $\hat{x}_{t+1} \in \mathbb{R}^{w \times h \times c}$. It receives the combined output (*i.e.,* the concatenation of I_t and H_t) and residual information (*i.e.,* r_t) from spatial-temporal encoder. We adopt the Deconvolution Network [38] as our decoder. It includes deconvolution layers, rectification layers and upsample layers. Then the future frame generated by the decoder can be formulated as:

$$\hat{x}_{t+1} = F^{\text{dec}}([I_t, H_t], r_t), \tag{4}$$

where $[I_t, H_t] \in \mathbb{R}^{w' \times h' \times 2c'}$ denotes the concatenation of the spatial-temporal encoder output, r_t is the residual feature from every layer of the spatial-temporal encoder before pooling, and F^{dec} is the operation of the predictive decoder. The formulation of Eq. (4) refers to that our proposed predictive decoder adopts I_t and H_t as input representations, as well as the residual feature r_t as the additional information to preserve motion-content in the frames. In addition, the top layer of our decoder employs a $tanh(\cdot)$ function for activation.

E. Training Objective of Future Frame Generation

Finally, we use a joint objective function to train the future frame generation model of our STC-GAN, as the following:

$$\mathcal{L}_{ST} = \alpha \mathcal{L}_{img} + \beta \mathcal{L}_{gen}, \tag{5}$$

where α and β are hyper-parameters during optimization. Inspired by [9], $\mathcal{L}_{img}$ is the loss in image space to guide the model to generate correct average sequence. It can be formulated by:

$$\mathcal{L}_{img} = \mathcal{L}_p(x_{t+k}, \hat{x}_{t+k}) + \mathcal{L}_{gdl}(x_{t+k}, \hat{x}_{t+k}), \tag{6}$$

where x_{t+k} and $\hat{x}_{t+k}$ are the ground truth and generated frames, respectively, and k refers to the number of future time step.

$$\begin{aligned}\mathcal{L}_p(y, z) &= \sum_{k=1}^{T} \|y - z\|_2^2,\\ \mathcal{L}_{gdl}(y, z) &= \sum_{i,j}^{h,w} |(|y_{i,j} - y_{i-1,j}| - |z_{i,j} - z_{i-1,j}|)|^{\lambda}\\ &\quad + |(|y_{i,j-1} - y_{i,j}| - |z_{i,j-1} - z_{i,j}|)|^{\lambda},\end{aligned} \tag{7}$$

where y and z denote the pixel value of x_{t+k} and $\hat{x}_{t+k}$, respectively, λ is a hyper-parameter, p refers to pixel value of a frame, gdl means the gradients of pixel values, and $\{i, j\}$ indicates the position of each pixel. Technically, $\mathcal{L}_p$ is utilized to match the average pixel values, and $\mathcal{L}_{gdl}$ is employed to match the gradients of pixel values between the ground-truth frame and the generated frame.

$\mathcal{L}_{gen}$ is the generator loss in adversarial training to guide our model to generate a more realistic frame, which is formulated as

$$\mathcal{L}_{gen} = -\log D([x_{1:t}, G(x_{1:t})]). \tag{8}$$

The discriminator loss in adversarial training is defined as

$$\mathcal{L}_{dis} = -\log D([x_{1:t}, x_{t+1:t+k}]) - \log(1 - D([x_{1:t}, G(x_{1:t})])), \tag{9}$$

where $x_{1:t}$ is the concatenation of the input images, $x_{t+1:t+k}$ is the concatenation of the ground-truth frames, $G(x_{1:t}) = \hat{x}_{t+1:t+k}$ is the concatenation of all predicted frames from the generator of STC-GAN, $D(\cdot)$ is the discriminator in our STC-GAN. The generator G tries to capture the underlying data density and confuse the discriminator D, while the optimization procedure of D aims to distinguish the ground-truth future frames from the generated frames by G.

F. Coupled Architecture

Our STC-GAN consists of two models: M_f for future frame generation and M_s for predictive scene parsing. As depicted in Fig. 3, we employ a weight sharing and feature adaptation transform mechanism to transfer the spatial-temporal representations learned by M_f to M_s, to address the dilemma of training with small-scale labeled data.

Specifically, our coupled structure is based on the existence of shared high-level representations between the image and the corresponding semantic parsing result. Generally speaking, the first several layers of the generative models tend to encode high-level semantic information and the last layers encode low-level details [34]. Therefore, for the first few layers we utilize a weight-sharing strategy to capture such high-level representations, while we adopt a feature adaptation transform technique to transfer knowledge details for the last few layers.[1]

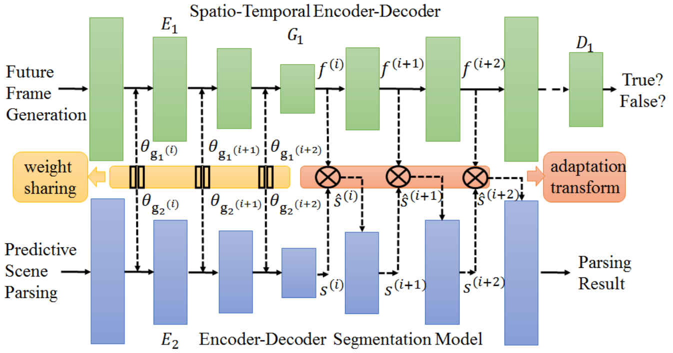

Fig. 3. Illustration of the coupled architecture in STC-GAN. $E_1/G_1/D_1$ are the encoding/generator/discriminator functions in the future frame generation model, and E_2 denotes the encoder in the predictive scene parsing model, respectively. $\theta_{g_1^{(i)}}/\theta_{g_2^{(i)}}$ and $f^{(i)}/s^{(i)}$ refer to the weight and feature in the i-th layer of two models. $\hat{s}^{(i)}$ denotes the adaptation transformed feature. The dotted lines denote using a weight sharing constraint and a feature adaptation transform between the first and last several layers of two models, respectively.

1) Generative Models: Let f and s be input from the marginal distribution of the two domains, $f \sim P_f$ and $s \sim P_s$, respectively. We define g_1 and g_2 as the generators of M_f and M_s:

$$\begin{aligned} g_1(f) &= g_1^{(m_1)}(g_1^{(m_1-1)}(\cdots g_1^{(2)}(g_1^{(1)}(f)),\\ g_2(s) &= g_2^{(m_2)}(g_2^{(m_2-1)}(\cdots g_2^{(2)}(g_2^{(1)}(s)),\end{aligned} \tag{10}$$

where $g_1^{(i)}$ and $g_2^{(i)}$ are the i-th layers of g_1 and g_2, and m_1 and m_2 are the numbers of layers in g_1 and g_2, respectively.

For the future frame generation model and the predictive scene parsing model, they share the same high-level concepts. We force the first layers of g_1 and g_2 to have identical structure and share the weights, that is $\theta_{g_1^{(i)}} = \theta_{g_2^{(i)}}$, for $i = 1, 2, \ldots, m$ where m is the number of shared layers, and $\theta_{g_1^{(i)}}$ and $\theta_{g_2^{(i)}}$ are the parameters of $g_1^{(i)}$ and $g_2^{(i)}$, respectively. Through this constraint, the high-level semantics will be encoded in the same way in g_1 and g_2. The weight-sharing constraint can help reduce the total number of parameters in the network. Note that the training of $g_1(f)$ is according to $\mathcal{L}_{ST}$ in Eq. (5).

2) Discriminator Models: Let d_1 be the discriminative models of M_f given by:

$$d_1(f) = d_1^{(n_1)}(d_1^{(n_1-1)}(\cdots d_1^{(2)}(d_1^{(1)}(f)), \tag{11}$$

where $d_1^{(i)}$ is the i-th layers of d_1 and n_1 is the number of layers. The discriminative model maps an input image to a probability score, estimating the likelihood that the input is drawn from a true data distribution. It is noteworthy that the training of $d_1(f)$ is based on $\mathcal{L}_{dis}$ in Eq. (9).

[1] In this paper, we choose the first three layers for weight-sharing and last three layers for feature adaptation transform in both models of our STC-GAN in experiments.

3) Learning: Formally, the STC-GAN model corresponds to a constraint minimax process given by:

$$\begin{aligned} &\max_{g_1,g_2} \min_{d_1} V(g_1, g_2, d_1), \\ &\text{subject to } \theta_{g_1^{(i)}} = \theta_{g_2^{(i)}}, \quad for \quad i = 1, 2, \cdots, m \end{aligned} \tag{12}$$

where the value function V is defined as the following, and $\mathbb{E}$ is the empirical estimate of the probability:

$$\begin{aligned} &V(g_1, g_2, d_1) \\ &\quad = \mathbb{E}_{f\sim P_f}[-\log d_1(f)] + \mathbb{E}_{f\sim P_f}[-\log(1 - d_1(g_1(f))] \\ &\qquad + \mathbb{E}_{s\sim P_s}[-\log g_2(s)]. \end{aligned} \tag{13}$$

4) Knowledge Distillation With Feature Adaptation Transform: Distillation knowledge transfer methods [13], [39], [40] suggest that employing the intermediate representation of the *teacher* network as *a hint* can benefit the training process and improve the final performance of the *student*. In our proposed coupled architecture, the generator of future frame generation model can be regarded as a *teacher*, while the predictive scene parsing model can be considered as a *student*. However, the norms of features in the two coupled layers maybe different. Therefore, in the last few layers in our proposed coupled architecture, we add a feature adaptation transform layer to match the number of channels of layers between the teacher and student networks, as shown in Fig. 3. Technically, we calculate the adaptation transformed feature $\hat{s}^{(i)}$ in the i-th layer of the student network by linearly combining with the representation $f^{(i)}$ and $s^{(i)}$ in the i-th layer of the teacher and student network, as the following:

$$\hat{s}^{(i)} = W_1 \otimes s^{(i)} + W_2 \otimes f^{(i)}, \tag{14}$$

where W_1 and W_2 refers to weight vectors, and $\otimes$ means per-channel scalar multiplication operation.

In summary, the future frame generation model tries to generate a future scene image, which confuses the discriminative model. The discriminative model tries to distinguish whether the image comes from real domain or generative model. Meanwhile, the encoder of predictive scene parsing model obtains the shared weights from future frame generation for initial training. Then we can employ a deconvolutional segmentation model as decoder in our predictive scene parsing model with the representation adaptation transform to improve scene parsing. Our architecture can be trained by back-propagation with the alternating gradient update steps. The coupled architecture has shown that a joint distribution of images can be learned through weight-sharing constraint and feature adaptation transform with adversarial training, for capturing the correspondences between image generation and semantic segmentation.

IV. Experimental Results

In this section, we describe the experimental evaluation of our framework. We apply and evaluate our framework to the task of predictive scene parsing on two public datasets, *i.e., Cityscapes* and *CamVid*.

A. Experimental Settings

Cityscapes [41] consists of 19 semantic categories and 5,000 images with a high resolution of 2048 × 1024 pixels. It is collected from 50 individual European cities with a diverse geographic and population distribution, at a frame rate of 17 fps. Each video sequence lasts for 1.8s and contains 30 frames, among which the 20th frame has fine-grained manual ground-truth of semantic segmentation. Following [41], we split the whole dataset into three parts: 2,975 training samples, 500 validation samples and 1,525 test samples in our experiments.

CamVid [42] contains over ten minutes' videos, and 701 color images with resolution 960 × 720 are pixel-level annotated on 11 semantic classes. These images are obtained from driving videos captured at daytime and dusk. Each video contains 5,000 frames on the average, amounting to 40K frames in total. Following [4], we choose 60%, 10% and 30% of the frames as training set, validation set (or development set) and test set, respectively.

Metrics: Following most of previous works, we adopt three standard performance metrics: per-pixel accuracy (PA), average per-class accuracy (CA), and mean Intersection-over-Union (mIoU). PA is defined as the percentage of all correctly classified pixels, while CA is the average of all category-wise accuracies. mIoU is defined as the pixel intersection-over-union averaged across all classes. Moreover, we utilized a pair-wise t-test [43] to compare the performance of our proposed method and other baseline models, in order to make the experimental results more convincing.

Baseline Methods: We compare our method with three types of approaches[2]:

- The state-of-the-art methods for image or video semantic segmentation, *i.e., FCN* [8], *SegNet* [10], *DilatedNet* [44], *DeepLab v2* [6], *FSO-CRF* [16], *Deep Structure* [45], *LRR-4X* [46], *RefineNet-Res101* [47], *Liu et al.* [48], *DAG-RNN* [49], *RTDF* [50], *MPF-RNN* [51], *VPN-Flow* [15], *NetWarp-Dilation* [13], *Low-Latency* [17], *DSTRF* [12] [3];
- Methods focused on predictive scene parsing, *i.e., S2S* [3], *PEARL* [1], SPMD [2], Bi-ConvLSTM [21], MTCE [22] and RPPNet [23]. *S2S* [3] employs an autoregressive model to predict the future frames, *PEARL* [1] adopts a two-stage GAN-based learning model to predictive video scene parsing, SPMD [2] incorporates motion dynamic information for predicting scene parsing, Bi-ConvLSTM [21] utilizes a bidirectional convolutional LSTM, MTCE [22] employs a multi-timescale context encoding method, and RPPNet [23] leverages a depth embedded recurrent network;

[2] In the experiments, the parameter setting of above-mentioned methods are adopted from the corresponding papers.

[3] Note that for the comparison with other single frame-wise segmentation models in our experiments, such as *SegNet* [10], the input of SegNet is the T-th frame, while the input of our proposed STC-GAN is a sequence of observed previous frames, *i.e.,* T-3, . . . , T-1, and eventually we compare their output parsing map of T-th frame.

- In addition, we report the performance with VGG16 [52] and Res101 [36] as the backbone of our STC-GAN for fair comparison, respectively.

B. Implementation Details

In this work, all the implementations are based on the PyTorch[4] framework. Generally, we choose VGG16 [52] as the backbone structure (up to the third pooling layer, and replace its 3×3 convolutions with $5 \times 5/5 \times 5/7 \times 7$ convolutions, respectively) as spatial-temporal encoder of our STC-GAN's future frame generation model. We also adopt the same architecture in the predictive scene parsing model. Meanwhile, we adopt three consecutive 3×3 and two consecutive 3×3 convolutions in the combination layers and in the multi-scale residual layers, respectively. We share the parameters for these first three layers, *i.e.,* $k = 3$ in Eq. (12). The predictive decoder of our STC-GAN is a three-layer Deconvolutional Network with the un-pooling operation. Hence, we transform the features of these three layers to the decoder in predictive scene parsing model. Our predictive scene parsing model employs an encoder-decoder architecture similar to SegNet [4], where the encoder is initially trained with the shared weights, and the decoder utilizes a deconvolutional segmentation model. Each encoder layer has a corresponding decoder layer, and the final decoder output is fed to a multi-class soft-max classifier to produce class probabilities for each pixel independently. In the experiments, we utilize the ADAM algorithm [53] for training. The learning rate is set to 2×10^{-4} and the momentum parameters are set to 0.5 and 0.999. Moreover, the mini-batch size is 128 with $25,000$ iterations. In practice, the hyper-parameters are set to $\alpha = 1$, $\beta = 0.02$, and λ=1 in the loss functions. All the framework is trained on a single NVIDIA 1080 Ti GPU.

Following [2], we obtain totally 35K and 8.8K sequences from Cityscapes and CamVid, respectively, by randomly choosing four preceding consecutive frames, *i.e.,* $k = 4$ in $x_{t-k:t-1}$, as the input data in future frame generation with enough motion changes (measured by the $\mathcal{L}_2$ distance between the raw frames). Moreover, we select four frames before the fine labeled frames as the input in the predictive scene parsing phrase. In addition, all input frames are normalized to $[-1, 1]$ in pixels level. Then we employ random cropping with the size of 256×256 and random mirroring for data augmentation.

Computational Efficiency: Our model can produce the parsing result of a frame within 0.6 seconds with resolution $1,024 \times 2,048$ on a single NVIDIA 1080 Ti GPU during test, which is faster than other existing works [1], [3].

C. Performance Comparisons

1) Results on Cityscapes: Quantitative performance comparisons of our STC-GAN with other methods following their evaluation protocol are shown in Table I. It is noted that our STC-GAN with ResNet101 achieves the best performance in terms of CA (47.3%), surpassing the strong baseline model (DeepLab v2 and DilatedNet) by 5%, showing that our method can distinguish and classify more accurate semantic class, while the representations extracted by more advanced model are beneficial to parsing. Even if comparing our proposed STC-GAN with other state-of-the-art predictive scene parsing approaches (*i.e.,* S2S [3], PEARL [1], SPMD [2], Bi-ConvLSTM [21], MTCE1 [22] and RPPNet [23]), our model still outperforms them demonstrating the effectiveness and superiority of the proposed coupled generative adversarial architecture. Furthermore, STC-GAN also obtains the best performance w.r.t. mIoU (76.1%), suggesting that our model is capable of capturing more consistent temporal information to help scene parsing. Note that our model focuses on generating future frames and learning better representations for video semantic segmentation, and thus adopts a basic structure similar to SegNet for segmentation. Our STC-GAN can be combined with the state-of-the-art segmentation models for further improved performance. When we employ Deep v3 [54] and v3+ [55] as our segmentation model, the results of STC-GAN can increase about 5% and 6% w.r.t mIoU in Cityscapes, respectively. Meanwhile, we also compare the results of our proposed model with other state-of-the-art methods in terms of per-class mIoU, and the results are reported in Table II. As shown in Table II, we can clearly observe that our proposed STC-GAN with Res101 achieves the best performance in terms of most of classes, *e.g.,* pole, vegetable and road, demonstrating the benefits of our approach. For example, our proposed model obtain 98.6% and 93.6% w.r.t mIoU in road and vegetable, showing our model is able to generate smoother parsing map and capture the space-time consistency between frames.

TABLE I

PERFORMANCE COMPARISONS OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES ON CITYSCAPES TEST SET. **Res101** [36] DENOTES THE METHODS ADOPTING RES101 AS THE BACKBONE, WHILE OTHER APPROACHES UTILIZE **VGG16** [52]. BEST RESULTS ARE IN BOLD

Methods	PA	CA	mIoU
FCN [8]	85.7	34.4	66.0
SegNet [4]	87.2	41.4	57.0
DilatedNet [44]	86.5	42.0	67.1
DeepLab v2 [6]	86.4	42.6	70.4
FSO-CRF [16]	N/A	N/A	70.3
Deep Structure [45]	N/A	N/A	71.6
LRR-4X [46]	N/A	N/A	69.7
RefineNet-Res101 [47]	N/A	N/A	73.6
Low-Latency [17]	N/A	N/A	75.3
S2S [3]	**91.8**	N/A	66.8
PEARL-VGG16 [1]	N/A	N/A	69.8
PEARL-Res101 [1]	N/A	N/A	74.9
SPMD-Res101 [2]	N/A	N/A	66.1
Bi-ConvLSTM-Res101 [21]	N/A	N/A	71.3
MTCE-Res101 [22]	N/A	N/A	72.2
RPPNet-Res101 [23]	89.8	N/A	42.3
STC-GAN VGG16	88.5	46.1	72.5
STC-GAN Res101	90.9	**47.3**	**76.1**

In addition, qualitative examples of future frame generation and scene parsing results produced by our STC-GAN with VGG16 are depicted in Fig. 4. Visually, the future frame

[4] http://pytorch.org

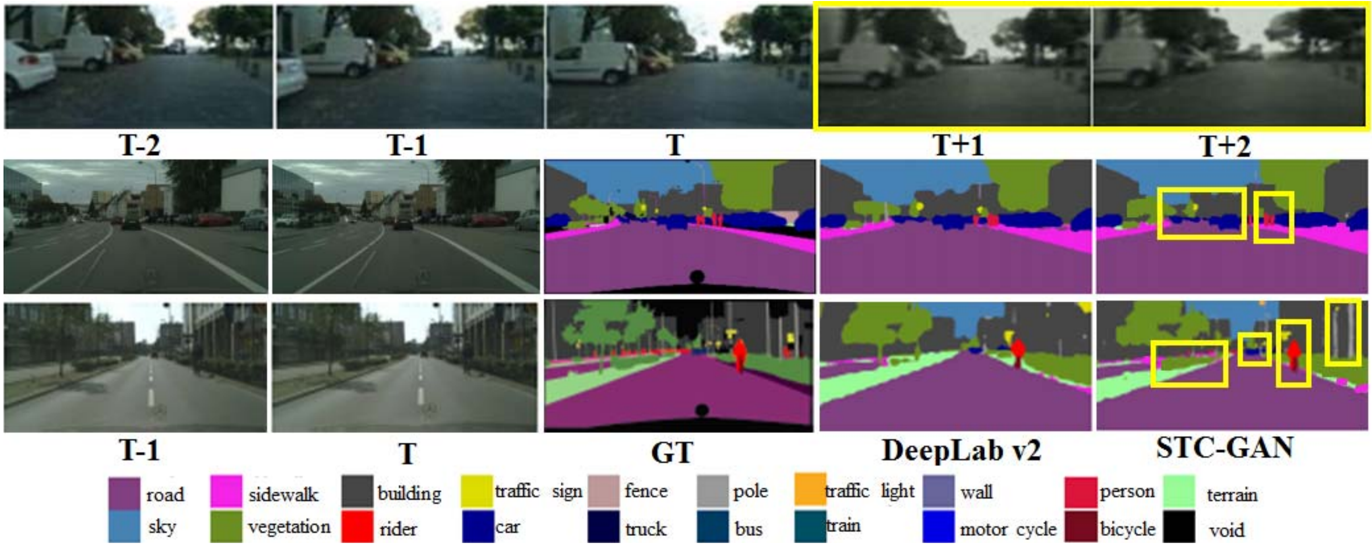

Fig. 4. Example of future frame generation and predictive scene parsing results on the Cityscapes dataset. In the first row, first three columns show the input prior frames of a video, *i.e.,* $x_{T-2:T}$, the last two columns are the future frames generated by our STC-GAN (denoted in yellow box), *i.e.,* $\hat{x}_{T+1:T+2}$. The second and third rows illustrate the predictive parsing results given observed frames, *i.e.,* $x_{T-1:T}$, and comparing the results of our STC-GAN with Ground-Truth (denoted as GT) and DeepLab v2. Quite a few small objects can be labeled accurately by our STC-GAN denoted in yellow box.

TABLE II

PER-CLASS RESULTS COMPARISONS W.R.T MEAN INTERSECTION-OVER-UNION (MIOU) OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES ON CITYSCAPES TEST SET. **RES101** [36] DENOTES THE METHODS ADOPT RES101 AS THE BACKBONE, WHILE OTHER APPROACHES UTILIZE **VGG16** [52]. BEST RESULTS ARE IN BOLD

Methods	Road	Sidewalk	Building	Wall	Fence	Pole	Trafficlight	Trafficsign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle
DilatedNet [44]	97.2	79.5	90.4	44.9	52.4	55.1	56.7	69.0	91.0	58.7	92.6	75.7	50.0	92.2	56.2	72.6	53.2	46.2	70.1
FCN [8]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8
DeepLab v2 [6]	97.3	77.6	87.7	43.6	40.4	29.7	44.5	55.4	89.4	67.0	92.7	71.2	49.4	91.4	48.7	56.7	49.1	47.9	58.6
SegNet [4]	96.4	73.2	84.0	28.4	29.0	35.7	39.8	45.1	87.0	63.8	91.8	62.8	42.8	88.3	38.1	43.1	44.1	35.8	51.9
LRR-4X [46]	98.0	81.5	91.4	50.5	52.7	59.4	66.8	72.7	92.5	70.1	95.0	81.3	60.1	94.3	51.2	67.7	54.6	55.6	69.6
RefineNet-Res101 [47]	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	**76.1**	**64.3**	62.2	70.0
RPPNet-Res101 [23]	97.4	79.9	89.6	47.0	46.6	50.3	57.9	69.7	90.1	58.4	92.1	72.5	46.0	92.3	60.7	68.1	46.4	41.9	67.4
STC-GAN VGG16	97.2	83.2	91.1	49.3	50.9	57.9	66.9	71.7	92.1	70.2	95.1	81.1	62.6	94.1	61.9	70.3	55.9	61.6	69.7
STC-GAN Res101	**98.6**	**85.5**	**92.6**	**51.6**	**53.5**	**60.2**	**67.7**	**73.9**	**93.6**	**72.5**	**96.2**	**83.2**	**65.5**	**95.9**	**65.7**	73.9	63.5	**65.2**	**71.3**

generated by our model is realistic and similar to the ground-truth, showing that our STC-GAN can capture the change of appearance and motion from prior frames. Furthermore, our model generates smoother parsing maps (*e.g.,* vegetation, traffic sign or pole), and successfully learn the video segmentation even when the frames change significantly. We think the reason why our proposed method outperforms the other approaches is that our STC-GAN can learn much better and consistent spatial-temporal features through the spatial-temporal encoder in generating future frames.

2) *Results on CamVid:* We also evaluate our model on the CamVid dataset against multiple state-of-the-art methods. It is worth noting that compared to the Cityscapes dataset, CamVid is a much smaller dataset, which may affect the power of deep architectures due to the lack of training examples. Nevertheless, our STC-GAN still outperforms all the baseline approaches in terms of CA and mIoU. In Table III, we report the results w.r.t. PA, CA and mIoU. Our STC-GAN with Res101 obtains the best performance, especially improving mIoU by 15% to 17% over SegNet, demonstrating the effectiveness and superiority of our model. Meanwhile, our STC-GAN with Res101 achieves the best performance across CA (85.2%), denoting that our model can recognize semantic classes more accurately and learn meaningful representation through the weight-sharing mechanism and feature adaptation transform between two models. Moreover, STC-GAN with VGG16 also obtains competitive performance compared with other state-of-the-art. In addition, our STC-GAN outperforms the optical flow based methods (*i.e.,* RTDF [50] and Liu *et al.* [48]), demonstrating STC-GAN can model spatio-temporal contextual information in a sequence. Fig. 5 illustrates the qualitative scene parsing maps generated by our STC-GAN on CamVid. We can see quite a few pixels are refined and several small objects (*e.g.,* pedestrian, tree or bicyclist) can be labeled correctly by our model. Additionally, we also show the per-class mIoU result of our proposed model and other approaches in Table IV. From the table, it is obvious that our proposed STC-GAN with Res101 outperforms other state-of-the-art methods across all classes, due to the spatial-temporal continuity captured by the predictive features

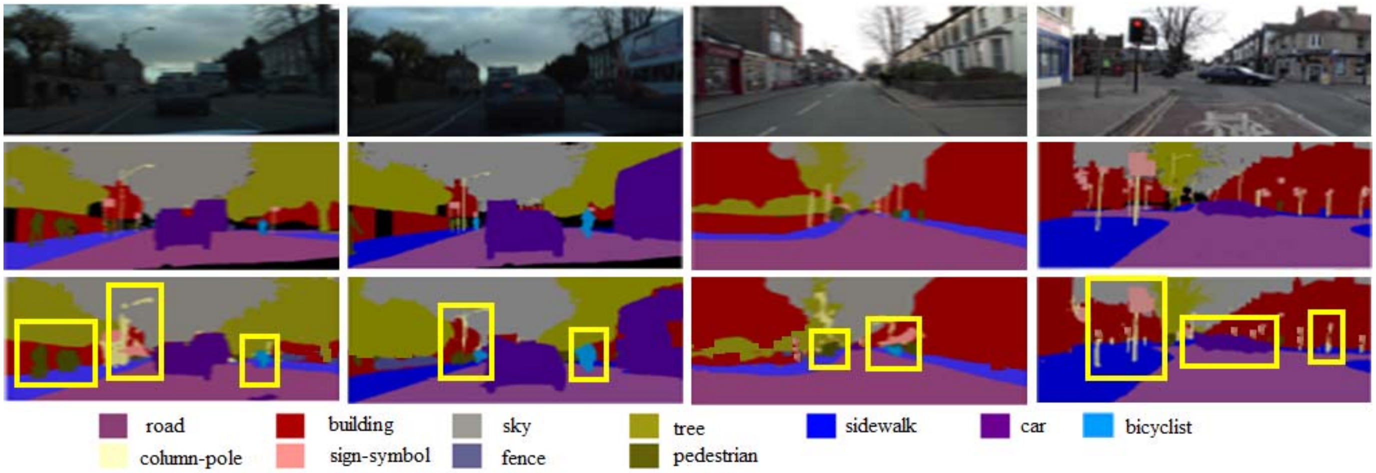

Fig. 5. Example scene parsing results of our model on the CamVid dataset. The first row is the last input frame, and the second row is the ground truth of segmentation map for the next frame, and the third row is the parsing map produced by our STC-GAN.

TABLE III

PERFORMANCE COMPARISONS OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES ON CAMVID TEST SET. **Res101** [36] DENOTES THE METHODS ADOPTING RES101 AS THE BACKBONE, WHILE OTHER APPROACHES UTILIZE **VGG16** [52]. BEST RESULTS ARE IN BOLD

Methods	PA	CA	mIoU
SegNet-Basic [4]	82.2	62.3	46.3
SegNet-Pretrained [4]	88.6	65.9	50.2
DeepLab v2 [6]	84.6	62.6	61.6
Liu et al. [48]	82.5	62.5	N/A
DAG-RNN [49]	91.6	78.1	N/A
RTDF [50]	89.9	80.5	N/A
MPF-RNN [51]	92.8	82.3	N/A
FSO-CRF [16]	N/A	N/A	66.1
VPN-Flow [15]	N/A	N/A	66.7
NetWarp-Dilation [13]	N/A	N/A	67.1
DSTRF [12]	N/A	N/A	65.7
Low-Latency [17]	**94.6**	82.9	N/A
S2S [3]	N/A	N/A	46.8
PEARL-Res101 [1]	94.4	83.2	N/A
STC-GAN VGG16	90.5	82.6	65.6
STC-GAN Res101	93.9	**85.2**	**67.3**

of STC-GAN. The consistently state-of-the-art performances on both datasets suggest that our framework is effective for both future frame generation and predictive scene parsing.

3) Statistical Analysis: Because current video segmentation methods utilized the small-size datasets, *i.e.,* the Cityscapes dataset and CamVid dataset for evaluation, it is necessary to examine whether the difference of performances is statistically significant or not. In our experiments, we use the pairwise t-test [43] to compare the performance w.r.t PA, CA and mIoU of our proposed method and two widely-adopted open-source baselines, *i.e.,* SegNet [4] and DeepLab v2 [6]. All of these models adopt VGG16 as backbone for fair comparison. Table V shows the experimental results (mean ± standard deviation) w.r.t PA, CA and mIoU obtained using our proposed STC-GAN and other baseline approaches on Cityscapes dataset and CamVid dataset. Meanwhile, Table VI reports the t-test results when comparing the performance of STC-GAN versus SegNet [4] and STC-GAN versus DeepLab v2 [6], where "$\gg$" refers to the p-value is lesser than 0.05, indicating a strong evidence that a method results in a greater value for the effectiveness measure than another method. From two tables, we can find that our proposed STC-GAN outperforms two baseline models in terms of quantitative results, as well as has statistically significant superior performances than SegNet [4] and DeepLab v2 [6] over three metrics (*i.e.,* PA, CA and mIoU). The reason can be explained by the proposed method is able to capture better spatial-temporal representation from observed frames to enhance semantic segmentation model.

D. Ablation Study

In the following, we will analyze and investigate the effect of each component of the proposed STC-GAN on Cityscapes, respectively.

1) Analysis of the Spatial-Temporal Encoder-Decoder: To provide evidence of the effectiveness of the spatial-temporal encoder-decoder in STC-GAN, we conduct further experiments for comparison. As can be seen from Table VII, we compare our proposed STC-GAN with two baselines including *Baseline-VGG16/Res101* that only utilizes the visual appearance features for scene parsing, and *OF-VGG16/Res101* that combines the optical flow features and appearance features for segmentation. *Baseline-VGG16/Res101* denotes inputting each frame to our segmentation model with VGG16 or Res101 and then merge the input frame's probability maps to achieve the final parsing map; *OF-VGG16/Res101* refers to concatenating optical flow features by epic flow [56] from x_{t-1} and x_t with the feature of frame x_t with VGG16 or Res101. Compared with *Baseline-VGG16/Res101*, our STC-GAN achieves significant performance improvements (gain of 9% and 4% w.r.t mIoU with VGG16 and Res101, respectively), suggesting that our STC-GAN obtains better performance as the result of our spatio-temporal encoder-decoder can capture distinctive and consistent representations in space-time than the appearance features or

TABLE IV

PER-CLASS RESULTS COMPARISONS W.R.T MEAN INTERSECTION-OVER-UNION (MIOU) OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES ON CAMVID TEST SET. **Res101** [36] DENOTES THE METHODS ADOPT RES101 AS THE BACKBONE, WHILE OTHER APPROACHES UTILIZE **VGG16** [52]. BEST RESULTS ARE IN BOLD

Methods	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist
FCN [8]	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0
SegNet [4]	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8
DilatedNet [44]	84.0	77.2	91.3	85.6	49.9	92.5	59.1	37.6	16.9	76.0	57.2
DeepLab v2 [6]	81.5	74.6	89.0	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1
STC-GAN VGG16	85.1	78.6	93.2	87.1	50.6	93.9	60.6	39.5	20.3	78.3	59.1
STC-GAN Res101	**86.7**	**80.2**	**95.0**	**88.6**	**52.3**	**94.6**	**63.1**	**41.2**	**22.6**	**80.2**	**62.3**

TABLE V

PERFORMANCE COMPARISONS OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES ON CITYSCAPES AND CAMVID TEST SET. THE RESULTS ARE SHOWN AS MEAN ± STANDARD DEVIATION. BEST RESULTS ARE IN BOLD

Dataset	Metrics	Methods		
		STC-GAN	SegNet [4]	DeepLab v2 [6]
Cityscapes	PA	**88.5 ± 0.9**	87.2 ± 1.5	86.4 ± 0.7
	CA	**46.1 ± 1.1**	41.4 ± 1.3	42.6 ± 1.6
	mIoU	**72.5 ± 0.7**	57.0 ± 1.2	70.4 ± 0.9
CamVid	PA	**90.5 ± 1.7**	82.2 ± 1.1	84.6 ± 1.7
	CA	**82.6 ± 1.5**	62.3 ± 1.3	62.6 ± 1.2
	mIoU	**65.6 ± 0.9**	46.3 ± 1.7	61.6 ± 1.1

TABLE VI

THE t-TEST RESULTS OF OUR METHOD COMPARED WITH SEGNET [4] AND DEEPLAB V2 [6] ON CITYSCAPES AND CAMVID TEST SET. "≫" REFERS TO THE p-VALUE IS LESSER THAN 0.05

Dataset	Metrics	Methods	
		STC-GAN vs SegNet	STC-GAN vs DeepLab v2
Cityscapes	PA	≫	≫
	CA	≫	≫
	mIoU	≫	≫
CamVid	PA	≫	≫
	CA	≫	≫
	mIoU	≫	≫

TABLE VII

COMPARATIVE STUDY TO ANALYZE THE SPATIAL-TEMPORAL ENCODER-DECODER. **VGG16** AND **Res101** DENOTE THE BACKBONE OF METHODS, AND **OF** MEANS THE OPTICAL FLOW MAPS. BEST RESULTS ARE IN BOLD

Methods	CA	PA	mIoU
Baseline-VGG16	37.5	85.9	63.4
OF-VGG16	40.6	86.7	64.5
STC-GAN-VGG16	**46.1**	**88.5**	**72.5**
Baseline-Res101	40.2	86.6	72.5
OF-Res101	42.6	87.2	72.7
STC-GAN-Res101	**47.3**	**90.9**	**76.1**
only $\mathcal{L}_{img}$-VGG16	39.5	86.2	63.9
only $\mathcal{L}_{gen}$-VGG16	42.3	87.5	65.2
$\mathcal{L}_{img}$+$\mathcal{L}_{gen}$-VGG16	**46.1**	**88.5**	**72.5**

optical flow features employed by the baselines. It reveals that the naively concatenating appearance feature and noisy probability maps would lead to worse performance. Furthermore, we observe that *OF-VGG16/Res101* obtain better performance than *Baseline-VGG16/Res101*, but still worse than our STC-GAN. Clearly, optical flow can be beneficial for representing the motional information, but our spatio-temporal encoder-decoder has much stronger capacity to capture temporal change and transfer spatial-temporal knowledge of observed frames to the predictive parsing task.

In addition, we have further conducted experiments on the efficacy of the adversarial loss, and the results are shown in Table VII. When we set $\alpha = 1$ and $\beta = 0$ in Eq. (5), it means we only use $\mathcal{L}_{img}$ without adversarial learning, while for $\alpha = 0$ and $\beta = 1$ we only use $\mathcal{L}_{gen}$. We find that our model with only $\mathcal{L}_{img}$ achieves the worst results, while only adopting $\mathcal{L}_{gen}$ can improve the performance obviously, suggesting the adversarial loss is more important and effective in STC-GAN. The discriminator is utilized for distinguishing the real frame and generated frames, which can help the generator create more realistic future frame and enhance the ability of learning distinctive representations through the adversarial training. When we utilize both loss functions in STC-GAN simultaneously, our model can obtain the best performance. This also demonstrates combining $\mathcal{L}_{img}$ and $\mathcal{L}_{gen}$ can enhance the frame prediction and make the generator better, by penalizing the differences of image gradient.

2) Ablation Study of the Coupled Architecture: We also evaluate the effectiveness of our proposed coupled architecture. We conduct further experiments to compare our proposed STC-GAN with two variants, *i.e.,* one variant only using the features of spatio-temporal encoder-decoder model for parsing, the other variant only using the features from predictive scene parsing model for parsing. Neither of these two variants employs the joint training strategy. The experimental results are reported in Table VIII. Specifically, the first variant utilizes the output feature of spatio-temporal encoder in STC-GAN with pixel classify layer for scene parsing, which only uses the spatio-temporal feature and the results are not satisfactory (denoted as 'S-T-Encoder' in Table VIII).

TABLE VIII

ABLATIVE STUDY TO ANALYZE THE EFFECTIVENESS OF THE COUPLED ARCHITECTURE. **VGG16** DENOTES THE BACKBONE OF METHODS, AND **S-T-Encoder** MEANS THE SPATIO-TEMPORAL ENCODER OF STC-GAN, AND **Parsing** IS THE PREDICTIVE SCENE PARSING MODEL OF STC-GAN. BEST RESULTS ARE IN BOLD

Methods	CA	PA	mIoU
S-T-Encoder-VGG16	42.5	86.8	68.5
Parsing-VGG16	43.9	87.7	69.1
STC-GAN-VGG16	**46.1**	**88.5**	**72.5**

The second variant leverages the output features of predictive scene parsing model in STC-GAN with the pixel classify layer (denoted as ‘Parsing’ in Table VIII), the performance of which is better than the first one. However, it lacks the jointly training, which limits its performance gain, and hence it is still inferior to our proposed method. Particularly, compared to these two variants, our STC-GAN achieves the best performance, which is credited to the strong complementary information learned by the weight-sharing and feature adaptation transform in the coupled architecture. The results reveal that the coupled architecture with joint training between the future frame generation model and the predictive scene parsing model in our proposed STC-GAN is quite effective.

V. CONCLUSION

In this paper, we present a novel Generative Adversarial Networks-based model (*i.e.,* STC-GAN) for predictive scene parsing. STC-GAN captures both spatial and temporal representations from the observed frames of a video through CNN and convolutional LSTM network. Moreover, a coupled architecture is employed to guide the adversarial training via a weight-sharing mechanism and a feature adaptation transform between the future frame generation model and the predictive scene parsing model. Extensive results on two widely-adopted benchmarks have demonstrated that our proposed STC-GAN model outperforms the state-of-the-art approaches and is capable of learning and producing consistent, robust and accurate anticipated semantic segmentation results. In the future, we will study how to introduce multi-modal information to help improve the performance of predictive scene parsing, *e.g.,* optical flow, audio and depth information. Furthermore, how to apply the proposed model into real autonomous driving scenario is also deserved to explore further.

REFERENCES

[1] X. Jin *et al.*, “Video scene parsing with predictive feature learning,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5580–5588.

[2] X. Jin *et al.*, “Predicting scene parsing and motion dynamics in the future,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6915–6924.

[3] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, “Predicting deeper into the future of semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 648–657.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[8] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” 2015, *arXiv:1511.05440*. [Online]. Available: http://arxiv.org/abs/1511.05440

[10] V. Badrinarayanan, A. Handa, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling,” 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

[11] M. Qi, Y. Wang, J. Qin, and A. Li, “KE-GAN: Knowledge embedded generative adversarial networks for semi-supervised scene parsing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5237–5246.

[12] S. Chandra, C. Couprie, and I. Kokkinos, “Deep spatio-temporal random fields for efficient video segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8915–8924.

[13] R. Gadde, V. Jampani, and P. V. Gehler, “Semantic video CNNs through representation warping,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4453–4462.

[14] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, “Efficient uncertainty estimation for semantic segmentation in videos,” in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 520–535.

[15] V. Jampani, R. Gadde, and P. V. Gehler, “Video propagation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 451–461.

[16] A. Kundu, V. Vineet, and V. Koltun, “Feature space optimization for semantic video segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3168–3175.

[17] Y. Li, J. Shi, and D. Lin, “Low-latency video semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5997–6005.

[18] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, “Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2125–2135.

[19] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, “Dynamic video segmentation network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6556–6565.

[20] Y. Zhu *et al.*, “Improving semantic segmentation via video propagation and label relaxation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8856–8865.

[21] M. Rochan *et al.*, “Future semantic segmentation with convolutional LSTM,” in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 137.

[22] X. Chen and Y. Han, “Multi-timescale context encoding for scene parsing prediction,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1624–1629.

[23] L. Zhou, H. Zhang, Y. Long, L. Shao, and J. Yang, “Depth embedded recurrent predictive parsing network for video scenes,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4643–4654, Dec. 2019.

[24] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[25] E. L. Denton *et al.*, “Deep generative image models using a Laplacian pyramid of adversarial networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[26] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[27] J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial network,” 2016, *arXiv:1609.03126*. [Online]. Available: http://arxiv.org/abs/1609.03126

[28] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 597–613.

[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[30] Y. Jang, G. Kim, and Y. Song, "Video prediction with appearance and motion conditions," 2018, *arXiv:1807.02635*. [Online]. Available: http://arxiv.org/abs/1807.02635

[31] N. Wichers, R. Villegas, D. Erhan, and H. Lee, "Hierarchical long-term video prediction without supervision," 2018, *arXiv:1806.04768*. [Online]. Available: http://arxiv.org/abs/1806.04768

[32] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 5040–5048.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[34] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.

[35] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," 2017, *arXiv:1706.08033*. [Online]. Available: http://arxiv.org/abs/1706.08033

[38] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[40] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: http://arxiv.org/abs/1412.6550

[41] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[42] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis.* Marseille, France: Springer, 2008, pp. 44–57.

[43] G. D. Ruxton, "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney u test," *Behav. Ecol.*, vol. 17, no. 4, pp. 688–690, Jul. 2006.

[44] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[45] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1352–1366, Jun. 2018.

[46] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 519–534.

[47] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[48] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1377–1385.

[49] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "DAG-recurrent neural networks for scene labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3620–3629.

[50] P. Lei and S. Todorovic, "Recurrent temporal deep field for semantic video labeling," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 302–317.

[51] X. Jin, Y. Chen, Z. Jie, J. Feng, and S. Yan, "Multi-path feedback recurrent neural networks for scene parsing," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4096–4102.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[54] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[55] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 801–818.

[56] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1164–1172.

Mengshi Qi (Member, IEEE) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012, and the M.S. and Ph.D. degrees in computer science from Beihang University, Beijing, in 2014 and 2019, respectively. He is currently a Postdoctoral Researcher with the CVLAB, EPFL. His research interests include machine learning and computer vision, especially scene understanding, 3D reconstruction, and multimedia analysis.

Yunhong Wang (Fellow, IEEE) is currently a Professor with the School of Computer Science and Engineering, Beihang University. She has published in excess of 200 academic articles in areas largely related to biometrics, surveillance, and image processing. Her research interests include biometrics, statistical pattern recognition, and digital image processing. She is a Fellow of IAPR. She has contributed to professional activities, on the Editorial Boards of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON BIOMETRICS, *Behavior and Identity Science*, and *Pattern Recognition*.

Annan Li (Member, IEEE) received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2003 and 2006, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He worked in Singapore as a Scientist with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR) and as a Postdoctoral Research Fellow at the National University of Singapore. He currently works at the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and statistical learning.

Jiebo Luo (Fellow, IEEE) joined the Department of Computer Science, University of Rochester, in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 400 technical articles and holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He is a Fellow of the ACM, AAAI, SPIE, and IAPR. He has served as the Program Chair of ACM Multimedia 2010, the IEEE CVPR 2012, the ACM ICMR 2016, and the IEEE ICIP 2017, and on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIG DATA, *Pattern Recognition*, *Machine Vision and Applications*, and the *ACM Transactions on Intelligent Systems and Technology*.