PAPER ANALYSIS

Presented by Yannis He

-

Paper: **TSIT: A Simple and Versatile Framework for Image-to-Image Translation**

Conference: ECCV 2020

Authors: Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, Chen Change Loy

- | Nanyang Technological University
- | University of California, Berkeley
- | SenseTime Research

https://arxiv.org/abs/2007.12072

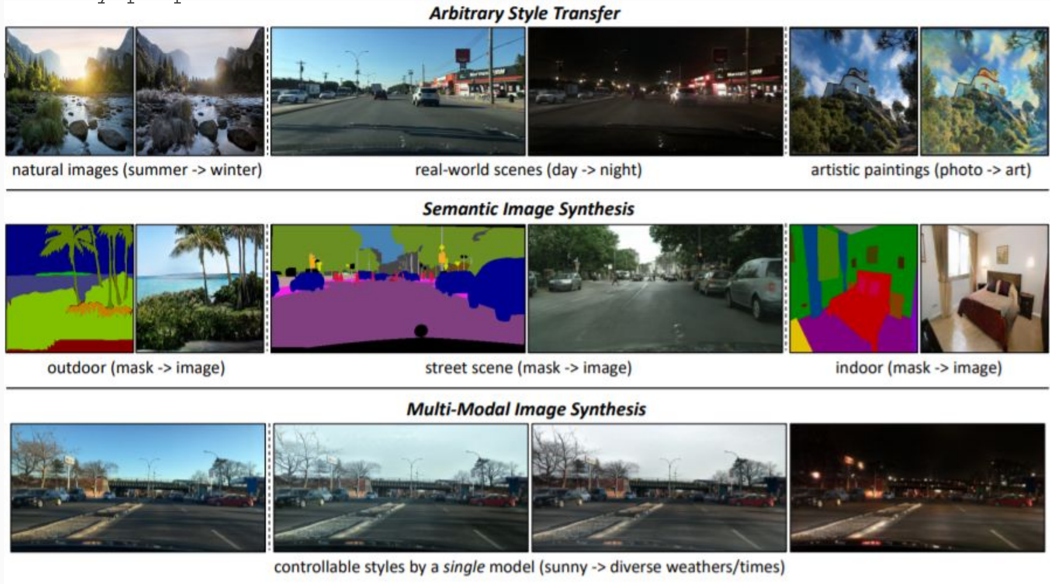
https://github.com/EndlessSora/TSIT

INTRODUCTION

- Background:
  - Image-to-image translation aims at translating one image representation to another
    - Generative Adversarial Networks (GANs) have made remarkable success in various of such tasks
  - Previous studies usually present specialized solutions for a specific from of applications, such as
    - arbitrary style transfer in unsupervised setting
    - Semantic images synthesis in supervised setting
- Goals:
  - Devise a general and unified framework that is applicable to different image-to-image translation tasks without degradation in synthesis quality
    - This is challenging, since it's difficult to cross-use these especially designed components from each specific task or integrate them into a unified framework
      - Certain conditional image synthesis tasks (e.g. arbitrary style transfer) do not have paired data available
        - In this unsupervised settings, translation task demands constraints on cycle consistency, semantic features, pixel gradients, or pixel values.
      - Semantic image synthesis (e.g. translation from segmentation labels to images) is more data-dependent and typically needs losses to minimize per-pixel distance between the generated sample and ground truth
      - Specialized structures are usually required to maintain spatial coherence and resolution

INTRODUCTION (CONT')

- Contribution: **T**wo-**S**tream **I**mage-to-image **T**ranslation (TSIT) framework.
  - A simple and versatile framework for image-to-image translation.
    - For unsupervised arbitrary style transfer: diverse scenarios can be handled (e.g. natural images, real-world, art paint)
    - For supervised semantic image synthesis: robust to different scenes (e.g. outdoor, street scenes, indoor).
  - Provided a two-stream generative model with newly proposed feature transformations in a coarse-to-fine fashion.
    - Allows multi-scale semantic structure information and style representation to be effectively captured and fused by the network
    - Permitting the proposed method to scale to various tasks in both unsupervised and supervised settings.
  - No additional condition (e.g. cycle consistency) are needed, contributing a clean and simple method
  - Possible to work with multi-modal image synthesis with arbitrary style control



- Contribution (cont'):
  - Differentiation:
    - Instead of only consider either semantic structure or style representation, both the structure and style in multi-scale feature levels are factorized, via a symmetrical two-stream network
    - The two streams jointly influence the new image generation in a coarse-to-fine manner via a consistent feature transformation scheme.
      - The content spatial structure is preserved by an element-wise feature adaptive denormalization (FADE) from the content stream
      - While the style information is exerted by feature adaptive instance normalization (FAdaIN) from the style stream
    - Standard loss functions such as adversarial loss and perceptual loss are used
    - No need additional constraints like cycle consistency
    - Pipeline is applicable to both unsupervised and supervised settings, easing the preparation of data

RELATED WORK

- Image-to-image translation:
  - Existing methods can be classified into two categories
    1. Unsupervised
       - Unsupervised image-to-image translation problem is inherently ill-posed*, where additional constraints are needed.
    2. Supervised
       - More data-depended, requiring well-annotated paired training samples
  - Limited by learning only one-to-one mapping between two domains, some GAN-based methods suffer from generating images with low diversity
  - Multi-domain translation and multi-modal translation significantly increase generation diversity.
  - Multi-mapping translation is defined in recent work. E.g. DMIT is designed to capture the multi-modal image nature in each domain
  - Existing methods lack the scalability to adapt to different tasks under diverse difficult settings
    - Which lead to suboptimality for cross-using these components due to either degradation in quality or introduction of additional constraints
- Arbitrary style transfer
  - Aim at retaining the content structure of an image, while manipulating its style representation adopted from others
  - Classical methods now can process in real-time as well as transfer multiple styles during inference
  - Many studies improve stylization via wavelet transform, graph cuts, or iterative error-correction
  - Some GAN-based methods show impressive results

* ill-posed: inverse of well-posed
well-posed: describe a problem has a uniquely determined solution that depends continuously on its data

- Semantic image synthesis
    - Aim at synthesizing a photorealistic image from a semantic segmentation mask
        - A special form of supervised image-to-image translation
        - The domain gap for this tasks is large
            - Keeping effective semantic information to enhance fidelity without losing diversity is challenging
    - *Pix2pix* first adopts conditional GAN in the semantic image synthesis task.
    - *pix2pixHD* contains a multi-scale generator and multi-scale discriminators to generate high-resolution images
    - SPADE takes a noise map as input, and resizes the semantic label map for modulating the activations in normalization layers by a learned affine transformation
    - Etc.

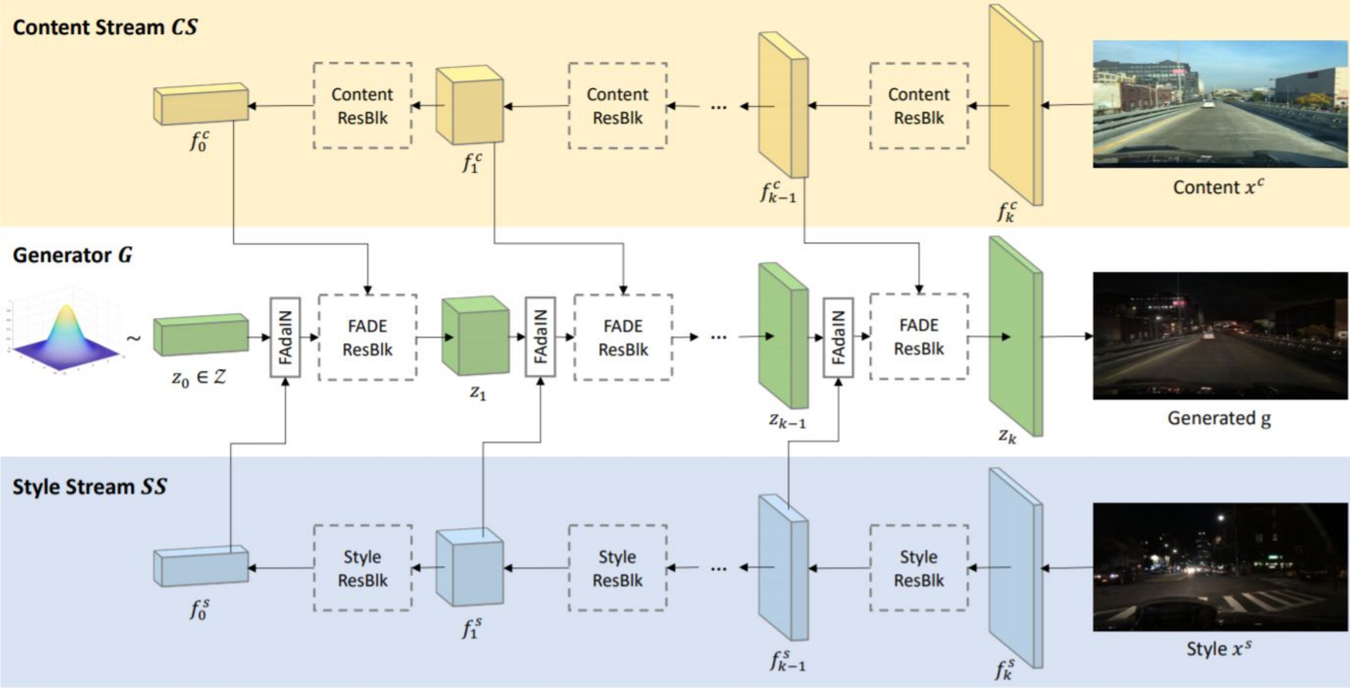
* ill-posed: inverse of well-posed
  well-posed: describe a problem has a uniquely determined solution that depends continuously on its data

METHODOLOGY - OVERVIEW

- Consider 3 key requirements in formulating a robust and scalable method to line various tasks:
  1. Both *semantic structure information* and *style representation* should be considered and fused adaptively
  2. The *content* and *style information* should be learned by networks in *feature level* instead of in *image level* to fit the nature of diverse semantic tasks
  3. The network structure and loss functions should be simple for easy training without additional constraints
- The methodology will be introduced in the following 3 sub-sections
  1. Network structure
  2. Feature transformation scheme
  3. Objective functions

METHODOLOGY - NETWORK STRUCTURE (COMPONENT LEVEL)

- As shown in the image, TSIT consists 4 component: content stream, style stream, generator, and discriminators (omitted)
    - First 3 components are fully convolutional and symmetrically designed
- Submodules (content residual block, style residual block, FAD residual block, FAD model in FAD residual block) are discussed in the next page
- Content / Style Stream:
    - Based on residual block
    - Two-stream network
    - Symmetrical with the same network structure
    - Aiming at extracting corresponding feature representations in different level
    - To extract features and feed them to the corresponding feature transformation layers in the generator.
    - Multi-scale content/style representation can be Learned by the stream, adaptively fitting different feature transformations



METHODOLOGY - NETWORK STRUCTURE (COMPONENT LEVEL)

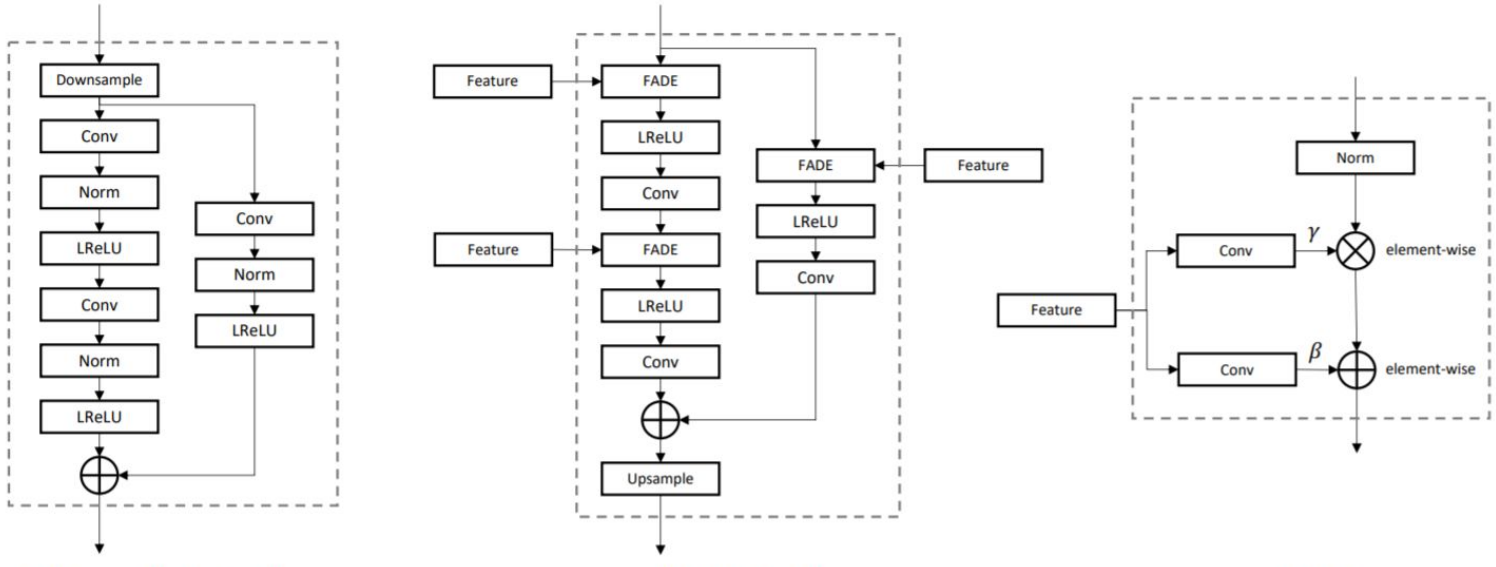
- Generator:
  - Generator has a inverse structure w.r.t. The content/style stream.
  - Designed to consistently match the level of semantic abstraction at different feature scales.
  - A noise map is sampled from a Gaussian distribution as the latent input
  - Feature maps from corresponding layers in content/style stream are taken as multi-scale feature inputs
  - The feature transformation are implemented by a FADE residual block
    - The FADE module, which is replaced the batch normalization layer in the FADE residual block, perform element-wise denormalization by modulating the normalized activation using a learned affine transformation defined by the modulation parameters γ and β
    - The FAdaIN module is used to exert style information through feature adaptive instance normalization
  - The entire generation process is performed in a coarse-to-fine manner.
    - Mutli-scale content/style features are injected to refine the generated image constantly from high-level latent code to low-level image representation
    - Semantic structure and style information are learnable and effectively fused in an end-to-end training
- Descriminators:
  - 3 regular discriminators with an identical architecture are included to discriminate images at different scales.
  - Patch-based training allows the discriminator operating at the coarsest scale to have the largest receptive field
    - Capturing global information of the image
  - Multi-scale patch-based discriminators further improve the robustness of our method for image-to-image translation task in different resolutions
  - The discriminators also serve as feature extracts fro the generator to optimize the feature matching loss

METHODOLOGY - NETWORK STRUCTURE (COMPONENT LEVEL)

- Generator:
  - Generator has a inverse structure w.r.t. The content/style stream.
  - Designed to consistently match the level of semantic abstraction at different feature scales.
  - A noise map is sampled from a Gaussian distribution as the latent input
  - Feature maps from corresponding layers in content/style stream are taken as multi-scale feature inputs
  - The feature transformation are implemented by a FADE residual block
    - The FADE module, which is replaced the batch normalization layer in the FADE residual block, perform element-wise denormalization by modulating the normalized activation using a learned affine transformation defined by the modulation parameters γ and β
    - The FAdaIN module is used to exert style information through feature adaptive instance normalization
  - The entire generation process is performed in a coarse-to-fine manner.
    - Mutli-scale content/style features are injected to refine the generated image constantly from high-level latent code to low-level image representation
    - Semantic structure and style information are learnable and effectively fused in an end-to-end training

METHODOLOGY - NETWORK STRUCTURE (SUBMODULE LEVEL)

- Content/Style residual blocks:
  - Each block has three convolutional layers, one of which is designed for the learned skip connection
  - Leaky ReLU is used as the activation function
- Feature transformation:
  - Implemented by a FADE residual block and a FAdaIN module
  - In the FADE residual block, the batch normalization layer are replaced with the FADE module to match the inverse architecture w.r.t. the content/style residual block



(a) Content/Style ResBlk (b) FADE ResBlk (c) FADE

- Overview:
  - A new feature transformation scheme is proposed
    - considering both semantic structure information and style representation
    - And fuse them adaptively
- Feature Adaptive Denormalization (FADE):
  - Inspired by spatially adaptive denormalization (SPADE)
  - Differentiation:
    - SPADE resize a semantic masks as its input, whereas we generalize the input to multi-scale *feature representation* of the content image
      - I.e. we can fully exploit semantic information captured by the content stream
  - The denomalization operation is element-wise, and the parameter, γ and β, are learned by one-layer convolutions from the *feature representation* in the FADE module
    - FADE experience more perceptible influence from coarse-to-fine *feature representations*
      - I.e. it can better preserve semantic structure information.
- Feature Adaptive Instance Normalization (FAdaIN):
  - Used to better fuse style representation
  - Inspired by adaptive instance normalization (AdaIN), with a generalization to enable the style stream SS to learn multi-scale feature-level style representation of the style image more effectively
  - Through FAdaIN, coarse-to-fine style features at different layers can be fused adaptively with the corresponding semantic structure features learned by FADE
    - Allowing the framework to be trained end-to-end and versatile to different tasks
    - Multi-modal image synthesis is made possible with arbitrary style control

METHODOLOGY - OBJECTIVE

- Standard losses are used in the objective function
- For generator: we applied hinge-based adversarial loss, perceptual loss, and feature matching loss
  - Perceptual loss minimized the difference between the feature representation extracted by VGG-19 network
  - Feature matching loss matches the intermediate features at different layers of multi-scale discriminator
- For multi-scale discriminators, only hinge-based adversarial loss is used to distinguish whether the image is real or fake
- The generator and discriminator are trained alternately to play a min-max game
- Due to the simple objective functions, our framework is stable and easy to train
- Because of the two-stream network, the typical KL loss for multi-modal image synthesis becomes optional

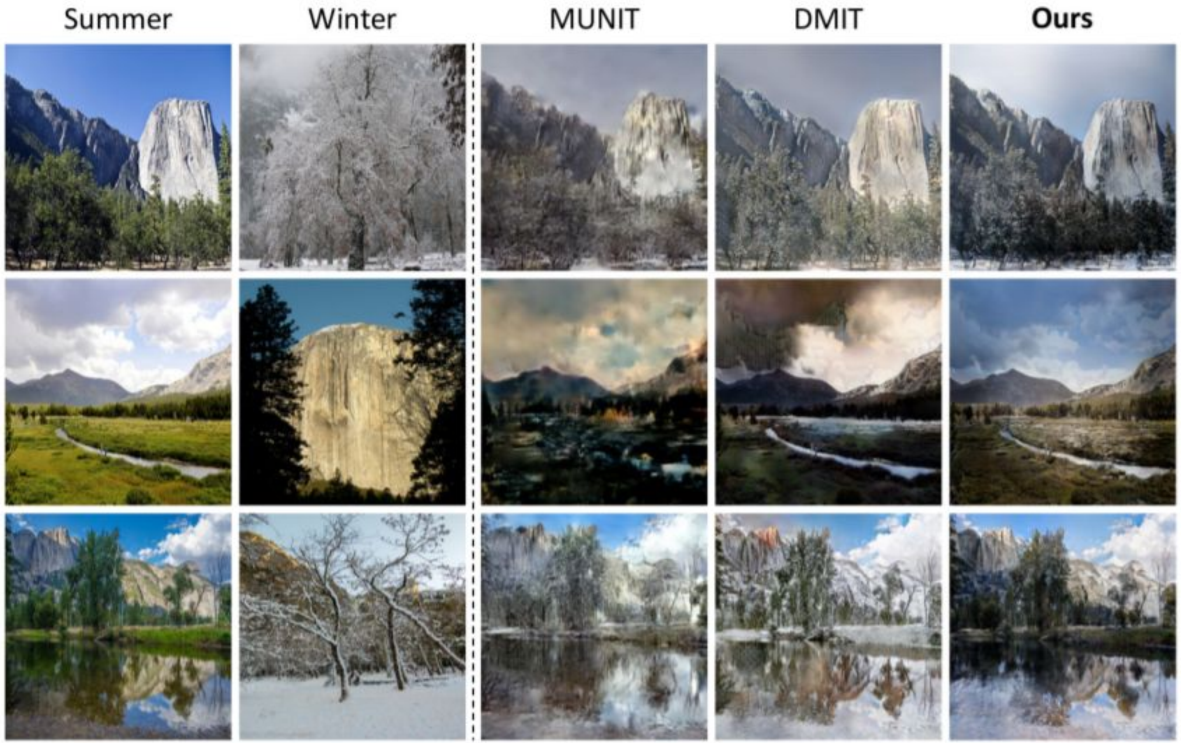


Fig. 4. Yosemite summer → winter season transfer results compared to baselines.

RESULTS

Fig. 5. BDD100K day → night time translation results compared to baselines.

Fig. 6. Photo → art style transfer results compared to baselines.

Table 1. The FID and IS scores of our method compared to state-of-the-art methods in arbitrary style transfer tasks. A lower FID and a higher IS indicate better performance.

Methods	summer → winter		day → night		photo → art	
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
MUNIT [14]	118.225	2.537	110.011	2.185	167.314	3.961
DMIT [50]	87.969	2.884	83.898	2.156	166.933	3.871
Ours	**80.138**	**2.996**	**79.697**	**2.203**	**165.561**	**4.020**

RESULTS

Fig. 7. Semantic image synthesis results compared to baselines.

RESULTS

Table 2. The mIoU, pixel accuracy (accu) and FID scores of our method compared to state-of-the-art methods in semantic image synthesis tasks. A higher mIoU, a higher pixel accuracy (accu) and a lower FID indicate better performance.

Methods	Cityscapes mIoU ↑	Cityscapes accu ↑	Cityscapes FID ↓	ADE20K mIoU ↑	ADE20K accu ↑	ADE20K FID ↓
CRN [4]	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [35]	47.2	75.5	**49.7**	N/A	N/A	N/A
pix2pixHD [42]	58.3	81.4	95.0	20.3	69.2	81.8
SPADE [34]	62.3	81.9	71.8	38.5	79.9	33.9
CC-FPSE [29]	65.5	82.3	54.3	**43.7**	**82.9**	31.7
Ours	**65.9**	**82.7***	59.2	38.6	80.8	**31.6**

Fig. 9. Cross validation of ineffectiveness of task-specific methods in inverse settings.