Exploration and Application of Deep Representation Editing in Computer Vision Tasks

Presented by

Chaoyue Wang (Postdoc,

Supervised by Prof. Daocheng Tao)

School of Computer Science

THE UNIVERSITY OF SYDNEY

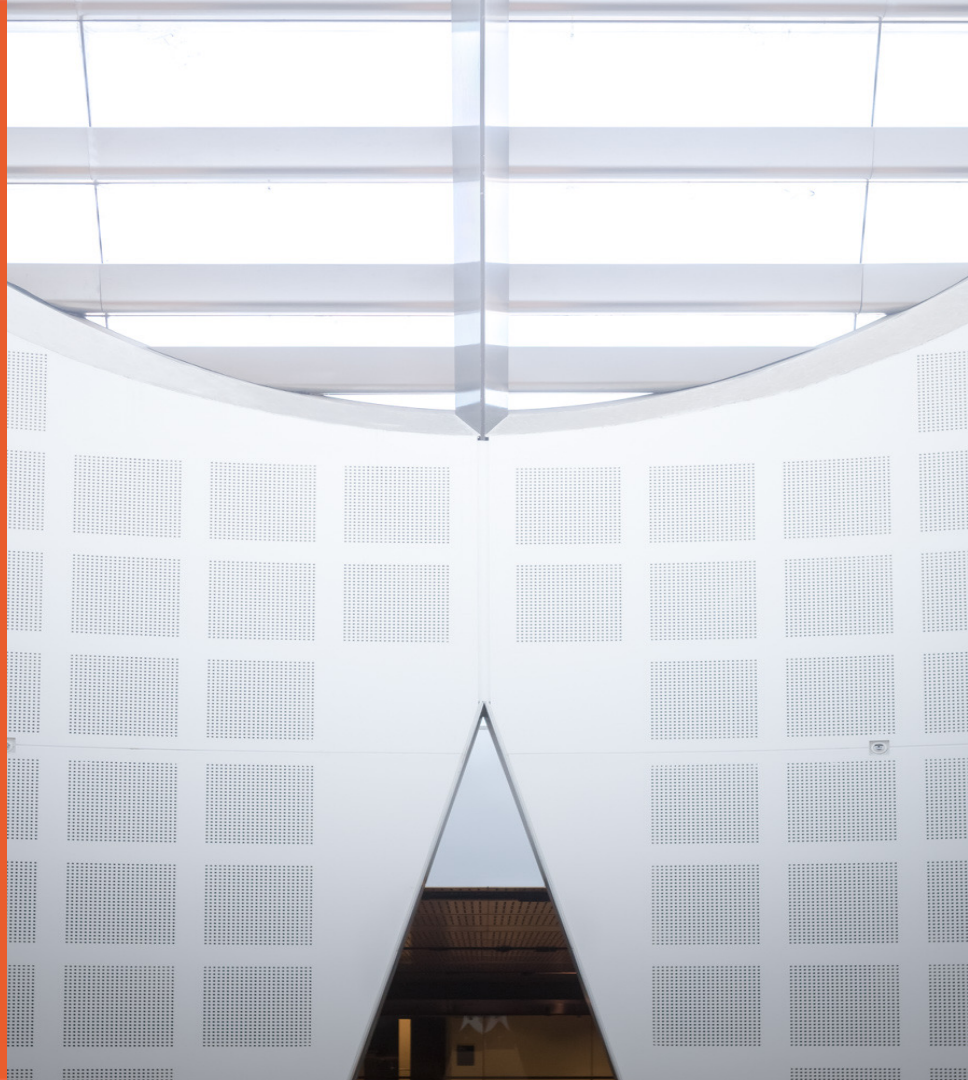
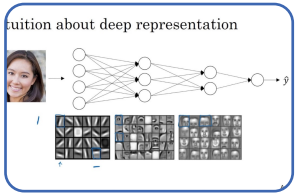
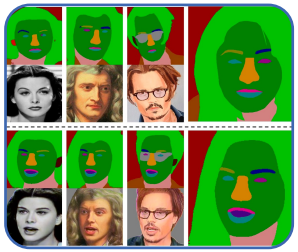


Table of Contents



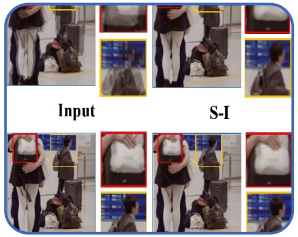
Deep Representations: Brief Introduction

- Deep Learning and Representation learning
- Why Deep Representations ?
- Deep Representation: Learning & Editing



PuppeteerGAN: Arbitrary Portrait Animation with Semantic-aware Appearance Transformation (CVPR-2020)

- Portrait Animation and Related Applications
- Existing Challenges and Our Motivation
- The Proposed method and Experimental results



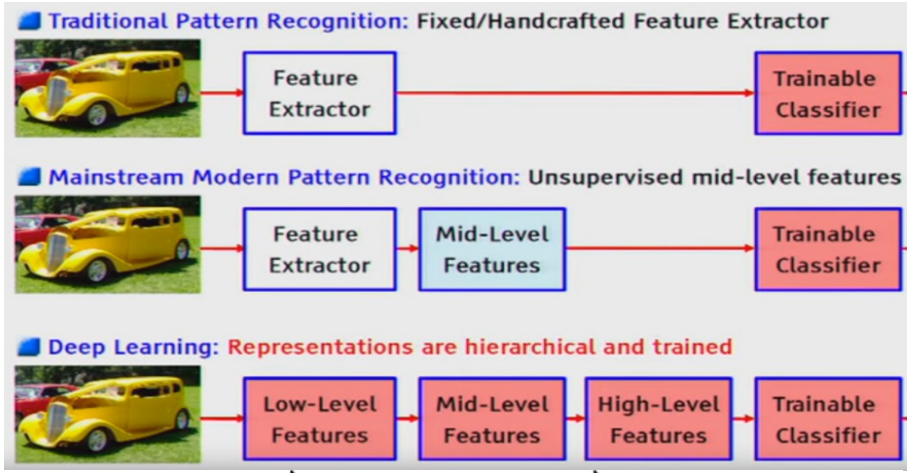
FeatureFlow: Robust Video Interpolation via Structure-to-texture Generation (CVPR-2020)

- Video Frame Interpolation and Related Applications
- Conventional Video Frame Interpolation: Optical Flow based Methods
- The proposed FeatureFlow and experimental results

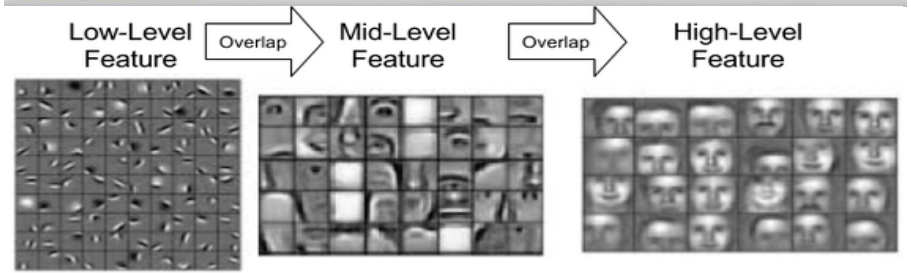
Deep Learning and Representation learning

Deep Learning:

- **Deep learning** (also known as deep structured learning) is part of a broader family of machine learning methods based on **artificial neural networks** with **representation learning**. (Wikipedia)
- **Representation learning:** The success of machine/deep learning algorithms generally depends on **data representation**, and we hypothesize that this is because different **representations can entangle and hide more or less the different explanatory factors of variation behind the data**.
(Bengio *et al*, 2014)



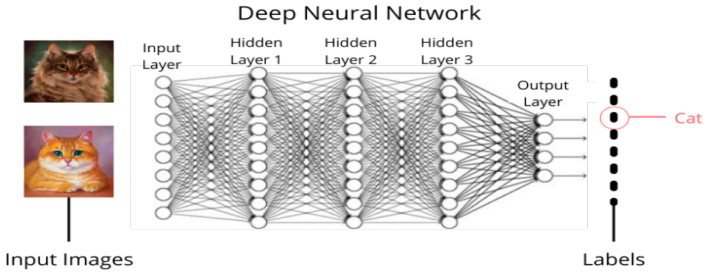
(Y LeCun)



Why should we care about deep representations ?

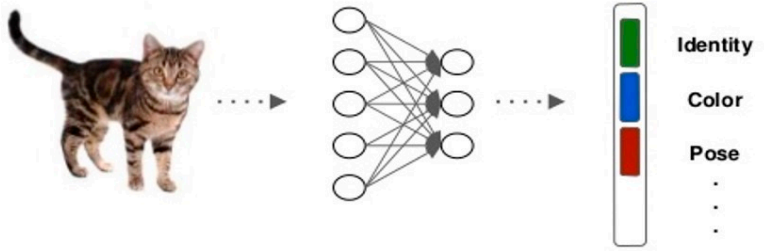
Perceiving tasks:

- Classification, regression, object detection, Speech Recognition, Natural Language Processing, etc.



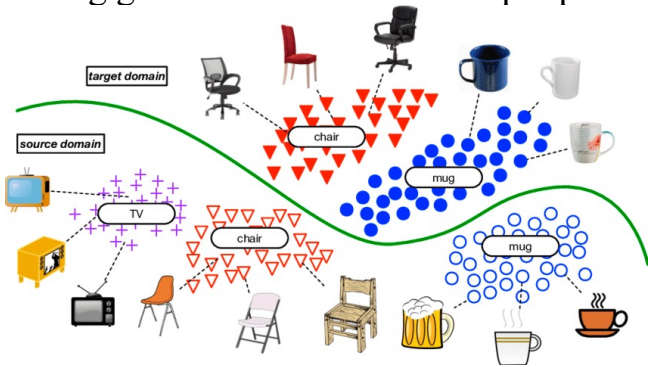
Disentangled representation learning:

- Disentangling the factors of variation. Different explanatory factors of the data tend to change independently.



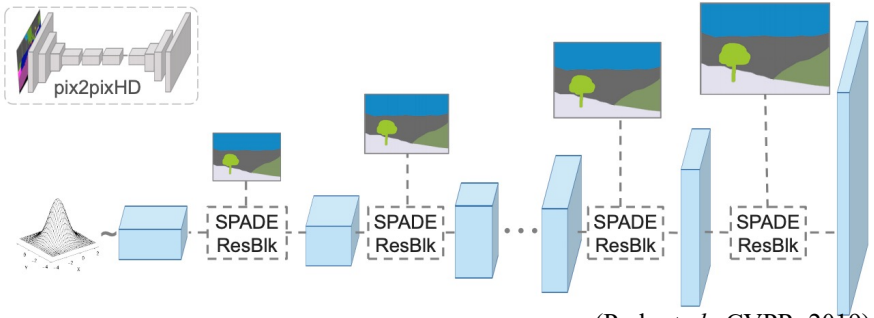
Multi-Task and Transfer Learning:

- Learning general and shareable deep representations.



Generative models & mapping to real data:

- Leaned to mapping low-dimensional, semantic representations to real-world data (e.g. images).



(Park *et al.*, CVPR, 2019)

Deep Representation/Feature

***Deep representation/Feature Learning & Editing*:**

- Semantic-aware; Hieratical structure; Spatial invariance (CNN feature/representation); Semantic Robust; ……

From

Perceiving

to

creation/ imaging

From

Learning

to

Editing

PuppeteerGAN: Arbitrary Portrait Animation with Semantic-aware Appearance Transformation

Zhuo Chen[1,2] Chaoyue Wang[2] Bo Yuan[1] Dacheng Tao[2]

[1]Shenzhen International Graduate School, Tsinghua University
[2]UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia

Portrait Animation and Related Applications

Portrait Animation:

- **Definition:**

  Animating a given still portrait image to life using poses (or expression information) extracted from target video frames.

- **Potential Applications:**

  1) The film industry;
  2) Art-making;
  3) Personalized media generation.

Existing methods:

1) Video-to-video / Conditional image synthesis based methods
2) Deformation/Warping based methods
3) Disentanglement learning based methods

Challenges:

1) Identity/personality mismatch.
2) Training data/domain limitations.
3) Low-efficiency in training/fine-tuning.

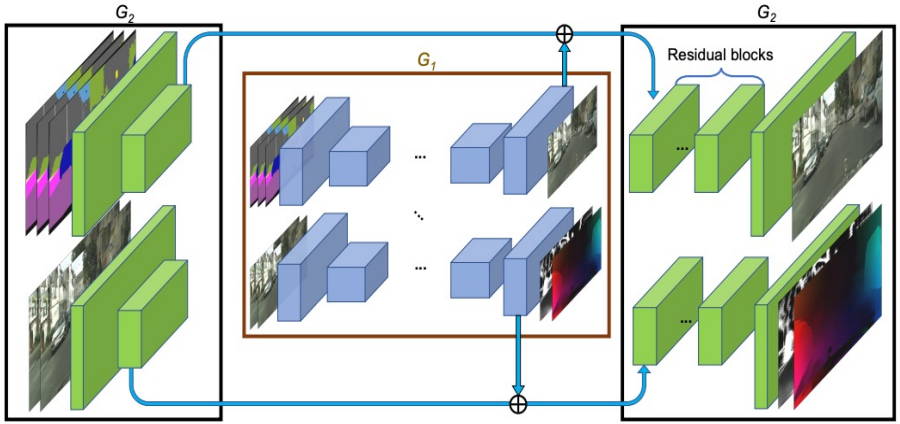
Source Target Output

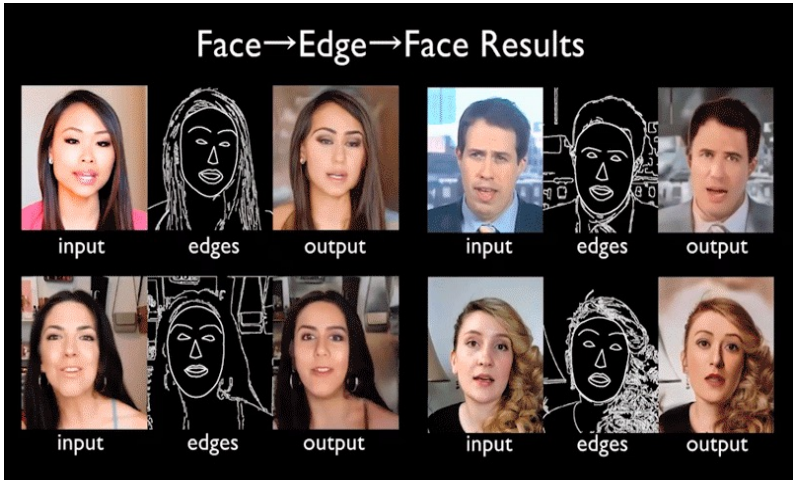
Source Target Output

Existing Methods and Challenges

Video-to-video / Conditional image synthesis based methods:

- Training **a specific network for every person** can boost the quality of the generated portraits significantly.
- Vid2Vid can generate temporally coherent videos conditioned on **sketches or poses**. Its few-shot adaptive variation learns to **synthesize videos of previously unseen subjects by leveraging few test examples**.
- Although video-to-video methods could generate realistic portrait, the identity specific training encounters the difficulty of **collecting training data, high computational cost and long inference time**.



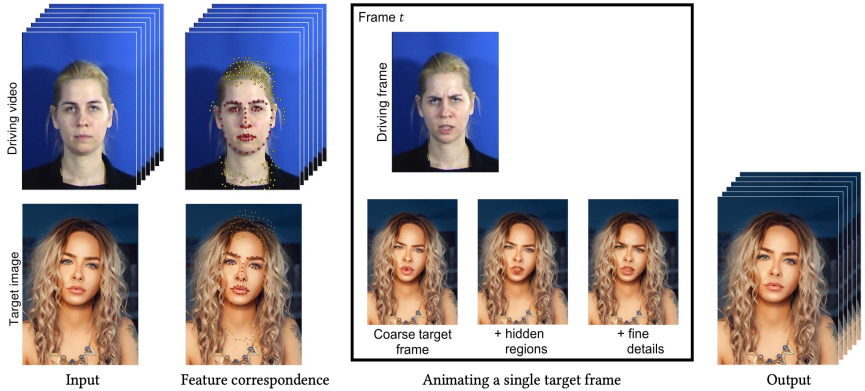


(Wang *et al.*, NeurIPS, 2018)

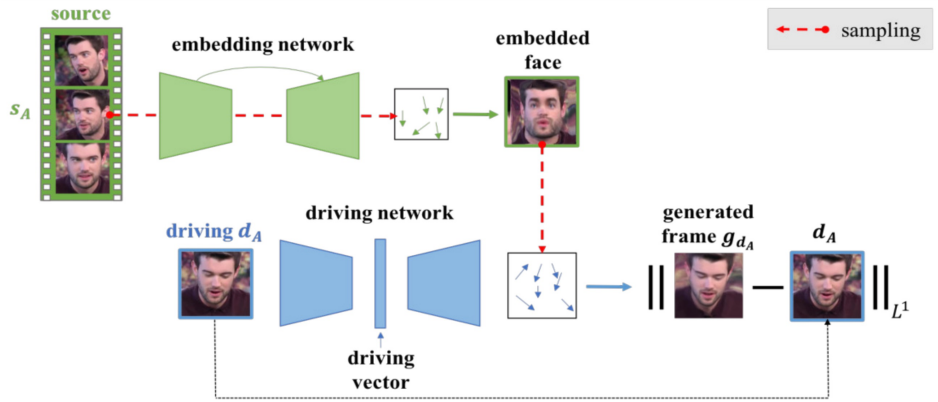
Existing Methods and Challenges

Deformation/Warping based methods :

- One detects **additional key points around the face** on the source and target image in order to control the deformation of the whole head. By copying the mouth region from the target, this method can manipulate facial expression **without any training or extra database on a still portrait**.
- X2face learns a pair **2D pixel-wise deformation** from the source pose to the frontal and the target pose which can be extracted from multiple optional media including video, audio and pose angle.
- Although these methods are efficient in transferring facial expression in both identity fidelity and face reality, they would **raise distortions when dealing with a large pose changes**.



(Averbuch-elor *et al.*, TOG, 2017)

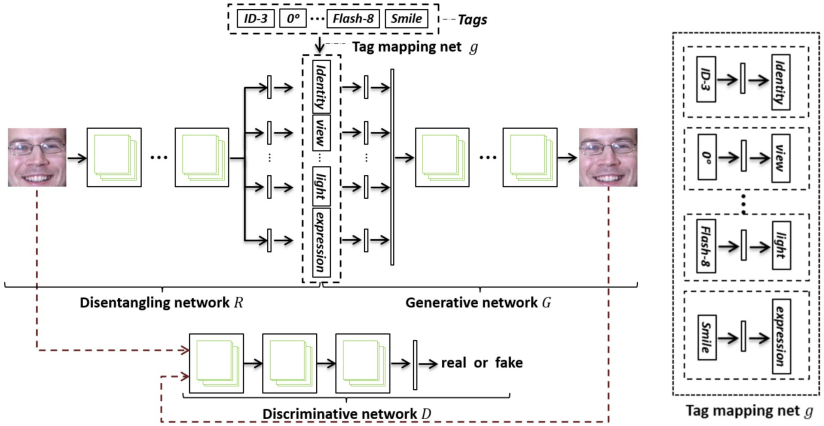


(Wiles *et al.*, ECCV, 2018)

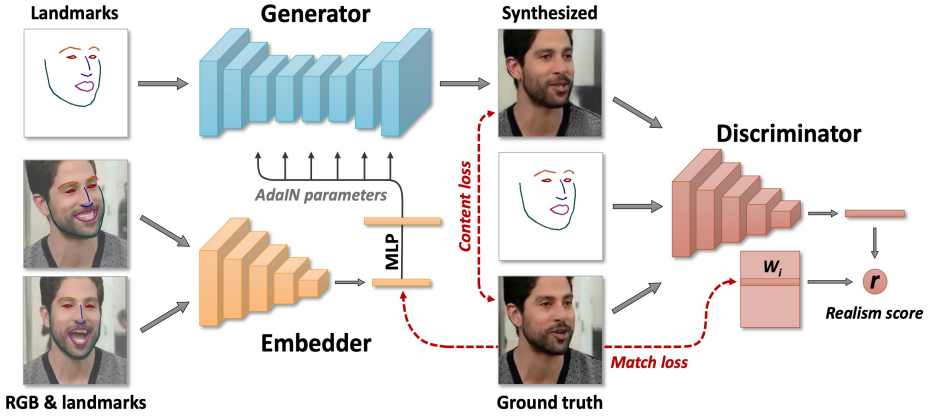
Existing Methods and Challenges

Disentanglement learning based methods :

- Decompose the **disentangled appearance and pose representations** from the input portrait image. Ideally, by **recombining the appearance feature of the source portrait with the pose feature of the target frames**, the generator/decoder is supposed to generate the desired outputs.
- Given the **landmarks** of the target frame as input, the network is trained to reconstruct the target frame by **conditioning on appearance information** extracted from other frames of the same video/person.
- Though promising progress has been made, these kind of methods may face 1) representation learning difficulties, 2) identity misalignment, 3) domain gap.



(Wang *et al.*, IJCAI, 2017)

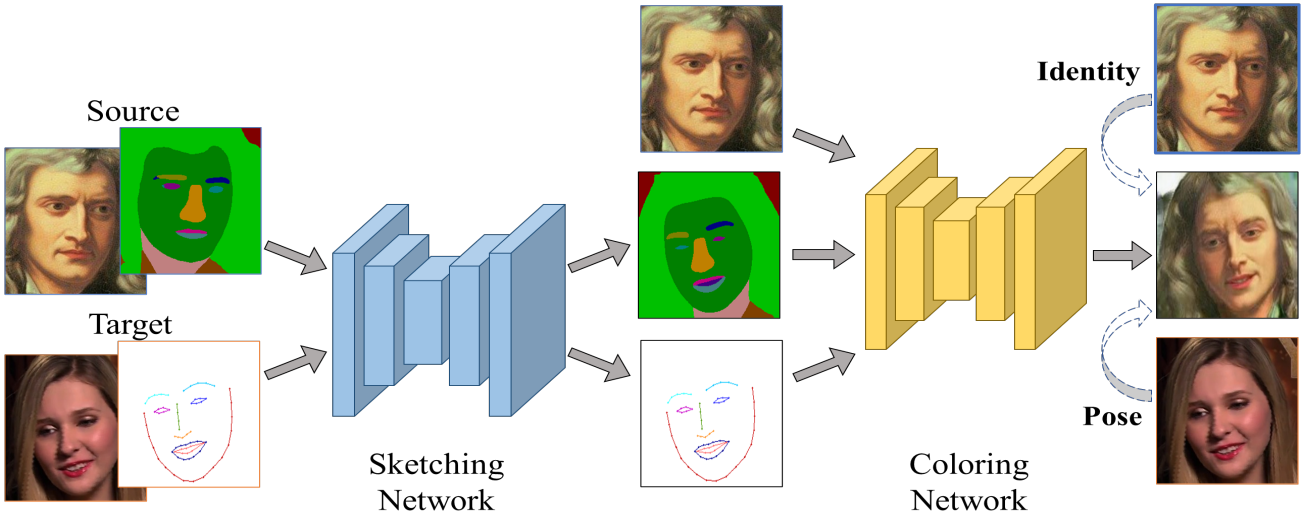


(Zakharov *et al.*, ICCV, 2019)

PuppeteerGAN Framework

Motivation and overall framework:

- We separate portrait animation into two stages: **pose retargeting** and **appearance transformation**.
- **Sketching Network** (Pose retargeting)
  Identity-preserved pose retargeting between the **semantic segmentation** of **any portraits**.
- **Coloring Network** (Appearance transformation)
  Fill in the animated semantic segmentation mask with the **appearance of the source portrait**. The coloring network **makes full use of the deep representation extracted by the encoder**.

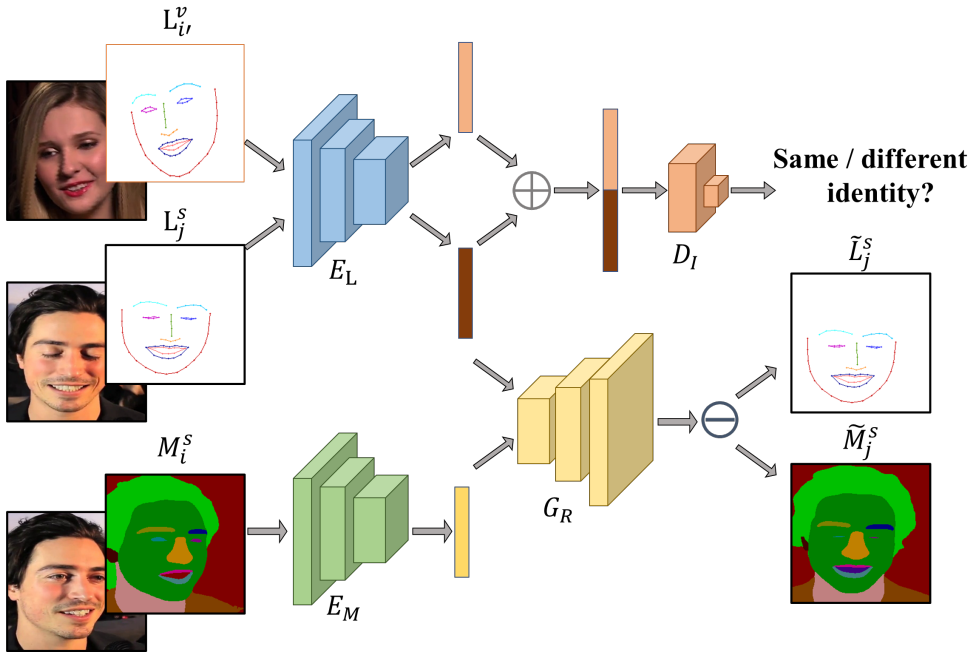


PuppeteerGAN Framework

Sketching Network:

- The sketching network is trained to synthesize the **animated segmentation masks and landmarks that keep characteristic details** (e.g. facial shape, hairstyle) of the source portrait yet with the same pose as the driven frame.
- Since the generated results could act as a general representation of different kinds of portraits, we can simply train the sketching network on a specific talking-head video dataset but perform **inference on arbitrary portraits**.
- **Identity-preserved pose retargeting:**

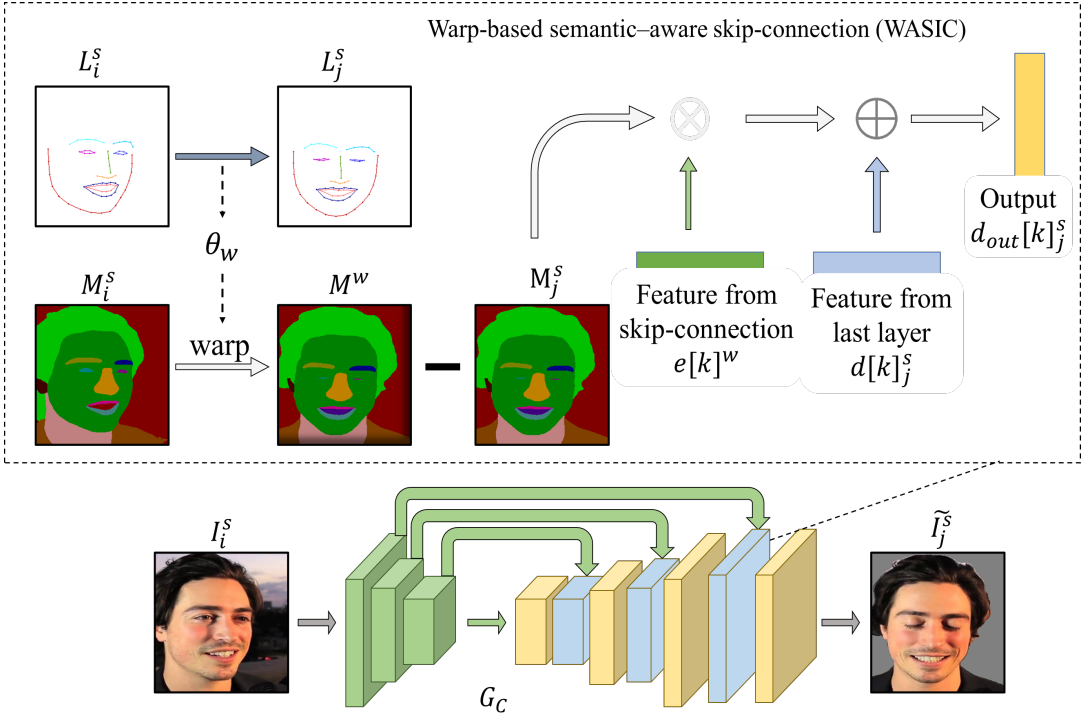
$$\mathcal{L}_{idt}(E_L, D_M) = \min_{E_L} \max_{D_I} D_I(E_L(L_i^s), E_L(L_j^s)) - D_I(E_L(L_{i'}^v), E_L(L_j^s)).$$



PuppeteerGAN Framework

Coloring Network:

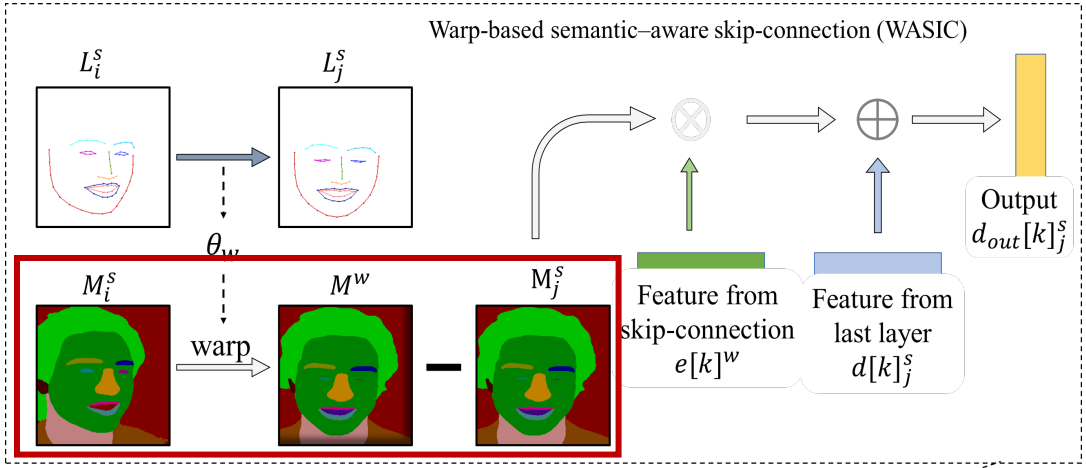
- In this stage, the challenge remains on the **appearance transformation** between different frames of the same person.
- We observe that, for the generated image , most of its appearance information could be directly found in the input image. Inspired by deformation based methods, we devised the **Warp-based semantic-aware skip-connection(WASIC)** for transforming these appearances.
- For unseen parts (*e.g.* open mouth), we hope that the coloring network could work as a **conditional generation network**, which is able to **imagine these parts based on the input images**.

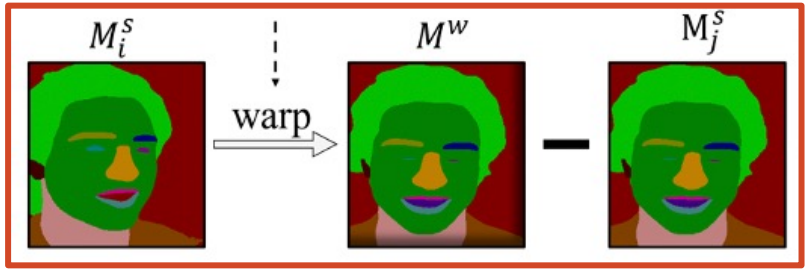


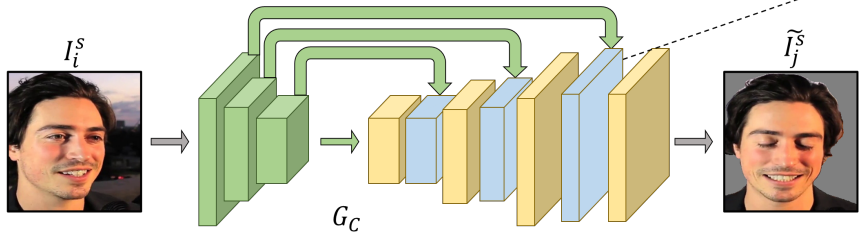
PuppeteerGAN Framework

Coloring Network - Geometry dropout strategy:

- We further expand the available data for training the coloring network **from video dataset to image dataset** through the proposed **geometry dropout strategy**.







Experimental Results

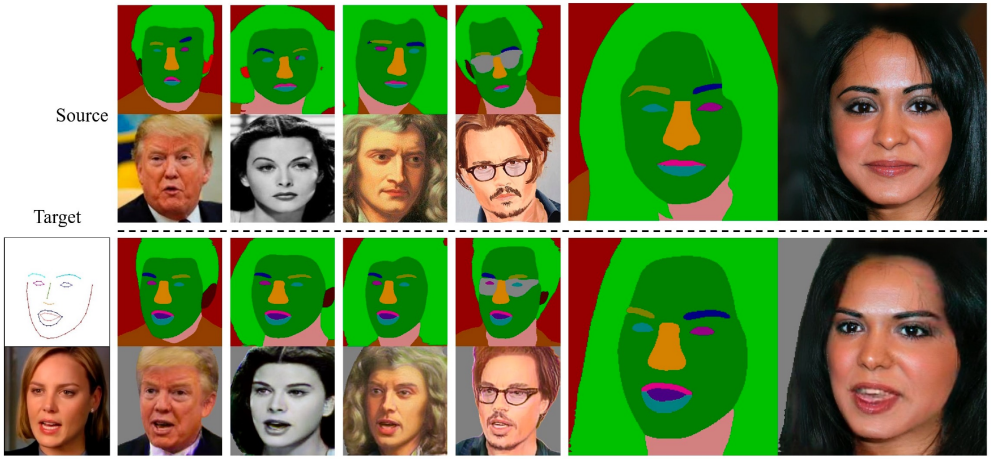
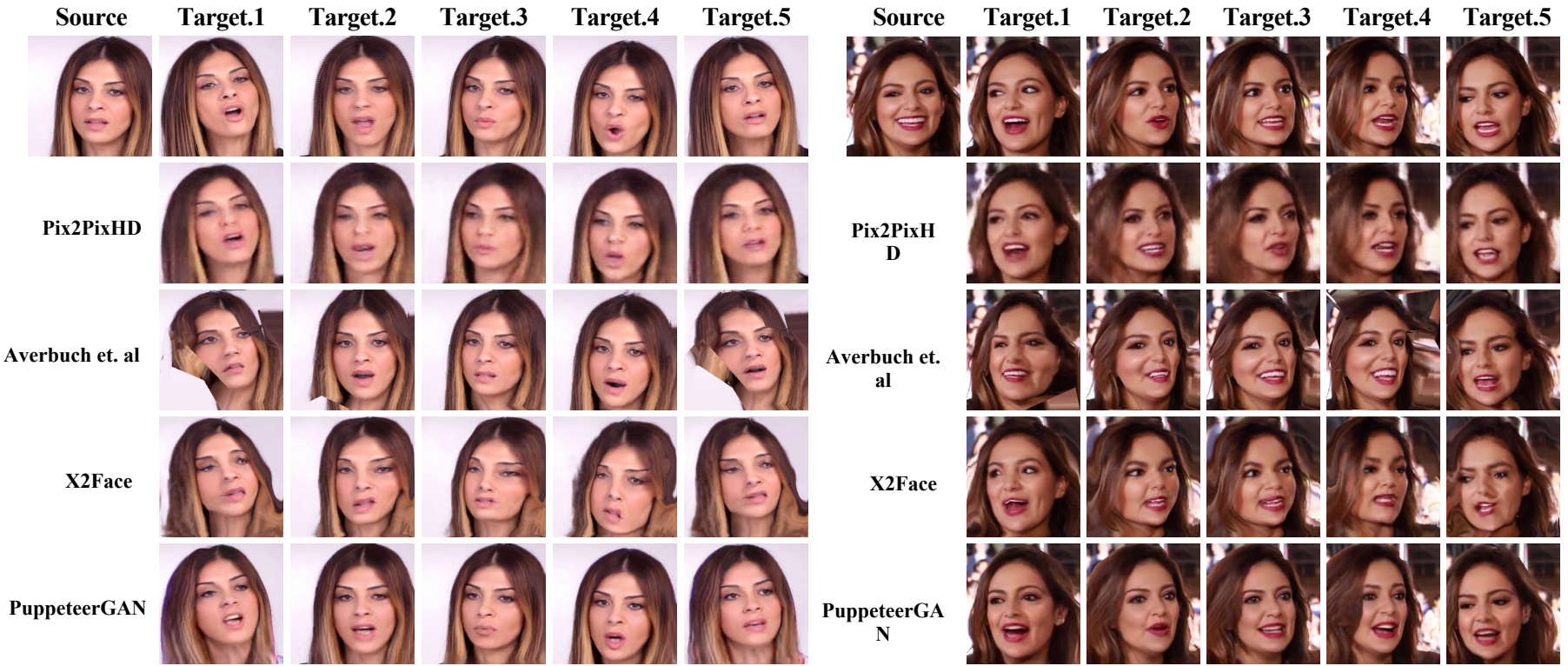


Figure 1: **Examples of animated portraits generated by the proposed PuppeteerGAN.** The results are at the same pose as the target frame (left column) while keeping the same appearance of the source (top row). As shown in the source images, our method can be applied to various portraits including color photos, black-and-white photos, paintings, cartoon characters and high-resolution images.

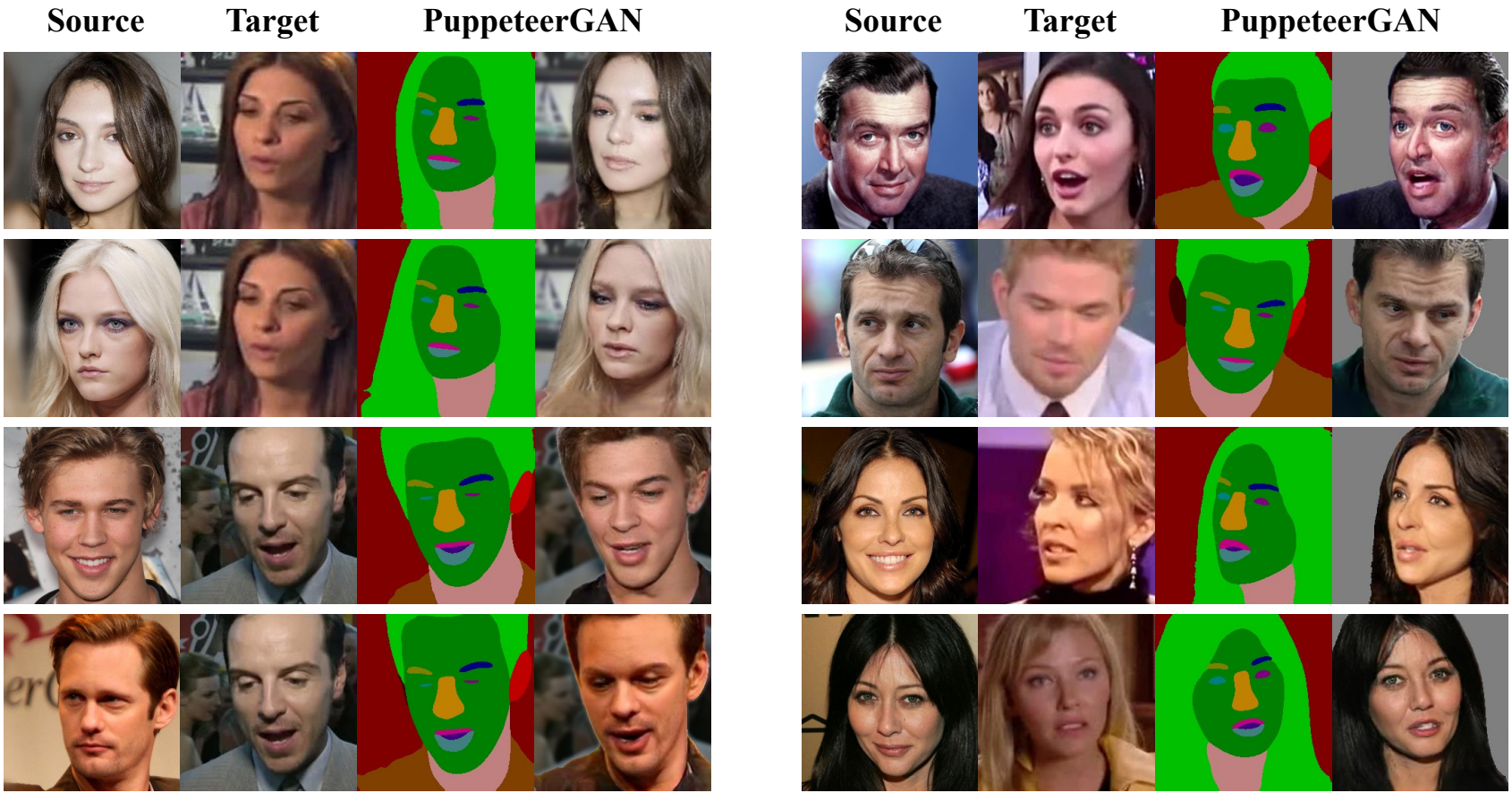
Methods	Vox [24]				
	SSIM↑	FID↓	PSNR↑	CSIM↑	MSE↓
Averbuch *et al.* [1]	0.6733	73.1115	**31.4702**	0.7600	0.1502
Pix2PixHD [41]	0.5500	70.3599	29.2691	0.4145	0.1876
PuppeteerGAN	0.7255	33.6119	31.3506	**0.8178**	**0.1033**
Zakharov *et al.* [45] (1)	0.6700	43.0000	-	-	-
X2Face [43] (1)	0.6800	45.8000	-	-	-
Pix2PixHD [41] (1)	0.5727	67.4887	29.6024	0.4789	0.1689
Zakharov *et al.* [41] (8)	0.7100	38.0000	-	-	-
X2Face [43] (8)	0.7300	51.5000	-	-	-
Pix2PixHD [41] (8)	0.5854	66.6279	29.8199	0.5117	0.1567
Zakharov *et al.* [41] (32)	0.7400	**29.5000**	-	-	-
X2Face [43] (32)	**0.7500**	56.5000	-	-	-
Pix2PixHD [41] (32)	0.6072	64.6087	30.2076	0.6082	0.1463
Vid2Vid [40]	0.6744	51.2171	31.4291	0.7715	0.1265

Methods	Time Cost (s)		
	Fine-tuning	Inference	Sum
Averbuch *et al.* [1]	-	0.9396	0.9396
X2Face [43]	4.3800	0.2257	4.6056
Pix2PixHD [41]	28.0415	0.2559	28.2974
Zakharov *et al.* [45]	48.7117	1.0220	49.7334
PuppeteerGAN	-	0.6117	**0.6117**

Experimental Results



Experimental Results

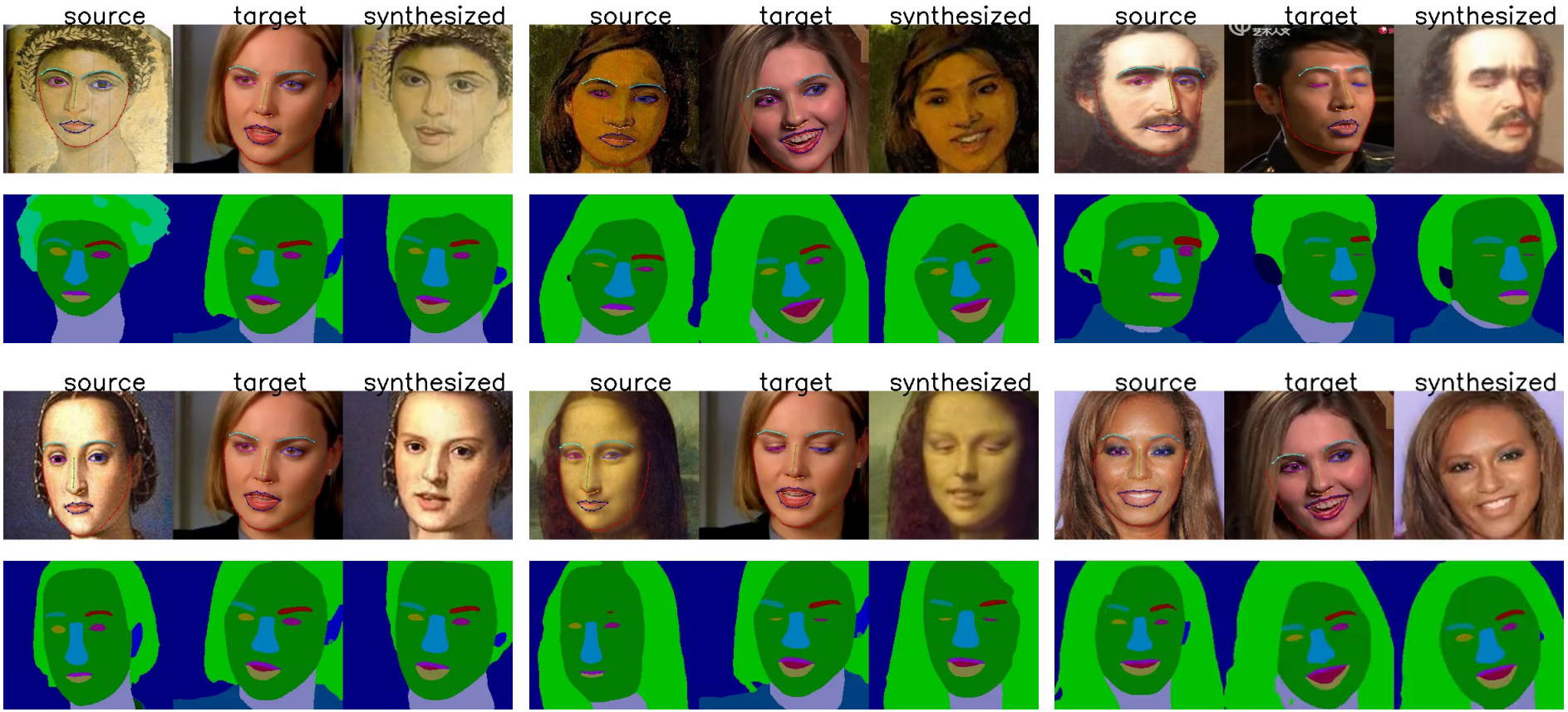


Experimental Results

Experimental Results

Experimental Results

Experimental Results



FeatureFlow: Robust Video Interpolation via Structure-to-texture Generation

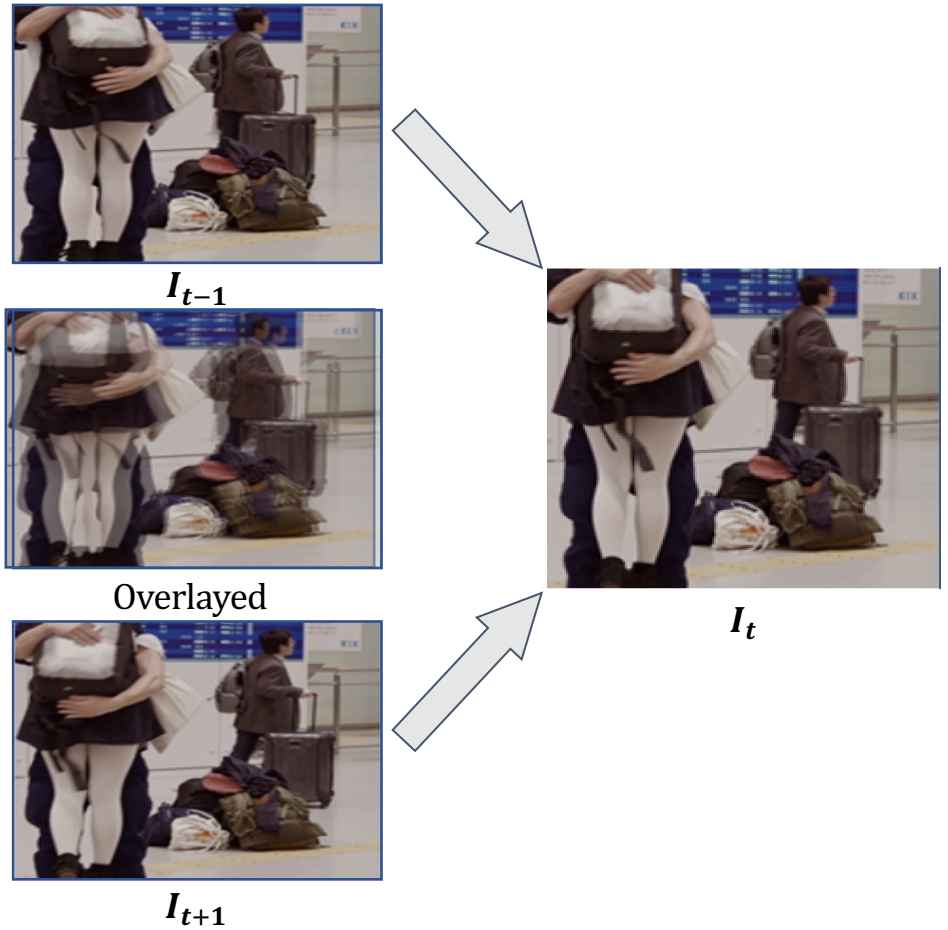
Shurui Gui*[1,2] Chaoyue Wang*[1] Qihua Chen[1,3] Dacheng Tao[1]

[1]UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia
[2]Department of Computer Science and Technology, University of Science and Technology of China
[3]School of Information Science and Technology, University of Science and Technology of China

Video Frame Interpolation and Related Applications

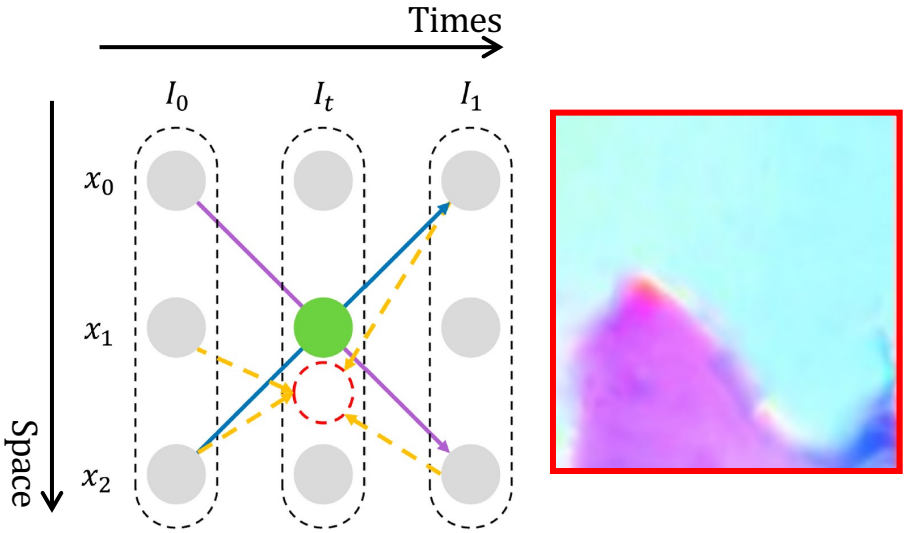
- **Video frame interpolation (VFI)** is an important research topic in the computer vision community. It aims to synthesize **unseen intermediate frames** between any two consecutive video frames.

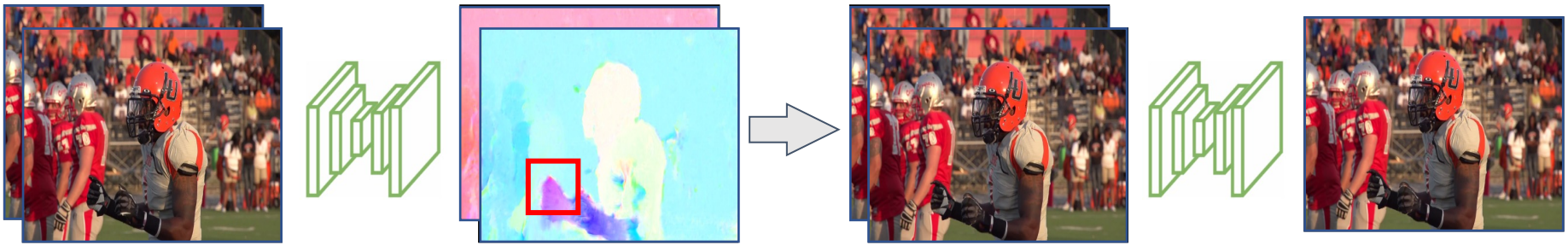
- Related techniques are wildly applied to real-world applications, such as **slow-motion production, frame rate upconversion, and video restoration.**



Conventional VFI Methods

- Optical Flow based methods can explicitly represent the dynamic motion and reach high fidelity in the details. However, due to the basic assumptions of the optical flow estimation, i.e., **smoothness and consistency**, optical flow based methods are inherently difficult to handle the interpolation of complicated dynamic scenes which include the regions suffering from **occlusion, blur or abrupt brightness change.**

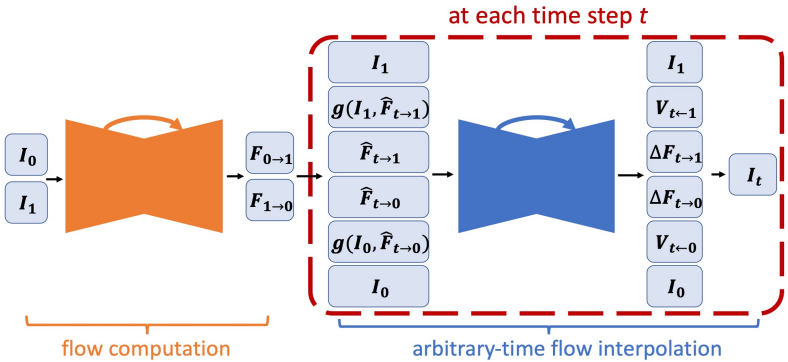




Input frames | Estimated Optical Flows | Warped frames | Interplation output

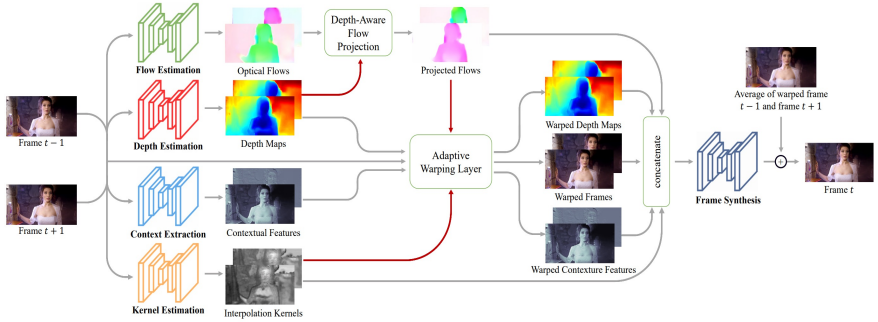
Conventional VFI Methods

Super SloMo:



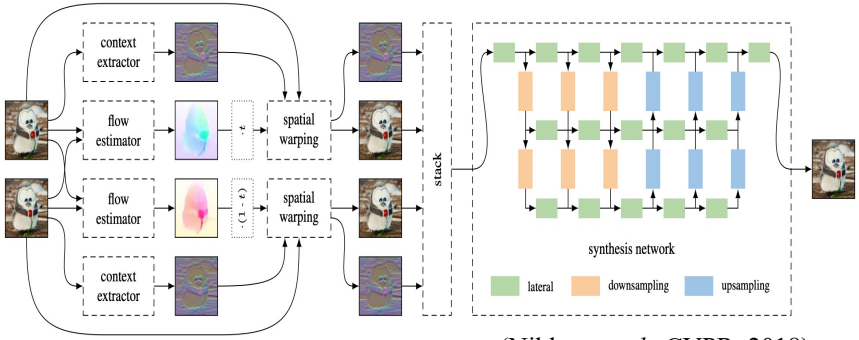
(Jiang *et al.*, CVPR, 2018)

Depth-aware VFI:



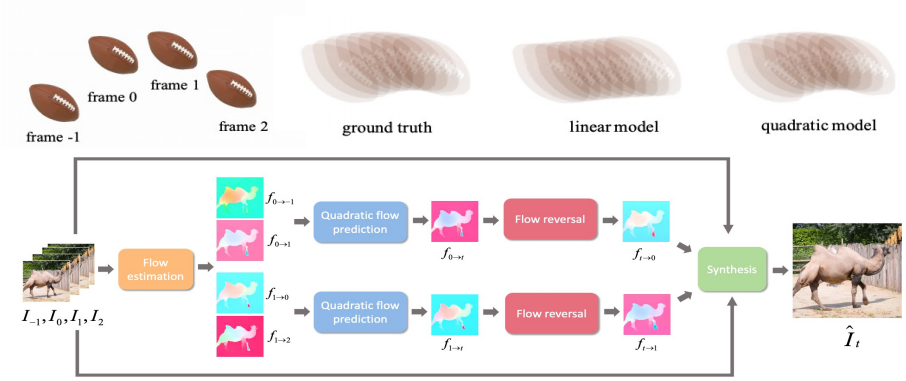
(Bao *et al.*, CVPR, 2019)

Context-aware Synthesis for VFI:



(Niklaus *et al.*, CVPR, 2018)

Quadratic Video Interpolation:



(Xu *et al.*, NeurIPS, 2019)

FeatureFlow Framework

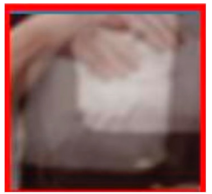
Motivation and overall framework:

- Instead of learning pixel-wise optical flow, our framework aims to explore **feature flows (FeFlow) in-between corresponding deep features**. To the best of our knowledge, this is **the first work that attempts to directly generate the intermediate frame** through blending deep features.
- We split the video interpolation task into two stages: **structure-guided interpolation** and **texture refinement**

Overlayed inputs

Stage – I

Stage – II



Inputs

GT

S – I

S – II

FeatureFlow Framework

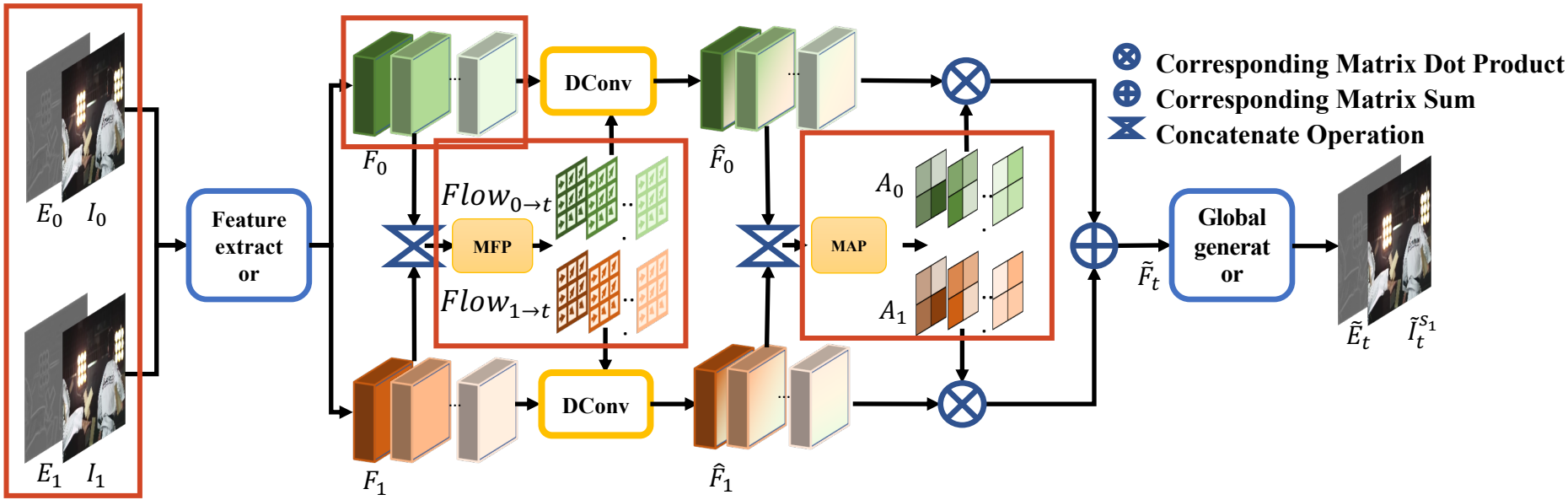
Motivation and overall framework:

- **Stage-I (structure-guided interpolation)**: deep structure-aware features are employed to predict feature flows from two consecutive frames to their intermediate result, and further generate the coarse result without detailed textures.
- **Stage-II (texture refinement):** through aligning original frames to the coarse result generated in the stage-I, a Frame Texture Compensator (FTC) is devised to synthesize the missed texture details.

FeatureFlow Framework

Structure-guided interpolation:

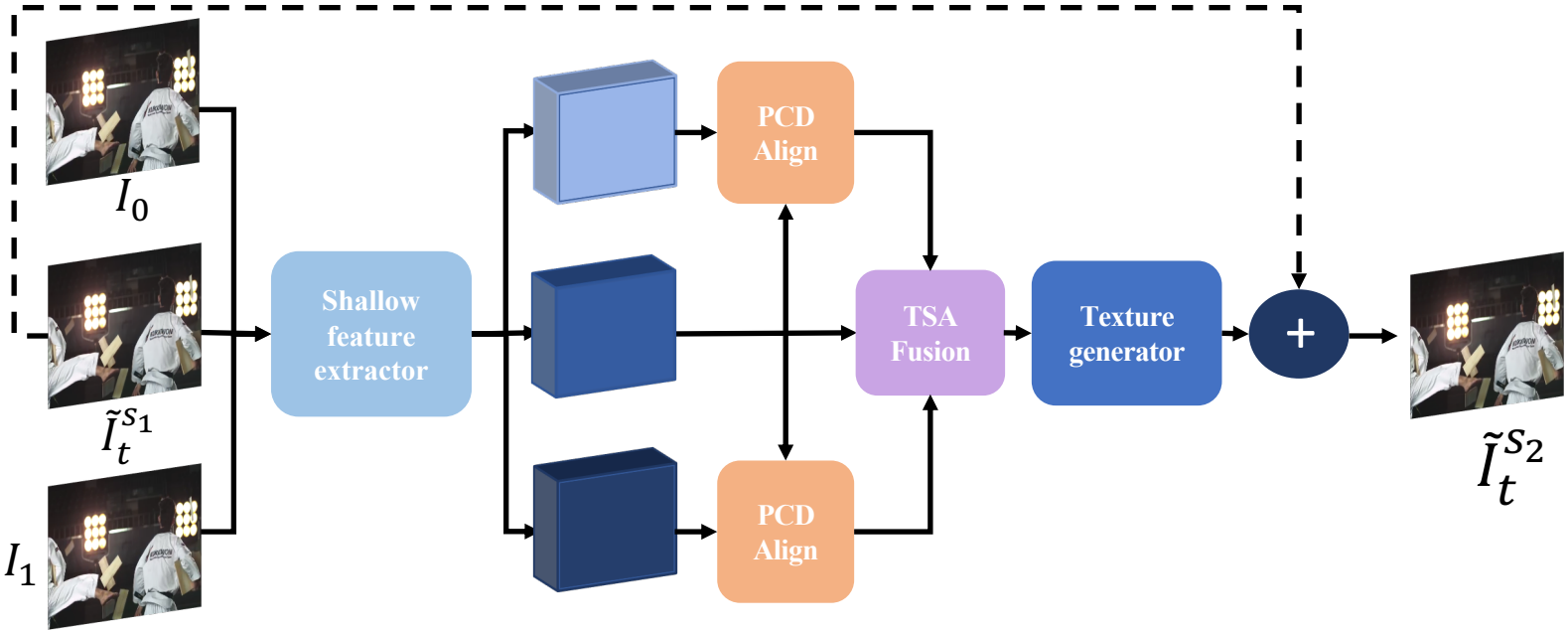
- **Multi-flow Multi-attention Generator (MMG)** is trained to estimate **feature flows** between both the input frames and the target middle one, and further synthesize the **coarse interpolation result which emphasizes the overall structure.**
- **Multi-Flow Predictor (MFP)** to generate the same number of **flow offsets** for extracted feature groups.
- **Multi-Attention Predictor (MAP)** is proposed for feature blending.



FeatureFlow Framework

Frame Texture Compensator (FTC):

- Inspired by TDAN and EDVR, our FTC aims to **locate and borrow the desired texture features from the original input frames**.
- Different from video restoration tasks, in our case, the original and reference frames haven't got the same image qualities.



Experimental Results

Structure-guided interpolation

Texture refinement

Structure-guided interpolation

Texture refinement

Method	Vimeo90K		Adobe240fps	
	PSNR	SSIM	PSNR	SSIM
SepConv -L_f [6]	33.45	0.9674	31.93	0.9492
SepConv -L_1[6]	33.79	0.9702	32.08	0.9512
MEMC-net [5]	34.40	0.9743	32.42	0.9537
DAIN [4]	34.71	0.9756	32.51	0.9539
FeFlow (Ours)	**35.28**	**0.9764**	**32.66**	**0.9550**

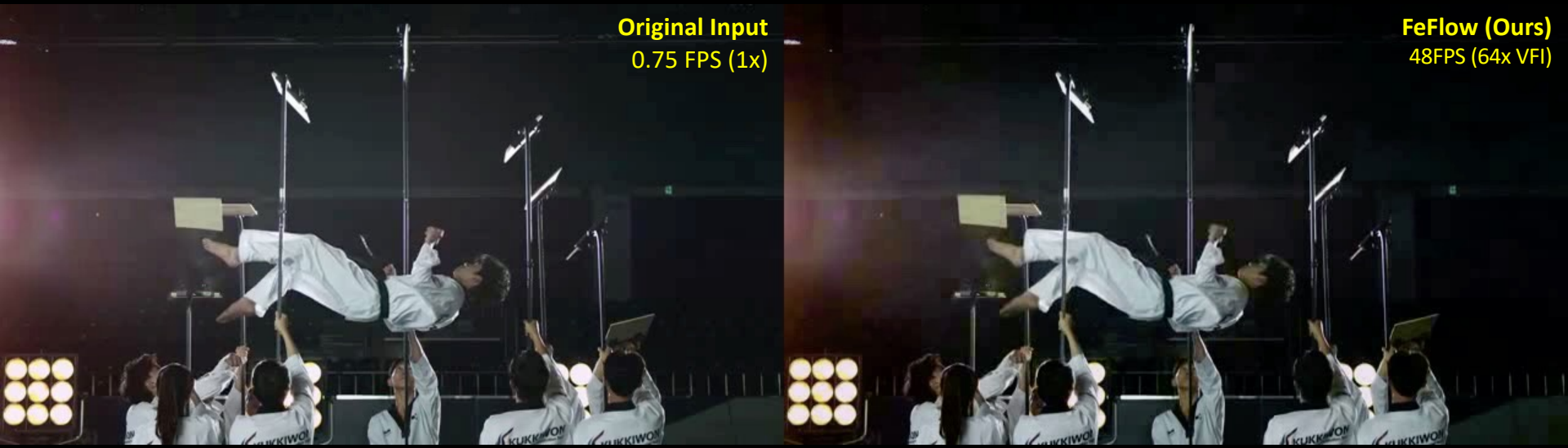
(Please refer the paper for more)

Up: Origin

Down: Ours (x4 interpolation)

Left: Origin

Right: Ours (x4 interpolation)


Original Input
0.75 FPS (1x)
FeFlow (Ours)
48FPS (64x VFI)


Super SloMo [1]


SeFlow (Ours)

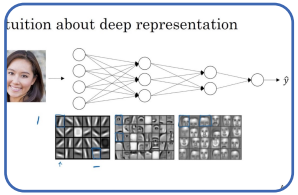
大王

七郎
のいぢ

高橋弥七郎
大塚　舞

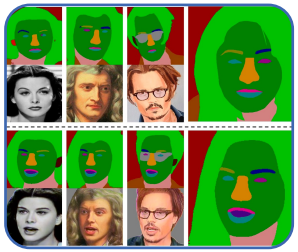

Snowy
DAIN
Snowy
Snowy
Ours
Snowy

Conclusion



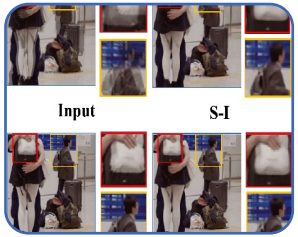
Deep Representations: Brief Introduction

- Deep Learning and Representation learning
- Why Deep Representations ?
- Deep Representation: Learning & Editing



PuppeteerGAN: Arbitrary Portrait Animation with Semantic-aware Appearance Transformation (CVPR-2020)

- Portrait Animation and Related Applications
- Existing Challenges and Our Motivation
- The Proposed method and Experimental results



FeatureFlow: Robust Video Interpolation via Structure-to-texture Generation (CVPR-2020)

- Video Frame Interpolation and Related Applications
- Conventional Video Frame Interpolation: Optical Flow based Methods
- The proposed FeatureFlow and experimental results

Reference

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1156, 2018.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):196, 2017.

Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018.

Chaoyue Wang, Chaohui Wang, Chang Xu, and Dacheng Tao. Tag disentangled generative adversarial network for object image re-rendering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI),* pages 2901–2907, 2017.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9459–9468, 2019.

Reference

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Tanaphol Thaipanich, Ping-Hao Wu, and C.-C. Jay Kuo. Low Complexity Algorithm for Robust Video Frame Rate Up-Conversion (FRUC) Technique *IEEE Transactions on Consumer Electronics, Vol. 55, No. 1, FEBRUARY 2009*

Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.*

Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* June 2019.

Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video inter- polation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. *In The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.