Learning to Forecast and Refine Residual Motion for Image-to-Video Generation

Long Zhao[1] Xi Peng[2] Yu Tian[1] Mubbasir Kapadia[1] Dimitris Metaxas[1]

[1] Rutgers University [2] Binghamton University

Motivation

We consider the problem of image-to-video translation, where a system receives one or more images as the input and translates it into a video containing realistic motions of a single object. We target at *conditional motion forecasting* and *realistic long-term video generation*.

- **Applications**
1) Facial Expression Retargeting
2) Human Motion Forecasting

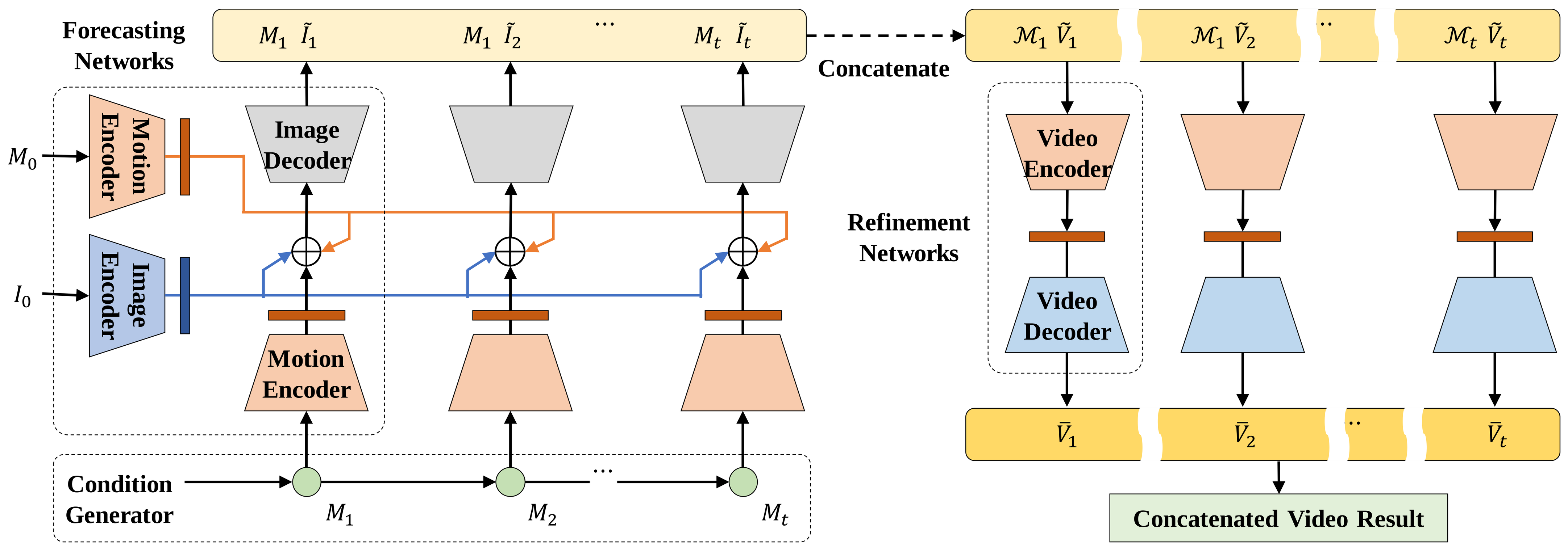
- **Challenges**
1) Preserve the identity consistency
2) Forecast conditional long-term motion
3) Maintain video coherence in pixel level

- **Contributions**
1) A novel two-stage generative framework
2) Investigate learning residual motion
3) Introduce dense connections for decoders

Framework

A **two-stage** generative framework

- Videos are (a) generated from conditions and then (b) refined.
- Three components: *a condition generator*, *motion forecasting network* and *motion refinement network*.

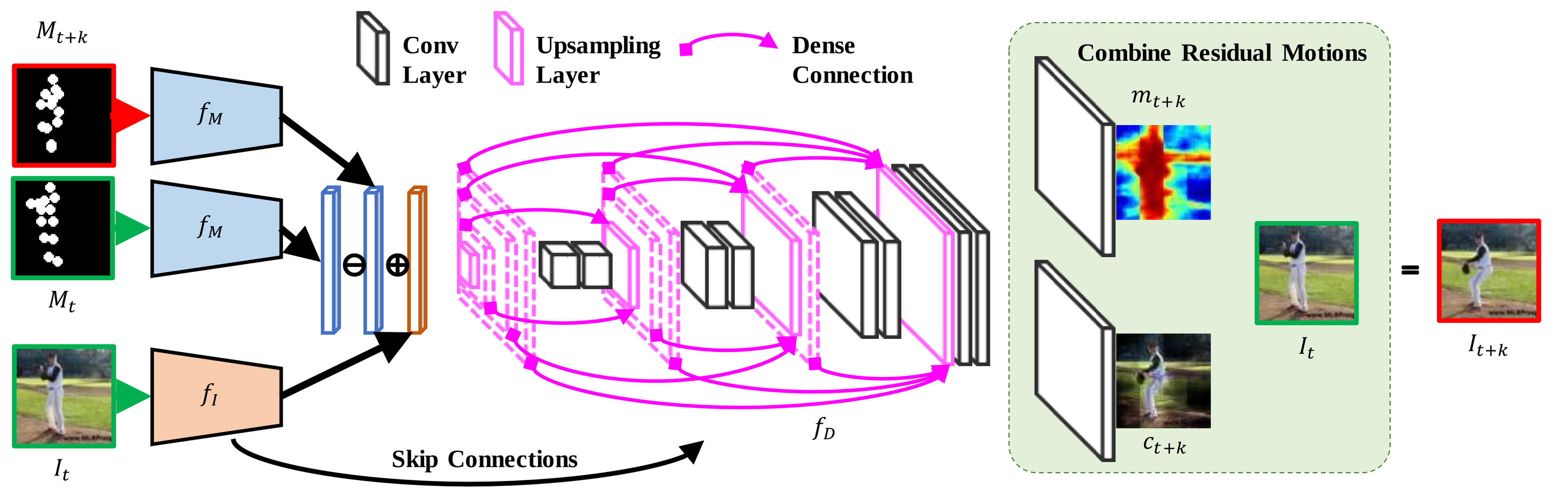


(a) Forecasting (b) Refinement

Stage 1: Motion Forecasting Network

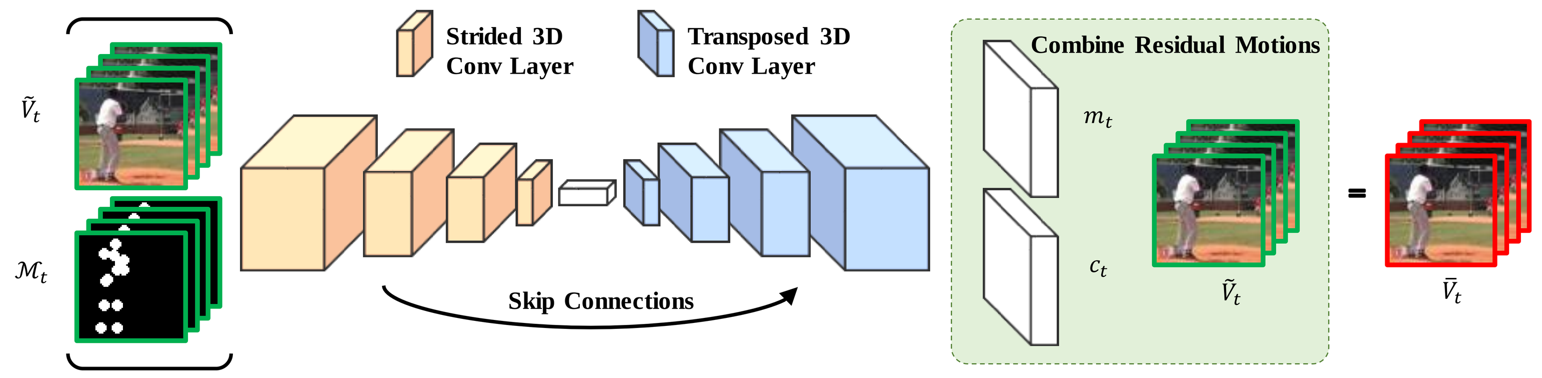
- **Motion disentangle, dense layers for decoders**
1) Generate motion guided by *domain knowledge*
2) Preserve the object identity
3) Ensure motion structures

- Face: 3D Morphable Model
- Pose: 2D Joints + LSTM



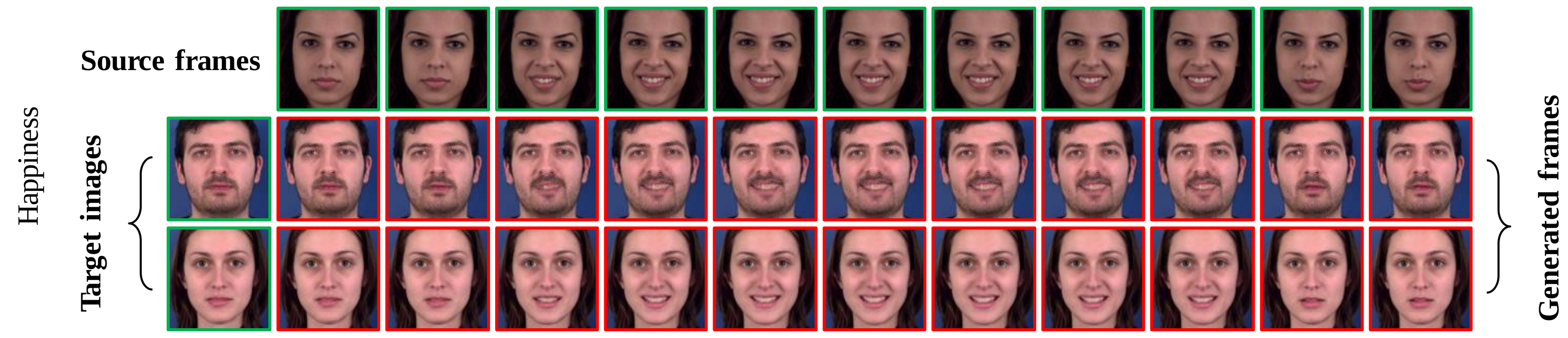
Stage 2: Motion Refinement Network

- **Learning for video refinement**
1) Refine videos with 3D convolutional networks
2) Model refinements in the residual space
3) Produce temporally coherent motions



Experiments

- **Evaluation on Facial Expression Retargeting**



Methods	ACD-I	ACD-C
MCNet	0.545	0.322
Villegas et al.	0.683	0.130
MoCoGAN	0.291	0.205
Ours	**0.184**	**0.107**

Table 1. Video generation quality comparison.

Methods	Preference (%)
Ours / MCNet	**84.2** / 15.8
Ours / Villegas et al.	**74.6** / 25.4
Ours / MoCoGAN	**62.5** / 37.5

Table 2. Average user preference score (%).

- **Evaluation on Human Pose Forecasting**

Methods	MSE	MSE (LSTM)
VGAN	0.047	-
Mathieu et al.	0.041	-
Villegas et al.	0.030	0.025
Ours	**0.023**	**0.011**

Table 3. MSE score on Penn Action Database.

Settings	ACD-I	ACD-C	MSE
G_M ~~(Dense)~~, $\cancel{G_R}$	0.459	0.155	0.027
G_M (Dense), $\cancel{G_R}$	0.252	0.140	0.014
G_M (Dense), G_R	**0.184**	**0.107**	**0.011**

Table 4. Quantitative results of ablation study.

- **References**

[MCNet] "Decomposing motion and content for natural video sequence prediction". ICLR'17.
[Villegas et al.] "Learning to generate long-term future via hierarchical prediction". ICML'17.
[MoCoGAN] "MoCoGAN: Decomposing Motion and Content for Video Generation". CVPR'18.
[Mathieu et al.] "Deep multi-scale video prediction beyond mean square error". ICLR'16.
[VGAN] "Generating videos with scene dynamics". NIPS'16.