Dual-MTGAN: Stochastic and Deterministic Motion Transfer for Image-to-Video Synthesis

ICPR 2020

1,2Fu-En Yang*, 1Jing-Cheng Chang*, 1Yuan-Hao Lee, 1,2Yu-Chiang Frank Wang

1Graduate Institute of Communication Engineering, National Taiwan University

2ASUS Intelligent Cloud Services (AICS)
What is image-to-video synthesis?

Synthesize videos from an input image with the motion of interest.
Deterministic Motion Transfer

- Transfer motion pattern across different videos

Source Video

Synthesized Videos with Different Identities
(Preserved Facial Expressions/Motions)

Huang et al. “Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment.” CVPR 2020
Stochastic Motion Generation

- Generate videos from an image or few frames
Can we jointly perform **deterministic** and **stochastic** motion transfer in a **unified** framework?
Method – Dual-Motion Transfer GAN (Dual-MTGAN)

Self-Supervised Content/Motion Disentanglement

Motion Consistency Guided Adversarial Learning
Method

$V = \begin{bmatrix} x^T \end{bmatrix}_1$
Method
– Self-Supervised Content/Motion Disentanglement

Temporal coherence across frames:
\[ \mathcal{L}_C = \| E_C(x^t) - E_C(x^{t+1}) \|_1 \]
Method
– Source Video Reconstruction

Reconstruction Loss:
\[ \mathcal{L}_{rec} = \| \tilde{V}_x - V \|_1 \]
Method
– Deterministic Motion Transfer

\[ V = [x^t]_1^T \]

Self-Supervised Content/Motion Disentanglement
Method

– Learning Motion Latent Space

$V = [x^t]_1^T$

Source Video

Target Image

$z_{c,y}$

$z_{M}$

$z_{c,x}$

$\tilde{V}_x$

$V_{y,d}$

Reconstruction

Deterministic Transfer

Self-Supervised Content/Motion Disentanglement
Method
– Stochastic Motion Transfer

Source Video

$V = [x^t]^T$

Motion Space

Target Image

$y$

$E_C$

$E_M$

$G_T$

$G_1$

$z_{c,x}^t$

$z_M^t$

$z_{m,d}^t$

$z_{m,s}^t$

$z_{c,y}$

$\hat{V}_x$

$\hat{V}_{y,d}$

$\hat{V}_{y,s}$

Reconstruction

Self-Supervised Content/Motion Disentanglement
Method
– Adversarial Learning

Image-level conditional adversarial training:
\[ \mathcal{L}_{GAN,x}^I = \mathcal{L}_{GAN,x}^I + \mathcal{L}_{GAN,y}^I, \]
where
\[ \mathcal{L}_{GAN,x}^I = \log(D_I(x^I, S_I(V))) + \frac{1}{2} [\log(1 - D_I(x^I, S_I(\tilde{V}_z))) + \log(1 - D_I(y, S_I(V)))] \]
\[ \mathcal{L}_{GAN,y}^I = \log(D_I(y, y)) + \frac{1}{2} [\log(1 - D_I(y, S_I(\tilde{V}_z))) + \log(1 - D_I(x^I, S_I(V_y)))] \]

Video-level adversarial training:
\[ \mathcal{L}_{GAN,x}^V = \mathcal{L}_{GAN,x}^V + \mathcal{L}_{GAN,y}^V, \]
where
\[ \mathcal{L}_{GAN,x}^V = \log(D_V(V)) + \log(1 - D_V(\tilde{V}_x)), \]
\[ \mathcal{L}_{GAN,y}^V = \log(D_V(V_y)) + \log(1 - D_V(\tilde{V}_y)). \]
Method
– Motion Consistency

\[ \mathcal{L}_M = ||E_M(\tilde{V}_x) - z_M||_1 + ||E_M(\tilde{V}_{y,d}) - z_M||_1 \]
Result

- **Deterministic** Face Reenactment and Motion Retargeting

- Facial expression & Human actions
Result

- **Stochastic** Robot Movement and Action Generation

- Robot pushing & Human actions
Result

– Quantitative Comparison & Ablation Study

• Quantitative Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM (↑)</th>
<th>LPIPS (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVG</td>
<td>0.815 ± 0.006</td>
<td>0.0308 ± 0.0005</td>
</tr>
<tr>
<td>Monkey-Net</td>
<td>0.783 ± 0.008</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours</td>
<td>0.827 ± 0.007</td>
<td>0.0422 ± 0.0003</td>
</tr>
</tbody>
</table>

• Ablation Study
Conclusion

• Given an input image, our proposed model allows transfer of motion patterns from video data, or synthesis of video sequences with motion diversity.

• By enforcing appearance coherence and motion consistency, our Dual-MTGAN factorizes visual latent representations into disjoint features describing content and motion information in a self-supervised manner.
Thanks for listening!