Dual-MTGAN:  
Stochastic and Deterministic Motion Transfer for Image-to-Video Synthesis  
Fu-En Yang*, Jing-Cheng Chang*, Yuan-Hao Lee, Yu-Chiang Frank Wang

Motivation

• **Image-to-Video Synthesis**
  Synthesize videos from an input image with the motion of interest.

• **Contributions**
  1. Given an input image, our proposed model allows transfer of motion patterns from video data, or synthesis of video sequences with motion diversity.
  2. By enforcing appearance coherence and motion consistency, our model factorizes visual latent representations into disjoint features describing content and motion features in a self-supervised manner.

Approach

• **Self-Supervised Content/Motion Disentanglement**
  - Content encoder $E_C$ aims to extract time-invariant content features $z_c$ by enforcing the temporal consistency across frames.
  - Video motion feature $z_m$ is derived from motion encoder $E_M$ and fits Gaussian prior for generating diverse outputs via sampling during testing.

• **Motion Consistency Guided Adversarial Learning**
  - Video-level adversarial learning ensures both video quality and temporal continuity.

  - Image-level adversarial learning guarantees the plausibility of synthesized frames, while ensures the appearance of output to match the conditioned image.

  - Motion consistency preserves motion information $z_M$ during training process.

Experiment Results

Comparisons with SOTAs

- **Deterministic Motion Transfer**
  ![Deterministic Motion Transfer](image1)

- **Stochastic Motion Generation**
  ![Stochastic Motion Generation](image2)

Comparisons with SOTAs

- **Quantitative Ablation Studies**
  ![Quantitative Ablation Studies](image3)

- **Comparisons with SOTAs**
  ![Comparisons with SOTAs](image4)

- **Robot pushing**
<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>0.0398 ± 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><strong>0.827 ± 0.007</strong></td>
<td><strong>0.0422 ± 0.0003</strong></td>
</tr>
</tbody>
</table>

- SSIM: measure the visual realism
- LPIPS: evaluate the visual diversity