A Recurrent Transformer Network for Novel View Action Synthesis
Task

- Given:
  - action video
  - appearance prior
- Goal: synthesize action from novel viewpoint

Input Action Video  Novel View/Desired Output
Related Work

- Mainly focused on video synthesis [1,2,3]
- Single view
- Most cross-view synthesis deals with images[4,5,6]
- One recent work in cross-view video synthesis[7]

Our Approach

- Transform action features to new viewpoint
- Predict action key-points
- Recurrently transform appearance features using
  - Transformed features
  - Action key-points
Network Architecture
Network Architecture
Network Architecture
Network Architecture
Network Architecture
Network Architecture
Network Architecture

**Prior**
- \( P^j \)

**Appearance encoder**
- \( g_E \)
- \( a^j \)

**Appearance transformation**
- \( t_A \)
- \( \tilde{a}^j \)

**Video synthesis**
- \( f_G \)
- \( \hat{V}^j \)

**Action encoder**
- \( f_E \)
- \( m^i \)

**Viewpoint predictor**
- \( h \)
- \( \theta_{ij} \)

**Action transformer**
- \( t_M \)
- \( \hat{m}^j \)

**Key-point predictor**
- \( k_G \)
- \( KP^j \)
Our Approach

• Predict key-points to localize the action in the video
Network Architecture

14
Our Approach

• Use Recurrent Appearance Transformation
Network Architecture
Network Architecture

Diagram showing the network architecture with labels for Prior, Appearance encoder, Action encoder, Appearance transformation, Video synthesis, Key-point predictor, and Synthesized video.
Network Architecture

Adversarial Loss & Perceptual Loss
Evaluation

- NTU-RGB+D Dataset
  - Over 56,000 videos
  - 40 different actors
  - 60 different actions
  - 80 different viewpoints

## Ablation Study

<table>
<thead>
<tr>
<th>Model</th>
<th>Pair-view PSNR Score</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$v_1 \rightarrow v_2$</td>
<td>$v_1 \rightarrow v_3$</td>
</tr>
<tr>
<td>basic model</td>
<td>23.8 ± 1.5</td>
<td>23.8 ± 1.4</td>
</tr>
<tr>
<td>w/ AC-Trans</td>
<td>24.7 ± 1.5</td>
<td>24.7 ± 1.4</td>
</tr>
<tr>
<td>w/ HI-Trans</td>
<td>26.7 ± 2.6</td>
<td>26.8 ± 2.5</td>
</tr>
<tr>
<td>w/ AP-Trans</td>
<td>27.6 ± 2.7</td>
<td>27.7 ± 2.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Pair-view SSIM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$v_1 \rightarrow v_2$</td>
</tr>
<tr>
<td>basic model</td>
<td>.939 ± .033</td>
</tr>
<tr>
<td>w/ AC-Trans</td>
<td>.950 ± .022</td>
</tr>
<tr>
<td>w/ HI-Trans</td>
<td>.967 ± .028</td>
</tr>
<tr>
<td>w/ AP-Trans</td>
<td>.974 ± .021</td>
</tr>
</tbody>
</table>
Quantitative Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Pair-view SSIM Score</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$v_1 \rightarrow v_2$</td>
<td>$v_1 \rightarrow v_3$</td>
</tr>
<tr>
<td>PG$^2$ [21]</td>
<td>.499 ± .071</td>
<td>.561 ± .060</td>
</tr>
<tr>
<td>VRNet [32]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VDNNet [17]</td>
<td>.789 ± .076</td>
<td>.791 ± .069</td>
</tr>
<tr>
<td>Proposed</td>
<td>.974 ± .021</td>
<td>.975 ± .021</td>
</tr>
</tbody>
</table>
Qualitative Results

Input Video

Synthesized Video

Ground Truth Video
Qualitative Results

Input Video

Synthesized Video

Appearance Prior
Thank you!

Project Page:  
![QR Code for Project Page]  

GitHub:  
![QR Code for GitHub]