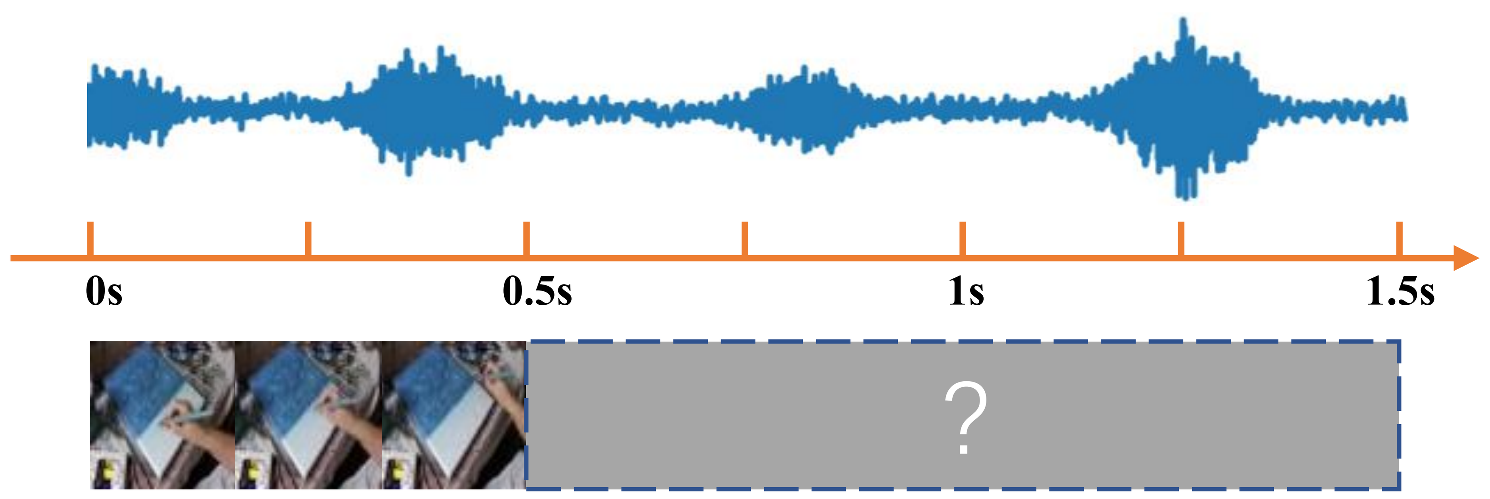
# Motion and Context-Aware Audio-Visual Conditioned Video Prediction


Yating Xu, Conghui Hu, Gim Hee Lee

National University of Singapore

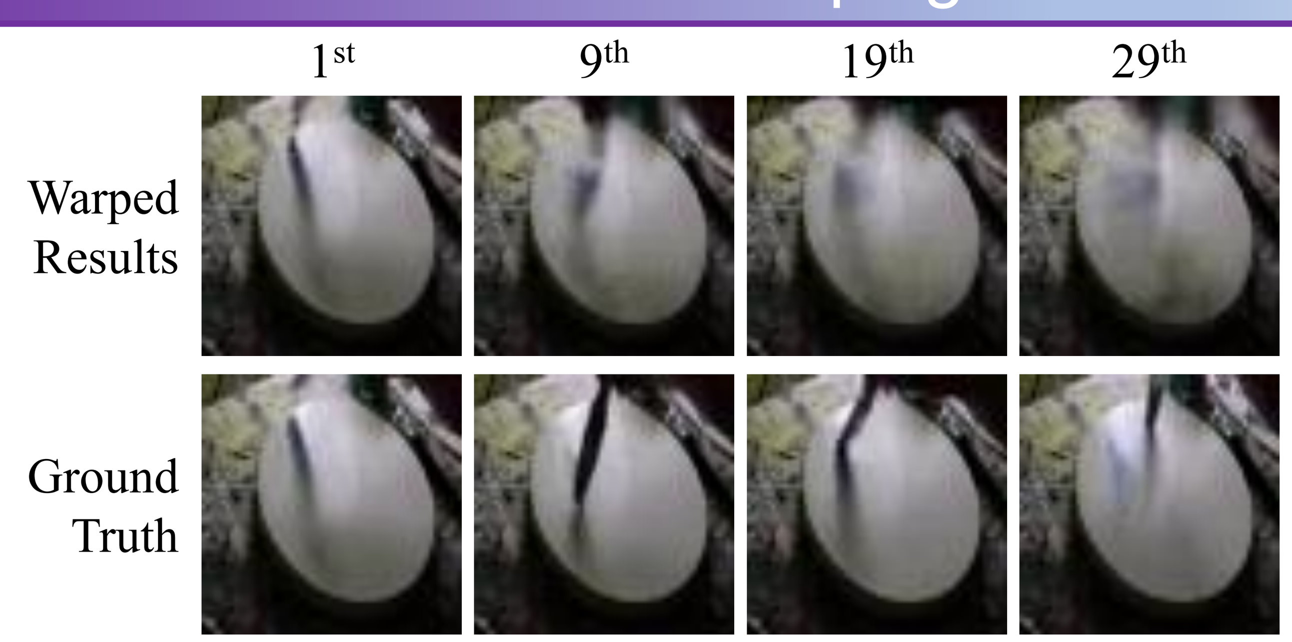

BMVC 2023

## 1. Introduction




- **Task:** Given a full audio clip and a short sequence of the past visual frames, the objective is to predict the missing future visual frames that are as close to the ground truth frames as possible
- **Problem:** Direct inference of per-pixel intensity is very challenging due to the high-dimensional image space.
- **Solution:** We decouple motion and appearance separately. We propose:
  - **multimodal motion estimation** to predict future optical flow.
  - **context-aware refinement** to refine the appearance of future images.

## 3. Problem with Recurrent Warping




The warped images become increasingly blurry and consequently lead to the loss of appearance context.

## 4. Optimization

- Stage 1: optimize MME only.

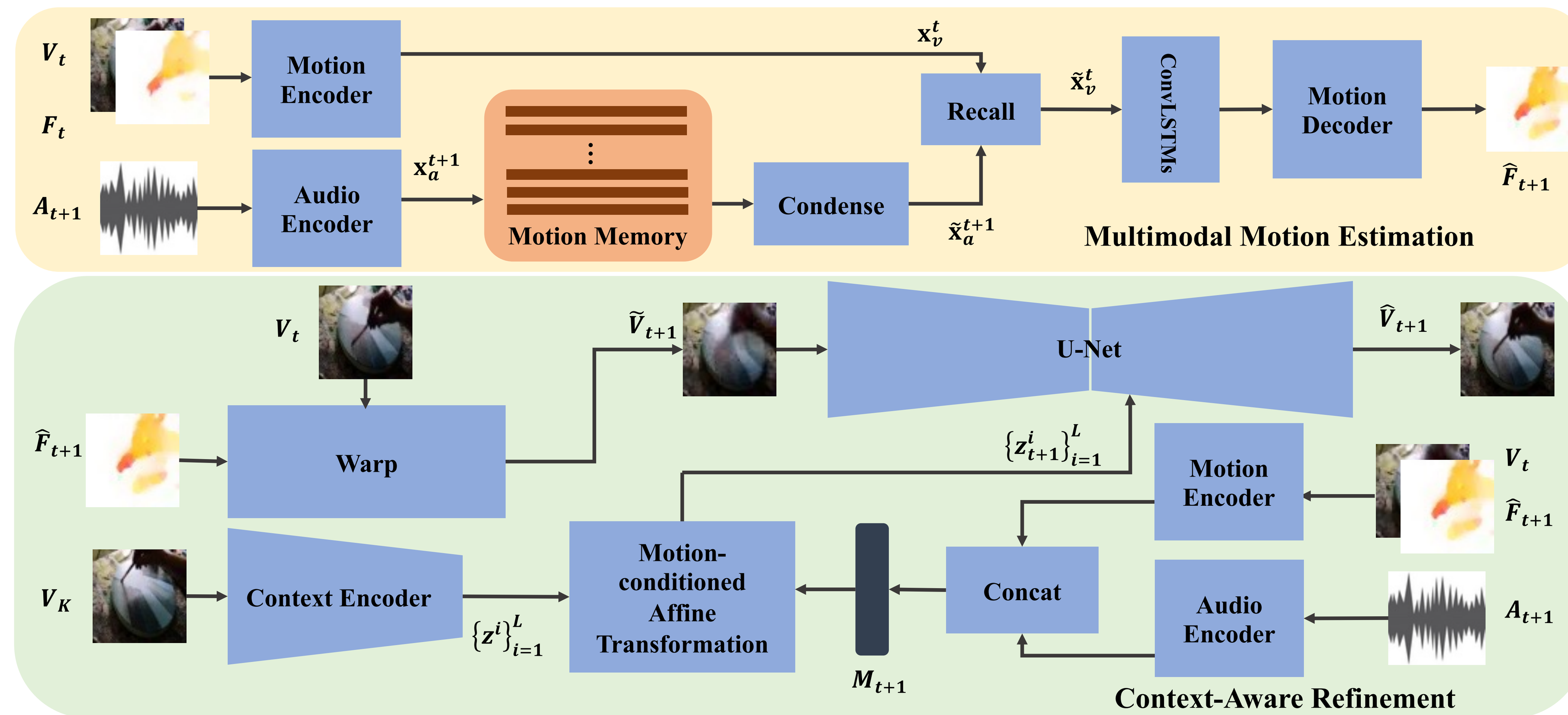
$$\mathcal{L}_{\text{MME}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}$$

$$\mathcal{L}_{\text{flow}} = \sum_{t=K+1}^{T} \|F_t - \hat{F}_t\|_2^2 \qquad \mathcal{L}_{\text{smooth}} = \sum_{t=K+1}^{T} \|\nabla \hat{F}_t\|_1 e^{-\|\nabla V_t\|_1}$$

- Stage 2: optimize CAR only.

$$\mathcal{L}_v = \sum_{t=K+1}^{T} \|V_t - \hat{V}_t\|_2^2$$

## 2. Overall Framework




- **Multimodal Motion Estimation (MME):**
  - An external **motion memory MM** stores the audio features as the long-term information:

$$\text{MM} = \{x_a^n\}_{n=1}^{t+1} \in \mathbb{R}^{(t+1)\times c}$$

  - **Condense**:

$$\tilde{x}_a^{t+1} = x_a^{t+1} + \text{Condense}(\text{MM})$$

$$\text{Condense}(\text{MM}) = \text{Atten}(\text{MM}, \text{MM}, \text{MM})[-1]$$

$$\text{Atten}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}$$

  - **Recall**:

$$\text{Recall}(x_v^t, \tilde{x}_a^{t+1}) = \text{Atten}(x_v^t, \tilde{x}_a^{t+1}, \tilde{x}_a^{t+1}), \quad \tilde{x}_v^t = x_v^t + \text{Recall}(x_v^t, \tilde{x}_a^{t+1})$$

  - Recurrent Prediction of future optical flow $\hat{F}_{t+1}$:

$$\{h_{t+1}, o_{t+1}\} = \text{ConvLSTM}\left(\tilde{x}_v^t, h_t\right), \quad \hat{F}_{t+1} = D_m(h_{t+1})$$

- **Context-aware Refinement (CAR):**
  - **Context encoder** extracts global appearance context $Z = \{z^i\}_{i=1}^{L}$ from the last given visual frame $V_K$.
  - **Motion-conditioned affine transformation** to adjust global appearance context:
  - ✓ Motion feature:

$$M_{t+1} = E_m\left(V_t, \hat{F}_{t+1}\right) || E_a\left(A_t\right)$$

  - ✓ Transformation parameter:

$$\gamma_{t+1}^i = \text{MLP}_1^i\left(M_{t+1}\right), \quad \beta_{t+1}^i = \text{MLP}_2^i\left(M_{t+1}\right)$$

  - ✓ Channel-wise scaling and shifting is performed on $z^i$ to obtain the adapted context $z_{t+1}^i$:

$$z_{t+1}^i = \gamma_{t+1}^i \cdot z^i + \beta_{t+1}^i$$

## 5. Experiments on Synthetic Dataset

| Method | Type | SSIM ↑ Fr 6 | SSIM ↑ Fr 15 | SSIM ↑ Fr 25 | SSIM ↑ Mean | PSNR ↑ Fr 6 | PSNR ↑ Fr 15 | PSNR ↑ Fr 25 | PSNR ↑ Mean |
|---|---|---|---|---|---|---|---|---|---|
| Denton and Fergus [ ] | V | 0.9265 | 0.8300 | 0.7999 | – | 18.59 | 14.65 | 13.98 | – |
| MSPred [ ] | V | 0.9400 | 0.8903 | 0.8060 | 0.8846 | 21.57 | 19.02 | 17.04 | 19.08 |
| Vougioukas *et al.* [ ] | M | 0.8600 | 0.8571 | 0.8573 | – | 15.17 | 14.99 | 15.01 | – |
| Sound2Sight [ ] | M | 0.9505 | 0.8780 | 0.8749 | 0.8910 | 22.22 | 18.16 | 17.84 | 18.70 |
| Our Method | M | **0.9608** | **0.9158** | **0.8990** | **0.9195** | **23.61** | **19.88** | **18.99** | **20.23** |

## 6. Experiments on Real-world Datasets

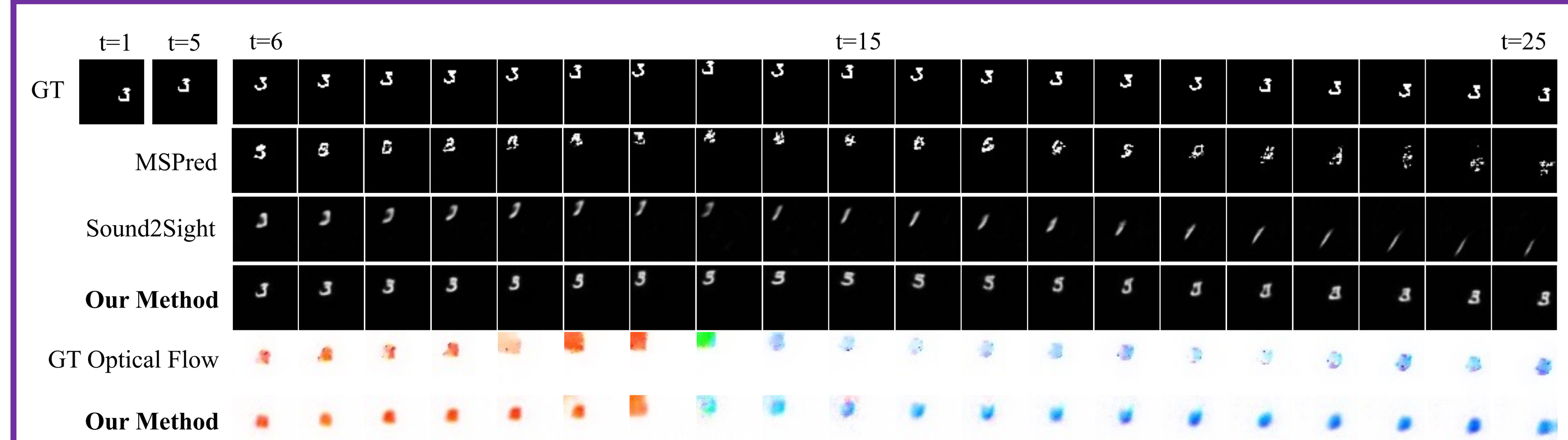
| Method | Type | SSIM ↑ Fr 16 | SSIM ↑ Fr 30 | SSIM ↑ Fr 45 | SSIM ↑ Mean | PSNR ↑ Fr 16 | PSNR ↑ Fr 30 | PSNR ↑ Fr 45 | PSNR ↑ Mean |
|---|---|---|---|---|---|---|---|---|---|
| Denton and Fergus [ ] | V | 0.9779 | 0.6654 | 0.4193 | – | 32.52 | 16.05 | 11.84 | – |
| MSPred [ ] | V | 0.9648 | 0.8991 | 0.8617 | 0.8965 | 33.42 | 26.42 | 24.25 | 26.65 |
| Vougioukas *et al.* [ ] | M | 0.9281 | 0.9126 | 0.9027 | – | 26.97 | 25.58 | 24.78 | – |
| Sound2Sight [ ] | M | 0.9716 | 0.9261 | 0.9074 | 0.9264 | 31.91 | 26.73 | 25.17 | 26.95 |
| Our Method | M | **0.9848** | **0.9284** | **0.9104** | **0.9313** | **35.12** | **27.19** | **25.53** | **27.70** |

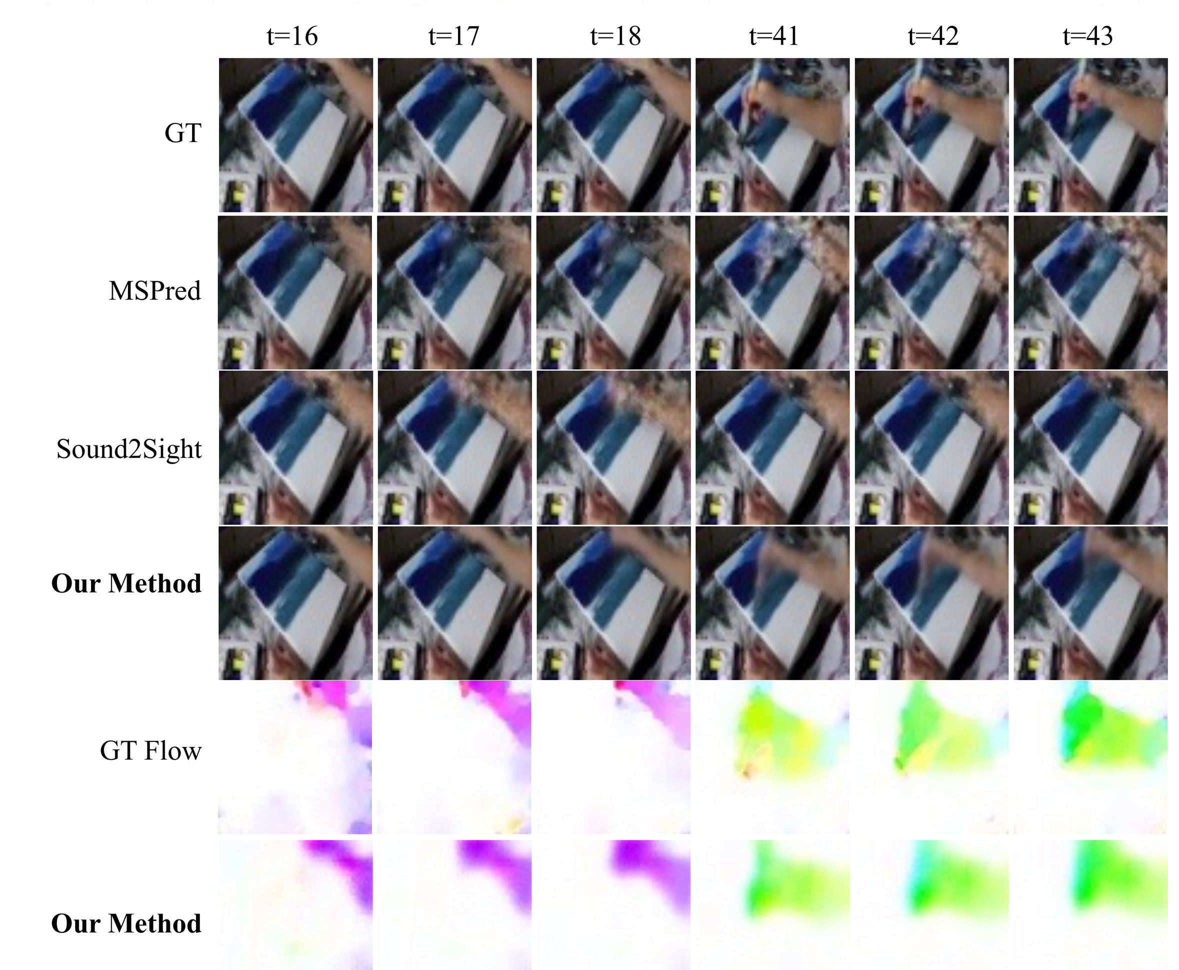
Table 2: Comparison on YouTube Painting with 15 seen frames.

| Method | Type | SSIM ↑ Fr 16 | SSIM ↑ Fr 30 | SSIM ↑ Fr 45 | SSIM ↑ Mean | PSNR ↑ Fr 16 | PSNR ↑ Fr 30 | PSNR ↑ Fr 45 | PSNR ↑ Mean |
|---|---|---|---|---|---|---|---|---|---|
| Denton and Fergus [ ] | V | 0.9706 | 0.6606 | 0.5097 | – | 30.01 | 16.57 | 13.49 | – |
| MSPred [ ] | V | 0.9799 | 0.9382 | 0.9214 | 0.9389 | 33.55 | 27.30 | 25.91 | 27.60 |
| Vougioukas *et al.* [ ] | M | 0.8986 | 0.8905 | 0.8866 | – | 23.62 | 23.14 | 22.91 | – |
| Sound2Sight [ ] | M | 0.9875 | 0.9524 | 0.9434 | 0.9544 | 34.23 | 27.73 | **26.71** | 28.13 |
| Our Method | M | **0.9896** | **0.9533** | **0.9437** | **0.9558** | **35.00** | **27.76** | 26.68 | **28.22** |

Table 3: Comparison on AudioSet-Drums with 15 seen frames.

## 7. Qualitative Results




## 8. Ablation Study

| Method | YouTube | MNIST | AudioSet |
|---|---|---|---|
| V | 11.25 | – | – |
| V + Recall | 10.43 | 6.40 | 4.63 |
| MME | **9.85** | **4.82** | **3.97** |

Table 4: Analysis of audio in MME. AEPE results are presented in $10^{-2}$ scale.

| Method | SSIM ↑ Fr 16 | SSIM ↑ Fr 30 | SSIM ↑ Fr 45 | SSIM ↑ Mean |
|---|---|---|---|---|
| MME+Unet | **0.9873** | 0.9105 | 0.8761 | 0.9143 |
| MME+Unet+ContextEnc | 0.9821 | 0.9266 | 0.9091 | 0.9291 |
| MME+CAR | 0.9848 | **0.9284** | **0.9104** | **0.9313** |

Table 5: Analysis of CAR on YouTube Painting.