Supplementary Material for "DAADiff: A Dual-Axis Adaptive Framework for Real-time Video Inpainting"

August 29, 2025

Contents

A. Extended Implementation Details

Our experimental framework begins by preprocessing both video frames and images. This step centers each frame or image on the infant's facial region and resizes it to 512×512 pixels. This operation is performed by a fine-tuned YOLOv8-based model, which achieved 100.0% accuracy in detecting the infant's face on our test set. All processed data are then standardized to this resolution for training, evaluation, and benchmark comparisons.

To train and evaluate our model's robustness under different occlusion scenarios, we implement two distinct masking approaches. The first employs the fine-tuned YOLOv8n model (Terven et al., 2023), trained on 96 annotated images from 4 babies in the ICOPEvid dataset and tested on 25 images from 2 babies. This model achieves 97.5% masking accuracy and an average IoU of 0.979, generating rectangular masks for occlusions. The second method leverages a fine-tuned custom segmentation model (Camporese et al., 2021) trained on 215 rigorously labeled images from 5 babies, reaching 96.4% accuracy and producing irregular-shaped masks with an average IoU of 0.930, better mimicking real-world occlusions.

We train with Adam optimizer, setting the learning rate 10^{-5}, with a batch size of 8, then trained for $420,000$ iterations. We implement our method using the PyTorch (version $v2.2.2$) framework. We use the following hyperparameter names, consistent with PyTorch conventions:

```
model:
  params:
    linear_start: 0.0008
    linear_end: 0.01450
    num_timesteps_cond: 1
    timesteps: 1000
    conditioning_key: fused crossattn
```

```
trainer:
  type: "Adam"
  base_learning_rate: e-5
  warm_up_steps: 1000
  batch_size: 8
  log_freq: 500
  val_log_freq: 2e3
  iterations: 420e3
```

B. Evaluation Metrics

B.1. Structural Similarity Index Measure (SSIM)

SSIM measures the structural similarity between ground truth frames $V = \{v_i\}_{i=1}^N$ and reconstructed frames $\hat{V} = \{\hat{v}_i\}_{i=1}^N$ by comparing luminance, contrast, and structural properties within an 11×11 Gaussian window W:

$$\text{SSIM}(v_i, \hat{v}_i) = \frac{(2\mu_{v_i}\mu_{\hat{v}_i} + C_1)(2\sigma_{v_i\hat{v}_i} + C_2)}{(\mu_{v_i}^2 + \mu_{\hat{v}_i}^2 + C_1)(\sigma_{v_i}^2 + \sigma_{\hat{v}_i}^2 + C_2)}, \tag{B.1}$$

where:

- μ_{v_i} and $\mu_{\hat{v}_i}$: mean pixel values over window W
- $\sigma_{v_i}^2$ and $\sigma_{\hat{v}_i}^2$: variances within the window
- $\sigma_{v_i\hat{v}_i}$: covariance between v_i and $\hat{v}_i$
- $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$: stabilization constants
- L: dynamic range of pixel values
- $k_1 = 0.01$ and $k_2 = 0.03$.

The mean SSIM across all frames is:

$$\text{Mean SSIM} = \frac{1}{N}\sum_{i=1}^{N} \text{SSIM}(v_i, \hat{v}_i). \tag{B.2}$$

B.2. Fréchet Inception Distance (FID)

FID quantifies the perceptual realism by comparing feature distributions of real and generated images in InceptionV3's feature space:

$$\text{FID}(V, \hat{V}) = \|\mu - \hat{\mu}\|_2^2 + \text{Tr}\left(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{1/2}\right), \tag{B.3}$$

where:

- μ and $\hat{\mu}$: mean feature vectors of real and generated images
- Σ and $\hat{\Sigma}$: covariance matrices of feature distributions.

B.3. Identity Retention Score

To ensure facial identity preservation, we employ ArcFace [1], a state-of-the-art face recognition model. Given ground truth frame v_i and reconstructed frame $\hat{v}_i$, we first detect and align faces using MTCNN [2]:

$$\mathbf{x}_i = \text{align}(\text{detect}(v_i)), \quad \hat{\mathbf{x}}_i = \text{align}(\text{detect}(\hat{v}_i)), \tag{B.4}$$

where $\mathbf{x}_i, \hat{\mathbf{x}}_i$ are aligned face crops. These are then passed through the pre-trained ArcFace model to extract normalized 512-dimensional embeddings:

$$\mathbf{e}_i = \frac{f_{\text{ArcFace}}(\mathbf{x}_i)}{\|f_{\text{ArcFace}}(\mathbf{x}_i)\|_2}, \quad \hat{\mathbf{e}}_i = \frac{f_{\text{ArcFace}}(\hat{\mathbf{x}}_i)}{\|f_{\text{ArcFace}}(\hat{\mathbf{x}}_i)\|_2}. \tag{B.5}$$

The identity retention score is computed as the cosine similarity:

$$\text{ID-Retention}(v_i, \hat{v}_i) = \mathbf{e}_i \cdot \hat{\mathbf{e}}_i = \sum_{j=1}^{p} e_{i,j} \cdot \hat{e}_{i,j}. \tag{B.6}$$

B.4. Fréchet Video Distance (FVD)

FVD extends FID to video by using I3D features that capture spatiotemporal patterns:

$$\text{FVD}(\mathbf{V}, \hat{\mathbf{V}}) = \|\boldsymbol{\mu}_{\mathbf{V}} - \boldsymbol{\mu}_{\hat{\mathbf{V}}}\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_{\mathbf{V}} + \boldsymbol{\Sigma}_{\hat{\mathbf{V}}} - 2(\boldsymbol{\Sigma}_{\mathbf{V}}\boldsymbol{\Sigma}_{\hat{\mathbf{V}}})^{1/2}\right), \tag{B.7}$$

where:

- $\mathbf{V}$ and $\hat{\mathbf{V}}$: sets of video clips
- $\boldsymbol{\mu}_{\mathbf{V}}$ and $\boldsymbol{\mu}_{\hat{\mathbf{V}}}$: mean I3D feature vectors
- $\boldsymbol{\Sigma}_{\mathbf{V}}$ and $\boldsymbol{\Sigma}_{\hat{\mathbf{V}}}$: covariance matrices of I3D features.

B.5. Temporal Consistency (TC)

TC measures frame-to-frame coherence through normalized pixel differences, where lower TC values indicate better temporal consistency:

$$\text{TC} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{\|\text{vec}(\hat{v}_{i+1}) - \text{vec}(\hat{v}_i)\|_2^2}{H \times W \times C \times 255^2}, \tag{B.8}$$

where:

- $\text{vec}(\cdot)$: vectorization operation
- H, W, C: height, width, and channels of the frame
- Normalization by 255^2 accounts for pixel value range.

B.6. VBench Fine-Grained Metrics

We adopt three key dimensions from the VBench, suite for a detailed generation quality analysis:

Temporal Flickering. Temporal Flickering quantifies unwanted brightness variations between consecutive frames:

$$\text{TF} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{1}{HW} \sum_{h,w} |L(\hat{v}_{i+1})_{h,w} - L(\hat{v}_i)_{h,w}|, \tag{B.9}$$

where $L(\cdot)$ converts RGB frames to grayscale luminance, lower values indicate less flickering:

$$L(v)_{h,w} = 0.299 \cdot v_{h,w,R} + 0.587 \cdot v_{h,w,G} + 0.114 \cdot v_{h,w,B}. \tag{B.10}$$

Motion Smoothness. Motion Smoothness evaluates trajectory plausibility through optical flow consistency:

$$\text{MS} = \frac{1}{N-2} \sum_{i=1}^{N-2} \exp\left(-\frac{\|\mathcal{F}_{i+1,i+2} - \mathcal{F}_{i,i+1}\|_2^2}{2\sigma^2}\right), \tag{B.11}$$

where $\mathcal{F}_{i,i+1}$ represents the dense optical flow field from frame i to $i+1$. For each pixel location (h, w):

$$\mathcal{F}_{i,i+1}(h, w) = [u_{h,w}, v_{h,w}]^T, \tag{B.12}$$

where $u_{h,w}$ and $v_{h,w}$ are horizontal and vertical displacement components.

Aesthetic Quality. Aesthetic Quality employs a pre-trained LAION aesthetic predictor:

$$\text{AQ} = \frac{1}{N} \sum_{i=1}^{N} \phi_{\text{aesthetic}}(\hat{v}_i), \tag{B.13}$$

where $\phi_{\text{aesthetic}}(\cdot)$ is the LAION-Aesthetics V2 predictor [3], a ViT-based model fine-tuned on human aesthetic preferences.

References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699, 2019.

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[3] C. Schuhmann, "Clip+ mlp aesthetic score predictor," Clip+ mlp aesthetic score predictor, vol. 6, p. S1, 2022.