Supplementary Material for Implicit Warping for Animation with Image Sets

Along with this document, we have also provided an HTML page for viewing the results, and a supplementary video for your perusal. Please view them to see high-resolution output videos and a summary of the framework proposed in this work. Below are additional details about our method, baselines, and training schemes for completeness and to aid reproducibility.

A Datasets

- TalkingHead-1KH [41]: https://github.com/deepimagination/TalkingHead-1KH
  512×512 face videos
  The videos were published on YouTube by their respective uploaders under the Creative Commons BY 3.0 license. The code and documentation files are under the MIT license.
  #Train videos: $185,517$, #Test videos: 120 (each of 1,024 frames)
- VoxCeleb2 [26]: https://github.com/AliaksandrSiarohin/video-preprocessing
  256×256 face videos
  #Train videos: $281,606$, #Total test videos: $36,237$, #Test videos used: $1,180$
  The test set has a total of 118 unique identities, and we choose 10 videos per identity. The minimum test video length is 99 frames, and the median is 151 frames.
- TED Talk [28]: https://github.com/AliaksandrSiarohin/video-preprocessing
  384×384 upper-body videos
  #Train videos: $1,092$, #Test videos: 128
  The minimum test video length is 129 frames, and the median is 186 frames.

B Metrics

B.1 Automated metrics

We use standard pixel-wise metrics for measuring image reconstruction quality, such as PSNR and the $\mathcal{L}_1$ loss. To measure the perceptual quality of the outputs, we use LPIPS [53] based on AlexNet and FID [12] based on Inception-v3. In order to directly measure the quality of pose transfer, we use average keypoint distance (AKD) measured on the keypoints predicted by OpenPose [4] on the TED talk dataset and by MTCNN [52] for the face datasets. OpenPose predicts the keypoint locations as well as a prediction confidence per keypoint in the $[0, 1]$ range. By thresholding the confidence, we can obtain keypoint visibilities. For the TED talk dataset, we also report the missing keypoint ratio (MKR), which is the fraction of keypoints missing in the prediction but present in the ground truth image. This metric was introduced by Siarohin *et al*. [28].

B.2 Human evaluation

We performed human preference evaluation on the TalkingHead-1KH and the TED Talk dataset results. As the TalkingHead-1KH evaluation videos are 1024 frames long, we extracted 3 4-second clips from each of the test videos, sampled from the start, the middle, and the end of the video. This gave a total of 360 and 128 videos, respectively. Each pair of videos was shown to 3 workers, leading to a total of 1080 and 720 questions for the respective datasets. The instructions and user interface shown to the workers is displayed in Fig. 7.

Each worker was allowed a total of 1 min to complete one comparison question, and the payment per response was $\$0.02$. Answering each question typically took less than 15s, leading to an approximate hourly rate of $\$5$. A total of approximately $\$150$ was spent on all the evaluations.

C Network architecture

Our framework can be divided into two main parts – the encoder, consisting of the keypoint detector, and the decoder, which performs source feature warping and outputs an image. The keypoint detector for $K = 20$ keypoints is visualized in Fig. 8. It follows the structure used in FOMM [26], with the

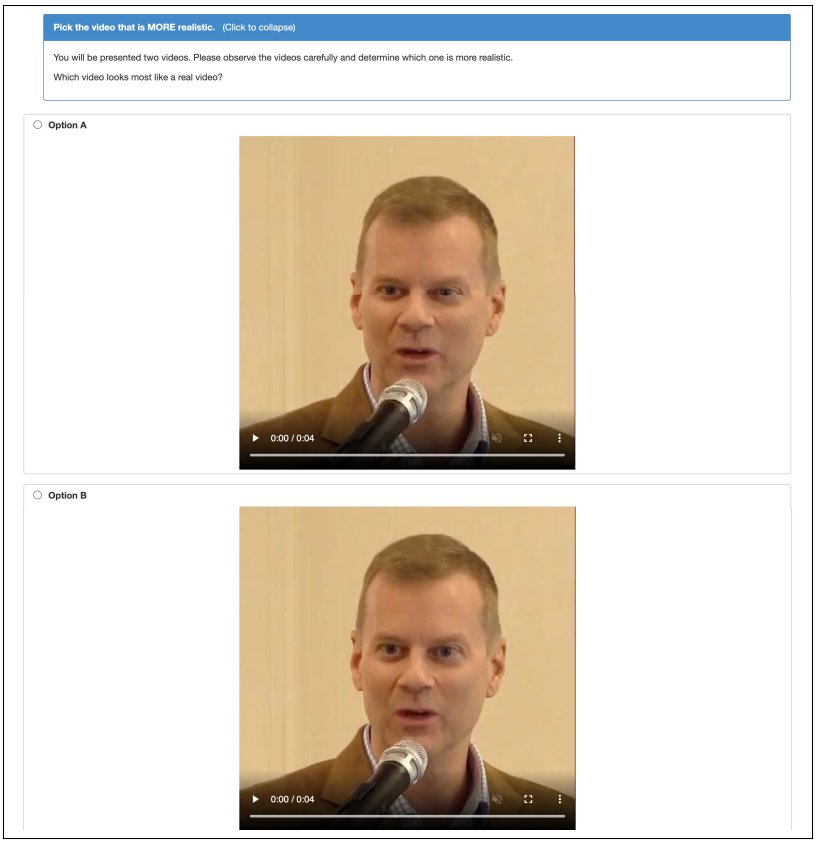

Fig. 7: User interface provided to MTurk users.

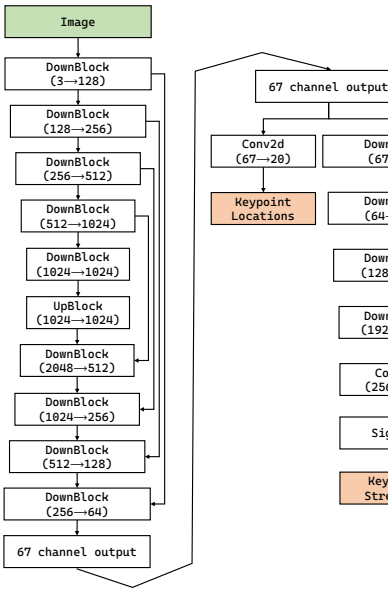

Fig. 8: Keypoint detector

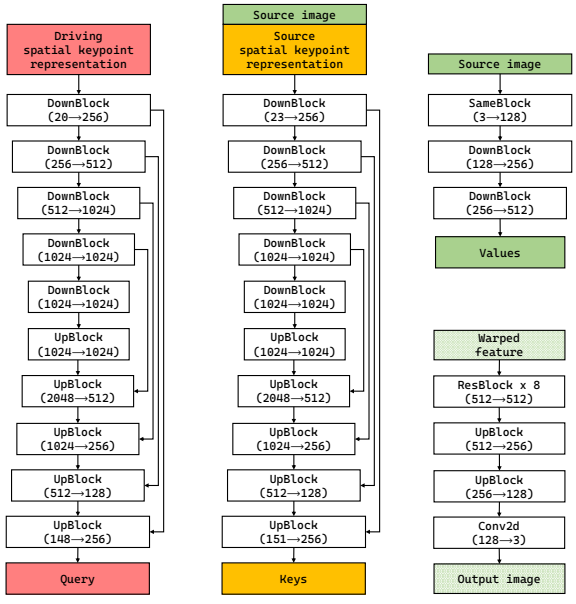

Fig. 9: Query, key, value, and decoder networks

Each `DownBlock` consists of one 3×3 convolution followed by a 2×2 max pooling layer. Each `UpBlock` consists of a 3×3 convolution followed by a $2\times$ bilinear upsampling. Each `ResBlock` consists of a pre-activation residual block with 3×3 convolution.

addition of a branch to also predict a scalar keypoint strength per keypoint. As mentioned in Section 3 of the main paper, we use U-nets to encode the spatial keypoint representations and obtain queries and keys. These U-nets are visualized in Fig. 9. The U-net for obtaining keys also takes the source image as input, along with the source keypoint representations. The keypoint detector and the two U-nets just described take inputs at $1/4$ the image resolution, *i.e.* 64×64 for 256×256 images.

We use a simple 3-layer network to encode the source image and obtain source features. Takes takes the full-resolution source image as input. After warping the source features using the cross-modal attention block discussed in Section 3 and Fig. 3 of the main paper, we pass it to a decoder network consisting of 8 residual blocks and 2 upsampling blocks.

D Experiments

Training details

Our models were trained on an internal server made up of NVIDIA DGX servers, each containing 8 A100 40GB GPUs. Training at 256×256 resolution took approximately 3.5 days. Finetuning at 512×512 took less than a day. In all experiments, we used a batch size of 8 with the Adam optimizer, with an initial learning rate of 0.0002, and $(\beta_1, \beta_2) = (0.5, 0.999)$. We reuse the discriminator architecture from face-vid2vid [41] for training with the GAN loss. The other losses used to train our network were the perceptual loss based on VGG-19, and the equivariance loss used in FOMM [26].

- TalkingHead-1KH [41]: Trained at 256×256 for a total of 120 iterations, with learning rate dropped by a factor of 10 after 100 iterations. Finetuned at 512×512 for 30 epochs.
- VoxCeleb2 [26]: Trained at 256×256 for a total of 100 iterations, with learning rate dropped by a factor of 10 after 70 iterations.
- TED Talk [28]: Trained for a total of $100,000$ iterations.

Ablation results

While developing our network, we ran ablations at 256×256 resolution on the TalkingHead-1KH dataset. Each network configuration was trained from scratch 3 times, and we report the mean and standard deviations of metrics obtained from those experiments in Table 5. For training the 512×512 model, we only chose and finetuned the best 256×256 model due to lack of computing resources.

Table 5: Ablations on the TalkingHead-1KH dataset at 256×256 resolution.

Residual connection	✘	✔	✔
Extra key-value	✘	✘	✔
PSNR ↑	$23.253_{\pm 0.21}$	$23.483_{\pm 0.12}$	$\mathbf{23.582}_{\pm 0.30}$
$\mathcal{L}_1$ ↓	$12.741_{\pm 0.30}$	$12.721_{\pm 0.26}$	$\mathbf{12.632}_{\pm 0.20}$
LPIPS ↓	$0.122_{\pm 0.005}$	$0.114_{\pm 0.02}$	$\mathbf{0.113}_{\pm 0.004}$
AKD (MTCNN) ↓	$1.969_{\pm 0.02}$	$1.914_{\pm 0.01}$	$\mathbf{1.895}_{\pm 0.08}$
FID ↓	$19.978_{\pm 0.83}$	$18.462_{\pm 0.15}$	$\mathbf{18.252}_{\pm 0.66}$