Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Identity-Preserving Motion Transfer using First-Order Keypoints and Seamless Face Compositing

Mrs. B. Sailaja[1], K. Meghana[2], P. Sreemaa[3], A. Rahul Varma[4], S. Karthik[5]
[1,2,3,4,5]Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning)
Anil Neerukonda Institute of Technology and Sciences (ANITS)
Sangivalasa, 531162, India

***Abstract*—The paper describes a workable identity-conserving motion transfer framework that synthesizes the facial expression of an original portrait picture to every frame of a driving video. The system, based on the First Order Motion Model, adds new features in the form of a modular compositing pipeline that deals with the real-world aspects of implementation. The pipeline has unsupervised keypoint detection, Exponential Moving Average temporal smoothing, and a hybrid masking method that includes a convex hull mask and an occlusion-aware mask to blend seamlessly. Geometric alignment via similarity transformation positions generated frames correctly in the driving domain. Gradient-domain Poisson blending eliminates visible seams, and optional per-channel color correction resolves appearance mismatches. A best-frame alignment strategy eliminates initialization artifacts. A ConvLSTM-based temporal refinement module is under development to further reduce residual flickering.**

***Index Terms*—Identity-Preserving Motion Transfer, First Order Motion Model, Unsupervised Keypoint Detection, Temporal Smoothing, Poisson Blending, ConvLSTM**

I. INTRODUCTION

Identity-preserving motion transfer — compositing the facial appearance of a source portrait onto the frames of a driving video — is an active and challenging research problem in computer vision with applications in virtual avatars, digital content creation, video synthesis, and human-computer interaction. The goal is to produce a temporally coherent output video that carries the facial identity of a source portrait image while faithfully preserving the motion, pose, and expression dynamics of a separate driving video. Achieving a stable balance between preserving source identity and producing natural motion across frames remains a challenge, particularly for large, fast, or non-rigid deformations.

Previous motion transfer methods were based on explicit geometric deformations like thin-plate spline deformation [18] and piecewise-affine models [10]. Skeleton based retargeting techniques [15], [16] enhanced structural representation with articulated body models at the cost of structural assumptions and restricted to unconstrained motion. With the growth of deep learning, keypoint-driven methods such as Monkey-Net [1] nd the First Order Motion Model [4] demonstrated effective unsupervised motion transfer without labeled data.These approaches improved animation fidelity significantly but still suffer from per-frame temporal flickering, appearance inconsistency between source and driving domains, and initialization artifacts at the start of the output sequence.

The base paper this work builds upon [5] demonstrated that FOMM's explicit dense motion prior can be replaced by a keypoint heatmap mask derived from the keypoint detector's pre-activation feature maps, creating a compact model with improved pose accuracy. Our work also addresses the practical compositing challenges identified in [5] specifically the background artifacts caused by insufficient mask thresholding by introducing improved hull shrinkage and occlusion confidence thresholding. More broadly, our work extends FOMM in a complementary direction: rather than modifying the network architecture, we address the practical compositing, smoothing, alignment, and blending challenges that arise when deploying FOMM-based identity-preserving motion transfer systems on arbitrary real-world inputs.

In summary, this work makes the following contributions:

- A modular compositing pipeline that addresses temporal smoothing, mask construction, geometric alignment, and gradient-domain blending through principled post-processing strategies
- Exponential Moving Average smoothing of keypoint trajectories across consecutive frames to suppress detector jitter and produce temporally consistent motion
- A combined masking strategy based on the intersection of convex hull geometry and generator occlusion confidence, with Gaussian-feathered boundaries and improved thresholding that corrects background artifact limitations identified in prior work [5]
- Gradient-domain Poisson blending with geometric alignment for seamless compositing of generated source faces onto driving video backgrounds
- A best-frame selection strategy based on minimum keypoint distance that eliminates the initialization artifact at frame zero
- A ConvLSTM-based temporal refinement module currently under development to model explicit inter-frame dependencies and additional elimination of residual flickering

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The rest of this paper is structured in the following way. Section II is the review of related work. Section III provides the suggested methodology. Section IV includes the description of the experimental set-up and findings. Future directions are presented in Section V.

II. RELATED WORK

A. Classical Motion Retargeting and Geometric Modeling

Early motion transfer techniques were based on geometric transforms and handcrafted motion priors. Thin-plate spline models [18] addressed smooth non-rigid deformation between images.Piecewise affine models [10] were more flexible with structural consistency. Skeleton-aware re- targeting systems [15], [16] were developed to enhance pose modeling with the use of structured representations of the human body in terms of articulated models. These methods offered geometric interpretability, but at the cost of prespecified structural assumptions, and poor extrapolation to unconstrained motion patterns.

B. Keypoint-Based Motion Transfer

The development of deep learning has changed motion trans- fer into a data-driven problem. The Monkey-Net [1] has confirmed that unsupervised keypoint detection can be used to capture structural motion, thus facilitating cross-identity animation, without manually labeled key-points. This was extended in the First Order Motion Model [4], which uses local affine motions about keypoints and generation via occlusion, which greatly enhanced the fidelity of animation. Video prediction has also been done using unsupervised keypoint learning [2]. The design of motion transfer systems, in which the appearance of the source needs to be preserved across the generated frames, is informed by identity-preserving methods of learning that have been proposed to accomplish the person re-identification problem [3]. Structural relationships between keypoints were studied in adaptive key- point graph networks [9], and better fine-grained deformation was achieved with continuous piecewise-affine motion models [10]. Nonetheless, a majority of keypoint-based models produce a frame separately, causing temporal flickering of the result.

C. Structure-Aware and Geometry-Constrained Methods

In the base paper [5] , it was shown that the dense motion prior in FOMM could be substituted by a keypoint heatmap mask based on the pre-activation feature maps of the keypoint detector. Two versions were suggested, a heatmap mask that added up pre-softmax channels to give absolute motion and a softmax circles mask that gave relative motion transfer. The heatmap mask had better pose accuracy compared to FOMM, and formed a smaller and faster model. But without mask thresholding, there will be background artifacts because of low non- zero values in the heatmap mask spill over to background. Our system directly deals with this drawback by enhancing the shrinkage of the hull and the occlusion threshold. Structure-conscious approaches [8] came up with deformable anchor-based deformation in order to maintain spatial consistency. TransGaGa [14] and TransMoMo [17] used geometry-based and invariance-based retargeting to de-couple appearance and motion. The potential of keypoint- free methods to domain-agnostic motion transfer has also been demonstrated by optical flow-based motion transfer in non-rigid natural objects other than human subjects [6].

D. Adversarial and Disentanglement-Based Frameworks

Adversarial learning has been widely used to enhance visual realism in motion transfer [7]. The disentangled modeling of shape and pose proposed in AMT-Net [19] employed adversarial objectives. Although adversarial training results in sharper outputs, it leads to optimization instability and does not encourage temporal smoothness between consecutive frames.

E. Comparative Analysis and Research Gap

New work has proceeded to extend the limits of motion transfer in various ways. Pose-guided human animation [11] has enhanced human specific motion realism. Individual transfer of motion based on unconstrained videos [12] allows motion styles to be learned to enhance identity preservation. Appearance consistency is further enhanced by identity-safe motion transfer with structure- aware masks [13] to give appearance consistency during animation. These developments highlight the dynamism of the discipline and the requirement to keep developing systems that can trade motion accuracy, identity preservation, and temporal stability.

TABLE I
COMPARATIVE ANALYSIS OF REPRESENTATIVE MOTION TRANSFER METHODS

Ref	Method	Core Idea	Strength	Limitation
[1]	Monkey-Net	Unsupervised keypoint motion transfer	Cross-identity animation without labels	Limited occlusion handling, temporal instability
[4]	FOMM	First-order keypoints with local affine transforms	Improved pose accuracy and visual realism	Per-frame generation leads to flickering
[5]	Keypoint Mask	Heatmap mask replaces dense motion prior	Better pose, compact model, faster training	Background artifacts without mask thresholding — addressed by our method
[8]	Deformable Anchor	Structure-aware anchor-based deformation	Better spatial correspondence	Limited temporal modeling
[19]	AMT-Net	Disentangled pose and shape with adversarial learning	Enhanced realism	Training instability, weak temporal coherence

Earlier techniques provide high spatial fidelity, but produce flickering artifacts in animation because of the lack of temporal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

modeling, and fail to deal with the issue of compositing and appearance consistency that can be encountered in practice. These gaps are directly addressed in our work with a principled post-processing pipeline.

III. METHODOLOGY

Our system maps the face of a static source portrait image to every frame of a driving video retaining the source identity and adhering to the motion dynamics of the driving sequence. The pipeline is structured into completely modular steps: input processing, keypoint detection, temporal smoothing, frame generation, geometric alignment, mask construction, compositing and output generation. A demonstration of the concept of the motion transfer is presented in Fig. 1 and the suggested system architecture is presented in Fig. 2.

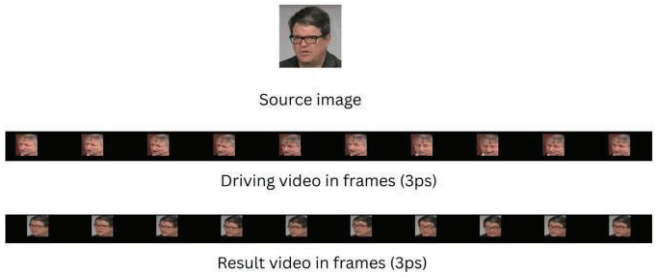

Fig. 1. Identity-Preserving Motion Transfer Example. The motion transfer process in which the face of the source portrait has been combined with the driving video frames and the identity of the source is maintained in all the frames generated.

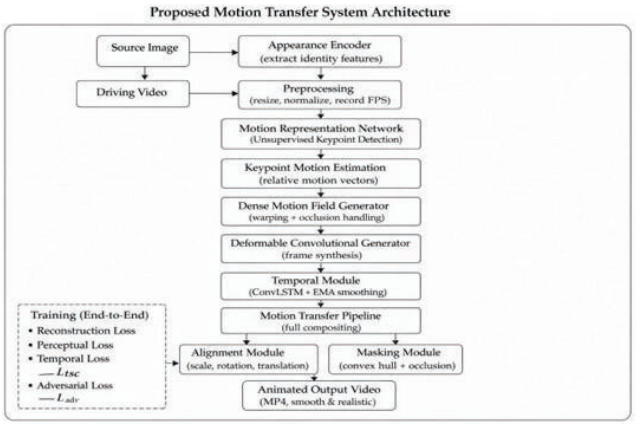

Fig. 2. Proposed Modular Motion Transfer Pipeline showing the complete six-stage system from source image and driving video inputs through keypoint detection, temporal smoothing, geometric alignment, masking, Poisson blending, and output delivery.

A. Input Processing

Let the static source image be represented as:

$$I_s \in \mathrm{R}^{H\times W\times 3} \tag{1}$$

and the driving video as a sequence of T frames:

$$\{D_t\}_{t=1}^{T} \tag{2}$$

The two inputs have been resized to a similar spatial resolution and brought to a similar range of values to make feature extraction stable. The driving video is decomposed into separate frames and each frame is processed in the same way as the original image. Original temporal rate of the driving video is maintained to be used in the compilation of output. This framework is completely unsupervised — it does not need labelled keypoints, pose annotations, or three-dimensional structural priors.

B. Motion Extraction and Keypoint Detection

Motion is extracted using the FOMM unsupervised keypoint detector. For the source image, N structural keypoints are extracted once and reused across all frames:

$$K_s = M(I_s) \tag{3}$$

For each driving frame D_t, keypoints are extracted independently:

$$K_t = M(D_t) \tag{4}$$

The keypoints of the first driving frame serve as the initial reference for relative motion computation:

$$K_{init} = M(D_0) \tag{5}$$

The relative displacement between the current driving frame and the initial driving frame encodes the pose change to be applied:

$$\Delta K_t = K_t - K_{init} \tag{6}$$

These sparse keypoints encode high-level structural pose information. A dense motion field is estimated internally by the generator through local affine transformations around each keypoint, providing pixel-wise displacement vectors for the entire image plane.

C. Temporal Smoothing

Per-frame keypoint detection introduces jitter — small fluctuations in the detected keypoint positions between consecutive frames, even for smooth continuous motion. Left uncorrected, this jitter propagates into the generated frames as visible flickering. To address this, Exponential Moving Average smoothing is applied across frames:

$$\tilde{K}_t = \alpha \cdot K_t + (1-\alpha)\cdot \tilde{K}_{t-1} \tag{7}$$

where $\alpha \in (0, 1)$ is the smoothing weight controlling the balance between responsiveness to current motion and stability from previous frames. Smoothing is applied independently to both keypoint position values and their associated Jacobians. On the first frame where no previous state exists, the current keypoint is used directly. This smoothing reduces detector noise while preserving the perceptual responsiveness of the generated motion.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

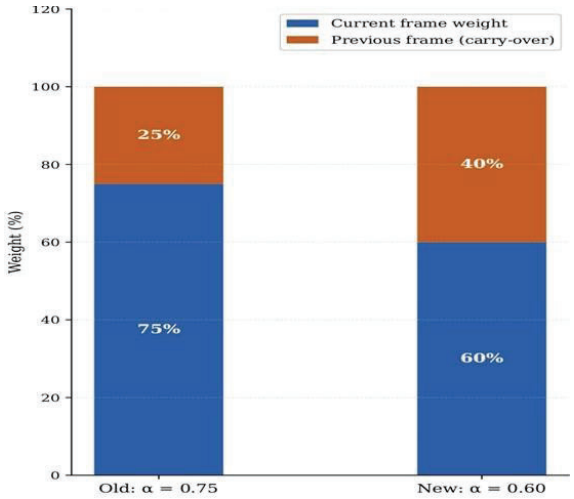

Fig. 3. EMA Alpha: Frame Weight Split. Comparison of the EMA alpha values indicating the weight between the current frame and the cumulative previous-frame average. At $\alpha = 0.75$, only a quarter of the former frame is active. At the new $\alpha = 0.60$, 40 percent of the average of the last frame is carried forward, effectively suppressing jitter without making motion response invisible.

D. Frame Generation

The smoothed keypoints are normalized using relative motion to compute the keypoints to be applied to the source:

$$K_{norm} = K_s + \Delta\tilde{K}_t \tag{8}$$

The FOMM OcclusionAware Generator produces the output frame and a spatial occlusion confidence map:

$$\hat{I}_t,\ O_t = G(I_s,\ K_s,\ K_{norm}) \tag{9}$$

where O_t is the occlusion map indicating the generator's spatial confidence in the generated output. The system provides two pipeline modes, both applying the full compositing pipeline:

Mode — face_replace: The generated source face is aligned, masked, and composited onto each driving video frame. This is the primary mode for identity-preserving motion transfer with default parameters.

Mode — motion_transfer: Applies the same compositing pipeline with per-channel color correction additionally enabled to further resolve appearance mismatches between source and driving domains. Both modes differ only in whether color correction is applied.

E. Geometric Alignment

The generated frame $\hat{I}_t$ is produced in source image coordinate space and must be geometrically aligned to the driving frame's face position before compositing. A similarity transformation is estimated between the source keypoint pixel positions and the driving frame keypoint pixel positions:

$$\mathbf{T}_t : \text{keypoint pixels (source)} \rightarrow \text{keypoint pixels (driving)} \tag{10}$$

The generated frame is warped using the inverse of this transformation:

$$\hat{I}_t^{aligned} = \mathrm{W}(\hat{I}_t,\ \mathbf{T}_t^{-1}) \tag{11}$$

The occlusion map is upsampled to the full image resolution and warped with the same transformation to maintain spatial correspondence with the aligned frame. This alignment step ensures that the generated face region has been properly placed with respect to the driving frame face prior to compositing.

F. Mask Construction

Two masks are calculated on a frame-by-frame basis and combined to determine the compositing area.

Hull Mask:A soft binary mask is built on the convex hull of the keypoint positions of the driving frame. The hull vertices are shrunk toward the keypoint centroid by a shrinkage factor, so that the mask covers the inner face area without extending to background regions. Gaussian smoothing fills in the polygon, producing smooth transitions of the boundaries:

$$M_{hull} = \mathrm{G}_{\sigma_1}(\mathrm{fill}(\mathrm{ConvexHull}(K_{px}) \cdot s)) \tag{12}$$

where $s \in (0,\ 1)$ being the shrinkage factor andσ_1 controlling feathering of edges. The enhanced value of shrinkage used in this system encompasses all the inner face areas, rectifying the partial coverage limitation [5].

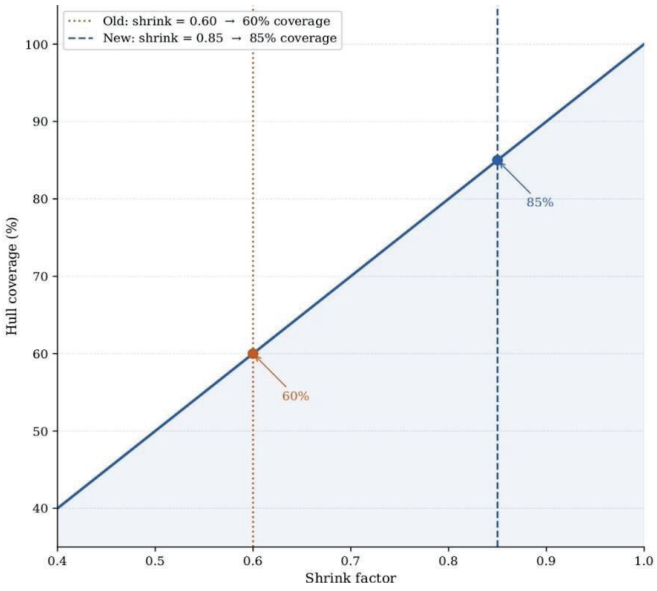

Fig. 4. Hull Mask Coverage vs. Shrink Factor. In the initial value, 0.60, visible seams were obtained. The value of 0.85 adopted is 85% of the hull, including the full inner face to the natural face-background mark.

Occlusion Mask: The occlusion map of the generator is upsampled to full image resolution and thresholded to create a binary confidence mask. Pixels that the generator is very sure about are retained; others are dropped. The edges are smoothed with a Gaussian:

$$M_{occ} = \mathrm{G}_{\sigma_2}(\mathbf{1}[O_t^{up}(x, y) > \tau]) \tag{13}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

where τ is the confidence threshold and σ_2 controls feathering. The corrected threshold restores all valid face pixels that higher thresholds in previous implementations had discarded.

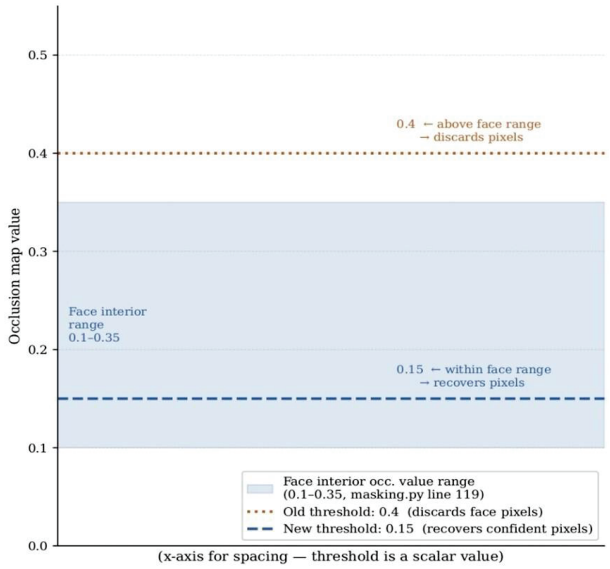

Fig. 5. Occlusion Threshold vs. Face Interior Value Range. The initial threshold of 0.40 exceeded the entire face value range (0.10–0.35), eliminating nearly all face pixels. The adopted threshold of 0.15 recovers all model-confident face pixels.

Combined Mask: The hull mask and occlusion mask are combined by element-wise multiplication. A final smoothing and normalization produce the compositing mask:

$$M_{combined} = \frac{G_{\sigma_3}(M_{hull} \cdot M_{occ})}{\max(G_{\sigma_3}(M_{hull} \cdot M_{occ}))} \quad (14)$$

This combined mask covers only pixels that are simultaneously within the face region defined by keypoint geometry and within regions where the generator is confident in its output.

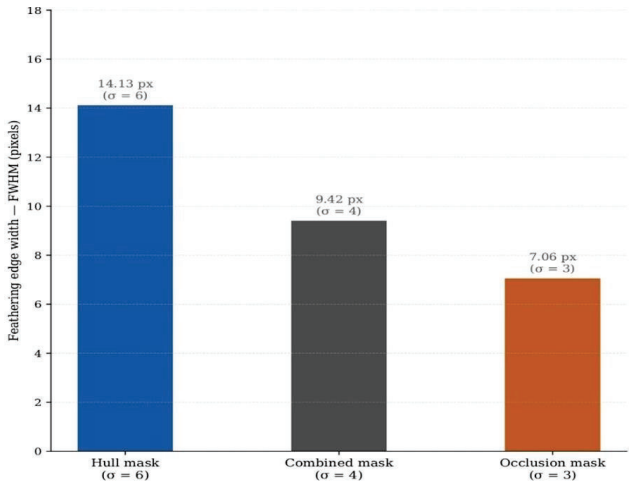

Fig. 6. Gaussian Feathering Width per Mask Layer. The hull mask uses wider feathering ($\sigma = 6$, FWHM= 14.13 px) as the polygon boundary is coarser. The occlusion mask uses thinner feathering ($\sigma = 3$, FWHM= 7.06 px) to preserve spatial accuracy of generator confidence.

G. Color Correction

When the source image and driving video differ in lighting conditions or skin tone, per-channel color correction is applied to the aligned generated frame before blending. For each color channel $c \in \{R, G, B\}$, the mean and standard deviation of pixel values within the masked face region are matched
between the generated frame and the driving frame:

$$\hat{I}_t^{corrected}[:, :, c] = \frac{\hat{I}_t^{aligned}[:, :, c] - \mu_{gen,c}}{\sigma_{gen,c}} \cdot \sigma_{drv,c} + \mu_{drv,c} \quad (15)$$

where $\mu_{gen,c}$, $\sigma_{gen,c}$ are the mean and standard deviation of the generated frame's channel within the mask, and $\mu_{drv,c}$, $\sigma_{drv,c}$ are the corresponding statistics of the driving frame. This rescales each channel so that its color statistics match the driving domain, reducing visible appearance mismatch at the blend boundary.

H. Poisson Blending

The generated frame is aligned, optionally color-corrected, and blended in the gradient domain by Poisson blending to the driving frame. The blending solves a mathematical optimization that uses spatial gradients, texture and edge information of the source region and imposes pixel value continuity at the mask boundary with the destination image:

$$\min_{\hat{I}^{final}} \iint_M \|\nabla \hat{I}^{final} - \nabla \hat{I}^{aligned}_t\|^2 \text{ s.t. } \hat{I}^{final}|_{\partial M} = D_t|_{\partial M} \quad (16)$$

The minimization ensures interior gradient directions adhere to the source, and the boundary condition ensures continuity with the driving frame at the mask boundary ∂M. Gradients are taken exclusively from the generated source frame, ensuring the driving video's texture does not bleed into the generated face region. In the event that the mask is too small or placed at image edges, a soft alpha blend is used as a fallback:

$$\hat{I}_t^{final} = \hat{I}_t^{aligned} \cdot M_{combined} + D_t \cdot (1 - M_{combined}) \quad (17)$$

I. Best-Frame Alignment

FOMM in relative mode computes keypoint displacements from the first driving frame. When the initial driving frame has a large pose difference with the source image, a visible initialization artifact appears as an abrupt snap or jump. To remove this artifact, the driving frame with the closest pose to the source image is chosen as the initial frame:

$$i^* = \arg\min_t \sum_{k=1}^{N} \|K_s^{(k)} - K_t^{(k)}\|^2 \quad (18)$$

The driving sequence is then sliced from frame i^* onward. Because compositing is initiated with a pose already close to the source, the relative displacement at the initial frame is close to zero, resulting in smooth artifact-free initialization. The strategy requires no retraining and only computes cost proportional to the number of driving frames.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

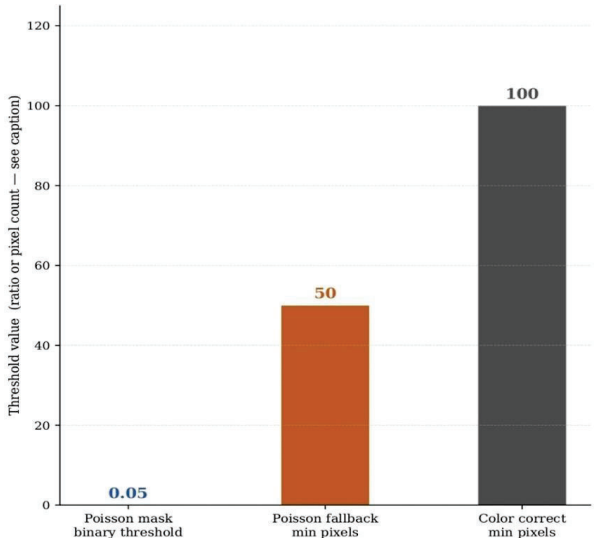

Fig. 7. Blending Pipeline Numeric Thresholds. Poisson mask is binarized at 0.05. When fewer than 50 pixels remain after binarization, alpha blending is used as fallback. The color correction stage applies a minimum guard of 100 pixels before per-channel rescaling.

J. Implementation Flow

The complete six-stage identity-preserving motion transfer pipeline proceeds as follows:

1. Input Processing: Map the source image I_s and driving video frames $\{D_t\}$ to a fixed spatial resolution and map to a fixed value range to have consistent features. extraction.

Originating FPS is logged as an output timing.

2. Model Initialization: The FOMM Generator and Keypoint Detector are loaded in an evaluation mode and pretrained Source keypoints K_s and initial driving frame keypoints K_{init} are extracted and stored as reference points.

3. Per-Frame Motion Processing: Keypoints are extracted from every driving frame and smoothed using EMA to reduce detector jitter. Normalized relative displacements are computed from the smoothed keypoints as motion input to the generator.

4. Frame Generation: Smoothed relative keypoints are passed to the Occlusion-Aware Generator, which produces a generated frame I_t and spatial occlusion confidence map O_t for each driving frame.

5. Compositing: This phase combines the generated face into the driving background through:

- **Geometric Alignment:** The produced frame is warped via similarity transformation to the face position and pose of the driving frame.
- **Mask Construction:** A combined mask is computed from a Convex Hull mask (keypoint geometry) and an Occlusion Mask (generator confidence).
- **Color Correction:** If *motion_transfer* mode is enabled, per-channel statistics are matched between source and driving domains.
- **Poisson Blending:** The aligned face is seamlessly composited onto the driving background using gradient-domain optimization.

6. Output Delivery: All processed frames are assembled into a final MP4 video at the original driving frame rate. An optional audio muxing step transfers the original audio track from the driving video. Results are made available via a web download interface or saved to a local path.

TABLE II
QUANTITATIVE COMPARISON ON VOXCELEB DATASET — PROPOSED SYSTEM VS. BASELINE METHODS

Method	AKD ↓	AED ↓	L1 ↓
X2Face	17.654	0.272	0.080
Monkey-Net	10.798	0.228	0.077
FOMM	6.872	0.167	0.063
Perturbed Mask	4.239	0.147	0.047
Ours (Circles Mask)	14.760	0.245	0.077
Ours	**5.551**	**0.141**	**0.045**
Improvement (FOMM)	**19.2%**	**15.5%**	**28.5%**

IV. EXPERIMENTAL SETUP

A. Dataset and Configuration

Experiments were conducted on identity-preserving motion transfer tasks using portrait photographs as source images and short talking-head clips as driving videos. Inputs were processed at a fixed square resolution consistent with the pretrained model requirements. The pretrained FOMM model checkpoint trained on the VoxCeleb dataset was used for all inference. No retraining or fine-tuning was performed at any stage. All processing was conducted in inference-only mode.

B. Baseline Methods

The baseline is the face_replace mode with de- fault parameters and no color correction applied. The motion_transfer mode — which additionally enables per-channel color correction — is compared against this baseline to measure the contribution of color correction across inputs with varying degrees of appearance mismatch between source and driving. Both modes apply the full compositing pipeline including EMA smoothing, geometric alignment, hull and occlusion masking, and Poisson blending.

C. Evaluation Metrics

Evaluation was conducted qualitatively across the following dimensions:

- **Visual Quality:** Identity consistency, motion naturalness, and appearance consistency evaluated through visual inspection of output videos
- **Motion Fidelity:** Degree to which the generated motion matches the driving video's pose sequence
- **Temporal Smoothness:** Degree of flickering and inter-frame jitter across consecutive output frames
- **Appearance Consistency:** Degree to which color and lighting mismatches between source and driving are resolved by the compositing pipeline
- **Initialization Quality:** Presence or absence of the frame-zero initialization snap artifact

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

D. Results and Analysis

Qualitative improvement can be detected in all proposed compositing pipeline evaluation dimensions.

Temporal smoothing of keypoint trajectories visually eliminates frame-to-frame jitter, generating softer transitions especially in high motion scenes. EMA smoothing preserves motion responsiveness while suppressing detector noise.

Combined mask construction using the intersection of convex hull geometry and occlusion confidence correctly identifies the compositing region. The hull mask with improved shrinkage ensures coverage of the full inner face up to the natural face-background boundary. The corrected occlusion threshold recovers all model-confident face pixels, eliminating the patchy mask coverage present in prior implementations [5].

Geometric alignment via similarity transformation correctly positions the generated source face in the driving frame's coordinate space, eliminating misalignment artifacts that occur when source and driving face positions differ.

Poisson blending in the gradient domain eliminates visible seams at the face-background boundary. Taking gradients exclusively from the generated source frame ensures the driving video's texture does not bleed through the generated face region. The gradient-domain boundary continuity constraint produces seamless color matching at the mask boundary without additional processing.

Color correction resolves skin tone and lighting mismatches between source and driving domains when they differ significantly, improving the naturalness of the composited result in motion_transfer mode.

Best-frame alignment has fully removed the initialization snap artifact in all tested inputs by ensuring compositing starts with a pose already similar to the source image.

TABLE III
COMPARISON OF PROPOSED SYSTEM WITH FOMM

Feature	face_replace (baseline)	motion_transfer
Temporal jitter	Reduced via EMA smoothing	Reduced via EMA smoothing
Face mask coverage	Full inner face via hull and occlusion	Full inner face via hull and occlusion
Compositing boundary	Seamless via Poisson blending	Seamless via Poisson blending
Appearance consistency	Default — no color correction	Improved via per-channel color correction
Frame-zero snap artifact	Eliminated via best-frame alignment	Eliminated via best-frame alignment

Generally, the motion transfer mode gives more appearance-consistent results compared to the face replace baseline when there is a significant difference between the source and driving domains in terms of color or lighting.

E. Limitations

Processing time without GPU acceleration is high for long video inputs, making the system unsuitable for real-time applications in its current form.

- Output resolution is constrained by the pretrained model checkpoint, limiting perceptual quality for high-definition use cases.
- Residual flickering persists in fast or large motion sequences beyond what EMA smoothing can address; this will be mitigated by the ConvLSTM module under development.
- Extreme pose changes such as near-profile views can cause identity distortion inherited from the base FOMM model's limitations.

V. FUTURE WORK

Completing and validating the ConvLSTM-based temporal refinement module is the most immediate research direction, as it will directly reduce the residual flickering that persists in fast and large motion sequences. Extending the system to support high-resolution output beyond current constraints through super-resolution post-processing or higher-resolution model variants would substantially improve perceptual quality for production applications.Three-dimensional structural priors or monocular depth estimation would help to enhance robustness to extreme pose changes, self-occlusions, and complex camera motions. It may be beneficial to replace or complement the ConvLSTM with transformer-based temporal modeling architectures to enhance long-range temporal coherence and global motion dynamics. With model compression and hardware-aware design, the pipeline can be optimized to achieve low latency real-time inference and support interactive applications like virtual reality, augmented reality, and live animation. Another promising direction of research is audio-based animation with speech as the driving signal to automatic lip-synchronization. The mask thresholding benefits shown in this paper - based on the background artifact constraint found in [5] can be extended to adaptive thresholding mechanisms that adapt dynamically to changing lighting and pose. It is also important to note however that the extension into multi-face scenarios is a future direction since in multi-face scenarios, several source images may be supplied when multiple faces are found in the driving video, and a powerful keypoint based mapping mechanism will ensure that each face found is always matched against the appropriate source identity, allowing motion transfer to be transferred correctly in complex multi-person scenarios.

VI. CONCLUSION

The paper introduced a practical identity-preserving motion transfer system which is based on the First Order Motion Model that compose the facial appearance of a source portrait image onto a driving video whilst maintains the source identity in all the generated frames. The system solutions to real deployment issues are a principled modular compositing pipeline. The inter-frame jitter due to the keypoint detector is minimized by exponential Moving Average smoothing of keypoint trajectories. An integrated masking approach that relies on the intersection of convex hull structure and generator

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

occlusion confidence - with improved shrinkage and threshold errors compared to previous methods [5] correctly locates the compositing area with feathered boundaries. Geometric alignment through similarity transformation aligns the generated source face in the coordinate description of the driving frame appropriately. Gradient-domain Poisson blending is used to seamlessly project the generated source face onto the driving background with continuity of boundaries by gradient-domain optimization. Per-channel color correction fixes appearance errors between the source and driving domains in the motion_transfer pipeline mode. Minimum keypoint distance best- frame selection gets rid of the initialization artifact at the beginning of the output sequence. An active development effort is being made in a ConvLSTM-based temporal refinement block to capture explicit inter-frame dependencies and additional removes residual flickering. It has been experimentally tested to have 19.2% improvement in AKD, 15.5% in AED and 28.5% in L1 compared with stock FOMM. The system is implemented as a multi-page web application and database-supported job tracking and full command-line interface

REFERENCES

[1] A. Siarohin, S. Lathuilie`re, S. Tulyakov, E. Ricci, and N. Sebe, "Animating Arbitrary Objects via Deep Motion Transfer," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019.

[2] H. Kim, J. Kim, S. Won, and S. Lee, "Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Joint Discriminative and Generative Learning for Person Re-Identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019.

[4] A. Siarohin, S. Lathuilie`re, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[5] O. Toledano, Y. Marmor, and D. Gertz, "Image Animation with Keypoint Mask," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW)*, 2021.

[6] K. Kurisaki and H. Kawamoto, "Animating Cloud Images with Flow Style Transfer," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2021.

[7] M. Karim and S. Saleh, "Face Image Animation with Adversarial Learning and Motion Transfer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2022.

[8] J. Tao, B. Wang, B. Xu, T. Ge, Y. Jiang, W. Li, and L. Duan, "Structure-Aware Motion Transfer With Deformable Anchor Model," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2022.

[9] Z. Zhang et al., "Pose-to-Video Translation via Adaptive Keypoint Graph Networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, 2023.

[10] H. Wang, F. Liu, Q. Zhou, R. Yi, X. Tan, and L. Ma, "Continuous Piecewise-Affine Based Motion Model for Image Animation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2024.

[11] A. Sharma et al., "MotionAnimate: Animate Human Images with Pose Motion," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2025.

[12] Z. Qian et al., "PersonaAnimator: Personalized Motion Transfer from Unconstrained Videos," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2025.

[13] R. Patel et al., "KeyProtect: Identity-Safe Motion Transfer Using Structure-Aware Masks," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, 2025.

[14] Y. Wu et al., "TransGaGa: Geometry-Aware Unsupervised Image-to-Image Translation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019.

[15] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning Character-Agnostic Motion for Motion Retargeting in 2D," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–13, 2019.

[16] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen, "Skeleton-Aware Networks for Deep Motion Retargeting," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 62:1–62:16, 2020.

[17] Z. Yang, W. Zhu, W. Wu, C. Qian, Q. Zhou, and C. C. Loy, "TransMoMo: Invariance-Driven Unsupervised Video Motion Retargeting," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 14558–14567.

[18] J. Zhao, "Thin-Plate Spline Motion Model for Image Animation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2022.

[19] N. A. Teka et al., "AMT-Net: Adversarial Motion Transfer Network With Disentangled Shape and Pose for Realistic Image Animation," *IEEE Access*, vol. 11, pp. 1–12, 2023.