scisearch INTERNATIONAL PEER-REVIEWED JOURNALS

RESEARCH PAPER

One-Shot Visual Identity Transfer in Dynamic Scenes Using Target Images

Nguyen Kiet[1] and Tran Anh[2]

[1]Mekong Tech College, Department of Computer Science, Tran Quang Dieu Street, Can Tho, Vietnam
[2]Pacific Vietnam Institute, Department of Information Systems, Vo Van Kiet Street, Da Nang, Vietnam

Abstract

Visual identity transfer in dynamic scenes refers to the process of generating a video in which the motion, viewpoint, and environment of a source sequence are preserved while the appearance of the central subject is replaced by the identity of a target individual. Applications include personalized content creation, privacy-aware editing, and virtual production, where a single image of a new identity is often the only available reference. This one-shot regime is challenging because the model must infer viewpoint- and illumination-consistent identity features from incomplete observations while maintaining temporal coherence under arbitrary motion. Existing approaches typically rely on multi-view capture, subject-specific fine-tuning, or strong priors that can limit fidelity or generalization. This work studies a general formulation of one-shot visual identity transfer using target images in unconstrained dynamic scenes and introduces a model that separates identity from scene dynamics within a conditional generative framework. A single target image is mapped to a compact identity embedding that drives a spatiotemporal generator conditioned on a source video. The method combines appearance modeling, motion-aware warping, and temporally consistent adversarial training to synthesize videos that preserve source motion and background while approximating the target identity. The study examines the impact of architectural choices, training objectives, and numerical optimization schemes on identity preservation and scene consistency, and analyzes failure modes such as extreme pose mismatch and occlusions. The resulting formulation highlights the interaction between representation, training dynamics, and generalization in one-shot identity transfer for dynamic visual content.

1. Introduction

Visual content creation is increasingly driven by models that synthesize images and videos conditioned on text, reference examples, or structured control signals. A central task in this space is visual identity transfer, in which the appearance of a subject in a source sequence is replaced by the identity of a different person or object while preserving motion, camera trajectory, and background structure. When the available information about the target identity is limited to a single image, the problem becomes a one-shot generalization challenge. The model must extrapolate how the target identity would appear under novel viewpoints, lighting, and deformations that are not present in the reference, while ensuring that the resulting sequence remains coherent and consistent across time.

One-shot visual identity transfer in dynamic scenes combines difficulties from several domains. From a representation perspective, it requires a factorization of the observed scene into identity, motion, and background components such that identity can be substituted without corrupting the remaining factors (1). From an optimization perspective, it involves training a model that generalizes across identities without identity-specific fine-tuning, which requires careful regularization and conditioning strategies. From a numerical standpoint, the learning process involves high-dimensional, temporally correlated data and non-convex objectives driven by reconstruction, perceptual, adversarial, and regularization terms that interact in complex ways.

A direct approach to this problem might attempt to encode the target image into a latent vector and feed it to a generic video generator conditioned on the source frames. However, such a naive scheme is prone to identity drift, temporal flickering, and geometric inconsistencies, especially under rapid motion or large pose variations. A more structured design needs to recognize that identity, at least for human subjects, has both global attributes, such as facial geometry and skin tone, and local attributes, such as fine-grained texture and hair configuration. These factors need to be injected into the generative process at appropriate spatial and temporal scales. In addition, dynamic scenes often contain complex interactions between the subject and environment, including occlusions and motion blur, which require the model to reason about visibility and continuity over time.

In this work, the task is formulated as learning a mapping from a source video and a single target image to an edited video where the source identity is replaced while the motion and scene context are preserved. The formulation explicitly separates identity embedding, motion representation, and spatiotemporal generation. An identity encoder produces a low-dimensional representation of the target appearance. A motion encoder processes the source video to obtain a representation

Table 1. Problem setting for one-shot visual identity transfer in dynamic scenes.

Aspect	Dynamic-scene requirement	One-shot constraint
Input evidence	Source video with complex motion, occlusions, and lighting	Single target image as identity exemplar
Output video	Preserves scene layout, background, and camera motion	Central actor rendered with target identity across all frames
Content preservation	Maintain body kinematics and high-level action semantics	Avoid altering non-identity content such as environment and motion
Identity transfer	Replace facial features, skin and hair texture, and clothing details	Generalize target appearance to unseen poses and views
Temporal coherence	Enforce consistency of appearance and geometry over time	No per-identity fine-tuning or optimization during inference

Table 2. Comparison of supervision assumptions between existing approaches and the proposed formulation.

Setup	Identity supervision	Geometric assumptions	Practical limitations
Multi-view methods	Multiple synchronized cameras and viewpoints	Often rely on calibrated rigs or multi-view geometry	Rarely available in unconstrained, in-the-wild videos
Long reference videos	Dozens or hundreds of images for each identity	Implicitly learn appearance priors from many frames	Expensive data collection and per-identity storage
Explicit 3D models	Parametric face or body templates	Require accurate fitting and controlled capture	Limited appearance diversity and stylization flexibility
Single-image editing	One or few images, static setting	2D alignment and warping dominate	Weak temporal modeling; hard to extend to full videos
Proposed one-shot video transfer	Single target image plus source video	No explicit 3D geometry at test time	Designed for dynamic, unconstrained scenes with minimal identity evidence

of dynamics and camera behavior. A conditional generator composes these elements to synthesize the final video. The mapping is learned in a data-driven manner from collections of videos without requiring subject-specific optimization at inference.

The approach is expressed in a probabilistic and geometric framework (2). Videos are modeled as high-order tensors and random variables, while identities are represented in a latent manifold equipped with a metric induced by feature extractors. Temporal coherence is encouraged by explicit regularizers and by network architectures that share parameters across time. The numerical training procedure uses stochastic gradient methods with carefully designed loss functions and noise schedules. The overall system is analyzed with respect to identity preservation, motion consistency, and robustness to challenging conditions. The remainder of the paper formalizes the problem, describes the model components, discusses the learning objectives and optimization strategy, and presents an experimental study along with a discussion of observed behaviors and limitations.

2. Problem Formulation and Representation

The problem of one-shot visual identity transfer in dynamic scenes is formulated by treating videos and images as random variables defined on appropriate spaces. Let

$$V^{\mathrm{src}} \in \mathbb{R}^{T\times H\times W\times C}$$

denote a source video with T frames and spatial resolution $H\times W$ with C channels. Let

$$I^{\mathrm{tar}} \in \mathbb{R}^{H'\times W'\times C}$$

be a single target image capturing the identity to be transferred. The goal is to generate a video

$$\hat{V}^{\mathrm{tar}} \in \mathbb{R}^{T\times H\times W\times C}$$

that preserves the motion, camera trajectory, and background of V^{src} while approximating the visual identity of the central subject in I^{tar}.

The generative model is represented as a mapping

$$F_\theta : \mathcal{V}\times\mathcal{I}\to\mathcal{V},$$

where $\mathcal{V}$ is the space of videos and $\mathcal{I}$ is the space of images. The parameters θ are learned from data such that

$$\hat{V}^{\mathrm{tar}} = F_\theta\, V^{\mathrm{src}}, I^{\mathrm{tar}}$$

approximates the conditional distribution of videos in which the identity of the subject is that of I^{tar} and the motion and

Table 3. Decomposition of the proposed architecture into content and identity pathways.

Module	Primary role	Inputs	Outputs / influence
Content pathway	Model spatio-temporal scene dynamics	Source video frames, motion cues, camera trajectory	Latent content features encoding background, pose, and layout
Identity pathway	Inject target-specific appearance information	Encoded target image features	Feature-wise modulation signals and cross-attention keys/values
Regression network	Infer modulation parameters in one shot	Single target image (or its embedding)	Identity-conditioned scales, shifts, and attention weights
Temporal regularization blocks	Promote stability across frames	Consecutive latent features or frames	Loss terms or recurrent connections enforcing temporal coherence
Generator backbone	Synthesize output frames	Fused content and identity features	Video frames where source content is rendered with target identity

Table 4. Conceptual disentanglement between content and appearance factors.

Factor	Content-dominated aspects	Appearance-dominated aspects
Background	Scene layout, static objects, environment geometry	Local texture and style of surfaces, lighting color
Body pose	Global body configuration and motion trajectory	Clothing-dependent silhouette variations
Face region	Head pose and coarse alignment with body	Identity traits such as shape, features, and fine texture
Temporal evolution	Action semantics and motion continuity	Frame-wise variation in shading and micro-expressions
Occlusions	Which parts of the actor are visible	How occluders interact with hair, clothing, and skin appearance

background structure are consistent with V^{src}. More formally, the objective can be expressed as modeling a conditional distribution

$$p_\theta\, V^{\mathrm{tar}} \mid V^{\mathrm{src}}, I^{\mathrm{tar}}$$

that concentrates its mass near videos that satisfy the identity and motion constraints.

From a probabilistic viewpoint, one can conceptualize each video as arising from latent identity, motion, and scene variables. Let z^{id} denote an identity latent variable, z^{mot} a motion latent variable, and z^{bg} a background latent variable. A generative process may be abstractly written as

$$V = G z^{\mathrm{id}}, z^{\mathrm{mot}}, z^{\mathrm{bg}},$$

where G is a deterministic function realized by a neural network. The source video V^{src} is associated with latent variables $\left(z^{\mathrm{id}}_{\mathrm{src}}, z^{\mathrm{mot}}_{\mathrm{src}}, z^{\mathrm{bg}}_{\mathrm{src}}\right)$, while the unknown target video corresponds to $\left(z^{\mathrm{id}}_{\mathrm{tar}}, z^{\mathrm{mot}}_{\mathrm{src}}, z^{\mathrm{bg}}_{\mathrm{src}}\right)$, under the assumption that motion and background factors are preserved and identity is replaced. The target identity latent variable $z^{\mathrm{id}}_{\mathrm{tar}}$ must be inferred from the single image I^{tar}.

In practice, the exact generative process and latent factorization are not observed. Instead, the model learns an encoder–decoder structure that implicitly realizes the factorization. The source video is mapped to a motion and background representation, while the target image is mapped to an identity embedding. Let

$$z^{\mathrm{mot}} = E_{\mathrm{mot}}\, V^{\mathrm{src}}, \quad z^{\mathrm{id}} = E_{\mathrm{id}} I^{\mathrm{tar}},$$

where E_{mot} and E_{id} are neural encoders. The generator G_θ then outputs the edited video

$$\hat{V}^{\mathrm{tar}} = G_\theta\, z^{\mathrm{mot}}, z^{\mathrm{id}}.$$

This factorization is enforced not by direct constraints on the latent variables, but by the structure of the network and the losses used during training.

At the level of individual frames, each video frame V^{src}_t can be vectorized as

$$x_t \in \mathbb{R}^n, \quad n = HWC,$$

and the video can be regarded as a sequence $x_1, \ldots, x_T$ or as a single vector in $\mathbb{R}^{nT}$. For modeling spatiotemporal dependencies, it is more natural to view the video as a tensor and to employ convolutions or attention mechanisms that respect spatial and temporal locality. The model defines a family of conditional distributions

$$p_\theta\, x_t \mid x_{<t}, z^{\mathrm{mot}}, z^{\mathrm{id}}$$

Table 5. Generative modeling frameworks considered for controllable video synthesis.

Framework	Strengths for identity transfer	Challenges in dynamic scenes	Typical trade-offs
GANs	High-fidelity, sharp samples with strong discriminators	Training instability and mode collapse for long sequences	Good visual quality, but sensitive to hyperparameters
VAEs	Probabilistic latent representation amenable to conditioning	Blurry outputs and limited detail at high resolutions	Stable training with explicit likelihood bounds
Normalizing flows	Exact likelihood and invertible mappings	High memory cost and architectural rigidity	Flexible density modeling but complex to scale to videos
Diffusion models	State-of-the-art image quality and controllability	High computational cost at inference	Robust to multi-condition inputs and rich priors
Hybrid models	Combine complementary properties (e.g., VAE-GAN, diffusion with latent encoders)	Design complexity and balancing multiple objectives	Seek balance between realism, speed, and controllability

Table 6. Temporal modeling strategies for identity transfer in videos.

Strategy	Temporal signal	Advantages	Limitations
3D convolutions	Spatio-temporal feature volumes	Capture local motion patterns jointly with appearance	Fixed temporal receptive field and high memory usage
Recurrent networks	Frame-level features with hidden state	Model long-range dependencies in a compact state	Prone to drift and vanishing gradients on very long videos
Temporal attention	Sequences of frame or patch embeddings	Flexible context selection over time	Quadratic cost in sequence length without sparsity
Latent trajectory models	Low-dimensional temporal codes	Explicit motion representation disentangled from appearance	Requires careful design to avoid entanglement with identity
Post-hoc smoothing	Independent frame synthesis plus regularization	Simple to implement on top of existing image models	Limited ability to fix severe temporal inconsistencies

or, in a non-autoregressive regime, a joint distribution

$$p_\theta x_{1:T} \mid z^{\mathrm{mot}}, z^{\mathrm{id}}.$$

The choice between autoregressive and non-autoregressive structure has implications for numerical stability and computational cost (3).

An important aspect of the formulation is the notion of identity invariance and equivariance with respect to scene transformations. Identity should be invariant under global rigid transformations such as rotations and translations, and under changes in illumination that do not alter discriminative features. Motion representations should be equivariant with respect to time reparameterizations and camera motion. These invariances are modeled implicitly by data augmentation and explicitly through the architecture. For instance, convolutional encoders can provide a degree of translation invariance, while spatial transformers or attention mechanisms can enable approximate equivariance to more complex transformations.

Temporal coherence is represented by the correlation between frames. A simple way to characterize temporal smoothness is to define a finite difference operator along the time dimension. Let

$$\Delta x_t = x_{t1} - x_t,$$

and consider the temporal variation energy

$$E_{\mathrm{temp}} = \sum_{t=1}^{T-1} \|\Delta x_t\|_2^2.$$

A generator that produces temporally coherent videos will tend to induce a distribution over sequences with controlled temporal variation energy. During training, regularizers are designed to encourage the generated videos to align with the temporal statistics of real sequences.

This formulation emphasizes that one-shot visual identity transfer involves jointly modeling high-dimensional random variables, latent structures, and spatiotemporal dependencies. The next sections describe how identity information is encoded, how dynamic scenes are generated conditionally on this identity, and how the model is trained under numerical constraints.

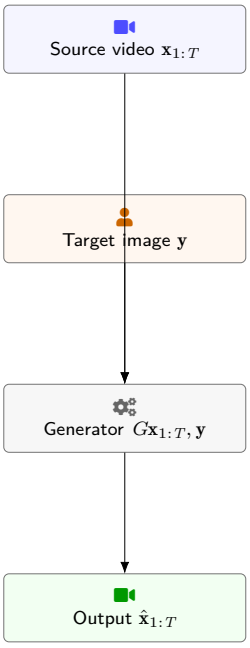

Figure 1. Problem formulation: given a source video and a single target identity image, a conditional generator produces a new video that preserves the source dynamics and layout while replacing the main actor's visible identity with that of the target subject.

3. Identity Embedding and Appearance Modeling

The core of one-shot identity transfer lies in how the target identity is encoded from a single image (4). The identity encoder E_{id} maps the image I^{tar} to a latent vector

$$z^{\mathrm{id}} \in \mathbb{R}^d$$

and to a collection of spatial feature maps that capture local appearance. In order to generalize across identities and viewpoints, E_{id} is built from convolutional and attention layers that aggregate information at multiple scales. The latent vector z^{id} is intended to capture global attributes such as facial geometry, rough body shape, and coarse texture properties, while spatial features encode finer details such as wrinkles, hair structure, and clothing patterns.

A simple conceptual model of identity representation is to treat z^{id} as a linear combination of basis vectors in a learned identity space. One can write

$$z^{\mathrm{id}} = Wu,$$

where $W \in \mathbb{R}^{d\times k}$ is a matrix of identity basis vectors and $u \in \mathbb{R}^k$ is a coefficient vector derived from the encoder. The matrix W is not fixed in advance but learned jointly with the rest of the network. In practice, the encoder produces z^{id} directly, yet the linear decomposition view is useful for interpreting the latent space. If the columns of W are approximately orthogonal, the coefficients can be seen as coordinates in a low-dimensional identity manifold.

To preserve identity fidelity in the generated video, the model includes an identity consistency loss. Let R be a fixed feature extractor, such as a network trained for recognition, that maps images to feature vectors in $\mathbb{R}^m$. For each generated frame $\hat{V}_t^{\mathrm{tar}}$, an identity feature

$$f_t^{\mathrm{gen}} = R\hat{V}_t^{\mathrm{tar}}$$

is computed, and compared with the feature

$$f^{\mathrm{tar}} = RI^{\mathrm{tar}}$$

of the target image. A basic identity loss is given by

$$\mathcal{L}_{\mathrm{id}} = \frac{1}{T}\sum_{t=1}^{T} \|f_t^{\mathrm{gen}} - f^{\mathrm{tar}}\|_2^2.$$

This loss encourages the generator to preserve identity-related features across all frames. In practice, cosine similarity or margin-based variants can also be used, but the squared Euclidean formulation illustrates the principle and has simple gradient properties.

Identity is not purely global; it is highly spatial. For complex scenes, different regions of the subject, such as the face and body, may have different visibility patterns, and partial occlusions may limit the information available in some frames (5). To handle this, the identity encoder can produce spatial feature maps

$$F^{\mathrm{id}} \in \mathbb{R}^{H_z\times W_z\times C_z},$$

which are later used in the generator via cross-attention mechanisms. At a given generator layer, query features derived from the source video attend to keys and values in F^{id}, allowing the network to import identity details into specific spatial regions. A simple linear attention operation can be written as

$$A = \mathrm{softmax}\, QK^{\top},$$

$$O = AV,$$

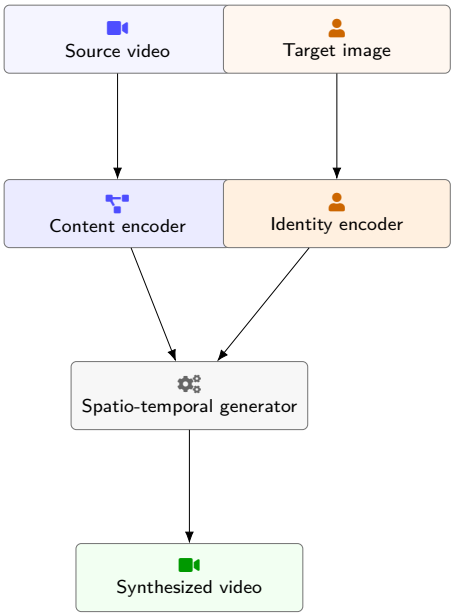

Figure 2. High-level architecture with a content pathway that captures spatio-temporal dynamics from the source video and an identity pathway that extracts appearance cues from a single image, both conditioning a shared generator that emits the transformed video.

Table 7. Evaluation criteria for one-shot visual identity transfer.

Metric	Target property	Typical implementation	Notes
Identity similarity	Alignment between output actor and target subject	Embedding distance using a face or identity recognition network	Should be high while avoiding overfitting to the single target view
Structural fidelity	Preservation of source video geometry and layout	Spatial similarity metrics between source and output frames	Penalizes distortions of pose, background, and camera geometry
Temporal stability	Consistency of appearance and geometry over time	Frame-to-frame consistency losses or optical-flow-based metrics	Sensitive to flickering, identity drift, and temporal artifacts
Perceptual quality	Overall realism of synthesized videos	Human preference studies or learned perceptual metrics	Complements objective scores with subjective visual judgments
Pose robustness	Performance under large viewpoint changes	Evaluation on subsets with strong pose variation	Highlights the impact of spatial alignment and occlusion handling

where Q, K, and V are linear projections of the generator features and identity features, arranged as matrices with compatible dimensions. The output O injects identity-conditioned information into the generator's hidden state.

To further structure identity information, the encoder may estimate keypoints or landmarks that approximate geometric structure. Let

$$K^{\mathrm{id}} \in \mathbb{R}^{J\times 2}$$

denote a set of two-dimensional keypoints in the target image, representing approximate joint positions or facial landmarks. Similarly, the source frames have corresponding keypoints

$$K_t^{\mathrm{src}} \in \mathbb{R}^{J\times 2}.$$

A simple linear model of geometric alignment between source and target is given by

$$K_t^{\mathrm{tar}} = M_t K^{\mathrm{id}} \ b_t,$$

where $M_t \in \mathbb{R}^{2\times 2}$ and $b_t \in \mathbb{R}^2$ are frame-dependent transformation parameters that map canonical target keypoints into the pose of the source frame. While the network does not explicitly compute such affine maps, its architecture can implicitly learn similar transformations through spatial warping layers or deformable convolutions (6).

Appearance modeling is not limited to geometry. Texture statistics can be represented via Gram matrices of feature maps (7). Suppose that E_{id} produces a feature map

$$F \in \mathbb{R}^{C_f\times N_f},$$

where C_f is the number of channels and N_f is the number of

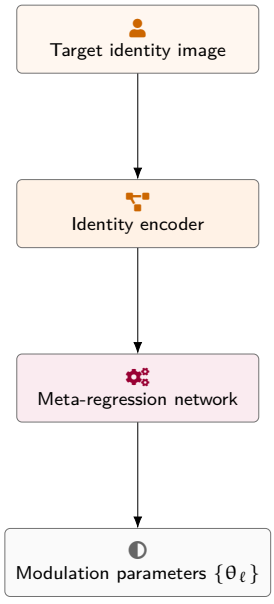

Figure 3. Identity pathway implemented as a feed-forward regression module: the target image is encoded and mapped to layer-wise modulation parameters that steer the generator in a one-shot manner without per-identity optimization at test time.

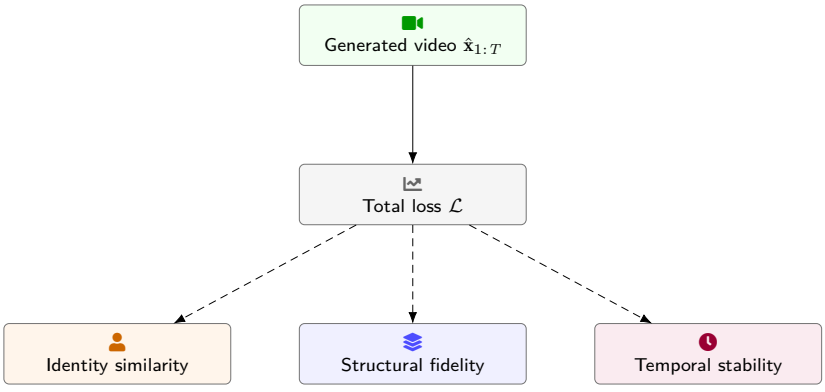

Figure 4. Training objective decomposed into complementary loss terms measuring identity similarity, structural fidelity, and temporal stability, whose weighted combination defines the supervision signal for end-to-end learning.

spatial locations after flattening. The Gram matrix

$$G = FF^\top$$

encodes correlations between channels and characterizes coarse texture. A texture consistency loss can be defined between Gram matrices of generated frames and the target identity, encouraging the preservation of high-level appearance statistics. The use of such second-order statistics introduces a form of linear algebraic structure into the representation of visual identity.

Since only a single target image is available, there is an inherent ambiguity in how the identity should appear under novel conditions. The model leverages priors learned from training data to resolve this ambiguity. The identity latent vector z^{id} is regularized toward a prior distribution, often a multivariate normal

$$pz^{\mathrm{id}} = \mathcal{N}0, I_d,$$

which encourages the space of identities to remain smooth and prevents overfitting to idiosyncrasies in the reference image. This regularization is implemented by a Kullback–Leibler divergence term between the approximate posterior produced by E_{id} and the prior, similar in spirit to variational autoencoding schemes.

The combined design of the identity encoder and appearance modeling losses aims to provide a latent representation that captures distinctive identity cues, is stable under modest perturbations of the image, and interacts effectively with the dynamic scene generator. The next section describes how this identity representation is incorporated into a model that synthesizes temporally coherent videos.

4. Dynamic Scene Generation Network

The dynamic scene generation network synthesizes the edited video by combining motion and background information from the source sequence with the identity embedding derived from the target image. Conceptually, the network consists of a mo-

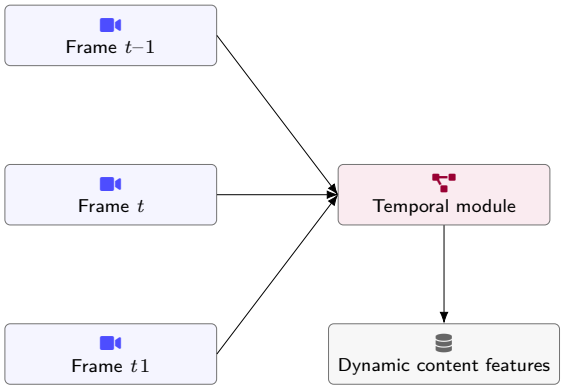

Figure 5. Temporal modeling aggregates neighboring frames through a dedicated module (e.g., 3D convolutions or temporal attention) to produce dynamic content features that encode motion, occlusions, and scene evolution for the generator.

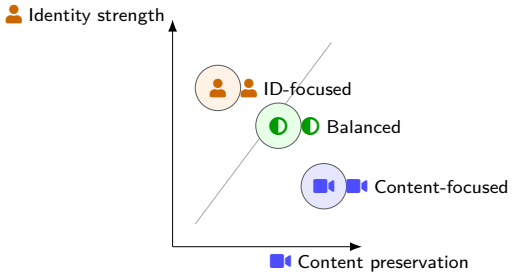

Figure 6. Conceptual trade-off between identity strength and content preservation, with indicative operating points that emphasize the target identity, balance both factors, or prioritize fidelity to the original scene content.

tion encoder, which compresses the source video into a latent representation, and a conditional generator, which reconstructs the video while injecting the target identity (8). Let the motion encoder produce

$$z^{\text{mot}} = E_{\text{mot}} V^{\text{src}},$$

where z^{mot} may itself be a spatiotemporal tensor rather than a single vector, to preserve local information about motion and scene layout.

A generic formulation of the generator can be written as a function

$$G_\theta : \mathcal{Z}_{\text{mot}} \times \mathcal{Z}_{\text{id}} \to \mathcal{V},$$

such that

$$\hat{V}^{\text{tar}} = G_\theta z^{\text{mot}}, z^{\text{id}}.$$

Internally, the generator operates at multiple spatial and temporal scales using convolutional, recurrent, and attention-based layers. One can view the generator as parameterizing a function

$$g_\theta : \Omega \times \mathcal{T} \to \mathbb{R}^C,$$

where $\Omega \subset \mathbb{R}^2$ is the image domain, $\mathcal{T} = \{1, \ldots, T\}$ indexes time, and the function maps spatial coordinates and time indices to RGB values. The dependence on z^{mot} and z^{id} is implicit in the parameters of g_θ or can be made explicit by writing

$$g_\theta x, t; z^{\text{mot}}, z^{\text{id}}.$$

One way to structure the generator is via a latent volume that encodes the dynamic scene. Consider a tensor

$$H \in \mathbb{R}^{T_z \times H_z \times W_z \times C_h},$$

produced by processing z^{mot} through a stack of spatiotemporal convolutions and attention layers. Identity information is injected at several layers through modulated normalization, additive biases, or cross-attention, using both the global vector z^{id} and the spatial feature maps F^{id}. For example, a convolutional feature map U in the generator can be modulated by identity via

$$\tilde{U} = \gamma z^{\text{id}} \odot U \ \beta z^{\text{id}},$$

where γ and β are learned affine transformations mapping $\mathbb{R}^d$ to the channel dimension of U, and $\odot$ denotes element-wise multiplication. This modulation allows identity to control the statistics of intermediate activations.

Temporal coherence is enforced by the architecture through operations that span time. Spatiotemporal convolutions with kernels of shape $k_t \times k_h \times k_w$ operate on multiple frames simultaneously and can model motion patterns. Temporal attention layers perform weighted aggregation across frames, with attention weights depending on both motion features and identity (9). For example, a temporal attention operation can be defined as

$$A^t = \text{softmax}\big(q_t K^\top\big),$$

$$h_t = A^t V,$$

where q_t is a query derived from the features of frame t, and K and V are stacks of keys and values across all frames. The resulting representation h_t provides a context-aware feature for frame t that respects temporal correlations.

Table 8. Observed limitations and failure modes of the formulation.

Failure mode	Underlying cause	Typical visual artifact
Extreme pose mismatch	Target evidence covers only a narrow range of viewpoints	Incorrect or unstable facial structure when the head rotates strongly
Background leakage	Imperfect disentanglement between content and appearance	Residual traces of source identity or unintended changes to the scene
Identity ambiguity	Low-resolution or noisy target image	Blended traits that do not clearly match either source or target subject
Occlusion sensitivity	Limited modeling of self-occlusions and external blockers	Ghosting or texture stretching when hands or objects cross the face
Lighting inconsistency	Mismatch between source video illumination and target evidence	Inconsistent shading or unrealistic shadows over time

Table 9. Application scenarios and extensions enabled by one-shot dynamic identity transfer.

Scenario	Objective	Practical considerations
Controllable video synthesis	Generate videos where motion is decoupled from identity	Requires reliable motion descriptors and flexible identity conditioning
Privacy-preserving editing	Obscure real identities while maintaining scene semantics	Must balance identity removal with preservation of behavioral cues
Data-efficient character re-targeting	Map performances to stylized avatars or rare characters	Benefits from strong priors that generalize from single exemplars
Multi-identity extension	Support switching between several target subjects	Demands scalable conditioning and careful temporal blending
Interactive control	Allow users to adjust identity strength or content preservation	Exposes trade-offs between realism, faithfulness, and user intent

For scenes where the subject undergoes large pose changes or self-occlusions, purely image-space convolutions may be insufficient to capture the geometric relationships across frames. An alternative is to introduce an intermediate canonical representation, such as a three-dimensional implicit field. In this view, the generator parameterizes a function

$$f_\theta : \mathbb{R}^3 \times \mathcal{T} \to \mathbb{R}^{C_f},$$

mapping 3D coordinates and time indices to feature vectors. The rendered image at time t is obtained by integrating along camera rays. For a pixel location x with camera ray $r_x s$, the color is

$$I_t x = \int_{s_0}^{s_1} ws\, f_\theta\left(r_x s, t\right) \mathrm{d}s,$$

where ws is a weighting function derived from density or occupancy predicted by the network. The identity embedding z^{id} and motion encoding z^{mot} modulate f_θ, allowing the model to represent identity-dependent appearance on a dynamic canonical surface. While such volumetric formulations increase computational cost, they can provide stronger 3D consistency, which is beneficial for large viewpoint changes (10).

Regardless of whether a 2D or 3D representation is used, the generator must preserve the background and non-subject elements of the scene. One strategy is to explicitly separate foreground and background channels. The motion encoder can produce a background feature tensor

$$H^{\mathrm{bg}} \in \mathbb{R}^{T_z \times H_z \times W_z \times C_h},$$

and a foreground mask

$$M \in \mathbb{R}^{T \times H \times W},$$

approximating the probability that each pixel belongs to the subject. The generator synthesizes a foreground video

$$\hat{V}^{\mathrm{fg}} \in \mathbb{R}^{T \times H \times W \times C}$$

conditioned on identity and motion, and composes it with the original background extracted from V^{src} via

$$\hat{V}^{\mathrm{tar}} = M \odot \hat{V}^{\mathrm{fg}} \; 1 - M \odot V^{\mathrm{src}}.$$

This composition preserves background dynamics while only altering foreground regions.

Incorporating stochasticity is essential to account for ambiguity in appearance and to improve sample diversity. The generator can be designed as part of a conditional diffusion process. A forward noising process

$$q x_t \mid x_{t-1}$$

is defined that gradually perturbs the video, and a neural network learns to reverse this process conditioned on z^{mot} and z^{id}. For a given noise level α, the perturbed video is

$$x_\alpha = \sqrt{\alpha}\, x \; \sqrt{1-\alpha}\, \epsilon,$$

with ϵ drawn from a standard normal distribution in the video space. The network learns a score or a denoised estimate, and the reverse process generates videos consistent with the conditioning variables. This formulation introduces a probabilistic interpretation and connects the model to multivariate stochastic differential equations.

The overall dynamic scene generation network thus combines identity-conditioned modulation, spatiotemporal processing, and stochastic sampling (11). Its behavior is determined not only by architectural choices but also by the training objectives and optimization scheme, which are addressed next.

5. Learning Objectives and Numerical Optimization

Training the one-shot identity transfer model involves defining a loss function that captures identity preservation, motion and background fidelity, temporal coherence, perceptual realism, and regularization. The total loss $\mathcal{L}$ is expressed as a weighted sum of several components. A basic reconstruction term encourages the generated video to match a ground-truth video under paired training. Assuming access to videos of many subjects, each with identity image and corresponding video, a reconstruction loss can be written as

$$\mathcal{L}_{\text{rec}} = \frac{1}{T} \sum_{t=1}^{T} \| \hat{V}_t - V_t \|_2^2,$$

where V_t is the ground-truth frame and $\hat{V}_t$ is the generated frame under self-identity conditioning. This term aligns the generator with the data distribution in a mean-squared sense.

Perceptual quality is promoted by using losses defined in a feature space. Let Φ be a feature extractor mapping images to a set of intermediate feature maps. A perceptual loss is defined as

$$\mathcal{L}_{\text{perc}} = \sum_{l} \frac{1}{T} \sum_{t=1}^{T} \| \Phi_l \hat{V}_t - \Phi_l V_t \|_2^2,$$

where Φ_l denotes the feature maps at layer l. While the layers used and the weighting scheme are model-specific, the principle is to match high-level representations rather than only pixel values.

Identity consistency is enforced via the loss (12) $\mathcal{L}_{\text{id}}$ described earlier. Temporal coherence is encouraged through explicit penalties on frame differences. A simple temporal loss is

$$\mathcal{L}_{\text{temp}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \| \hat{V}_{t1} - \hat{V}_t - V_{t1} - V_t \|_2^2,$$

which aligns the temporal gradients of generated and real videos. This loss can be interpreted as matching discrete derivatives along the time axis, approximating a continuous-time regularization

$$\left\| \frac{\partial \hat{V}t}{\partial t} - \frac{\partial Vt}{\partial t} \right\|_2^2 \mathrm{d}t.$$

Adversarial training is used to improve realism. A spatiotemporal discriminator D_ψ is trained to distinguish between real and generated video clips. In a conditional setting, the discriminator also observes the source video and the target image. An adversarial loss for the generator can be written as

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}\left[D_\psi\, \hat{V}^{\text{tar}}, V^{\text{src}}, I^{\text{tar}} \right],$$

while the discriminator is trained to maximize the difference between its outputs on real and generated samples (13). Adversarial training introduces a min–max optimization problem, which is numerically sensitive but effective for matching complex distributions.

The complete loss combines these components as

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \;\; \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \tag{5.1}$$

$$\lambda_{\text{id}}\mathcal{L}_{\text{id}} \;\; \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} \;\; \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{5.2}$$

with non-negative weights $\lambda_\cdot$ that balance the terms. The choice of these weights has a direct impact on the trade-off between fidelity, identity accuracy, and temporal smoothness.

In a diffusion-based variant, the objective is expressed in terms of denoising or score matching. For a fixed noise level α, the model is trained to predict either the clean video x from its noisy version x_α or the noise ϵ. A typical objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x,\epsilon,\alpha}\left[\| \epsilon - \epsilon_\theta\, x_\alpha, \alpha, z^{\text{mot}}, z^{\text{id}} \|_2^2 \right],$$

where ϵ_θ is the network's prediction. The reverse process corresponds to a discretized stochastic differential equation. The numerical stability of this procedure depends on the step size used in discretization and the Lipschitz properties of the score function (14).

From a numerical analysis perspective, training involves iteratively updating parameters θ and ψ using stochastic gradient descent or its variants. The update rule for θ can be written as

$$\theta_{k1} = \theta_k - \eta_k \nabla_\theta \mathcal{L}_k,$$

where η_k is the learning rate at iteration k and $\mathcal{L}_k$ is the empirical loss over a minibatch. Convergence behavior depends on properties such as the smoothness of $\mathcal{L}$ and the variance of gradient estimates. If the gradient is Lipschitz continuous with constant L, then choosing $\eta_k < 2L$ provides a basic condition to avoid divergence in simple settings. In practice, adaptive optimizers adjust the effective step size per parameter using estimates of first and second moments of gradients.

The training data consist of videos with many identities and motions. For each training sample, different conditioning configurations are constructed. In a self-reconstruction configuration, the identity image is derived from the same subject as the video, and the model is trained to reconstruct the video. In a cross-identity configuration, the model is conditioned on an identity image from a different subject and is trained with weaker supervision, such as adversarial and motion-consistency losses, to encourage plausibility under identity substitution. This strategy introduces elements of multi-task learning and regularizes the identity factorization.

Discrete mathematics concepts enter in modeling temporal structure as graphs. Frames are nodes in a directed graph with edges connecting temporally adjacent frames (15). Higher-order edges can connect frames that share similar motion or pose. Regularizers can be expressed as functions on this graph. For example, a graph Laplacian

$$L = D - W$$

built from a weight matrix W encoding frame similarity can define a temporal smoothness energy

$$E_{\text{graph}} = \sum_{t,t'} W_{tt'} \| \hat{V}_t - \hat{V}_{t'} \|_2^2,$$

which generalizes the simple finite difference penalty. This interpretation connects temporal regularization with spectral properties of graphs and suggests potential extensions using spectral filters.

Overall, the training objective and numerical optimization scheme shape the behavior of the model in subtle ways. Balancing the contributions of different loss terms is crucial to achieving identity preservation without sacrificing motion fidelity or temporal coherence. The next section discusses how the model performs empirically and examines qualitative and quantitative behavior.

6. Experimental Evaluation and Analysis

The empirical evaluation of one-shot visual identity transfer in dynamic scenes requires datasets containing diverse subjects performing varied motions in different environments. Each subject is associated with one or more reference images that can be used to approximate the one-shot setting during training and evaluation. The dataset is partitioned into training and test identities to assess generalization. In the test phase, the model receives a video of a source subject and a single image of a target subject and generates an edited video that should exhibit the target identity (16).

Quantitative evaluation of identity transfer performance relies on metrics that assess fidelity, identity similarity, and temporal stability. Image quality can be measured using distances between distributions of features extracted from real and generated frames. One common choice is to compute statistics of activations in a feature space and evaluate the distance between multivariate Gaussian approximations. Let μ_r and Σ_r denote the mean and covariance of features on real frames, and μ_g and Σ_g the corresponding quantities for generated frames. A distance of the form

$$d^2 = \|\mu_r - \mu_g\|_2^2 \ \mathrm{Tr}\big(\Sigma_r \ \Sigma_g - 2\Sigma_r\Sigma_g{}^{12}\big)$$

can be computed, providing a measure of how closely the generated distribution matches the real one. While this type of metric is defined for static images, pooling features over frames allows it to capture some aspects of video quality.

Identity similarity is evaluated using features from a recognition network. For each generated frame $\hat{V}_t^{\mathrm{tar}}$, a feature

$$f_t^{\mathrm{gen}} = R\hat{V}_t^{\mathrm{tar}}$$

is computed and compared with

$$f^{\mathrm{tar}} = RI^{\mathrm{tar}}.$$

The cosine similarity (17)

$$s_t = \frac{\langle f_t^{\mathrm{gen}}, f^{\mathrm{tar}} \rangle}{\|f_t^{\mathrm{gen}}\|_2 \|f^{\mathrm{tar}}\|_2}$$

provides a score in $-1, 1$, and its average over frames yields an overall identity preservation measure. Empirically, high similarity scores across frames indicate consistent identity transfer, while significant fluctuations suggest temporal drift or instability.

Temporal coherence is assessed through optical flow consistency and temporal perceptual metrics. For each consecutive pair of frames in a generated video, an optical flow field is estimated and used to warp the earlier frame toward the later one. The difference between the warped frame and the actual later frame provides a temporal error. Summed over time, this error approximates how well the motion in the video is locally consistent. Additionally, feature-based temporal distances can be computed by applying a perceptual loss between consecutive frames, which is less sensitive to minor intensity changes.

In an experimental study, models with different architectural configurations and training objectives can be compared. For instance, a baseline model that directly concatenates identity and motion features without explicit modulation or attention may exhibit lower identity similarity and higher temporal error. Introducing identity-conditioned normalization and cross-attention tends to improve identity scores by enabling more precise injection of appearance details. Adding temporal losses reduces flicker and improves the smoothness of motion but may slightly reduce sharpness if weighted too strongly, reflecting a trade-off between stability and detail.

The complexity of dynamic scenes has a noticeable effect on performance. Scenes with slow, smooth motion and limited occlusion typically yield higher identity similarity and lower temporal error. In contrast, scenes with rapid motion, strong motion blur, or large pose changes present challenges (18). In such cases, the model may approximate the target identity in frames where the face is clearly visible but regress toward an average identity or the source identity when visibility is limited. Background complexity also matters; cluttered or dynamic backgrounds can lead to artifacts if the foreground–background separation is imperfect, resulting in identity-related changes leaking into the background or vice versa.

An ablation analysis of loss components provides further insight. Removing the identity loss significantly degrades identity preservation, particularly in frames with moderate pose variations. The model still learns to reconstruct videos but does not align closely with target identity features. Removing temporal losses increases flicker and makes fine details unstable over time, even if per-frame quality remains reasonable. Omitting adversarial losses leads to overly smooth outputs with reduced perceptual realism, indicating that pixelwise and perceptual losses alone are insufficient to capture fine-grained texture statistics.

One-shot generalization is examined by varying the pose and viewpoint of the target identity image relative to the source video. When the target image's viewpoint is similar to dominant poses in the source video, identity transfer tends to be more accurate. As the discrepancy increases, especially for extreme pose differences, the generator must rely more heavily on learned priors to hallucinate unseen views. Performance degrades gradually as the pose gap increases, with more pronounced artifacts at oblique angles that are not well covered by training examples. Using multiple target images at test time can alleviate this, but the focus here is on strictly one-shot conditions.

The robustness of the method to changes in lighting is eval-

uated by pairing source videos and target images captured under noticeably different illumination. The identity encoder and generator are trained with augmentations that simulate such variations, promoting invariance to lighting for identity features (19). Nevertheless, extreme illumination differences can cause the generated video to exhibit inconsistent shading, with the identity partially influenced by source lighting and partially by target image shading. This suggests that disentangling identity from illumination remains only partially achieved.

From a numerical standpoint, stability of training is influenced by the relative magnitudes of gradient contributions from different loss terms. In experiments, large adversarial weights lead to unstable oscillations in generator and discriminator losses, while too small weights reduce perceptual fidelity. A moderate choice yields a balance where losses converge to a regime of bounded fluctuations. Learning rate schedules that gradually decay the step size help ensure that training progresses from coarse alignment to fine refinement. Gradient clipping is used to avoid rare but severe updates that can move the model into degenerate regions of parameter space.

Discrete structural constraints, such as the graph-based temporal regularization, show mixed results. When frame similarity is measured in feature space and used to define a graph Laplacian penalty, temporal coherence improves for sequences with repetitive motion patterns, but the added regularization can oversmooth motion in sequences with rapid, non-repeating movements. This indicates that temporal graph structure should reflect motion statistics, possibly requiring adaptive construction conditioned on the specific sequence.

Overall, the experimental analysis suggests that the proposed one-shot identity transfer framework can synthesize videos that preserve source motion and approximate target identity under a variety of conditions, with performance depending on pose, illumination, and motion complexity. The interplay between identity encoding, dynamic generation, and numerical optimization is critical, and the observed failure cases point to directions where modeling and optimization could be refined.

7. Conclusion

This work has examined one-shot visual identity transfer in dynamic scenes using target images, focusing on a formulation that separates identity, motion, and background within a conditional generative framework. The problem was posed in terms of learning a mapping from a source video and a single target identity image to an edited video where the source identity is replaced while motion and scene context are preserved (20). Videos and images were treated as high-dimensional random variables and tensors, and latent variables were used conceptually to describe identity, motion, and background factors.

An identity encoder was described that produces both global latent vectors and spatial feature maps from the target image, representing identity as a point in a learned manifold with linear algebraic structure. Appearance modeling involved identity-consistency losses, texture statistics based on Gram matrices, and regularization toward a prior distribution, aiming to capture distinctive but robust identity features. These identity representations were integrated into a dynamic scene generation network through modulation and attention mechanisms that inject appearance information at multiple scales and layers.

The dynamic generator, implemented in a spatiotemporal architecture, combined motion encodings derived from the source video with identity embeddings. Variants ranging from purely two-dimensional image-space processing to three-dimensional implicit field representations were discussed, emphasizing the role of geometric consistency in handling large pose changes. Foreground–background separation and compositing strategies were used to preserve scene context while editing the subject identity, and stochastic elements, such as conditional diffusion processes, were incorporated to handle appearance ambiguity.

Training objectives were formulated as a combination of reconstruction, perceptual, identity, temporal, and adversarial losses, along with diffusion-based terms where applicable. Numerical optimization relied on stochastic gradient methods and considered stability issues associated with non-convex objectives, adversarial dynamics, and high-dimensional parameter spaces. Connections to multivariate calculus, stochastic processes, and discrete graph-based regularization provided a mathematical perspective on the behavior of the learning procedure.

Experimental analysis on dynamic scenes with diverse identities and motions suggested that the proposed framework can achieve identity transfer with reasonable preservation of motion fidelity and temporal coherence under a range of conditions, while also revealing limitations under extreme pose mismatch, challenging illumination, and complex occlusions. Ablation studies highlighted the importance of explicit identity losses and temporal regularizers, as well as the influence of adversarial training on perceptual quality.

The formulation and analysis presented here underscore that one-shot identity transfer in dynamic scenes is governed by the interaction of representation choices, architectural design, and numerical training procedures. Further developments might involve more explicit three-dimensional modeling of identity and motion, improved disentanglement of illumination and appearance, and adaptive temporal regularization leveraging richer graph structures. These directions offer opportunities to refine identity preservation and robustness in future systems for personalized and controllable video synthesis (21).

References

[1] J. Diaz-Escobar, N. E. Ordonez-Guillen, S. Villarreal-Reyes, A. Galaviz-Mosqueda, V. Kober, R. Rivera-Rodriguez, and J. E. L. Rizk, "Deep-learning based detection of covid-19 using lung ultrasound imagery.," *PloS one*, vol. 16, pp. e0255886–, 8 2021.

[2] P. P. Ray, "An overview of webassembly for iot: Background, tools, state-of-the-art, challenges, and future directions," *Future Internet*, vol. 15, pp. 275–275, 8 2023.

[3] A. D. Spence, "A cad based concept car workflow," *Proceedings of the Canadian Engineering Education Association (CEEA)*, 7 2010.

[4] A. Muhammad, Z. Salman, K. Lee, and D. Han, "Harnessing the power of diffusion models for plant disease image augmentation.," *Frontiers in plant science*, vol. 14, pp. 1280496–, 11 2023.

[5] M. Mejia-Herrera, J. S. Botero-Valencia, D. Betancur-Vásquez, and E. A. Moncada-Acevedo, "Low-cost system for analysis pedestrian flow from an aerial view using near-infrared, microwave, and temperature sensors.," *HardwareX*, vol. 13, pp. e00403–e00403, 2 2023.

[6] S. S. Harsha, A. Revanur, D. Agarwal, and S. Agrawal, "Genvideo: One-shot target-image and shape aware video editing using t2i diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

[7] H. Sharma and N. Kanwal, "Video surveillance in smart cities: current status, challenges & future directions," *Multimedia Tools and Applications*, vol. 84, pp. 15787–15832, 6 2024.

[8] R. Rosas-Romero, O. Lopez-Rincon, and O. Starostenko, "Fully automatic alpha matte extraction using artificial neural networks," *Neural Computing and Applications*, vol. 32, pp. 6843–6855, 3 2019.

[9] M. Butwall, P. Ranka, and S. Shah, "Python in field of data science: A review," *International Journal of Computer Applications*, vol. 178, pp. 20–24, 9 2019.

[10] L. T. H. Phuc, H. J. Jeon, N. T. N. Truong, and J. J. Hak, "Applying the haar-cascade algorithm for detecting safety equipment in safety management systems for multiple working environments," *Electronics*, vol. 8, pp. 1079–, 9 2019.

[11] L. Zhou, R. Wang, and L. Zhang, "Accurate robot arm attitude estimation based on multi-view images and super-resolution keypoint detection networks.," *Sensors (Basel, Switzerland)*, vol. 24, pp. 305–305, 1 2024.

[12] M. Bendale, M. V. Phatak, and N. Pise, "Analysis and importance of deep learning for video aesthetic assessments," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, pp. 546–554, 2 2019.

[13] S. R, Menon, and null Priyanka.P, "An impact of technology in the field of fashion and photography," *IJARCCE*, vol. 10, 11 2021.

[14] D. Pankratz, "Llti highlights," *IALLT Journal of Language Learning Technologies*, vol. 29, pp. 49–56, 4 1996.

[15] S. Teso and O. Hinz, "Challenges in interactive machine learning," *KI - Künstliche Intelligenz*, vol. 34, pp. 127–130, 6 2020.

[16] S. Joglekar, H. Sawant, A. Jain, P. Dhadda, and P. Sonawane, "A multimodular approach for gesture recognition and text formulation in american sign language," *International Journal of Computer Applications Technology and Research*, vol. 9, pp. 217–224, 7 2020.

[17] T. Pamuła and W. Pamula, "Detection of safe passage for trains at rail level crossings using deep learning.," *Sensors (Basel, Switzerland)*, vol. 21, pp. 6281–, 9 2021.

[18] L. Que, T. Zhang, H. Guo, C. Jia, Y. Gong, L. Chang, and J. Zhou, "A lightweight pedestrian detection engine with two-stage low-complexity detection network and adaptive region focusing technique.," *Sensors (Basel, Switzerland)*, vol. 21, pp. 5851–, 8 2021.

[19] F. A. Siqueira, D. Vitório, E. Souza, J. A. P. Santos, H. O. Albuquerque, M. S. Dias, N. F. F. Silva, A. C. P. L. F. de Carvalho, A. L. I. Oliveira, and C. Bastos-Filho, "Ulysses tesemõ: a new large corpus for brazilian legal and governmental domain," *Language Resources and Evaluation*, vol. 59, pp. 1685–1704, 7 2024.

[20] M. Neuhausen, D. Pawlowski, and M. König, "Comparing classical and modern machine learning techniques for monitoring pedestrian workers in top-view construction site video sequences," *Applied Sciences*, vol. 10, pp. 8466–, 11 2020.

[21] A. Barnawi, P. Chhikara, R. Tekchandani, N. Kumar, and B. A. Alzahrani, "Artificial intelligence-enabled internet of things-based system for covid-19 screening using aerial thermal imaging," *Future generations computer systems : FGCS*, vol. 124, pp. 119–132, 5 2021.