

Restaurarea Ghidată a Imaginilor și Secvențelor Video Faciale prin Injectarea Informației de Referință

Alexandru-Gabriel Lefterache¹ Anamaria Rădoi¹

¹Facultatea de Electronică, Telecomunicații și Tehnologia Informației,
Universitatea Națională de Știință și Tehnologie POLITEHNICA București
alefterache@upb.ro anamaria.radoi@upb.ro

Rezumat

Restaurarea imaginilor și secvențelor video faciale degradate rămâne o problemă deschisă, în special în scenarii reale, unde modelele generative produc reconstrucții plauzibile, dar pot altera identitatea subiecților. Lucrarea propune un sistem modular construit incremental în trei etape, pornind de la un model de bază de restaurare facială: (i) o extensie cu regularizare identitară și discriminator local pe regiuni faciale perceptual importante, pentru îmbunătățirea consistenței identitare; (ii) condiționare cu imagini de referință ale aceleiași identități, prin modulație spațială multi-rezoluție și agregare adaptivă; (iii) extinderea către restaurare video prin mecanisme de coerență temporală și un set de date propriu pentru antrenare și evaluare. Rezultatele arată că metodele propuse ating sau depășesc performanța soluțiilor existente, menținând un număr redus de parametri și latență scăzută.

Cuvinte cheie: restaurare facială, condiționare prin referință, VQGAN, restaurare video, coerență temporală.

1. Introducere

Restaurarea imaginilor și a secvențelor video este o problemă fundamentală în domeniul procesării imaginilor, având ca obiectiv reconstrucția unor imagini de înaltă calitate din observații degradate de estompare, rezoluție redusă, zgomot sau artefacte de compresie. Deși modelele generative recente obțin rezultate vizuale remarcabile, ele rămân limitate în scenarii reale, unde degradările sunt complexe și dificil de modelat.

O provocare majoră o reprezintă restaurarea facială, unde este necesară nu doar calitatea perceptuală, ci și preservarea identității. Modelele generative pot produce imagini plauzibile, dar tind să modifice subtil trăsăturile faciale, afectând consistența identitară. Acest aspect este critic în aplicații precum restaurarea imaginilor și secvențelor video de arhivă, teleconferințe sau editare semantică, unde chiar și variații minore ale trăsăturilor distinctive (ochi, nas, buze, pistrui, cicatrice) sunt ușor perceptibile de către utilizatorul uman.

În acest context, lucrarea propune un sistem de restaurare facială orientat spre preservarea identității, vizând atât îmbunătățirea calității vizuale, cât și fidelitatea față de subiect. Abordarea este structurată pe trei niveluri progresive, începând cu restaurarea generică bazată pe un prior generativ pre-antrenat, continuând cu ghidarea prin referință statică pentru condiționarea procesului

de restaurare și extinzându-se ulterior la video, unde sunt introduse constrângeri de coerență temporală între cadre.

Contribuțiile principale ale lucrării includ:

- extinderea antrenării printr-un cost bazat pe similaritatea reprezentărilor vectoriale identitare și un termen adversarial aplicat regiunilor faciale critice (ochi, gură), pentru îmbunătățirea consistenței structurale și preservarea identității;
- integrarea unui mecanism de condiționare prin referință pentru ghidarea restaurării pe informației de identitate;
- extinderea metodei la restaurare video prin constrângeri explicite de consistență temporală;
- alcătuirea unui set de date propriu cu 668 secvențe video și 1500 imagini de referință aparținând a 388 de identități, utilizat pentru antrenare și evaluare.

2. Metodologii existente în literatură

Metodele de restaurare facială pot fi organizate într-o schemă ierarhică, în funcție de tipul datelor și de modul în care este integrată informația de identitate în procesul de restaurare. În primul rând, acestea se împart în metode dedicate **imaginilor statice** și metode dedicate **secvențelor video**. În cadrul fiecărei categorii, abordările diferă în funcție de tipul și gradul de utilizare a informației de identitate, incluzând metode generice, metode asistate prin referință și metode personalizate.

2.1. Metode de restaurare a imaginilor faciale

În cazul imaginilor statice, metodele recente de restaurare se bazează predominant pe priori generativi puternici pentru reconstrucția detaliilor faciale în condiții de degradare severă. Principala distincție între abordări constă în utilizarea informației de identitate: metodele generice operează exclusiv pe imaginea degradată, în timp ce metodele asistate prin referință utilizează una sau mai multe imagini ale aceleiași identități pentru a ghida reconstrucția și a îmbunătăți consistența identitară.

2.1.1. Metode generice de restaurare

Metodele de restaurare generică (engl., *blind restoration*) reprezintă cea mai studiată categorie din literatură, iar evoluția lor poate fi urmărită prin prisma tipului de prior generativ utilizat: de la modele pur discriminative, la modele bazate pe rețele generative adversariale (GAN), dicționare de reprezentări faciale discrete și modele de difuzie.

O direcție majoră o constituie metodele bazate pe priori generativi adversariali, în care un generator pre-antrenat modelează distribuția fețelor de înaltă calitate și este utilizat ca spațiu de regularizare pentru reconstrucție. Metode precum *GFP-GAN* [1] și *GPEN* [2] se bazează pe *StyleGAN2* [3], integrat într-o arhitectură codor-decodor. Ambele metode proiectează imaginea degradată în spațiul latent al generatorului, însă diferă în mecanismul de condiționare: *GFP-GAN* injectează reprezentări extrase din intrare prin modulație spațială a activărilor, în timp ce *GPEN* utilizează codul latent z pentru a condiționa generatorul *StyleGAN2* prin modulația reprezentării stilului la nivelul fiecărui bloc, controlând astfel structura globală a feței.

În aceeași linie, *CodeFormer* [4] adoptă o formulare de tip *VQGAN* [5], utilizând un dicționar discret de reprezentări faciale. Restaurarea este formulată ca o problemă de predicție secvențială a codurilor latente corespunzătoare imaginii degradate, modelată printr-o arhitectură de tip *Transformer*. Un avantaj important îl constituie controlul explicit al compromisului dintre fidelitatea structurală și realism, realizat printr-un factor de ponderare care modulează contribuția informației de intrare în decodor.

O altă direcție majoră este reprezentată de modelele de difuzie, care reconstruiesc imaginea printr-un proces iterativ de eliminare a zgomotului, modelând inversul unui proces de degradare progresivă. Metode precum *DiffBIR* [6] sau *OSDFace* [7] demonstrează capacitatea acestor modele de a genera detalii realiste și consistente cu distribuția naturală a fețelor, însă cu un cost computațional ridicat datorat inferenței iterative.

2.1.2. Metode asistate prin referință

Metodele asistate prin referință introduc imagini suplimentare ale aceleiași identități la inferență, cu scopul de a îmbunătăți consistența identitară în condiții de degradare severă. Aceste metode nu necesită adaptarea parametrilor modelului și sunt, în general, agnostice față de subiect.

O direcție dominantă este reprezentată de modelele de difuzie condiționate de referință. *ReF-LDM* [8] extinde modelele de difuzie în spațiu latent printr-un mecanism de stocare internă a reprezentările cheie-valoare ale imaginilor de referință, pe care le reutilizează în toate etapele de eliminare a zgomotului. În paralel, o funcție de cost de identitate ponderată în funcție de pasul de difuzie stabilizează reconstrucția, crescând gradual influența referinței pe măsură ce procesul tinde să devină determinist.

InstantRestore [9] reduce costul computațional printr-o formulare într-un singur pas, bazată pe transfer de atenție dintr-un codor de referință înghețat. Informația de identitate este injectată direct în straturile de atenție ale modelului principal, evitând alinierea explicită și permițând un transfer semantic robust între referință și imaginea degradată.

2.2. Metode video de restaurare facială

Restaurarea video adaugă, pe lângă consistența spațială, constrângerea coerenței temporale. În lipsa acesteia, pot apărea variații inconsistente între cadre consecutive, cunoscute sub denumirea de artefacte de tip „flickering”.

Metodele clasice, precum *BasicVSR++* [10] și *EDVR* [11], abordează această problemă prin aliniere și agregare temporală a caracteristicilor, utilizând flux optic și/sau convoluții deformabile. Deși eficiente în restaurarea video generică, acestea nu includ un prior facial explicit, ceea ce limitează capacitatea de recuperare a detaliilor identitare fine.

O direcție mai recentă este reprezentată de modelele video bazate pe difuzie, care integrează explicit dependențele spațio-temporale în procesul generativ. *SVFR* [12], *DVFace* [13] și *StableB-FVR* [14] modelează simultan consistența temporală și realismul vizual prin mecanisme de atenție și propagare a informației între cadre, obținând reconstrucții mai stabile și mai naturale.

În paralel, metodele video asistate prin referință introduc explicit informația de identitate în procesul de restaurare. *IP-FVR* [15] adaptează paradigma IP-Adapter [16] în context video, utilizând mecanisme de atenție încrucișată pentru a propaga caracteristicile unei imagini de referință la nivelul întregii secvențe.

3. Seturi de date și degradări sintetice

Seturile de date utilizate sunt organizate după două criterii principale: tipul datelor (imagini statice, respectiv secvențe video) și structura de identitate (absența sau prezența grupării pe identitate).

3.1. Seturi de imagini statice

Seturi generice (fără referință de identitate). *FFHQ* [3] (Flickr-Faces-HQ) conține 70 000 de imagini faciale la rezoluția 1024×1024 , colectate de pe platforma Flickr și pre-procesate prin aliniere și decupare automată. Setul include exclusiv imagini de înaltă calitate și este utilizat pentru antrenarea modelului de restaurare generică.

The Second Challenge on Real-World Face Restoration at NTIRE 2026 [17] propune un set de 450 de imagini degradate real, provenite din surse *in-the-wild* (CelebA-Test, WebPhoto-Test, Wider-Test), fără imagini de referință de calitate înaltă. Este utilizat exclusiv pentru evaluarea restaurării în condiții reale, utilizând metrici agnostice de referință, precum NIQE (engl., *Naturalness Image Quality Evaluator*), FID (engl., *Fréchet Inception Distance*) sau CLIP-IQA (engl., *CLIP-Image Quality Assessment*).

Seturi cu referință de identitate. *FFHQ-Ref* [8] extinde FFHQ prin asocierea imaginilor aparținând aceleiași identități folosind un model de similaritate facială; perechile cu similaritate cosinus peste pragul $r = 0.4$ sunt considerate valide. Setul rezultat constă în 20 405 imagini corespunzătoare a 6 523 de identități și utilizat pentru antrenarea și validarea modelului de restaurare condiționată prin referință.

CelebA-Test-Ref [8] derivă din subsetul de test CelebA și conține 2 533 perechi LQ–HQ cu referințe de identitate corespunzătoare. Este utilizat ca set standard pentru evaluarea restaurării condiționate prin referință.

3.2. Setul de date video:

Datorită absenței unor seturi publice adecvate paradigmei video cu referință, este construit un **set de date propriu**, realizat prin asocierea la nivel de identitate (similaritate cosinus ArcFace, prag $r = 0.4$) a imaginilor HQ din CelebRef-HQ [18] și CelebAMask-HQ [19], cu secvențe video din VFHQ [20]. Corespondențele au fost **verificate manual** pentru eliminarea cazurilor fals pozitive.

Cadrele video sunt extrase și aliniate conform protocolului FFHQ [3], pe baza celor cinci repere faciale furnizate de metadatele VFHQ. Anterior alinierii, coordonatele reperelor sunt netezite temporal printr-un filtru Gaussian 1D.

Setul final conține **668 clipuri video** și **1 501 imagini de referință**, corespunzătoare **388 de identități**, partiționat în antrenare (533 clipuri, 310 identități), validare (66, 39) și testare (69, 39), cu stratificare pe gen și număr de secvențe/video per identitate.

3.3. Generarea perechilor LQ–HQ prin degradări sintetice

Pentru antrenare, imaginile de calitate înaltă sunt degradate dinamic pentru a genera perechi LQ–HQ, strategie standard ce permite acoperirea unei game largi de degradări realiste.

Imagini statice. Schema de degradare urmează Real-ESRGAN [21] și constă în aplicarea succesivă a operațiilor de estompare, redimensionare, zgomot și compresie JPEG:

$$\mathcal{D}_i(I) = \text{JPEG}_{q_i}((I \otimes k_i) \downarrow_{r_i} + n_i), \quad (1)$$

unde k_i reprezintă nuclul de estompare (Gaussian izotrop/anizotrop, Gaussian generalizat sau de tip platou), $\downarrow_{r_i}$ redimensionarea, n_i zgomot Gaussian sau Poisson, iar q_i factorul de calitate JPEG.

Se aplică N etape succesive, urmate opțional de un filtru sinc 2D pentru simularea artefactelor de tip *ringing*:

$$I_{LQ} = ((\mathcal{D}_N \circ \dots \circ \mathcal{D}_1)(I_{HQ})) \uparrow_{r_o} \otimes k_{\text{sinc}}. \quad (2)$$

Secvențe video. Pentru video, compresia JPEG este înlocuită cu standardul de compresie H.264, pentru a modela artefacte temporale reale (macroblocuri persistente și fluctuații inter-cadru). Compresia este aplicată la nivel de secvență, cu un factor CRF eșantionat o singură dată per clip:

$$I_{LQ}^{(t)} = \text{Codec}_{\text{CRF}}(\mathcal{D}_1(I_{HQ}^{(t)})) \uparrow, \quad \text{CRF} \sim \mathcal{U}[25, 45]. \quad (3)$$

Seturile video de validare și test sunt generate pe trei niveluri de dificultate (redus/moderat/sever),

definite prin intensitatea estompării, zgomot și compresie video, cu parametrii detaliați în Secțiunea 5.

4. Soluția propusă

4.1. Model generic de reconstrucție

Punctul de plecare este modelul CodeFormer [4], utilizat în varianta pre-antrenată pe setul FFHQ [3]. Arhitectura este construită în jurul unui spațiu latent discret și include un **codor pentru imagini de înaltă calitate** E_{HQ} , utilizat pentru învățarea unui **dictionar de coduri latente** $\mathcal{C}$, un **codor pentru imagini degradate** E_{LQ} , un **Transformer** T_w ce prezice secvența de coduri discrete și un **decoder** D ce reconstruiește imaginea restaurată. Antrenarea sistemului a fost realizată în trei etape: (i) definirea unei arhitecturi VQGAN pentru construirea dictionarului discret de reprezentări faciale și a decodurului, (ii) antrenarea codorului LQ și a *Transformerului* pentru predicția codurilor din imagini degradate, după înghețarea componentelor anterioare, și (iii) introducerea blocurilor de modulație spațială CFT (*Controllable Feature Transform*), ce condiționează activările decodurului pe baza trăsăturilor extrase din imaginea degradată, permițând reinjectarea informației structurale și controlul compromisului dintre fidelitate și realism. Configurația utilizată în această lucrare păstrează arhitectura originală a modelului, incluzând un dictionar discret de 1024 coduri, un modul *Transformer* cu 9 straturi de tip *self-attention*, precum și conexiuni CFT aplicate la mai multe niveluri de rezoluție ale decodurului.

Pornind de la această bază, sunt propuse **trei extensii metodologice** destinate îmbunătățirii preservării identității: (i) un cost identitar bazat pe reprezentări identitare ArcFace, (ii) un cost adversarial local aplicat regiunilor faciale cu importanță perceptuală ridicată și (iii) o schemă de degradare în doi pași pentru simularea mai realistă a degradărilor întâlnite în practică. Aceste modificări nu alterează arhitectura de bază a modelului, ci vizează exclusiv procesul de antrenare și formularea funcției obiectiv.

4.2. Model de restaurare asistat prin referință

Deși modelul de bază poate genera reconstrucții plauzibile, acesta nu dispune de un mecanism explicit de condiționare a identității. În cazul degradărilor severe, acest lucru conduce la reconstrucții corecte perceptual, dar inconsistente identitar, deoarece selecția codurilor latente este ghidată exclusiv de distribuția datelor de antrenare, fără un semnal identitar explicit.

Pentru a rezolva această problemă, se introduce condiționare explicită pe baza uneia sau mai multor imagini de referință. Obiectivul este dublu: (1) injectarea unui semnal identitar expresiv, (2) menținerea stabilității structurale a reconstrucției. Pentru aceasta sunt explorate două strategii:

Varianta A: condiționare în spațiul intern al reprezentărilor modelului

Această variantă valorifică direct reprezentările extrase de codorul de înaltă calitate, aliniate cu spațiul latent al modelului de restaurare. Pentru fiecare nivel de rezoluție l , se consideră: trăsăturile F_{LQ}^l extrase de codorul imaginii degradate, ce conțin informația structurală disponibilă în imaginea de intrare; trăsăturile F_{Dec}^l corespunzătoare stării curente a decodurului; respectiv trăsăturile de referință F_{Ref}^l , extrase din imaginile de referință și utilizate pentru furnizarea informației de identitate.

Informația de referință este fuzionată cu cea structurală printr-un bloc rezidual, rezultând o reprezentare condiționată utilizată pentru modulație spațială. Blocul extins introduce un semnal suplimentar de condiționare identitară peste mecanismul standard de modulație:

$$\hat{F}_{Dec}^l \leftarrow F_{Dec}^l + w \cdot (F_{Dec}^l \odot \gamma^l + \beta^l), \quad (4)$$

unde γ^l și β^l sunt generate din reprezentarea fuzionată LQ-referință.

În cazul mai multor referințe, acestea sunt agregate adaptiv în funcție de contextul curent al restaurării (caracteristicile LQ și starea decodului), rezultând o reprezentare de identitate ca sumă ponderată a trăsăturilor referințelor, cu selectarea automată a celor mai informative imagini. Mecanismul este ilustrat în Figura 1.

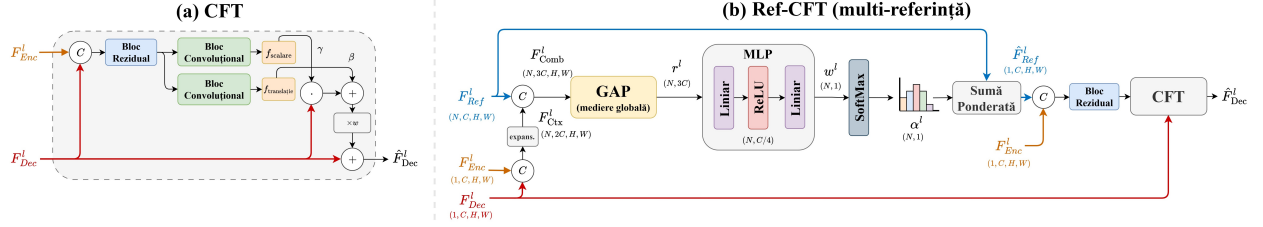


Figura 1: Comparație între mecanismul CFT original din CodeFormer și extensia propusă pentru integrarea mai multor imagini de referință. (a) Blocul CFT standard, condiționat exclusiv pe trăsăturile extrase din imaginea degradată. (b) Blocul Ref-CFT extins cu agregare adaptivă a referințelor, care estimează ponderi de relevanță și construiește o reprezentare identitară agregată utilizată pentru modulația decodului.

În varianta extinsă este introdusă o injecție opțională a informației de identitate la nivelul *transformerului*, complementar modulației din decodor. Aceasta constă într-o ramură paralelă, inițializată identic cu cea principală, utilizată exclusiv pentru extragerea reprezentărilor cheie-valoare din referințe. În fiecare strat, aceste reprezentări sunt integrate în *transformerul* principal prin atenție încrucișată, unde starea latentă a imaginii degradate acționează ca interogare. Astfel, informația de identitate este propagată strat cu strat în spațiul latent al restaurării.

Varianta B: condiționare într-un spațiu de identitate pre-antrenat

Această variantă utilizează exclusiv reprezentări extrase din modele pre-antrenate pentru recunoaștere facială, separând explicit modelarea identității de procesul de restaurare. Se folosesc două tipuri complementare de semnale: vectori de identitate discriminativi (ArcFace [22]), ce oferă robustețe la variații de iluminare și expresie, și descriptori structurali și semantici (FaRL [23]), ce codifică informație facială la nivel global și local. Contextul de identitate este construit prin agregarea acestor reprezentări într-un spațiu comun de condiționare, inspirat din metoda [24]. Spre deosebire de aceasta, unde condiționarea este injectată prin mecanisme de atenție încrucișată distribuite la nivelul arhitecturii U-Net de difuzie, abordarea propusă realizează condiționarea exclusiv în decodor, utilizând același mecanism de modulație spațială introdus anterior.

Pentru mai multe imagini de referință, reprezentările ArcFace sunt agregate printr-un mecanism de atenție condiționat pe caracteristicile imaginii degradate, în timp ce descriptorii FaRL sunt agregați prin mediere. Cele două surse sunt apoi proiectate într-un spațiu comun și concatenate pentru a forma contextul de identitate:

$$\text{ctx} = [\phi(\tilde{v}_{id}) \parallel \Psi(\tilde{T}_{FaRL})], \quad (5)$$

unde $\tilde{v}_{id}$ reprezintă vectorul de identitate agregat, $\tilde{T}_{FaRL}$ descriptorii faciali agregați, iar $\phi(\cdot)$ și $\Psi(\cdot)$ sunt transformări liniare antrenabile ce proiectează cele două reprezentări în spațiul caracteristicilor decodului.

Contextul astfel construit este utilizat pentru condiționarea caracteristicilor intermediare ale decodului printr-un mecanism de atenție încrucișată:

$$F^l \leftarrow F^l + \tanh(g) \cdot \text{CA}(F^l, \text{ctx}), \quad (6)$$

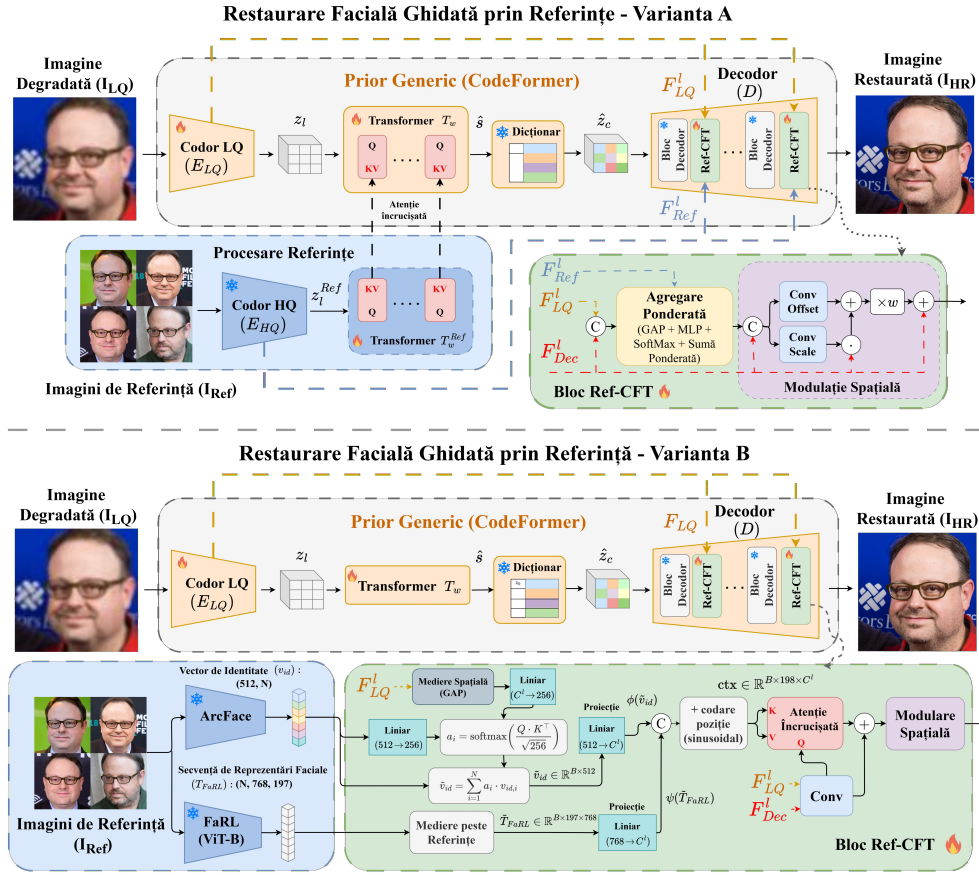


Figura 2: Strategiile investigate pentru integrarea informației de identitate prin imagini de referință. (a) Condiționare bazată pe reprezentările interne ale modelului și agregare adaptivă a referințelor. (b) Condiționare bazată pe reprezentări faciale pre-antrenate și injectarea contextului identitar în decodor.

unde F^l reprezintă caracteristicile decodurului la nivelul de rezoluție l , $CA(\cdot)$ denotă operația de atenție încrucișată, iar g este o poartă scalară învățabilă care controlează contribuția semnalului identitar. Caracteristicile rezultate sunt ulterior rafinate prin mecanismul de modulație spațială descris în Varianta A.

Comparație. Cele două variante diferă atât prin sursa semnalului de identitate, cât și prin modul de integrare în arhitectura de restaurare. Prima variantă operează în spațiul intern al modelului, utilizând reprezentări învățate intrinsec și oferind o integrare compactă și complet aliniată cu procesul de restaurare. În schimb, a doua variantă se bazează pe reprezentări externe pre-antrenate, ceea ce permite un semnal de identitate mai expresiv și mai bine structurat semantic, însă introduce dependență de modele auxiliare și de spații de reprezentare predefinite. O ilustrare schematică a celor două mecanisme de condiționare prin referință este prezentată în Figura 2.

4.3. Extensia video

Restaurarea independentă a cadrelor introduce inconsistențe temporale, manifestate prin variații de culoare, textură și identitate între cadre succesive.

Pentru reducerea acestor efecte sunt propuse două mecanisme complementare, ilustrate în Figura 3: (a) un modul de propagare temporală în spațiul latent, ce introduce informație din cadrele anterioare înaintea procesului de predicție a codurilor discrete, și (b) un modul de fuziune

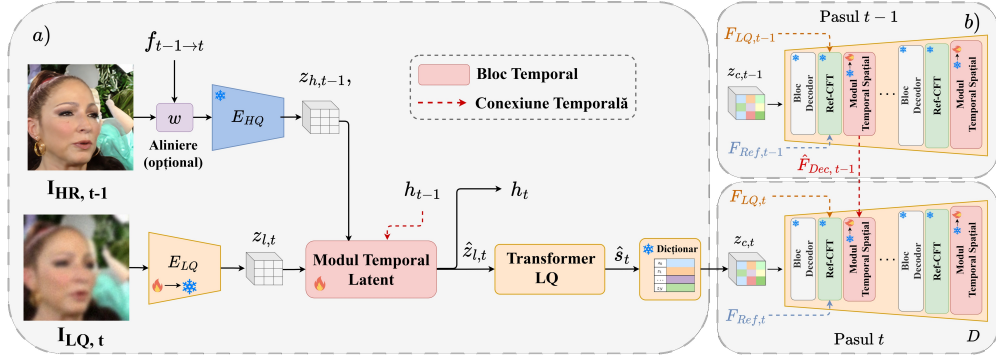


Figura 3: Extensie video pentru modelarea coerenței temporale: (a) modul temporal latent cu fuziune recurentă a informației din cadrele anterioare; (b) integrare în decodor pentru rafinarea reprezentărilor curente și reducerea inconsistențelor temporale.

temporală la nivelul decodurului, destinat reducerii fluctuațiilor locale de textură și identitate în etapa de reconstrucție. Ambele componente sunt formulate într-o manieră modulară, ceea ce permite integrarea lor atât în arhitectura de bază CodeFormer, cât și în extensia ghidată prin referință.

Procesarea recurentă a secvenței temporale. Această variantă procesează secvența video cadru cu cadru, permițând inferență online cu consum constant de memorie. Metoda este inspirată de abordări de tip KEEP, însă simplifică modelarea temporală prin utilizarea unui modul ConvGRU în spațiul latent și a unui modul eficient de fuziune în decodor.

La momentul temporal t , starea latentă este actualizată recurent:

$$h_t = \text{ConvGRU}(z_{h,t}, h_{t-1}), \quad (7)$$

iar reprezentarea utilizată ulterior de model este obținută prin combinarea observației curente cu informația temporală acumulată:

$$\hat{z}_{l,t} = f(z_t, h_t). \quad (8)$$

Astfel, coerența temporală este propagată prin starea ascunsă, iar corecția este implementată fie prin **interpolare adaptivă**, fie printr-o **conexiune reziduală** controlată de o poartă învățabilă. Complementar, la nivelul decodurului, caracteristicile cadrului anterior sunt fuzionate cu cele ale cadrului curent la rezoluțiile de injectare a identității, prin convoluții separabile pe adâncime (DWS).

Procesarea secvențelor video pe ferestre temporale. În această variantă, secvența video este procesată pe ferestre temporale de dimensiune fixă T , în locul unei procesări recurente cadru-cu-cadru. Această strategie permite modelarea simultană a dependențelor temporale locale, exploatând contextul disponibil în fereastra curentă, fără costurile asociate propagării stării pe secvențe lungi.

Dependențele temporale sunt modelate prin mecanisme de tip *self-attention*, aplicate de-a lungul dimensiunii temporale, atât în spațiul latent, cât și în decodor. Pentru fiecare poziție spațială, reprezentările celor T cadre sunt reorganizate și procesate printr-un bloc Transformer aplicat pe axa temporală:

$$\begin{aligned} \tilde{X} &= X + \text{MSA}(\text{LN}(X)), \\ X^{\text{ctx}} &= \tilde{X} + \text{MLP}(\text{LN}(\tilde{X})), \end{aligned} \quad (9)$$

unde $X \in \mathbb{R}^{B \times T \times C}$ reprezintă secvența de trăsături pentru o poziție spațială fixă, MSA reprezintă mecanismul *multi-head self-attention* pe axa temporală, LN este operația de normalizare *LayerNorm* [25], iar MLP este un bloc alcătuit din două straturi liniare. Acest mecanism agregă informația din contextul temporal al ferestrei, reducând simultan inconsistențele inter-cadru în reprezentările din spațiul latent și decodor.

5. Metodologie experimentală și analiza rezultatelor

Modelele propuse sunt implementate în PyTorch și antrenate cu optimizatorul AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.99$, $\eta = 10^{-5}$). Rezoluția de procesare a imaginilor, respectiv cadrelor temporale este 512×512 pixeli, cu o dimensiune a loturilor la antrenare de egală cu 4. Antrenarea, precum și raportarea metricilor și a timpilor de inferență sunt realizate pe un GPU RTX 4090.

5.1. Experiment I – Restaurare facială generică

5.1.1. Metodologie de antrenare

Modelul este inițializat cu ponderile modelului de bază CodeFormer [4] și rafinat pe setul **FFHQ** [3] pentru încă 250 000 iterații, excluzând imaginile din FFHQ-Ref [8], acestea fiind rezervate pentru validare și testare. Imaginile de antrenare sunt degradate sintetic conform schemei în **doi pași** descrise în Secțiunea 3, cu parametrii sintetizați în Tabelul 1.

Tabelul 1: Parametrii schemei de degradare în doi pași.

	Pasul 1 ($\mathcal{D}_1$)	Pasul 2 ($\mathcal{D}_2$)
Nucleu k_i : dim. (px); σ_k	[7, 21]; [0.2, 3.0]	[7, 21]; [0.2, 1.5]
Factor redimensionare r_i	[0.15, 1.5]	[0.1, 1.2]
Zgomot σ_n	[1, 30]	[1, 25]
JPEG q_i	[30, 95]	[30, 95]
Filtru sinc final (prob.)	0.0	0.2

Funcția de cost a modelului de bază este extinsă prin două componente principale: (i) un termen de identitate ArcFace [22], ce minimizează distanța în spațiul de recunoaștere facială dintre imaginea restaurată și referința HQ (echivalent cu maximizarea similarității cosinus), și (ii) un discriminator pe regiuni de interes (ochi și gură), care impune coerență locală și consistență structurală a trăsăturilor faciale relevante perceptual. Funcția de cost totală este definită astfel:

$$\begin{aligned}
\mathcal{L}_{\text{total}} = & \underbrace{\lambda_{CE}\mathcal{L}_{CE} + \lambda_{feat}\mathcal{L}_{feat}}_{\text{cod latent}} + \underbrace{\lambda_{Charb}\mathcal{L}_{Charbonnier} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{GAN}\mathcal{L}_{GAN}}_{\text{reconstrucție globală}} \\
& + \underbrace{\lambda_{Charb}^{ROI}\mathcal{L}_{Charbonnier}^{ROI} + \lambda_{GAN}^{ROI}\mathcal{L}_{GAN}^{ROI}}_{\text{regiuni de interes}} + \lambda_{ID}\mathcal{L}_{\text{ArcFace}}, \tag{10}
\end{aligned}$$

unde $\mathcal{L}_{CE}$ reprezintă costul de tip entropie încrucișată categorică pentru secvențele de cod, $\mathcal{L}_{feat}$ eroarea MSE în spațiul codului latent, termenii globali controlează fidelitatea și realismul reconstrucției, iar termenii ROI impun consistență locală în zonele faciale critice. Coeficienții sunt setați $\lambda_{CE} = \lambda_{feat} = \lambda_{Charb} = \lambda_{LPIPS} = \lambda_{ID} = 1.0$, $\lambda_{Charb}^{ROI} = 0.5$, iar $\lambda_{GAN} = \lambda_{GAN}^{ROI} = 0.1$. Se utilizează doi discriminatori de tip *PatchGAN* [26], iar costul adversarial este formulat în varianta *non-saturating*.

5.1.2. Analiza performanțelor

Evaluarea cantitativă utilizează metricile LPIPS [27] ($\downarrow$), IDS ($\uparrow$, similaritate cosinus în spațiul ArcFace față de imaginea de referință), NIQE ($\downarrow$) și CLIP-IQA [28] ($\uparrow$), cu o valoare a factorului de control a restaurării $w = 1.0$ pentru degradări moderate, respectiv $w = 0.5$ pentru degradări severe. Pe setul de evaluare reală **NTIRE 2026** [17], care nu dispune de imagini de referință (*ground truth*), LPIPS nu poate fi raportat, iar fidelitatea identitară este evaluată conform protocolului competiției prin similaritatea cosinus dintre imaginea degradată și reconstrucție, ca măsură a fidelității față de structura identitară a intrării.

Analiza rezultatelor din Tabelul 2 evidențiază îmbunătățiri consistente ale metodei propuse pe toate seturile de evaluare. În scenariile de degradare sintetică se observă un câștig constant în ceea ce privește calitatea perceptuală, reflectat prin reducerea scorului NIQE și creșterea CLIP-IQA, indicând reconstrucții mai naturale și mai apropiate de distribuția imaginilor reale. În același timp, metricile de identitate arată un comportament dependent de severitatea degradării: în regim moderat, se obține o creștere consistentă a fidelității identitare, în timp ce în scenariu sever apare o ușoară degradare, explicată prin pierderea informației structurale din intrarea LQ.

Pe setul real de imagini degradate, NTIRE 2026 [17], metoda propusă generalizează favorabil, obținând îmbunătățiri clare în fidelitatea identitară și scoruri competitive pe metricile perceptuale, confirmând robustețea abordării în condiții non-sintetice și absența unei referințe de calitate.

Tabelul 2: Comparatie cantitativă pe seturi sintetice (FFHQ-Ref-Test și CelebA-Test-Ref), respectiv reale (NTIRE 2026 [17]) în regim de degradare moderată și severă. Valorile evidențiate indică cel mai bun (**bold**) și al doilea cel mai bun (underline) rezultat pentru fiecare metrică.

Set de test	CodeFormer [4]				Metoda Propusă			
	LPIPS $\downarrow$	IDS $\uparrow$	NIQE $\downarrow$	CLIP-IQA $\uparrow$	LPIPS $\downarrow$	IDS $\uparrow$	NIQE $\downarrow$	CLIP-IQA $\uparrow$
FFHQ-Ref-Test (Moderat)	0.1588	0.7644	4.5102	0.6766	0.1570	0.7875	4.3110	0.6969
FFHQ-Ref-Test (Sever)	0.2649	0.3358	4.6514	0.7003	0.2866	0.3043	3.3680	0.7344
CelebA-Test-Ref	0.1750	0.7505	4.6524	0.6404	0.1929	0.7725	4.1998	0.6927
NTIRE 2026-Test	–	4.499	0.684	0.726	–	4.463	0.689	0.787

Figura 4 prezintă comparații calitative. În situații de degradare moderată, metoda propusă conservă semnalmamente identitare fine – pistrui, diastemă, nevi pigmentari – în cazul în care acestea **sunt prezente în imaginea degradată**, în timp ce CodeFormer le atenuază. Totuși, în condiții severe de degradare, ambele metode apelează preponderent la priori statistici, diferențele identitare fiind irecuperabile.

5.2. Experiment II – Restaurare asistată prin referință (CodeRefFormer)

5.2.1. Metodologie de antrenare

Metoda propusă pentru restaurare ghidată prin referință este inițializată cu ponderile modelului din etapa anterioară și antrenată pe subsetul de antrenare al **FFHQ-Ref** [8] pentru 200 000 de iterații. Pentru fiecare identitate sunt utilizate cinci imagini de referință per iterație, supuse unor augmentări de culoare, luminozitate și perspectivă, pentru a crește robustețea la variații extreme de poziție și iluminare.

Schema de degradare este definită printr-un singur pas de corupere, intenționat mai sever decât în etapa anterioară, cu $\sigma_k \in [0, 16]$, redimensionare până la un factor de $32\times$ și fără filtrare trece-jos de tip sinc. Această configurație reduce dependența de semnalul rezidual al imaginii de intrare și încurajează utilizarea informației din referință pentru reconstrucție.

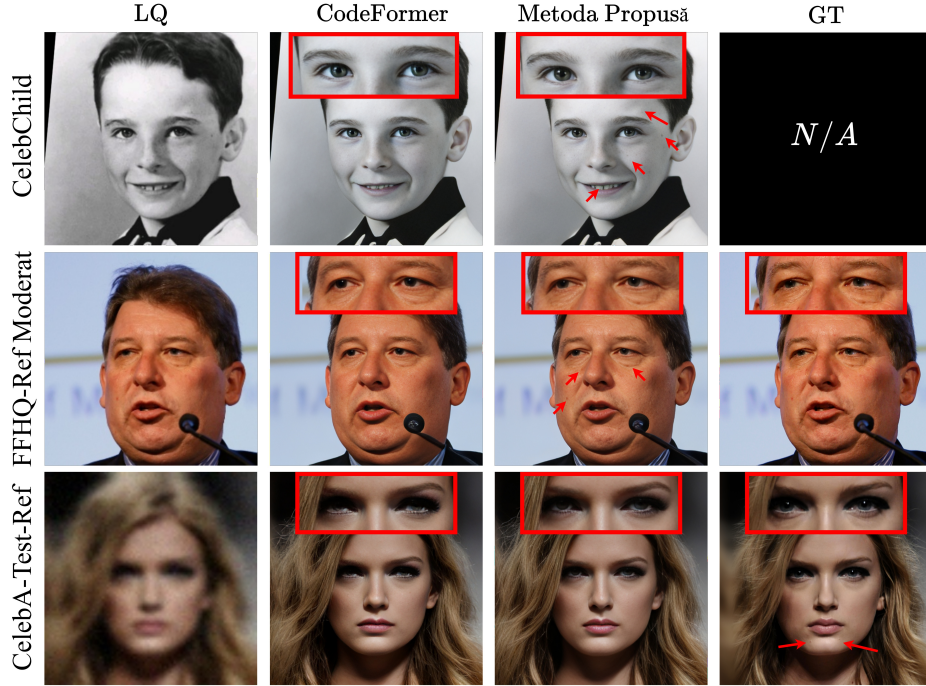


Figura 4: Comparație CodeFormer vs. metoda propusă pe CelebA-Test-Ref. De la stânga la dreapta: LQ, CodeFormer, metoda propusă, HQ. În degradare moderată, metoda propusă păstrează mai bine semnalmente identitare, în timp ce CodeFormer le atenuează. În degradare severă, ambele metode se bazează pe priori statistici, iar diferențele identitare dispar.

Funcția de cost extinde formularea din etapa anterioară prin introducerea unui mecanism de ancorare bazat pe referință, în care $\phi(\cdot)$ reprezintă extractorul de caracteristici faciale ArcFace [22], iar termenul de referință optimizează similaritatea cosinus față de media reprezentărilor vectoriale celor cinci imagini de referință, reducând sensibilitatea la variațiile scenei:

$$\mathcal{L}_{\text{ArcFace}}^{\text{ref}} = \lambda_{ID}^{GT} (1 - \cos(\phi(\hat{I}), \phi(I_{HQ}))) + \lambda_{ID}^{\text{Ref}} \left(1 - \cos\left(\phi(\hat{I}), \frac{1}{5} \sum_{i=1}^5 \phi(I_i^{\text{ref}})\right) \right), \quad (11)$$

unde $\lambda_{ID}^{GT} = \lambda_{ID}^{\text{Ref}} = 1.0$, iar ceilalți coeficienți rămân identici cu cei din etapa anterioară.

5.2.2. Analiza performanțelor

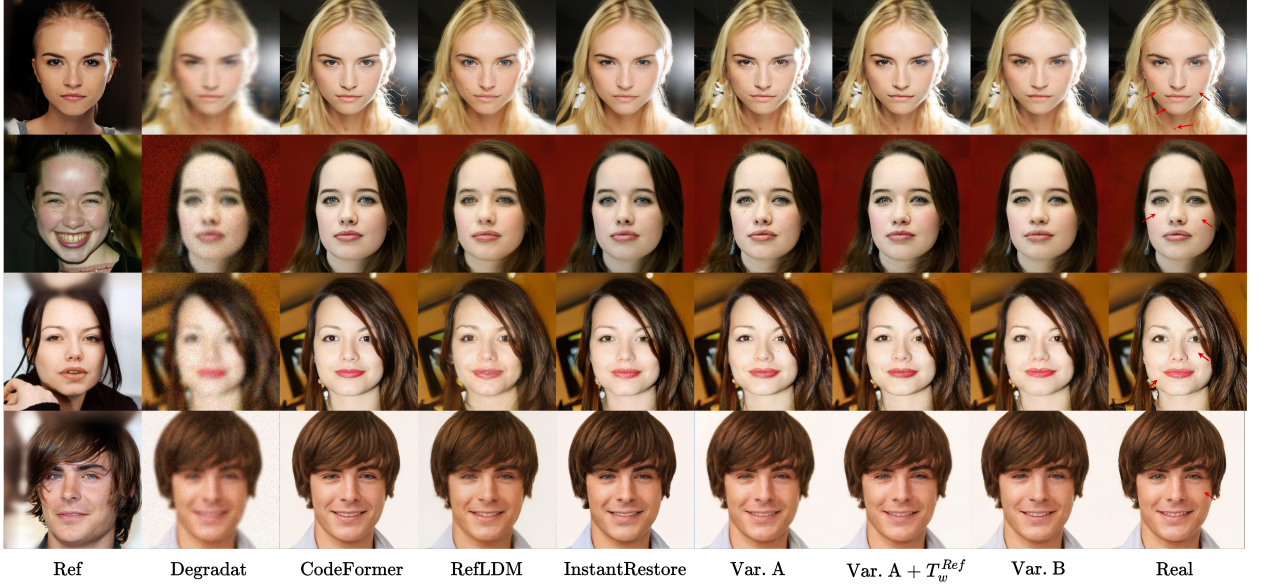
Evaluarea este realizată pe seturile **FFHQ-Ref-Test** și **CelebA-Test-Ref**, utilizând 3 referințe per subiect pentru fiecare model la momentul inferenței. Tabelul 3 prezintă comparația cantitativă, incluzând numărul de parametri și timpul de inferență.

Notă privind posibila contaminare a datelor. Setul **FFHQ-Ref-Test** nu este utilizat în antrenarea metodei propuse, fiind rezervat exclusiv evaluării. Totuși, modelul de inițializare CodeFormer [4] a fost pre-antrenat pe întregul FFHQ, ceea ce poate introduce un posibil efect de *data leakage* în comparațiile directe. Prin urmare, rezultatele pe acest subset trebuie interpretate cu precauție și coroborate cu evaluări pe seturi independente, precum CelebA-Test-Ref.

Analizând rezultatele obținute, se observă că introducerea informației de referință îmbunătățește semnificativ fidelitatea identitară față de varianta generică, efectul fiind mai pronunțat în regim de degradare severă. Dintre variantele propuse, Varianta A se evidențiază prin cel mai bun compromis între conservarea identității și calitatea perceptuală, depășind în majoritatea cazurilor ReF-LDM [8]

Tabelul 3: Comparație cantitativă (3 referințe). **Bold:** cel mai bun. **Roșu:** al doilea. Subliniat: al treilea.

Metodă	Ref.	Param.	Timp	FFHQ-Ref-Test (Moderat)				FFHQ-Ref-Test (Sever)				CelebA-Test-Ref			
				LPIPS	IDS	NIQE	CLIP-IQA	LPIPS	IDS	NIQE	CLIP-IQA	LPIPS	IDS	NIQE	CLIP-IQA
CodeFormer [4]	–	94M	32 ms	0.159	0.764	4.510	0.677	0.265	0.336	4.651	0.700	0.175	0.751	4.652	0.640
Metoda propusă (generică)	–	94M	32 ms	0.157	0.787	4.311	0.697	0.287	0.304	3.368	<u>0.734</u>	0.193	0.773	<u>4.200</u>	0.693
ReF-LDM [8]	✓	175M	1.90 s	0.185	0.837	4.444	0.662	0.270	<u>0.665</u>	4.409	0.681	0.202	0.774	4.615	0.625
InstantRestore [9]	✓	1.93B	0.41 s	0.167	0.872	6.006	0.577	0.276	0.696	6.300	0.586	0.175	0.829	6.197	0.559
Var. A (<i>propus</i>)	✓	128M	59 ms	<u>0.158</u>	<u>0.867</u>	4.174	0.697	<u>0.266</u>	0.521	3.989	0.737	0.198	<u>0.798</u>	4.186	0.667
Var. A + T_w^{Ref} (<i>propus</i>)	✓	157M	64 ms	0.162	0.850	<u>4.204</u>	0.707	0.268	0.536	4.095	0.723	0.202	0.787	4.297	<u>0.678</u>
Var. B (<i>propus</i>)	✓	289M	39 ms	0.159	0.862	4.353	0.684	0.270	0.544	4.289	0.706	<u>0.193</u>	0.795	4.510	0.667

**Figura 5:** Comparație calitativă pe CelebA-Test-Ref (1 referință). Coloanele prezintă, de la stânga la dreapta: imaginea degradată (LQ), CodeFormer, ReF-LDM, InstantRestore, Varianta A (cu și fără extensia T_w^{Ref}), Varianta B și referința HQ.

atât din perspectiva metricilor de restaurare, cât și a eficienței computaționale.

Este important de menționat că rezultatele ReF-LDM sunt obținute utilizând configurația recomandată de autori, cu 100 pași de restaurare și $\lambda_{cfg} = 1.5$, ceea ce conduce la un timp de inferență de aproximativ 1.90 s per imagine. În aceste condiții, Varianta A necesită doar 59 ms pentru inferență, menținând în același timp valori NIQE și IDS comparabile sau superioare. Extensia T_w^{Ref} conduce la îmbunătățiri suplimentare ale calității perceptuale, evidențiate prin scorurile CLIP-IQA, precum și la creșteri ale fidelității identitare în regim de degradare severă. Deși InstantRestore [9] obține cele mai ridicate valori IDS, acest avantaj este însoțit de o degradare semnificativă a metricilor perceptuale și de o complexitate computațională considerabil mai mare. În ansamblu, metodele propuse oferă un echilibru favorabil între fidelitate identitară, calitatea restaurării și eficiența inferenței.

Figura 5 prezintă o comparație calitativă pe exemple din setul de test sintetic CelebA-Test-Ref, utilizând o singură referință HQ la inferență. Varianta A se remarcă prin capacitatea de a reconstitui semnalmamente identitare subtile, precum nevi pigmentari sau pistrui, absente în rezultatele modelului generic și, în anumite cazuri, ReF-LDM. În schimb, Varianta B conservă cu precădere attribute faciale globale, precum forma și culoarea ochilor. InstantRestore poate reproduce ocazional astfel de detalii, însă cu o consistență mai redusă decât Varianta A.

Condiționare cu referință eronată. Un caz limită îl reprezintă utilizarea unor referințe aparținând unei identități diferite de cea din imaginea degradată. Acest experiment evidențiază atât

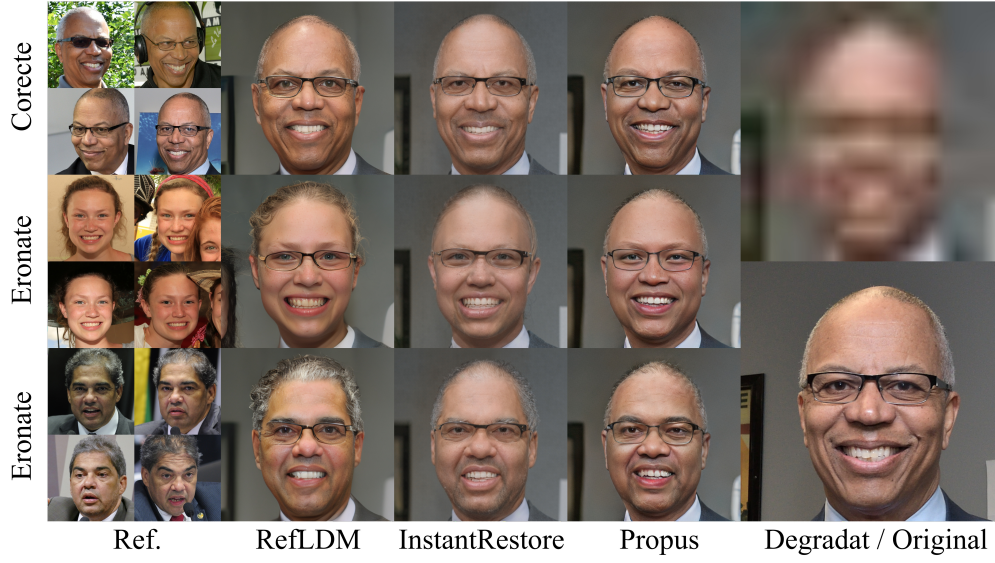


Figura 6: Analiza efectului unei referințe eronate. Coloanele prezintă: referințe HQ (corectă/eronată), rezultatele ReF-LDM [8], InstantRestore [9], metoda propusă (Varianta A + T_w^{Ref}), respectiv pe ultima coloană imaginea degradată și cea originală.

influența ramurii de condiționare identitară asupra restaurării, cât și riscul utilizării unor astfel de metode pentru transferul de caracteristici identitare, fenomen asociat generării pasive de conținut *deepfake*.

Figura 6 compară metoda propusă (Varianta A + T_w^{Ref}), ReF-LDM și InstantRestore în acest scenariu, folosind pentru fiecare metodă numărul maxim de referințe permis de configurația experimentală.

5.3. Experiment III – Extensie video

5.3.1. Metodologie de antrenare

Extensia video pornește de la Varianta A a metodei propuse cu ghidaj pe bază de referință și introduce mecanisme de coerență temporală antrenate în două etape, fiecare pentru 50 000 de iterații pe **subsetul de antrenare propriu** (533 clipuri, 310 identități), cu 3 referințe per identitate și secvențe video degradate alcătuite din 10 cadre consecutive. Schema de degradare utilizează un singur pas sever: estompare ($\sigma_k \in [0, 16]$, nuclee 21×21 px), reducerea rezoluției până la $10\times$ și compresie video H.264 cu CRF $\sim \mathcal{U}[25, 45]$, aplicată la nivel de secvență.

Antrenarea se desfășoară în două etape. *Etapa 1* rafinează codorul LQ și modulul temporal la nivelul codului latent:

$$\mathcal{L}_I = \lambda_{CE} \mathcal{L}_{CE} + \lambda_{feat} \mathcal{L}_{feat}, \quad \lambda_{CE} = 0.5, \quad \lambda_{feat} = 1.0. \quad (12)$$

Etapa 2 adaugă modulele de aliniere temporală în decodor și extinde funcția de cost cu termeni de reconstrucție, identitate și coerență temporală bazată pe flux optic, cu mască de ocluzie:

$$\mathcal{L}_{II} = \mathcal{L}_I + \lambda_{Charb} \mathcal{L}_{Charbonnier} + \lambda_{LPIPS} \mathcal{L}_{LPIPS} + \lambda_{ID} \mathcal{L}_{ArcFace}^{ref} + \lambda_{temp} \mathcal{L}_{temp}, \quad (13)$$

unde $\lambda_{CE} = 0.5$ și toți ceilalți coeficienți = 1.0. Costul adversarial este omis pentru a reduce instabilitatea indusă de procesarea temporală, iar pentru a asigura coerența temporală, se introduce un termen de regularizare $\mathcal{L}_{temp}$, formulat ca eroarea pătratică medie (MSE), aplicată doar în regiunile

Tabelul 4: Rezultate cantitative pe setul video propus (3 referințe pentru condiționarea identității). **Bold:** cel mai bun. Subliniat: al doilea. **Roșu:** al treilea.

Metodă	ms/ cadru	Redus				Moderat				Sever			
		IDS↑	σ_{IDS} ↓	E_w ↓	CLIP↑	IDS↑	σ_{IDS} ↓	E_w ↓	CLIP↑	IDS↑	σ_{IDS} ↓	E_w ↓	CLIP↑
CodeFormer [4]	32	0.887	2.36	7.53	0.672	0.633	4.85	8.02	<u>0.691</u>	0.561	4.46	6.58	<u>0.690</u>
Var. A propusă (cadru cu cadru)	32	<u>0.941</u>	1.85	8.00	<u>0.669</u>	0.688	4.39	8.55	0.708	0.585	4.32	7.01	0.706
<i>Extensii video propuse (pornind de la Var. A):</i>													
GRU-Interp	46	0.902	1.94	7.09	0.642	0.685	4.06	7.20	0.652	0.600	4.00	6.15	0.658
GRU-Interp+DWS	46	0.908	1.88	<u>6.12</u>	0.535	0.691	4.07	<u>6.34</u>	0.552	0.603	3.82	5.33	0.565
GRU-Rez	46	0.927	1.87	<u>7.13</u>	0.554	0.691	4.14	7.32	0.580	0.604	3.83	6.10	0.591
GRU-Rez+DWS	46	0.932	1.78	6.46	0.512	0.698	4.14	6.55	0.544	<u>0.609</u>	3.80	5.43	0.555
Atenție-Temp. (Latent)	38	0.928	1.85	7.16	0.555	0.696	4.11	7.30	0.578	0.606	3.83	6.04	0.591
Atenție-Temp. (Latent+DCD)	38	0.942	1.38	5.41	0.458	0.737	3.42	5.29	0.493	0.637	3.48	4.42	0.501
KEEP [29]	45	0.871	<u>1.65</u>	6.36	0.493	0.683	<u>3.87</u>	6.42	0.490	0.597	<u>3.49</u>	<u>5.25</u>	0.502

lipsite de ocluzii, determinate pe baza unei măști de consistență din fluxul optic bidirecțional.

Sunt comparate nouă configurații: (i) CodeFormer și Varianta A propusă aplicate cadru cu cadru, ca soluții de referință; (ii) patru variante recurente cu propagare temporală prin ConvGRU, diferențiate prin mecanismul de fuziune la nivelul codului latent (GRU-Interp, GRU-Interp+DWS, GRU-Rez, GRU-Rez+DWS); (iii) două variante cu procesare pe fereastră, bazate pe atenție aplicată exclusiv în spațiul latent, respectiv extinsă și la nivelul decodurului (Latent+DCD). Ca referință externă este inclus KEEP [29], un model consacrat de restaurare video facială generică.

5.3.2. Analiza performanțelor

Performanța este evaluată prin metrici de fidelitate identitară (IDS), consistență identitară (σ_{IDS}), temporală (E_{warp}), calitate perceptuală (CLIP-IQA) și latență per cadru. Tabelul 4 prezintă rezultatele pe setul video propus, la cele trei niveluri de degradare sintetizate, incluzând latența per cadru. Rezultatele evidențiază un compromis constant între consistența temporală și calitatea perceptuală. Mecanismele de propagare temporală reduc variația inter-cadru a identității (σ_{IDS}) și eroarea de consistență (E_{warp}), dar determină o scădere a CLIP-IQA, indicând o limitare a detaliilor fine în favoarea stabilității temporale.

Dintre variantele recurente, *GRU-Interp*. oferă cel mai bun echilibru între fidelitate identitară, coerență temporală și calitate perceptuală. Alinierea explicită la nivelul decodurului îmbunătățește suplimentar stabilitatea, însă accentuează degradarea CLIP-IQA, sugerând o pierdere în generarea detaliilor locale.

Cea mai bună performanță este obținută de varianta cu atenție temporală (Latent+DCD), care maximizează IDS și stabilitatea temporală pe toate nivelurile de degradare, depășind variantele recurente și KEEP [29], dar cu cele mai scăzute valori CLIP-IQA, confirmând compromisul între realism perceptual și coerență temporală.

În cazul metodelor de restaurare bazate pe procesarea independentă a fiecărui cadru, se observă fluctuații temporale vizibile (*flickering*), manifestate prin variații ale caracteristicilor faciale locale între cadre consecutive și a identității. Acest fenomen este ilustrat în Figura 7, unde anumite trăsături, precum culoarea irisului, forma ochilor, caracterul identitar sau detaliile din regiunea bucală, prezintă modificări temporare asociate schimbării bruște a poziției capului.

6. Concluzii

Lucrarea a propus un sistem modular și incremental de restaurare facială orientat către prezervarea identității, construit pe baza modelului CodeFormer [4] și extins în trei direcții principale.

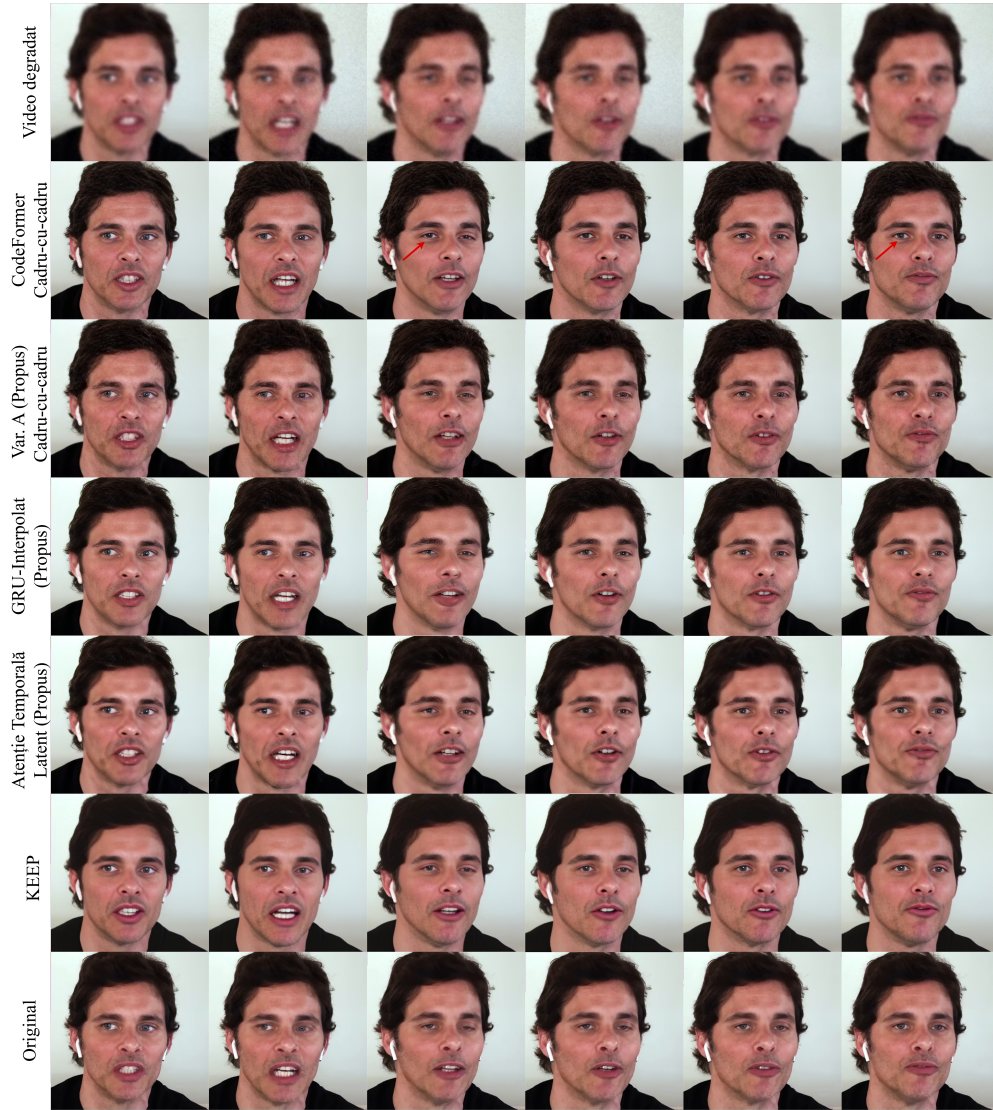


Figura 7: Comparație a metodelor de restaurare video pe o secvență cu degradare moderată. Metodele fără consistență temporală prezintă *flickering* și instabilitate identitară, vizibile prin variații ale trăsăturilor faciale odată cu schimbarea poziției capului.

Restaurare generică cu supervizare identitară. Introducerea unui termen de identitate bazat pe similaritate cosinus în spațiul ArcFace, împreună cu un discriminator pe regiuni faciale relevante (ochi și gură), conduce la îmbunătățiri consistente față de modelul de bază. Rezultatele confirmă că supervizarea în spațiul de recunoaștere facială stabilizează reconstrucția identitară, în special în regim de degradare moderată.

Restaurare ghidată prin imagini de referință. Modulul Ref-CFT, bazat pe modulație spațială multi-rezoluție și agregare adaptivă a referințelor, permite injectarea eficientă a informației identitare fără reantrenare per identitate. Varianta A depășește ReF-LDM la nivelul metricilor IDS și NIQE, în scenarii de degradare moderată, utilizând 26% mai puțini parametri și având un cost de inferență de aproximativ $32\times$ mai redus. Varianta B, condiționată de reprezentări faciale ArcFace și FaRL, are un cost de inferență slab dependent de numărul de referințe, fiind adecvată scenariilor cu seturi extinse de referințe. Ambele variante depășesc metoda generică în regim de degradare severă, unde semnalul identitar din imaginea de intrare devine insuficient pentru ghidare.

Extensie la restaurare video. Mecanismele de coerență temporală propuse – propagare în spațiul latent prin ConvGRU și fuziune în decodor prin DWS, respectiv atenție temporală – reduc eroarea de consistență inter-cadru față de abordările ce procesează cadru cu cadru. Varianta cu atenție temporală obține cele mai bune valori IDS și σ_{IDS} pe toate nivelurile de degradare, depășind KEEP [29]. De asemenea, setul de date propus constituie o resursă dedicată restaurării video ghidate prin referință, cu potențial de utilizare și extindere în cercetări viitoare din această direcție.

Limitări și direcții viitoare. O limitare importantă este sensibilitatea la referințe eronate, în absența unui mecanism explicit de verificare a identității, ceea ce poate conduce la transfer nedorit de trăsături și riscuri de impersonare. Direcții viitoare includ integrarea unor mecanisme de protecție precum inscripția digitală perceptuală (*perceptual watermarking*), precum și rafinarea modelelor temporale existente pentru o modelare mai fidelă a dinamicii și expresivității faciale în secvențe video.

Referințe

- [1] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards real-world blind face restoration with generative facial prior,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9164–9174.
- [2] T. Yang, P. Ren, X. Xie, and L. Zhang, “Gan prior embedded network for blind face restoration in the wild,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 672–681.
- [3] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.
- [4] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, “Towards robust blind face restoration with codebook lookup transformer,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS ’22, 2022.
- [5] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 868–12 878, 2020.
- [6] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, and C. Dong, “Diffbir: Toward blind image restoration with generative diffusion prior,” Berlin, Heidelberg, p. 430–448, 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-73202-7_25
- [7] J. Wang, J. Gong, L. Zhang *et al.*, “Osdface: One-step diffusion model for face restoration,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 12 626–12 636.
- [8] C.-W. Hsiao *et al.*, “ReF-LDM: A latent diffusion model for reference-based face image restoration,” in *Advances in Neural Information Processing Systems*, 2024.
- [9] H. Zhang, Y. Alaluf, S. Ma, A. Kadambi, J. Wang, and K. Aberman, “InstantRestore: Single-step personalized face restoration with shared-image attention,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [10] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, “On the generalization of BasicVSR++ to video deblurring and denoising,” *arXiv preprint arXiv:2204.05308*, 2022.
- [11] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, “Edvr: Video restoration with enhanced deformable convolutional networks,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1954–1963, 2019.
- [12] Z. Wang, X. Chen, C. Xu, J. Zhu, X. Hu, J. Zhang, C. Wang, Y. Liu, Y. Zhou, and R. Ji, “Svfr: A unified framework for generalized video face restoration,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 7406–7415.
- [13] Z. Chen, B. Chai, R. Gao, M. Nie, X. Li, B. Duan, J. Fang, X. Liu, L. Kong, and Y. Zhang, “Dvface:

- Spatio-temporal dual-prior diffusion for video face restoration,” *arXiv preprint arXiv:2604.14560*, 2026.
- [14] J. Tan, H. Park, Y. Zhang *et al.*, “Blind face video restoration with temporal consistent generative prior and degradation-aware prompt,” in *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, 2024, p. 1417–1426.
 - [15] W. Han, W. Lin, Y. Zhou, Q. Liu, S. Wang, C. Yao, and J. Chen, “Show and polish: Reference-guided identity preservation in face video restoration,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM ’25. New York, NY, USA: Association for Computing Machinery, 2025, p. 10315–10324.
 - [16] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.06721>
 - [17] J. Wang, J. Gong, Z. Chen, K. Liu, J. Li, R. Timofte, Y. Zhang *et al.*, “The second challenge on real-world face restoration at ntire 2026: Methods and results,” in *CVPRW*, 2026.
 - [18] X. Li, S. Zhang, S. Zhou, L. Zhang, and W. Zuo, “Learning dual memory dictionaries for blind face restoration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, p. 5904–5917, May 2023.
 - [19] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
 - [20] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan, “Vfhq: A high-quality dataset and benchmark for video face super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.
 - [21] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution with pure synthetic data,” in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021, pp. 1905–1914.
 - [22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.
 - [23] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, “General facial representation learning in a visual-linguistic manner,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 676–18 688.
 - [24] M. Zhou, K. Ye, V. Shah *et al.*, “Reference-guided identity preserving face restoration,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.21905>
 - [25] J. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *ArXiv*, vol. abs/1607.06450, 2016.
 - [26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.
 - [27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
 - [28] J. Wang, K. C. Chan, and C. C. Loy, “Exploring CLIP for assessing the look and feel of images,” in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, ser. AAAI’23. AAAI Press, 2023. [Online]. Available: <https://doi.org/10.1609/aaai.v37i2.25353>
 - [29] R. Feng, C. Li, and C. C. Loy, “Kalman-inspired feature propagation for video face super-resolution,” in *ECCV 2024: 18th European Conference, Milan, Italy, Sept. 29–Oct. 4, 2024*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 202–218. [Online]. Available: https://doi.org/10.1007/978-3-031-73347-5_12