DOI: https://doi.org/10.15276/ict.02.2025.39
УДК 004.932:004.382:004.056.5

Генеративна модель MST-GAN для покращення відео: якість, стабільність, ефективність

Максимів Микола Романович
Аспірант каф. Електронних обчислювальних машин
ORCID: https://orcid.org/0009-0004-4915-6265; mykolamaksymivua@gmail.com
Національний університет «Львівська політехніка», вул. С. Бандери, 12. Львів, 79000, Україна

АНОТАЦІЯ

У роботі розглянуто аналіз вдосконалення архітектури MST-GAN (Multi-Scale Temporal GAN) для задач відео-суперрезолюції, спрямоване на забезпечення високої візуальної якості, міжкадрової стабільності та ефективної роботи в реальному часі. Модель поєднує багатомасштабне вирівнювання ознак, часову агрегацію та генеративну генерацію кадрів з використанням гібридної функції втрат.

Особливу увагу приділено забезпеченню стабільності навчання: попереднє навчання генератора без дискримінатора, поетапне включення часової узгодженості, застосування перцептивних критеріїв та регуляризаційних технік дозволили уникнути типових проблем генеративного навчання, зокрема нестабільної динаміки та втрати різноманіття. У процесі тренування модель адаптована до реалістичних сценаріїв відео, що дозволяє їй зберігати якість навіть у складних сценах з динамічними об'єктами.

Також описано низку апаратних оптимізацій, що включають структурне проріджування моделі, квантування ваг у формат INT8, компіляцію в TensorRT та організацію потокової обробки кадрів. У результаті MST-GAN досягла значного прискорення інференсу без помітної втрати якості.

Якісні приклади на авторських відеоданих (зокрема, сцена з активним рухом) демонструють переваги моделі у збереженні текстур і плавності руху, порівняно з класичними методами збільшення зображення. На відміну від конкурентних підходів, MST-GAN дозволяє уникнути характерного «миготіння» та забезпечує природну передачу динаміки сцени. Отримані результати свідчать про придатність MST-GAN для використання у практичних системах відеопокращення, зокрема у відеострімах, моніторингових системах, AR-застосунках, де поєднання якості, стабільності та швидкодії є критично важливим.

Ключові слова: відео високої чіткості; відео-суперрезолюція; генеративні мережі; оптимізація на GPU; перцептивна якість; покадрова стабільність; продуктивність моделі; структурне проріджування; часовий дискримінатор; якість відео

Актуальність. Сучасна динаміка розвитку відеоконтенту, починаючи від онлайн-трансляцій і відеоконференцій, закінчуючи системами відеоспостереження й генеративних мультимедійних платформ – вимагає високої якості відео за умов обмежених каналів зв'язку, низької вихідної роздільної здатності або поганих умов зйомки. Традиційні алгоритми масштабування (як-от бікубічна інтерполяція) не здатні відновлювати втрачений візуальний зміст, бо вони створюють розмиті зображення, спотворені артефактами, особливо при наявності руху або текстурних деталей [1].

Натомість сучасні моделі глибокого навчання, зокрема генеративні нейронні мережі на основі GAN (Generative Adversarial Networks), демонструють здатність фотореалістично відновлювати відеоряд завдяки навчанню на великих наборах даних (рис. 1). Такі моделі формують детальні й реалістичні зображення з низькороздільного відео, зберігаючи текстурні структури [2]. Проте процес їхнього навчання супроводжується численними викликами: нестабільна збіжність, зменшення різноманітності результатів (mode collapse) і коливання значень функції втрат [2]. Для відеоданих проблема ускладнюється ще й потребою зберігати часову узгодженість між кадрами: навіть якщо поодинокі зображення виглядають якісно, їхнє послідовне відтворення може спричинити ефект «мерехтіння» або розсинхронізації [2].

Потреба в ефективних моделях відео-суперрезолюції (Video Super-Resolution, VSR) особливо актуалізується в прикладних задачах, де важливими є не лише якість, а й швидкодія, стабільність руху та здатність до виконання на апаратурі з обмеженими ресурсами. У цьому контексті завдання одночасного забезпечення високої якості відновлення та ефективності моделі залишається одним із ключових напрямів розвитку сучасних VSR-методів.

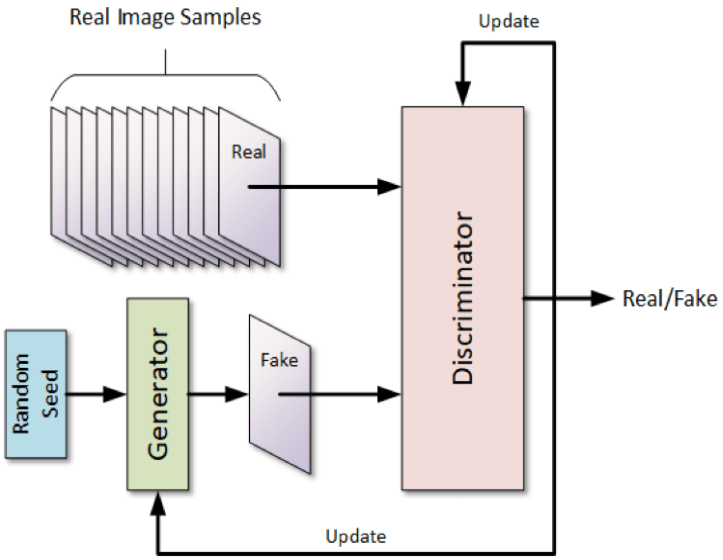

Рис. 1. **Типова архітектура GAN з генератором та дискримінатором**

Метою роботи є аналіз розширеної архітектури MST-GAN (Multi-Scale Temporal GAN), запропонованої у [1], з позиції її адаптації для практичного застосування. Основною метою є поєднання стабільного навчання, високої якості відновлення та апаратної ефективності для досягнення продуктивності, придатної до реального часу.

Зокрема, робота ставить за ціль:

1. проаналізувати поведінку MST-GAN при навчанні та описати методи усунення нестабільностей (колапс мод, артефакти часу);
2. оцінити апаратні оптимізації (структурне проріджування, квантування, компіляція з TensorRT), що дозволяють значно зменшити обчислювальну складність моделі без суттєвої втрати якості.

Крім цього, особливу увагу приділено демонстрації якісних результатів покадрової та міжкадрової реконструкції, які візуально підтверджують переваги MST-GAN на складних відеосценах.

Попередні роботи. На сьогодні розроблено низку ефективних моделей для відео-суперрезолюції, які демонструють високі показники відновлення зображень за класичними метриками точності. Однією з базових є модель BasicVSR, яка реалізує двонаправлену рекурентну схему з вирівнюванням кадрів на основі оптичного потоку. Її подальший розвиток, BasicVSR++, включає покращене розповсюдження ознак та точніше вирівнювання, що дозволяє досягати високої якості при помірній обчислювальній складності [3]. Модель IconVSR вводить ієрархічне накопичення прихованих станів для поліпшення узгодженості між кадрами.

Модель EDVR використовує каскад деформованих згорток і просторово-часовий модуль уваги для злиття інформації між кадрами [4]. Така архітектура забезпечує хороші результати у задачах з високим рівнем деформацій і складних рухів, але супроводжується великою обчислювальною вартістю. Натомість модель RSDN (Recurrent Structure-Detail Network) запроваджує окрему обробку структурних і детальних компонент зображення, що дозволяє зменшити дублювання обчислень завдяки рекурентній природі та знизити використання ресурсів [5].

Усі зазначені методи демонструють високі результати за метриками PSNR та SSIM на стандартних тестових наборах (Vimeo-90K-T, Vid4, REDS тощо), проте ці метрики часто не відображають суб'єктивну якість зображення. У дослідженнях [6] наголошується, що PSNR/SSIM можуть бути погано узгоджені з візуальним сприйняттям користувача. Тому все частіше в роботах використовується додаткове оцінювання за перцептивними метриками, такими як LPIPS (Learned Perceptual Image Patch Similarity) і VMAF (Video Multi-Method

Assessment Fusion), які краще враховують візуальні особливості та сприйняття послідовностей відео [6].

У попередній роботі авторів [1] було запропоновано модель MST-GAN, що поєднує багатомасштабне вирівнювання ознак, резидуальне підсилення, регуляризацію оптичного потоку та часовий дискримінатор.

Огляд архітектури MST-GAN та процес навчання. У попередній роботі авторів [1] було запропоновано модель MST-GAN (Multi-Scale Temporal GAN) - багатокадрову генеративну нейронну мережу, побудовану для покращення відео високої роздільної здатності зі збереженням стабільності руху. Архітектура поєднує кілька ключових компонентів, що працюють узгоджено:

1. Багатомасштабне вирівнювання ознак (MSFA). Вирівнювання між сусідніми кадрами виконується на кількох масштабах за допомогою або згорткових мереж, або оптичного потоку. Така пірамідальна структура забезпечує узгодження як глобальних змін у кадрі (рухи камери, великі об’єкти), так і локальних деталей (краї, текстури).
2. Резидуальний модуль підсилення деталей. Після початкового вирівнювання деякі дрібні структури можуть бути згладжені або втрачені. Відповідний модуль повторно підсилює різницю між базовим передбаченням і справжнім зображенням, додаючи втрачену текстурну інформацію.
3. Регуляризація оптичного потоку. Під час тренування в мережу вбудовується регуляризаційний компонент, який зменшує різкі стрибки у передбаченому русі між кадрами. Це дозволяє уникати артефактів типу «рваний рух».
4. Часовий дискримінатор. На відміну від класичних GAN, що оцінюють лише окремі кадри, MST-GAN містить дискримінатор, який аналізує послідовності. Це стимулює генератор забезпечити не тільки візуальну якість, але й узгодженість між кадрами, унеможливлюючи мерехтіння чи стрибки об’єктів.

Крім того, MST-GAN використовує перцептивну функцію втрат - порівняння вихідних кадрів не тільки з точки зору піксельної точності, а й у просторі ознак попередньо натренованої моделі (наприклад, VGG). Це дозволяє моделі генерувати зображення, ближчі до людського сприйняття, навіть якщо PSNR дещо нижчий.

Особливості процесу навчання. Для навчання MST-GAN було сформовано мікс із кількох загальноприйнятих відео-наборів: Vimeo-90K-T, Vid4, а також додаткових власноруч зібраних послідовностей високої роздільної здатності. Як вхід подавалися фрагменти з 5-7 послідовних кадрів, що дозволяє моделі навчитися відновлювати деталі в різних контекстах руху.

Процес навчання моделі MST-GAN було організовано поетапно з метою досягнення стабільної збіжності, уникнення типових проблем генеративного навчання та забезпечення високої якості відновленого відео. На початковому етапі здійснювалося попереднє тренування генератора без участі дискримінатора, де оптимізація проводилась виключно за піксельною та перцептивною функціями втрат. Такий підхід дозволив моделі сформувати базову реконструкцію високої роздільності, уникнувши передчасного домінування дискримінатора, яке часто призводить до руйнування процесу навчання на ранніх етапах.

Після цього модель переходила до повноцінного змагального навчання у складі GAN, в якому додатково вводилася темпоральна втрата, що штрафувала за непослідовні зміни між сусідніми кадрами відеоряду. Цей компонент сприяв збереженню часової узгодженості та зменшенню артефактів руху, які можуть виникати при невдалому вирівнюванні або генерації нестабільних текстур.

Фінальна функція втрат моделі була комплексною та адаптивною, включаючи піксельну втрату (L1), перцептивну втрату (LPIPS), часову втрату, а також класичну змагальну компоненту, пов’язану з відповіддю дискримінатора. Ваги кожної з цих складових підбирались емпірично відповідно до мети досягнення балансу між точністю реконструкції, візуальною привабливістю та плавністю руху у відео.

Для запобігання зменшенню різноманіття згенерованих результатів, що часто спостерігається у GAN-моделях (так званий mode collapse), у процес було включено низку регуляризаційних засобів, зокрема додавання шуму до вхідних зображень у дискримінаторі (instance noise) та постійний моніторинг LPIPS на валідаційному наборі. Це дозволило не лише зберігати різноманітність вихідного відео, а й забезпечити його високу суб'єктивну якість та узгодженість.

У результаті навчання вдалося отримати стабільну модель, яка не лише демонструє конкурентні результати на загальновідомих метриках, а й забезпечує перцептивну перевагу: якість відновлення візуально вища, менше артефактів і «запізнень» між кадрами, ніж у деяких традиційних моделей типу BasicVSR чи EDVR.

Оптимізація та прискорення MST-GAN. Нейромережеві моделі для відео-суперрезолюції, зокрема багатокадрові генеративні архітектури, характеризуються значним обчислювальним навантаженням та високими вимогами до ресурсів апаратного забезпечення. Це значно ускладнює їх практичне застосування в реальному часі, особливо на пристроях з обмеженою відеопам'яттю або енергоспоживанням. З метою підвищення ефективності MST-GAN було реалізовано комплекс заходів на рівні структури моделі, точності обчислень та апаратної реалізації [7].

Одним із ключових напрямів стало структурне проріджування (structured pruning), спрямоване на зменшення кількості параметрів моделі без суттєвого зниження якості результату. Аналіз важливості фільтрів у згорткових шарах дозволив ідентифікувати ті компоненти, які мають мінімальний вплив на фінальну реконструкцію. Видалення таких фільтрів та подальше донавчання мережі (fine-tuning) дало змогу зменшити розмір MST-GAN на понад 25 % та скоротити обчислення на 20-30 %. Важливо, що проріджування виконувалося на рівні цілих каналів, що забезпечує реальне прискорення при виконанні на GPU, на відміну від нерегулярних форм [7].

Наступним етапом стала квантизація моделі до нижчої чисельної точності. Зокрема, було застосовано пост-тренувальне квантування ваг до формату INT8 з використанням калібрування на валідаційному наборі. Це дозволило не лише зменшити розмір моделі майже вчетверо, а й активувати спеціалізовані тензорні ядра сучасних GPU (наприклад, NVIDIA Tensor Cores), що значно підвищило швидкість інференсу. Як показали експерименти, приріст продуктивності у режимі INT8 становив понад 2×, а зменшення точності оцінювалося в межах 0.05 дБ PSNR, що практично непомітно при візуальному порівнянні [7].

Для подальшої оптимізації було використано фреймворк TensorRT, який компілює модель у низькорівневий граф з урахуванням апаратної специфіки обраного GPU. Серед застосованих оптимізацій: об'єднання суміжних операцій (layer fusion), розподіл пам'яті з урахуванням залежностей, автоматичне перемикання на FP16/INT8 тощо.

Окремо було впроваджено конвеєрну обробку (pipelining) кадрів: декілька кадрів оброблялися паралельно з різними стадіями генерації, передобробки та запису. Такий підхід дозволив мінімізувати затримки введення/виведення та забезпечити стабільний потік навіть у режимі онлайн-обробки. Експерименти з розміром пакету (batch size) показали, що оптимальним значенням для MST-GAN є обробка 4 кадрів одночасно, що забезпечує максимальне завантаження GPU без переповнення пам'яті.

Крім основних заходів, додатково було реалізовано зменшення дублювання обчислень (наприклад, кешування попередньо розрахованого оптичного потоку для симетричних кадрів), спрощення післяобробки та використання глибинно-роздільних згорток (depthwise separable convolutions) у неголовних шарах генератора. Усі ці модифікації спрямовані на те, щоб зробити MST-GAN придатною для застосування в реальному часі при збереженні високої якості відновленого відео [7].

Дослідження у [1] базувалося на експериментальному порівнянні базової та оптимізованої версії MST-GAN із сучасними моделями відеосуперроздільності. Для об'єктивної оцінки застосовано класичні метрики точності PSNR та SSIM, а також і

перцептивні показники LPIPS та VMAF, які краще корелюють із візуальним сприйняттям. Навчання MST-GAN здійснювалося поетапно: спочатку проводилось попереднє навчання генератора на сумі піксельної L1-втрати та перцептивної LPIPS без участі дискримінатора (стабілізація стартової фази), далі повне змагальне навчання із доданою часовою складовою втрат. Для запобігання збідненню різноманіття результатів застосовувались регуляризаційні прийоми (зокрема instance noise) та постійний моніторинг валідованих метрик. Дані для навчання формувалися як мікс публічних наборів (Vimeo-90K-T, REDS, Vid4) та підготовлених високоякісних фрагментів.

Для обрахунку результатів використовувались такі формули:

$$\text{MSE} = \frac{1}{|\Omega|}\sum_{x\in\Omega}\left(I(x) - \hat{I}(x)\right)^2, \quad (1)$$

де I – еталонне (ground-truth) зображення або кадр відео (HR);

$\hat{I}$ – відновлене/згенероване зображення або кадр (результат моделі).

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right), \quad (2)$$

де MAX – максимально можливе значення пікселя (зазвичай (255) для 8-бітних зображень), MSE середньоквадратична помилка (формула 1).

Формула SSIM визначається через: $u_I u_y$ середні інтенсивності двох порівнюваних зображень, σ_I^2, σ_Y^2 дисперсія даних зображень і σ_{IY} їхня коваріація:

$$SSIM(I,Y) = \frac{2u_I u_y + k_1}{u_I^2 + u_Y^2 + k_1} * \frac{2\sigma_{IY} + k_2}{\sigma_I^2 + \sigma_Y^2 + k_2}, \quad (3)$$

де C_1, C_2 є константами стабілізації.

$$\text{LPIPS}\left(I, \hat{I}\right) = \sum_l w_l \cdot \left\|\phi_l(I) - \phi_l\left(\hat{I}\right)\right\|_2^2, \quad (4)$$

де ϕ_l ознаки (feature maps), отримані з шару (l) попередньо навченої мережі (наприклад, VGG); w_l – ваговий коефіцієнт для шару (l), а $\|\phi_l\|_2^2$квадрат L2-норми (евклідова відстань між ознаками).

$$\text{VMAF} = f(\text{VIF,ADM,TI}), \quad (5)$$

де $f-$ агрегована регресійна модель, навчена за суб'єктивною оцінкою (MOS); VIF – точність візуальної інформації; ADM – метрика втрати деталей; TI – тимчасова інформація кадру.

Окрім кількісних метрик, для наочного підтвердження переваг MST-GAN наведено фрагменти реального відео з YouTube-сцени "Me at the zoo" [8] (послідовність кадрів – 1 кадр в секунду, з 3 по 6 секунду включно), що стало відомим прикладом складного відеоконтенту з природним рухом. На першому колажі зображено результат апскейлу до 4K за допомогою типової VSR-моделі на основі ланцюгової обробки (наприклад, з використанням BasicVSR++ або подібної рекурентної архітектури). Видно, що зображення має помітну втрату дрібних деталей, зниження чіткості обличчя та "змазування" волосся та фону. Особливо це проявляється при поворотах голови та русі тварини на задньому плані.

На другому колажі продемонстровано результат MST-GAN на тій самій сцені. Модель відновлює значно чіткіші деталі обличчя, фактуру одягу та конструкцій позаду, зменшуючи артефакти руху. Також збережено природну плавність переходу між кадрами: рух голови, положення вух слона, а також освітлення не мають спотворень, характерних для неконсистентної генерації. Це свідчить про ефективність поєднання перцептивної та часової втрат, а також наявність в MST-GAN спеціального дискримінатора, що працює із послідовностями, а не окремими кадрами.

Рис. 2. **Приклад сцени з 3 по 6 секунду [8] покращеною ланцюгової обробки**

Загалом MST-GAN демонструє візуально вищу відповідність до очікуваної реальності та кращу стабільність руху, що має ключове значення для практичного застосування в умовах реального відео, зокрема стрімінгу, відеозв'язку та обробки відеоархівів.

Висновки. У роботі представлено комплексне вдосконалення архітектури MST-GAN для задач відео-суперрезолюції. Підвищення якості та стабільності досягнуто шляхом впровадження гібридної функції втрат (піксельна, перцептивна, темпоральна, змагальна) і прийомів стабілізації генеративного навчання. Додатково, для зниження обчислювальної складності було застосовано структурне проріджування, пост-тренувальне квантування ваг до INT8, компіляцію графу обчислень у TensorRT, а також потокову організацію обробки кадрів.

Експериментальні результати підтверджують, що оптимізована MST-GAN досягає рівня якості, співставного або вищого за сучасні VSR-моделі. За метриками PSNR/SSIM модель зберігає близько 98 % точності оригінальної версії, при цьому перевершує аналоги за перцептивними показниками LPIPS і VMAF. Процес роботи було прискорено більш ніж у 4 рази без втрати візуальної якості.

Рис. 3. **Приклад сцени з 3 по 6 секунду [8] покращеної за допомогою MST-GAN моделі**

Окрему увагу приділено якісному порівнянню MST-GAN з конкурентами на складних сценах (зокрема, з набору Vid4). Візуальні приклади демонструють здатність моделі зберігати дрібні деталі, узгодженість між кадрами та відсутність характерного для GAN мерехтіння. Це дозволяє розглядати MST-GAN як придатне рішення для практичного застосування в системах відеопокращення в реальному часі.

СПИСОК ЛІТЕРАТУРИ

1. Maksymiv M., Rak T. “Multi-scale Temporal GAN-based Method for High-resolution and Motion Stable Video Enhancement.” *Radio Electronics, Computer Science, Control*. 2025; 3 (74): 86–94. DOI: https://doi.org/10.15407/jai2023.03.047.

2. Maksymiv M., Rak T. “Methods of video quality-improving.” *Artificial Intelligence*. 2023; 3 (97): 47–62. DOI: https://doi.org/10.15407/jai2023.03.047.

3. Chan K. C., Zhou X., Xu X., Loy C. C. “BasicVSR: The search for essential components in video super-resolution and beyond.” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2021. p. 4947–4956. DOI: https://doi.org/10.1109/CVPR46437.2021.00491.

4. Wang X., Chan K. C., Yu K., Dong C., Loy C. C. “EDVR: Video restoration with enhanced deformable convolutional networks.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 2019. DOI: https://doi.org/10.48550/arXiv.1905.02716.

5. Isobe T., Li J., Jia X., Harada T. “Recurrent Structure-Detail Network for Video Super-Resolution”. *European Conference on Computer Vision (ECCV).* 2020. p. 645–660. DOI: https://doi.org/10.1007/978-3-030-58536-5_38.

6. Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. “Image Quality Assessment: From Error Visibility to Structural Similarity”. *IEEE TIP.* 2004; 13 (4): 600–612.

7. Maksymiv M., Rak T. “Hardware optimization of video quality improvement methods based on deep neural networks”. *Computer Systems and Networks*. 2025; 7 (2) (in press).

8. “Me at the zoo”. *YouTube. Uploaded by jawed,* 2005. – URL: https://www.youtube.com/watch?v=jNQXAC9IVRw.

DOI: https://doi.org/10.15276/ict.02.2025.39
UDC 004.932:004.382:004.056.5

MST-GAN generative model for video enhancement: quality, stability, efficiency

Mykola R. Maksymiv
Postgraduate Student of the Department of Electronic Computing
ORCID: https://orcid.org/0009-0004-4915-6265; mykolamaksymivua@gmail.com.
Lviv Polytechnic National University, 12, S. Bandery St. Lviv, 79000, Ukraine

ABSTRACT

The paper considers the improvement of the MST-GAN (Multi-Scale Temporal GAN) architecture for video super-resolution tasks, aimed at ensuring high visual quality, inter-frame stability and efficient operation in real time. The model combines multi-scale feature alignment, temporal aggregation and generative frame generation using a hybrid loss function.

Particular attention is paid to ensuring the stability of training: pre-training the generator without a discriminator, the gradual inclusion of temporal consistency, the use of perceptual criteria and regularization techniques allowed to avoid typical problems of generative learning, in particular unstable dynamics and loss of diversity. During training, the model is adapted to realistic video scenarios, which allows it to maintain quality even in complex scenes with dynamic objects.

A number of hardware optimizations are also described, including structural thinning of the model, quantization of weights to INT8 format, compilation in TensorRT and organization of frame streaming processing. As a result, MST-GAN achieved significant inference acceleration without noticeable loss of quality.

Qualitative examples on original video data (in particular, a scene with active motion) demonstrate the advantages of the model in preserving textures and smoothness of motion, compared to classical methods of image augmentation. Unlike competing approaches, MST-GAN allows avoiding the characteristic "flicker" and provides a natural transfer of scene dynamics. The obtained results indicate the suitability of MST-GAN for use in practical video enhancement systems, in particular in video streams, monitoring systems, AR applications, where the combination of quality, stability and speed is critically important.

Keywords: high-definition video; video superresolution; generative networks; GPU optimization; perceptual quality; frame-by-frame stability; model performance; structural thinning; temporal discriminator; video quality