# Сравнительный анализ методов сверхразрешения в системе сжатия видеопотока при FPV управлении беспилотными системами

А. И. Козлова, Н. А. Облаков, А. А. Березкин

*Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича*

kozlovap.alyona@yandex.ru, d.5.u.5.d.5.e@gmail.com, berezkin.aa@sut.ru

***Аннотация.*** **Одним из методов минимизации задержек в управлении беспилотными системами от первого лица является сжатие информации. Компрессия видеоданных, передаваемых по сети от беспилотной системы к оператору, позволяет уменьшить время на передачу данных, что в свою очередь способствует снижению общей задержки в системе управления. Однако, сжатие с потерями ухудшает итоговое качество кадров. В данной статье проведен анализ нейросетевых методов и моделей повышения разрешения кадров FPV видеопотока, которые обеспечивают приемлемую скорость и высокий уровень качества выходного изображения.**

***Ключевые слова: сверхразрешение; декодирование; беспилотные летательные аппараты***

## I. Введение

В настоящее время одним из наиболее эффективных методов уменьшения задержек в управлении беспилотными системами от первого лица является применение методов компрессии данных. Подробнее про методы сжатия видеопотока с потерями с сохранением возможности беспилотным системам выполнять поставленные задачи освещается в статье [1]. Особенно важным является сжатие видеопотока с беспилотной системы к станции внешнего пилота (СВП), так как это позволяет существенно сократить время передачи видео и, как следствие, уменьшить задержку рассинхронизации в системе управления.

Стоит отметить, что использование механизмов сжатия с потерями может отрицательно сказаться на качестве кадров видеопотока на стороне СВП [2]. По этой причине, на этапе декодирования должно осуществляться улучшение качества кадров за счет нейросетевых моделей повышения разрешения (*Super Resolution, SR*) [3]. Используемый при этом диффузионный нейросетевой декодер представлен на рис. 1, где на финальном этапе декодирования применяется модель SR.

Для повышения качества кадров видеопотока используются методы сверхразрешения на основе диффузионных моделей и нейросетевых моделей различной конфигурации. На данный момент сравнение подходов сверхразрешения в реальных условиях остается проблемой [4].

Научная статья подготовлена в рамках прикладных научных исследований СПбГУТ, регистрационный номер 1023031600087-9 в ЕГИСУ НИОКТР.

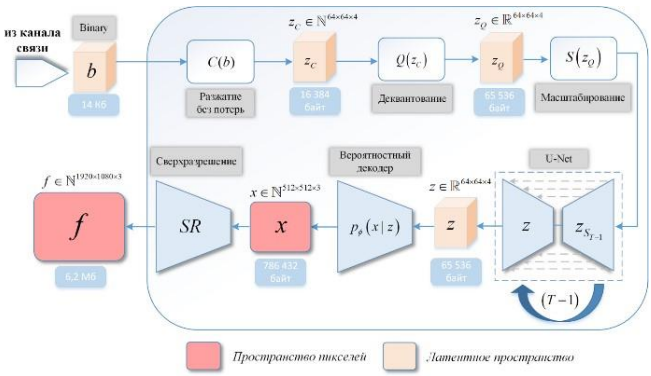


Рис. 1. Схема нейросетевого диффузионного декодера

Данное исследование направлено на сравнительный анализ различных методов сверхразрешения в задаче управления беспилотными системами от первого лица. Рассмотрены такие модели, как модели скрытой латентной диффузии (LDM) [5], сверхразрешение изображений с помощью итеративного уточнения (ISRR) [6], нейросетевая модель трансформера SwinIR [7] и генеративно-состязательная сеть (ESRGAN) [8]. Для сравнительного анализа использованы следующие метрики качества изображений: индекс структурного сходства (SSIM) [9] и пиковот отношение сигнал/шум (PSNR) [10].

## II. Методы сверх-разрешения

### A. Модель скрытой латентной диффузии

Модель скрытой латентной диффузии (*Latent Diffusion Model, LDM*) [11] объединяет диффузионные процессы и глубокие генеративные модели для получения изображений сверхвысокого разрешения.

Пусть $I_{LR}$ – изображения низкого разрешения, которые кодируются в латентное пространство $z$ с помощью функции кодирования $E$, такой, что $z = E(I_{LR})$.

Латентное представление $z$ подвергается процессу диффузии, характеризующемуся последовательностью преобразований $T_t$ где $t$ – индекс итерации. Каждое $T_t$ обновляет $z$, постепенно накапливая информацию из изображения с низким разрешением:

$$z_t = T_t\left(z_t - 1\right),\ z_0 = z\,.$$

После завершения процесса диффузии латентное представление $z_t$ отображается обратно в пространство изображений сверхвысокого разрешения с помощью

генеративной сети *G*, в результате чего получается оценка изображения со сверхвысоким разрешением: $I_{HR} = G(z_t)$, где $I_{HR}$ – изображение высокого разрешения.

Модель обучается путем оптимизации комбинации функций потерь: перцептивной и адверсарной. Перцептивная потеря измеряет перцептивное сходство между $I_{HR}$ и истинным изображением сверхвысокого разрешения, часто вычисляемым с помощью предварительно обученной нейронной сети. Адверсарные потери при сгенерированном изображении стремятся к тому, чтобы оно было неотличимо от реальных изображений сверхвысокого разрешения.

Итерационный процесс уточнения разворачивается следующим образом:

- на каждой *i*-й итерации (шаге) улучшается оценка (*enchance*) изображения высокого разрешения:

$$I_{HR}^{i} = enhance\left(I_{HR}^{i-1}\right).$$

- на этапе обновления "улучшенная" информация интегрируется с исходным изображением низкого разрешения для создания обновленной оценки $I^{i+1}$:

$$I_{HR}^{i+1} = update\left(I_{HR}^{i}, I_{LR}\right).$$

Итерационный процесс продолжается до тех пор, пока не будет достигнут заданный порог итераций, либо выполнятся заранее определенные критерии сходимости.

### *B. Нейросетевой трансформер SwinIR*

SwinIR – это современная модель сверхразрешения, использующая архитектуру *SWIN Transformer*. Изображение разбивается на сектора малого размера, каждое из которых смещено относительно предыдущего. Такой проход по всему изображению позволяет анализировать его более детально. Использование трансформера *SWIN* позволяет более эффективно повышать качества входных изображений.

Процесс сверхразрешения при использовании модели *SwinIR* можно описать выражением:

$$I_{HR} = SwinIR\left(I_{LR}\right).$$

Существует модификация модели *SwinIR*: *SwinIR_large,* которая представляет собой версию *SwinIR* с большим количеством параметров и более глубокими слоями, что позволяет достигать более высокого качества восстановления изображений.

### *C. Усовершенствованная генеративная сеть сверхразрешения ESRGAN*

ESRGAN использует возможности GAN для повышения разрешения изображений и состоит из:

- сеть генератора *G*, которая принимает входное изображение низкого разрешения $I_{LR}$ и стремится сгенерировать соответствующее изображение высокого разрешения $I_{HR}$: $I_{HR}=G(I_{LR})$.
- сеть дискриминатора D, которая призвана отличать реальные изображения высокого разрешения от изображений, сгенерированных генератором *G*. Она оценивает вероятность *D(I)* того, что изображение *I* является реальным.

Общая функция потерь *L(G, D)* при обучении *ESRGAN* можно представить следующим выражением [8]:

$$L(G,D) = \mathrm{E}_{I_{HR}\sim P_{data}(I_{HR})}\left[\log\left(D\left(I_{HR}\right)\right)\right] + \mathrm{E}_{I_{LR}\sim P_{data}(I_{LR})}\left[\log\left(1-D\left(G\left(I_{LR}\right)\right)\right)\right],$$

где *D* – дискриминатор, *G* – генератор, *P* – вероятность того, что изображение оригинально или сгенерировано.

Данная функция потерь побуждает генератор создавать изображения высокого разрешения, которые дискриминатор с трудом отличает от реальных. ESRGAN работает по принципу состязательного обучения, где генератор и дискриминатор находятся в постоянном соперничестве.

### *A. Модель сверхразрешения с помощью итеративного уточнения*

Модель *ISSR* (*Image Super-Resolution via Iterative Refinement*) основывается на итеративных шагах для постепенного улучшения качества изображения с низким разрешением.

Модель использует цикл итераций для постепенного улучшения качества изображения. На каждой итерации применяется операция уточнения, которая корректирует исходное изображение, учитывая информацию о текстуре и деталях изображения. После каждой итерации полученное изображение с высоким разрешением обновляется в соответствии с результатами операции уточнения:

$$I_{HR}^{i+1} = f\left(I_{HR}^{t}, I_{LR}\right),$$

где $I_{LR}$ – входное изображение низкого разрешения, $I_{HR}^{t}$ – изображение высокого разрешения после *t*-ой итерации, *f* – функция уточнения, которая использует текущее изображение с высоким разрешением.

## III. Оценки качества работы моделей

### *A. Пиковое отношение сигнал/шум (PSNR)*

Метрика *PSNR* измеряет искажения изображения, вызванные сжатием, шумом или другими факторами. Вычисление *PSNR* осуществляется путем сравнения средней квадратичной ошибки (*MSE*) между оригинальными и восстановленными изображениями:

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

где *MAX* – максимально возможное значение пикселя изображения, *MSE* – средняя квадратичная ошибка между исходным и восстановленным изображениями, которая рассчитывается в соответствии с выражением:

$$MSE = \frac{1}{m\times n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(I_o(i,j) - I_r(i,j)\right)^2,$$

где $m\times n$ – размер изображений, а $I_o(i,j)$ и $I_r(i,j)$ – значения пикселей в позиции $(i,j)$ на оригинальном и реконструированном изображениях, соответственно.

### B. *Индекс структурного сходства (SSIM)*

Метрика *SSIM* измеряет степень схожести между изображениями, охватывая яркость, контрастность и структурные аспекты. Комбинированный индекс SSIM рассчитывается на основе этих трех компонентов и определяется следующим образом:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

где $x$ и $y$ – эталонное и искаженное изображения, соответственно, $\mu_x$ и $\mu_y$ – средняя интенсивность пикселей изображений $x$ и $y$, $\sigma_x$ и $\sigma_y$ – стандартные отклонения интенсивности пикселей изображений $x$ и $y$, $\sigma_{xy}$ – перекрестная ковариация интенсивностей пикселей между изображениями $x$ и $y$, $c_1$ и $c_2$ – небольшие значения, чтобы избежать деления на ноль.

## IV. Эксперимент

Для проведения эксперимента были собраны кадры *FullHD* видеопотока, пример которых представлен на рис. 2. Сжатие исходных кадров проводилось с помощью нейросетевого диффузионного кодера [1], а декодирование с помощью нейросетевого диффузионного декодера [2]. В результате декодирования на вход блока SR подавался кадр размера 512×512 пикселей.

Далее к декодированным кадрам применялись методы сверхразрешения для восстановления первоначального размера. Оценка качества восстановления проводилась на основе метрик *PSNR* и *SSIM*. Результаты представлены на рис. 3 и 4.

В результате эксперимента около 80 % исследуемых моделей сверхразрешения получили результат по метрике *SSIM* свыше 0.9 (рис. 3), что говорит о высокой схожести оригинального и восстановленного изображений. Наибольшее значение *SSIM* равное 0.94 показала усовершенствованная модель *SwinIR_large*.

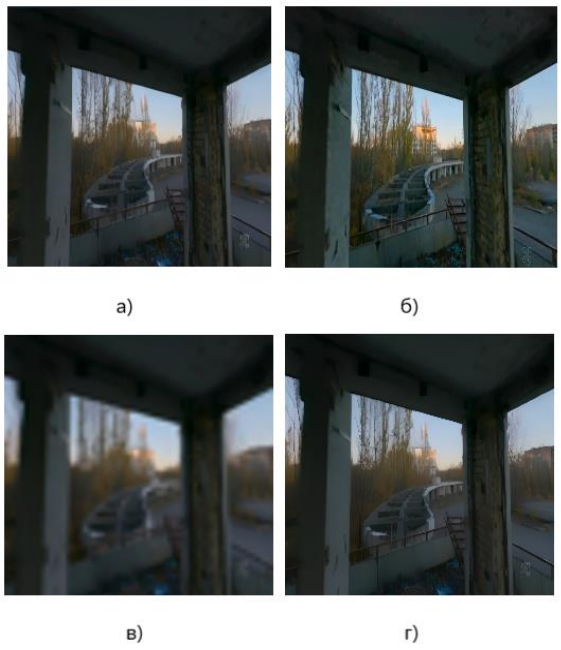


Рис. 2. Примеры работы моделей сверхразрешения: а) оригинальное изображение; б) LDM; в) ISSR; г) SwinIR

ТАБЛИЦА I.

| Название модели | SSIM | PSNR |
|---|---|---|
| LDM | 0.928514 | 35.551792 |
| ISSR | 0.820297 | 33.982809 |
| SwinIR | 0.928649 | 37.379892 |
| SwinIR_large | 0.940866 | 37.571892 |
| ESRGAN | 0.925347 | 37.123174 |

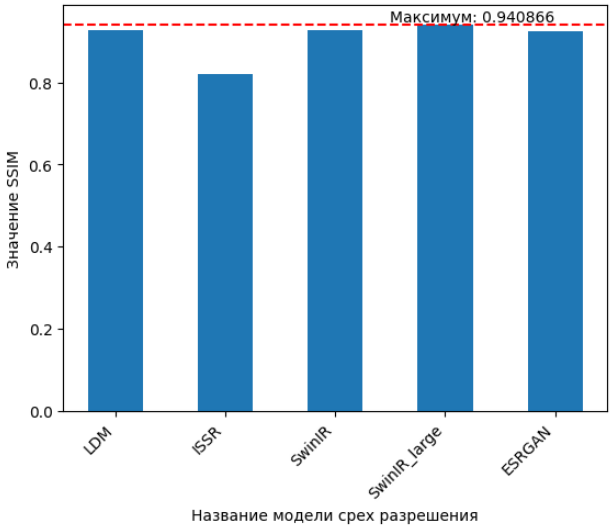


Рис. 3. Результаты оценки качества по метрике SSIM

Высокое значение *PSNR* указывает на то, что потери информации в изображении с повышенной четкостью очень незначительны. Лидирующие позиции по данной метрике заняли модели семейства SwinIR (рис. 4).

Также были проведены исследования скорости работы моделей (рис. 5). Наибольшее время работы имеет модель *ISSR*. Самыми быстрыми оказались модели семейства *SwinIR*, показав результат менее 10 секунд.

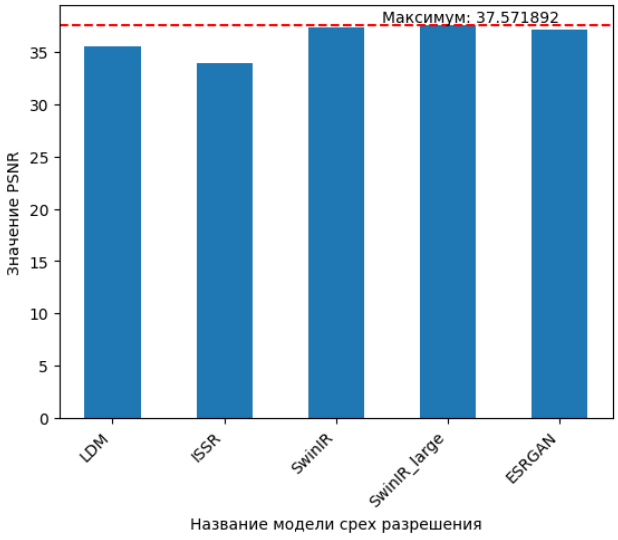


Рис. 4. Результаты оценке качества по метрике PSNR

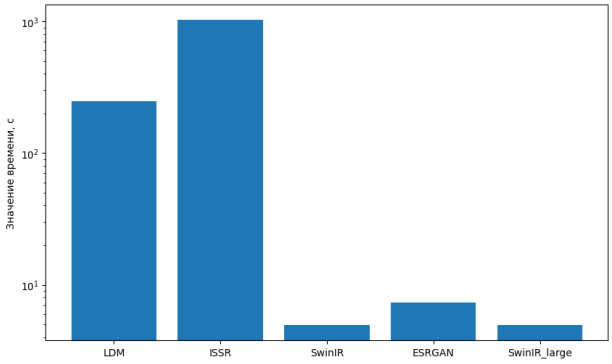


Рис. 5. Длительность работы моделей сверх разрешения

## V. Заключение

В результате эксперимента было выявлено, что модели семейства *SwinIR* показывают высокое качество восстановления кадров видеопотока, параллельно с этим обладая большим быстродействием. Модель *LDM*, объединяющая диффузионные процессы и глубокие генеративные модели для получения изображений сверхвысокого разрешения, показала хорошие результаты. Однако она уступает по скорости работы моделям *SwinIR*.

Модель *ISSR*, в основе которой лежит итеративное уточнение, показала наихудшие результаты как по оценкам качества, так и по скорости своей работы.

## Список литературы

[1] Березкин А.А., Вивчарь Р.М., Слепнев А.В., Киричек Р.В., Захаров А.А. Метод сжатия видеопотока при управлении беспилотными системами в гибридных орбитально-наземных сетях связи // Электросвязь. 2023. №10. С. 48-56.

[2] Березкин А.А., Вивчарь Р.М., Киричек Р.В., Захаров А.А. Метод декомпрессии FPV-видеопотока от беспилотных систем на основе латентной диффузионной нейросетевой модели // Электросвязь. 2024. №1. С. 42-53.

[3] Sacoto-Martins R. et al. Multi-purpose low latency streaming using unmanned aerial vehicles // 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP). IEEE. 2020. P. 1-6.

[4] Maral B.C. Single Image Super-Resolution Methods: A Survey // arXiv preprint arXiv:2202.11763. 2022.

[5] Rombach R. et al. High-resolution image synthesis with latent diffusion models // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. P. 10684-10695.

[6] Saharia C. Image super-resolution via iterative refinement / C. Saharia, J. Ho, W. Chan et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. Vol. 45, Issue 4. P. 4713-4726.

[7] Saharia C. et al. Image super-resolution via iterative refinement //IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. Т. 45. №. 4. P. 4713-4726.

[8] Wang Xintao, et al. Esrgan: Enhanced super-resolution generative adversarial networks // Proceedings of the European conference on computer vision (ECCV) workshops. 2018.

[9] Sara U., Akter M., Uddin M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR – a comparative study // Journal of Computer and Communications. 2019. Т. 7. №. 3. P. 8-18.

[10] Ashry I. Normalized differential method for improving the signal- to-noise ratio of a distributed acoustic sensor / I. Ashry, Y. Mao, M.S. Alias et al. // Applied Optics. 2019. Vol. 58, Issue 18. P. 4933-4938.

[11] Zhou D. et al. Magicvideo: Efficient video generation with latent diffusion models // arXiv preprint arXiv:2211.11018. 2022.

[12] Croitoru F.A., Hondru V., Ionescu R.T., Shah M. Diffusion models in vision: A survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023.