Анализ моделей предсказания кадров FPV-видеопотока в каналах информационного обмена беспилотных систем

В. Г. Гончаров

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

pottergp@mail.ru

А. А. Березкин

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

berezkin.aa@sut.ru

Аннотация. **В статье проведен анализ и сравнение моделей предсказания кадров для решения задачи компенсации FPS при потере кадров видеопотока от беспилотной системы к станции внешнего пилота при управлении от первого лица. Рассмотрены различные архитектуры и методы, ориентированные на предсказание будущих кадров с целью адаптивного управления интенсивностью видеопотока на стороне станции внешнего пилота. Оцениваются точность и качество предсказания моделей. Полученные в ходе исследования результаты предоставляют научные и практические рекомендации для борьбы с проблемой потери кадров в FPV-видеопотоке, что способствует плавности восприятия полетной обстановки и к своевременным управленческим воздействиям.**

Ключевые слова: беспилотные системы, управление от первого лица, предсказание видео, нейронные сети

I. Введение

В условиях растущего объема передаваемого трафика, особенно в гибридных орбитально-наземных гибридных сетях связи (ГОНСС), существует проблема потерь данных при передаче на большие расстояния, вызванная как техническими ограничениями, так и особенностями самой системы. В связи с этим, восстановление потерянных данных является важной составляющей обеспечения качества передачи.

ГОНСС, характеризующиеся высокими задержками передачи данных в спутниковых сегментах, представляют собой технически сложную среду, где неустойчивость каналов информационного обмена сильно влияет на непрерывность видеопотока. В связи с этим, передача видеопотока от беспилотных летательных аппаратов (БПЛА) приобретает особо важное значение в системах FPV-управления (управление от первого лица), где внешним пилотам необходимо получать высококачественный видеопоток в режиме реального времени.

Решение проблемы можно найти в применении различных нейросетевых моделей для восстановления кадров. Нейросетевые методы показали себя эффективными в решении задач восстановления изображений, и их применение в данной задаче перспективно для улучшения качества видеопотока. Предсказательные модели могут адаптироваться к различным условиям передачи данных и обеспечивать достаточно точное и надежное восстановление кадров, что способствует улучшению пользовательского опыта и эффективности системы в целом.

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

II. Различные методы генерации видеокадров

Количество методов предсказания кадров растет с каждым годом. В [1] исследованы решения, основанные на методе оптического и глубокого воксельного потоков. В данной главе рассмотрены четыре ключевых метода: интерполяция, генеративные архитектуры (GAN) [2], модели на основе долгосрочной краткосрочной памяти (LSTM) [3] и диффузионные модели (LDM) [4].

A. Интерполяция кадров

Интерполяция кадров – метод, призванный создавать промежуточные изображения между двумя имеющимися кадрами. Он основывается на взвешенном среднем значении двух кадров, где веса определяются параметром интерполяции, изменяющимся от 0 до 1. Математически, это представляется формулой:

$$I_t = (1-\alpha)\cdot I_{t-1} + \alpha \cdot I_{t+1},$$

где t – текущий момент времени, I_t – текущий кадр, I_{t-1} – предыдущий кадр, I_{t+1} – следующий кадр, α – коэффициент интерполяции.

Однако в контексте передачи видеопотока существует фундаментальная проблема. Для эффективной интерполяции кадра I_t, требуется доступ к «кадру из будущего» I_{t+1}. Это становится недопустимым в условиях реального времени и ограниченной пропускной способности ГОНСС, что делает метод интерполяции неэффективным и неприменимым в данных условиях.

Таким образом, интерполяция не соответствует требованиям эффективной передачи видеопотока в условиях динамических изменений и ограниченных ресурсов сети.

B. GAN (Generative Adversarial Networks)

Генеративно-состязательные сети (GAN) представляют метод генерации, который включает две основные компоненты: генератор и дискриминатор. Генератор стремится создать реалистичные данные, в то время как дискриминатор направлен на различие между реальными и сгенерированными данными. Процесс обучения GAN оформляется в виде состязания между этими двумя сетями.

При передаче видеопотока через ГОНСС, GAN может столкнуться с трудностями из-за нестабильности

обучения и необходимости большого объема данных для достижения качественных результатов. Однако, если успешно применено, GAN может создавать реалистичные видеопоследовательности, сохраняя при этом динамичность и сложность сцен.

Таким образом, GAN представляет мощный инструмент для генерации видеопотока, однако его применение в реальных условиях требует тщательной настройки и учета особенностей передачи данных через ограниченные сети связи.

C. *LSTM (Long Short-Term Memory)*

Модели на основе долгосрочной краткосрочной памяти (LSTM) представляют собой разновидность рекуррентных нейронных сетей (*recurrent neural networks*, RNN), разработанных для работы с последовательными данными. В отличие от стандартных RNN [5], которые сталкиваются с проблемой затухания градиентов во время обучения на долгих последовательностях, LSTM обладают способностью сохранять информацию в течение длительных периодов времени.

В задаче передачи видеопотока, LSTM могут быть использованы для предсказания будущих кадров на основе предыдущих. Они могут улавливать долгосрочные зависимости в видеопоследовательностях, эффективно моделировать динамические изменения в сценах и обеспечивать стабильность воспроизведения что делает их привлекательным вариантом для решения задач передачи видеопотока в реальном времени через ограниченные сети связи.

D. *Диффузионные модели*

Диффузионные модели представляют собой подход, основанный на концепции диффузии, где пиксели в изображении распространяют информацию на близлежащие пиксели с течением времени. Эти модели стремятся воссоздать динамические изменения в видеопоследовательности, предсказывая значения пикселей на основе их окружения и предыдущих кадров.

Диффузионные модели позволяют учитывать локальные зависимости между пикселями, что делает их эффективными для моделирования динамических изменений в видеопоследовательностях. Они могут быть особенно полезны в сценариях, где объекты могут двигаться с непредсказуемой траекторией.

В контексте передачи видеопотока через ГОНСС, диффузионные модели могут предоставлять адаптивные решения для предсказания кадров с учетом динамических условий и ограничений сети, поэтому диффузионные модели представляют собой перспективный подход для решения задач передачи видеопотока в реальном времени, обеспечивая баланс между точностью предсказаний и эффективностью использования ресурсов сети.

ТАБЛИЦА I. СРАВНЕНИЕ МЕТОД ПРЕДСКАЗАНИЯ КАДРОВ

Метод / Параметр	Интерполяция	GAN	LSTM	Диффузия
Качество предсказания	-	+	+	+
Легкость обучения	+	-	+	-
Адаптивность	-	+	+	+
Вычислительная эффективность	+	-	-	-

Несмотря на свою простоту, интерполяция кадров ограничена в способности воспроизводить динамические изменения и поддерживать высокое качество предсказания кадров в ГОНСС. GAN, несмотря на потенциально высокое качество, требуют тщательного обучения и большой вычислительной мощности. LSTM и диффузионные модели, с другой стороны, представляют собой перспективные подходы.

LSTM обеспечивают высокое качество воспроизведения и эффективно снижают объем передаваемых данных. Диффузионные модели предоставляют адаптивные решения, особенно в условиях изменяющихся сцен. В ходе сравнения, использование LSTM и диффузионных моделей представляется более перспективным для решения задач передачи видеопотока в реальном времени в беспилотных системах и ГОНСС.

На основе выводов было решено взять LSTM и диффузионную модель как основу для дальнейшего сравнения. Для более глубокого анализа были выбраны две модели: *SwinLSTM* (*Shifted Windows Interaction Network LSTM*) [6] и *VDT* (*Video Diffusion Transformer*) [8].

III. МОДЕЛЬ SWINLSTM

SwinLSTM – это *RNN*, сочетающая в себе преимущества *Swin Transformer* [7] и упрощенного LSTM. Она была предложена в 2023 году и показала лучшие результаты по сравнению с другими методами на нескольких наборах данных, таких как Moving MNIST, Human3.6m, TaxiBJ и KTH.

На вход сети подается последовательность изображений, каждое из которых разбивается на непересекающиеся области (патчи) фиксированного размера $P \times P$ (где P обычно равно 2 или 4). Размерность признаков каждого патча составляет $C \times P^2$, где C – количество каналов в изображении. Полученные патчи преобразуются слоем *Patch Embedding*, который осуществляет линейное отображение исходных признаков патчей в пространство произвольной размерности.

SwinLSTM-B. Подготовленные патчи, вместе с скрытым состоянием H_{t-1} и состоянием ячейки C_{t-1} с предыдущего временного шага, поступают на слой *SwinLSTM*. *SwinLSTM* вычисляет новое скрытое состояние H_t и состояние ячейки C_t для текущего шага. Скрытое состояние H_t дублируется: одна копия направляется на слой реконструкции, другая – вместе с C_t передается на слой *SwinLSTM* следующего временного шага (рис. 1).

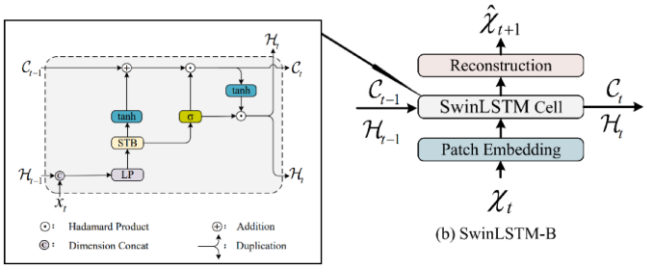


Рис. 1. Структура ячейки модели SwinLSTM-B. STB – Swin Transformer Block, LP - Linear Projection

SwinLSTM-D. Имеет аналогичную структуру *SwinLSTM-B*, однако отличается наличием дополнительных слоев:

- Patch Merging – отвечает за понижение разрешения (downsampling) патчей.
- Patch Expanding – отвечает за повышение разрешения (upsampling) патчей.

Данные слои позволяют модели обрабатывать изображения с различным разрешением (рис. 2).

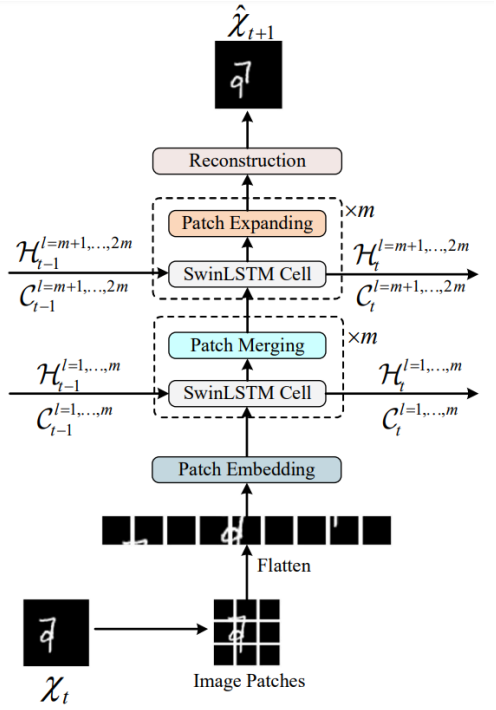


Рис. 2. Структура ячейки модели SwinLSTM-D

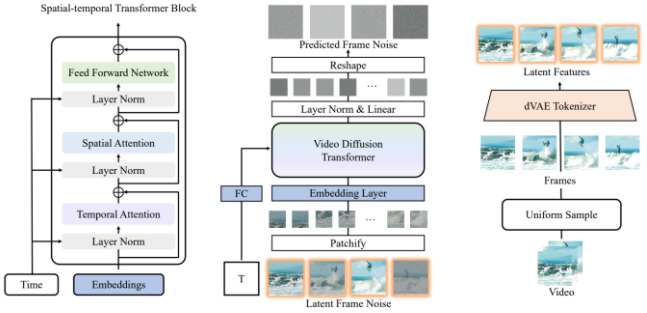
IV. Модель VDT

VDT – это модель, основанная на методе диффузии с масками, предназначенная для генерации, предсказания, интерполяции и управления стилем видео. Она была разработана в 2023 году и является развитием модели MCVD.

VDT проецирует видео в латентное пространство с помощью предварительно обученного токенизатора *VAE* (*Variational AutoEncoder*) из *LDM* (*Latent Diffusion Model*). Данный подход существенно сокращает размерность входных и выходных данных, представляя их как скрытые признаки/шум ($F \times H/8 \times W/8 \times C$), где F – количество кадров в видео, $H/8$ и $W/8$ – пониженное разрешение по высоте и ширине соответственно (коэффициент понижения – 8) и C – размерность скрытого признака.

Подобно модели ViT (Visual Transformer), VDT разделяет представление скрытых признаков на неперекрывающиеся патчи фиксированного размера N × N в пространственной области. Для явного обучения как пространственным, так и временным взаимосвязям к каждому патчу добавляются пространственные и временные позиционные кодировки (рис. 3).

Принцип работы VDT: сеть U-Net получает зашумленные кадры и условную информацию (прошлые/будущие кадры, уровень шума). Residual block обрабатывает эту информацию. Сеть предсказывает шум в кадрах. Шум вычитается, восстанавливая исходное изображение.



(a) блок VDT (b) Video Diffusion Transformer (c) Проецирование в латентное пространство

Рис. 3. Архитектура модели VDT

V. Эксперимент

В качестве моделей для сравнения были выбраны предобученные модели *SwinLSTM-D* и *VDT*. В качестве набора данных был выбран *Moving MNIST* [9]. Основная цель эксперимента – определить модель, которая максимально точно предсказывает движения цифр. Для определения точности использовались технические метрики оценки качества изображений [10] *PSNR* (*Peak Signal-to-Noise Ratio*) и *SSIM* (*Structural Similarity Index Measure*). Итоговые результаты представлены в табл. II и на рис. 4 и 5, где k – количество входных кадров, *pred* – количество предсказанных кадров на основе k.

ТАБЛИЦА II. Итоговые метрики

Модель	*k*	*pred*	PSNR↑	SSIM↑
SwinLSTM	10	**10**	**38.80**	**0.962**
VDT	**8**	4	30.12	0.905

Рис. 4. Результат работы SwinLSTM: 1 строка – входные данные, вторая строка – ожидаемые предсказания, третья строка – предсказания модели

Рис. 5. Результат работы VDT: первая строка – входные данные, вторая строка – предсказания модели

Из табл. II следует, что наиболее эффективной в задаче предсказания видео является *SwinLSTM*, основанная на архитектуре *RNN*. Она предсказывает 10 кадров на основе 10 предыдущих, что позволяет восстанавливать как минимум каждый второй кадр.

VI. Заключение

Нейросетевые модели являются наиболее перспективным подходом в решении проблемы

предсказания кадров в FPV-видеопотоке в контексте управления беспилотными системами в ГОНСС. Они показывают достаточную точность предсказания при достаточно небольшом количестве входных данных, что делает их эффективным инструментом для обеспечения стабильности трансляции видеопотока на стороне внешнего пилота.

В частности, модель *SwinLSTM* на основе механизма долгосрочной краткосрочной памяти показала лучший результат, что свидетельствует о том, что диффузионные модели еще недостаточно адаптированы для решения задачи предсказания кадров.

Список литературы

[1] Березкин А.А., Савелов Д.Ю., Суходоева А.В., Туманов И.А., Киричек Р.В. Исследование нейросетевых моделей предсказания видеопотока при управлении беспилотными системами от первого лица // Труды Научно-исследовательского института радио. 2023. № 3-4. С. 40-56.

[2] Wang K. et al. Generative adversarial networks: introduction and outlook // IEEE/CAA Journal of Automatica Sinica. 2017. Т. 4. №. 4. С. 588-598.

[3] Schmidhuber J. et al. Long short-term memory // Neural Comput. 1997. Т. 9. №. 8. С. 1735-1780.

[4] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models // Advances in neural information processing systems. 2020. Т. 33. С. 6840-6851.

[5] Schmidt R. M. Recurrent neural networks (rnns): A gentle introduction and overview // arXiv preprint arXiv:1912.05911. 2019.

[6] Tang S. et al. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. С. 13470-13479.

[7] Liu Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows // Proceedings of the IEEE/CVF international conference on computer vision. 2021. С. 10012-10022.

[8] Lu H. et al. Vdt: General-purpose video diffusion transformers via mask modeling // The Twelfth International Conference on Learning Representations. 2023.

[9] Srivastava N., Mansimov E., Salakhudinov R. Unsupervised learning of video representations using lstms // International conference on machine learning (PMLR). 2015. С. 843-852.

[10] Березкин А.А., Вивчарь Р.М., Киричек Р.В. Модель системы управления мобильными роботизированными комплексами различного назначения // Электросвязь. 2023. № 8. С. 12-18.