

УДК 621.397.33:004.932.72

Улучшение интерполяции кадров видео с помощью адаптивных ядер

Али Салман

Спасенов Алексей Юрьевич (научный руководитель)

МГТУ им. Н.Э. Баумана, Москва, Россия

Интерполяция кадров видео (VFI) генерирует промежуточные кадры между последовательными кадрами видео, обычно для замедленного воспроизведения или сжатия. Традиционные методы, такие как оптический поток или свертки с фиксированным ядром, плохо справляются с окклюзией и большими движениями. Современные решения используют глубокое обучение, сочетая оценку потока и адаптивные ядра. ST-MFNet — это двухступенчатая модель для уточнения потока и синтеза адаптивных ядер. Изучено влияние аппаратных ограничений и размера данных на качество интерполяции. Адаптивные ядра повысили PSNR/SSIM на 2,28 дБ/0,037 для динамических текстур (VFITex), но оригинальная модель лучше работает с большими наборами данных и мощными GPU. Это подчеркивает зависимость VFI от разнообразия данных и вычислительной мощности.

Ключевые слова: VFI, оптический поток, основанные на ядре, гибридные методы интерполяции, ST-MFNet, PSNR, SSIM

Введение. Интерполяция кадров видео (VFI) генерирует промежуточные кадры для замедления видео, повышения частоты кадров и компенсации потерь при передаче [1]. Традиционные методы использовали оптический поток, что приводило к артефактам из-за ошибок оценки [2]. Современные подходы на основе глубокого обучения включают оптический поток, пространственно-временные свертки и гибридные методы. В статье предлагается модификация архитектуры ST-MFNet с адаптивными ядрами, оптимизированными под ограничения ресурсов (малые наборы данных, малая память GPU), что позволяет эффективно обрабатывать сложные текстуры и движения, что подтверждается ростом PSNR и SSIM на небольших наборах данных.

Предыдущее исследование. В связи с растущим спросом на VFI и ее применением для сложных задач обработки движения, рассмотрим последние достижения в этой области, включая динамический синтез текстур, ставшие основой предлагаемого подхода. Существующие методы VFI можно разделить на две основные категории.

Методы, основанные на оптическом потоке, вычисляют соответствие пикселей между кадрами, интерполируя их по траекториям движения. Традиционные алгоритмы (вариативные, блочное сопоставление) сталкиваются с проблемами окклюзий, крупных движений и изменений яркости [3]. Современные подходы используют глубокое обучение, например DVF с трилиней-

ной выборкой 3D-потока или модель с масками видимости на основе U-Net [4]. Наиболее продвинутые методы, такие как DAIN [1], объединяют предобученные сети оптического потока с картами глубины и признаками ResNet для эффективной обработки окклюзий.

Методы, основанные на ядрах, формируют промежуточные кадры с помощью пространственно-изменяющихся сверток, неявно моделируя движение и окклюзии. Ранние подходы, такие как AdaConv и SepConv [2], имели ограничения по размеру ядра, что снижало эффективность при больших смещениях. AdaCoF [5] предложили деформируемую выборку и гибкие перемещения точек, улучшив качество, но увеличив вычислительную сложность.

Гибридные методы, сочетают оптический поток с адаптивными ядрами для точной интерполяции кадров видео. MEMC-Net [6] и DAIN [1] грубо выравнивают кадры с помощью потока, затем ядрами уточняют детали, устраняя артефакты и обрабатывая окклюзии. Подходы Zooming Slow-Mo [7] дополнительно используют семантический контекст и суперразрешение. Современные трансформерные модели, такие как VFI-Transformer [8], используют механизмы внимания для улучшения адаптивности свертки. Несмотря на преимущества, такие методы требуют значительных вычислительных ресурсов, и поиск оптимального баланса между компонентами остается актуальной задачей.

Адаптивные ядра в кадровой интерполяции. Адаптивные — это фильтры, динамически меняющие размер, форму и веса в зависимости от входных данных. Малые ядра позволяют точно передать движение и мелкие текстуры, а крупные — эффективно подавляют шум в статичных зонах. Такой подход, в отличие от фиксированных сверток, обеспечивает гибкость и повышает качество интерполяции. Первоначальная концепция адаптивных сверток была предложена Никлаусом и др. [2] и развита в модели Super SloMo [4].

ST-MFNet (Пространственно-временная многопоточная сеть для интерполяции кадров видео). ST-MFNet [9] — это двухэтапная модель для интерполяции кадров видео, состоящая из этапов оценки движения и синтеза кадров.

Этап I (мультиинтерполяция и двунаправленный поток). Сеть MIFNet оценивает оптические потоки между входными кадрами, создавая двунаправленные потоки движения. Затем сеть BLFNet уточняет эти потоки для более точного выравнивания кадров. После этого выполняется деформация исходных кадров по полученным потокам, а результаты объединяются модулем многомасштабного слияния.

Этап II (синтез кадра). Полученные промежуточные кадры передаются в сеть TENet, которая улучшает их временную и визуальную согласованность. На выходе TENet формируется плавная последовательность интерполированных кадров с высоким качеством, подтверждающая эффективность архитектуры ST-MFNet. На рисунке 1 представлены архитектурный обзор сети ST-MFNet.

В рамках исходной реализации модель обучалась с размером партии до 4 и размером ядра до 5 на GPU NVIDIA P100 и RTX 3090. Обучение проводилось на наборах Vimeo-90k [10] (91701 септуплет, 448×256) и BVI-DVC [11], содержащих различные текстуры и движения в разрешениях от 270p до 2160p. Дополнительно было создано 100 000 септуплетов путем случайных трансформаций исходных данных. Для оценки использовались наборы квинтуплетов кадров из UCF-101 [12], DAVIS [13] и VFITex. Качество оценивалось с помощью метрик PSNR и SSIM [14], результаты приведены в табл. 1.

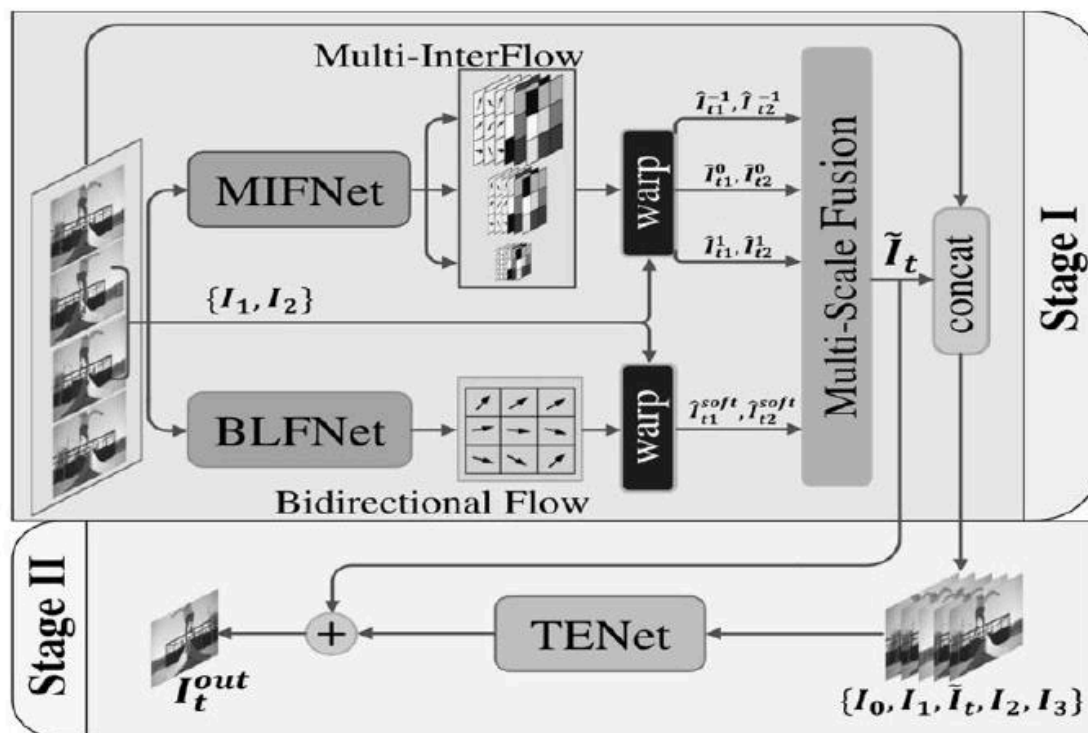


Рис. 1. Обзор архитектуры ST-MFNet

Таблица 1

Результат оценки оригинальной модели

	UCF101	DAVIS	VFITex
(PSNR/SSIM)	33.384/0.970	28.287/0.895	29.175/0.929

Обучение по индивидуальному эталону. Для учета ограничений по памяти (RTX 3050, 6 ГБ) выбран компактный набор данных Media Lab Video Group МГУ (объем ~6 ГБ) [15], подходящий для экспериментов в условиях ограниченных ресурсов. В отличие от объемных наборов Vimeo-90k и BVI-DVC, он обеспечил достаточное разнообразие при приемлемой нагрузке.

Подготовка данных. Видео в формате Full HD были масштабированы до 480×270, затем сжаты с разными битрейтами (100–6000 кбит/с) с помощью

пяти кодеков. Далее кадры разбивались на группы по пять для подачи на вход модели, где третий кадр использовался как целевой. Оценка производилась на высокобитрейтных версиях всех кодеков (рис. 2).

Аппаратные ограничения и модификации параметров из-за нехватки памяти GPU (RTX 3050 с 6 ГБ оперативной памяти). Решили, что использование исходных параметров (размер партии — 4, размер ядра — 5) не подходит. Вместо этого уменьшили параметры до Batch size-2, Kernel size-3.

Результаты оценки. В таблице 2 представлены результаты PSNR/SSIM модели с измененными параметрами на нашем наборе данных.

Анализ. Обзор производительности оригинальной модели: как видно из табл. 2, изменения привели к очевидному падению производительности. Наиболее важные выводы таковы.

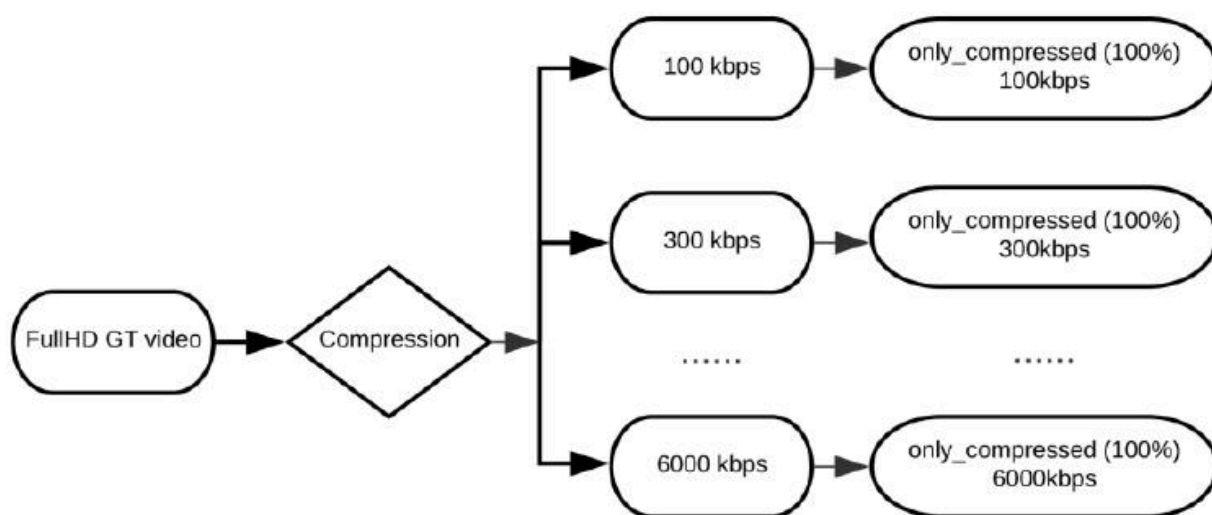


Рис. 1. Рабочий процесс подготовки набора данных

Таблица 2

Результат обучения в модели с измененными параметрами

	UCF101	DAVIS	VFITex
(PSNR/SSIM)	31.034/ 0.947	25.651/0.848	26.119/0.881

Снижение PSNR и SSIM: на UCF101 (-2,35 дБ), DAVIS (-2,64 дБ) и VFITex (-3,06 дБ) из-за уменьшения размера ядра (5→3), что ухудшило детализацию движения. Уменьшение размера партии (4→2) и объема данных (80→6 ГБ) снизило стабильность обучения и обобщение модели. Влияние на наборы: VFITex (сложные текстуры/движения) и DAVIS (высокоинтенсивные сцены) показали наибольшее падение. UCF101, с более структурированными движениями, оказался устойчивее. Причины: малые ядра и данные ограничили способность модели к точной интерполяции, особенно в динамичных сценах.

Ключевое усовершенствование: адаптивные размеры ядра. Авторы немного модифицировали ST-MFNet, чтобы обеспечить улучшенную обработку многомасштабных и многопоточных потоков с помощью адаптивных ядер. Это позволило внедрить адаптивные ядра для прогнозирования многомасштабных потоков. Когда увеличиваем рецептивное поле, наблюдаем его влияние на оценку ядра. Это полезно для точной оценки адаптивных ядер, так как помогает модели изучать крупномасштабные особенности. Чтобы повысить эффективность нашей модели, необходимо адаптивно изменять размеры ядра в зависимости от содержания сайта. Авторы хотят использовать более крупные ядра в плоских областях (например, на статичных фонах) и более мелкие ядра в областях с оживленным движением (например, на быстро движущихся объектах). Хотя фактический размер ядра остается неизменным, эффективный размер ядра изменяется, поскольку распределение веса начинает меняться, а более крупное рецептивное поле позволяет лучше оценивать пространственно меняющиеся веса. Как отмечают Никлаус и др., это изменение согласуется с предложением Enhanced Multi-Scale and Multi-Flow Handling with Adaptive Kernels [2], в которой показано, что адаптивные ядра могут улучшить обработку различных типов движений за счет локализации свертки.

Результаты оценки. Эти значения PSNR/SSIM для изменения модели в нашем наборе данных приведены в табл. 3.

Таблица 3

Результаты обучения в модифицированной модели

	UCF101	DAVIS	VFITex
(PSNR/SSIM)	32.76/0.968	27.324/ 0.879	28.402/ 0.918

Основные наблюдения и выводы о производительности. Улучшение с адаптивными ядрами: модель с адаптивными ядрами показала прирост PSNR (UCF101: +1,73 дБ, DAVIS: +1,67 дБ, VFITex: +2,28 дБ) и SSIM, особенно на динамичных (DAVIS) и текстурных (VFITex) данных. Сравнение с оригиналом: оригинал превосходит адаптивную модель на ~0,62–0,96 дБ из-за большего объема данных (80 ГБ и 6 ГБ) и ресурсов GPU. Однако адаптивная модель демонстрирует сопоставимую эффективность в условиях ограничений.

Заключение. Использование адаптивных ядер в ST-MFNet значительно улучшает интерполяцию видеок кадров, особенно при сложных движениях и текстурах, что подтверждается ростом PSNR и SSIM. Несмотря на ограниченные ресурсы и малый объем данных, модифицированная модель показала результаты, близкие к оригинальной. Это подтверждает эффективность адаптивных ядер, хотя для достижения максимального качества по-прежнему требуются большие датасеты и мощное оборудование.

Заглядывая в будущее, важно оптимизировать методы адаптивных ядер для эффективной работы при ограниченных ресурсах. Также планируется

модифицировать второй этап ST-MFNet (TENet), добавив функцию суперразрешения. Это позволит объединить интерполяцию и апсемплинг в единой архитектуре, что сделает модель более универсальным решением для повышения качества видео.

Литература

- [1] Bao W., Lai W.-S., Zhang X., Gao Z., Yang M.-H. Depth-aware video frame interpolation. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). Long Beach, USA, 2019, pp. 3698–3707.
- [2] Niklaus S., Mai L., Liu F. Video frame interpolation via adaptive separable convolution. Proc. IEEE Int. Conf. Comput. Vis (ICCV). Venice, Italy, 2017, pp. 261–270.
- [3] Cheng X., Chen Z. Video frame interpolation via deformable separable convolution. Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 10607–10614.
- [4] Jiang H., Sun D., Jampani V., Yang M.-H., Learned-Miller E., Kautz J. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). Salt Lake City, USA, 2018, pp. 9000–9008.
- [5] Bao W., Lai W.-S., Zhang X., Gao Z., Yang M.-H. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. CoRR, 2018, vol. abs/1810.08768, arXiv: 1810.08768v2.
- [6] Xiang X., Tian Y., Zhang Y., Fu Y., Allebach J.P., Xu C. Zooming Slow-Mo: Fast and accurate one-stage space-time video super-resolution. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). Seattle, USA, 2020, pp. 3370–3379.
- [7] Lee H., Kim T., Chung T.-Y., Pak D., Ban Y., Lee S. AdaCoF: Adaptive collaboration of flows for video frame interpolation. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). Seattle, USA, 2020, pp. 5313–5322.
- [8] Lu L., Wu R., Lin H., Lu J., Jia J. Video frame interpolation with transformer. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). New Orleans, USA, 2022, pp. 3532–3542.
- [9] Danier D., Zhang F., Bull D. ST-MFNet: A spatio-temporal multi-flow network for frame interpolation. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR). New Orleans, USA, 2022, pp. 3521–3531.
- [10] Xue T., Chen B., Wu J., Wei D., Freeman W.T. Video enhancement with task-oriented flow. Int. J. Comput. Vis., 2019, vol. 127, no. 8, pp. 1106–1125.
- [11] Ma D., Zhang F., Bull D. BVI-DVC: A training database for deep video compression. IEEE Trans. Multimedia, 2021, pp. 1–1. Arxiv:2003.13552v2.
- [12] Soomro K., Zamir A.R., Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, 2012, vol. abs/1212.0402. Arxiv:1212.0402v1.
- [13] Perazzi F., Pont-Tuset J., McWilliams B., Van Gool L., Gross M., Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR). Las Vegas, USA, 2016, pp. 724–732.
- [14] Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process, 2004, vol. 13, no. 4, pp. 600–612.
- [15] Bogatyrev E., Molodetskikh I., Vatolin D. SR+Codec: A benchmark of super-resolution for video compression bitrate reduction. CoRR, 2024. Arxiv:2305.04844v3.

Enhancing video frame interpolation with adaptive kernels

Ali Salman

Spasenov Alexey Yurievich (*scientific advisor*)

BMSTU, Moscow, Russia

Video frame interpolation (VFI) creates intermediate frames for slow-motion playback or video compression. Traditional methods such as optical flow and fixed convolutions perform poorly in occlusion and large motion. ST-MFNet is a two-stage model that improves flow estimation and adaptive kernel synthesis. The adaptive kernels improve PSNR/SSIM by 2.28 dB/0.037 for dynamic textures (VFITex), compensating for small datasets. However, on larger datasets (80 GB vs. 6 GB) and powerful GPUs, the original model outperforms, showing the dependence of VFI on data and compute resources.

Keywords: VFI, optical flow, kernel-based, hybrid interpolation techniques, ST-MFNet, PSNR, SSIM