

УДК 621.397.33:004.932.72

Обзор методов интерполяции кадров

Али Салман slmanalii90@gmail.com

Карпенко Анатолий Павлович apkarpenko@mail.ru

Спасёнов Алексей Юрьевич a.spasenov@mail.ru

МГТУ им. Н.Э. Баумана, Москва, Россия

Представлен обзор на тему интерполяции видеокадров. Прослежено ее развитие от базовых методов оптического потока 1981 года, через классические подходы компенсации движения, до современных методов глубокого обучения, включая методы на основе ядер, потоков, гибридные и фазовые методы, методы, основанные на GAN, трансформеры, модели Matba и диффузионные модели. Представлены ключевые проблемы интерполяции кадров, такие как окклюзии и вычислительная эффективность. Также рассмотрены области применения интерполяции кадров, включая системы на основе событий, мультипликацию, медицинскую визуализацию и интегрированные задачи компьютерного зрения. Показана возможность использования комплекснозначных нейронных сетей при интерполяции видеокадров. Предложены направления разработки новых архитектур и теории интерполяции кадров.

Ключевые слова: интерполяция видеокадров, комплекснозначные нейронные сети, глубокое обучение, окклюзия

Введение. Интерполяция видеокадров VFI (Video Frame Interpolation) — это фундаментальная задача низкоуровневого компьютерного зрения, суть которой состоит в синтезе промежуточных кадров между соседними по времени входными кадрами для повышения временного разрешения видео [1]. VFI решает такие проблемы, как неоднозначность движения и помехи. VFI имеет решающее значение для таких приложений, как генерация замедленной съемки и преобразование частоты кадров [2].

В эпоху, предшествовавшую глубокому обучению, доминировали подходы к интерполяции кадров, основанные на компенсации движения MCFI (Motion-Compensated Frame Interpolation), которые использовали явную оценку движения с помощью сопоставления блоков кадров или их параметрических моделей [3].

С появлением глубокого обучения появились методы, использующие пространственно-адаптивные сверточные ядра, которые неявно кодируют информацию о движении. Эти методы коренным образом изменили VFI. Система AdaConv стал пионером в этом направлении, используя архитектуры, подобные U-Net [4]. Методы деформируемого ядра еще больше повысили адаптивность за счет обучаемых смещений, позволяя выполнять выборку за пределами обычных сверточных сеток [5].

Методы, основанные на потоках, явно оценивают оптический поток для выравнивания во времени. Различают методы прямого искажения, при кото-

рых исходные пиксели проецируются на целевые позиции, и методы обратным искажением, когда производится выборка из источника на основе целевых координат [6, 7].

Гибридные методы сочетают в себе преимущества подходов, основанных на ядрах и потоках, используя поток для глобального выравнивания, а ядра для локального уточнения [8]. Методы, основанные на фазе, используют информацию о фазе в частотной области, обеспечивая устойчивость к изменениям освещения, но имеют ограниченную эффективность при больших перемещениях [9].

Недавняя эволюция привела к появлению архитектур на основе трансформаторов, использующих механизмы внимания для моделирования долгосрочных зависимостей [10].

Совсем недавно предложены методы, основанные на Mamba и использующие модели структурированного пространства состояний [11]. Подходы, основанные на диффузии, переосмысливают VFI как процесс условного устранения шума, позволяя выполнять интерполяцию с учетом неопределенности [12].

Современные проблемы VFI. В VFI существуют следующие проблемы.

Управление большими объемами движения остается проблематичным, поскольку значительные перемещения объектов между кадрами часто превышают восприимчивость стандартных сетей, что затрудняет установление точного соответствия [13].

Фундаментальные неопределенности создает окклюзия, когда области в промежуточных кадрах не видны ни в одном из входных кадров, что создает проблемы с синтезом контента, особенно вблизи границ движения [8].

Устойчивость к изменению освещения нарушается, когда временные изменения освещенности нарушают предположение о постоянстве яркости, необходимое для многих методов оценки движения [9].

Моделирование нелинейных движений оказывается сложным, поскольку сценарии реального мира, включающие ускорение и криволинейные траектории, нарушают предположение о линейном движении, используемое во многих методах [1].

Вычислительная эффективность становится все более важной по мере того, как приложения VFI расширяются до сценариев реального времени, требуя тщательного баланса между производительностью и практической возможностью развертывания [2].

Комплекснозначные нейронные сети и их классификация. Комплекснозначные нейронные сети CVNNs (Complex-Valued Neural Networks) — это специализированный класс нейронных сетей, в которых входные данные, выходные данные, веса и, возможно, функции активации являются комплексными числами. Это отличает их от повсеместно распространенных действительностнозначных нейронных сетей RVNN (Real-Valued Neural Network). Основным аргумент в пользу использования CVNN следует из их способности обрабатывать данные, которые по своей сути являются ком-

плексными, такие как электромагнитные волны, сигналы МРТ и данные радара [14].

Классификация CVNNs, в первую очередь, основана на их архитектуре и способах обработки комплекснозначных данных. Существуют два основных типа CVNNs: разделенные CVNNs и полные CVNNs. Разделенные CVNNs обрабатывают сложные входные данные, разделяя их на действительную и мнимую части (или амплитуду и фазу), которые затем передаются в сеть, которая обычно использует функции активации с вещественными значениями. Хотя такой подход проще в реализации, он может привести к искажению фазы и увеличению числа параметров. Напротив, полные CVNNs обрабатывают комплексные числа как самостоятельные объекты, используя комплекснозначные веса и комплексные функции активации. Это сохраняет алгебраическую структуру и фазовые соотношения в данных, делая такие сети более мощными, хотя они сталкиваются с трудностями при разработке подходящих комплексно-дифференцируемых функций активации [14, 15].

CVNNs могут быть классифицированы на основе следующих ключевых характеристик:

а) *функции активации*. Основной проблемой CVNNs является разработка функций активации в соответствии с теоремой Лиувилля, которая утверждает, что ограниченная комплексно-дифференцируемая функция должна быть постоянной. Функции активации CVNNs в широком смысле подразделяются на следующие;

- голоморфные (полностью комплексные) — дифференцируемые в комплексной области, например, комплексный гиперболический тангенс [16];

- неголоморфные (расщепленные) — дифференцируемые по действительной и мнимой частям отдельно. Этот класс функций включает в себя тип А (реальное и мнимое разделение, например, раздельное применение ReLU (Rectified linear unit)) и тип В (амплитудно-фазовое разделение, например, уменьшение амплитуды при сохранении фазы) [17, 15];

- другие типы, например, многозначные нейроны MVNs (Multi-Valued Neurons), которые сопоставляют выходные сигналы с дискретными точками на единичной окружности или фазовые нейронные сети и кардиоидные (cardioid) активации [18];

б) *методы обучения и оптимизации*:

- методы, основанные на градиентах. Используется сложное обратное распространение на основе математики Виртингера для вычисления градиентов как для голоморфных, так и для неголоморфных функций активации [19];

- не основанные на градиенте. Используются правила исправления ошибок без производных, как, например, при обучении для MVNs [20];

в) *представления входных данных*: Входные данные могут быть сложными по своей природе (например, результаты преобразования Фурье) или по смыслу (например, представление направления ветра). Выходные данные, как правило, сложны для регрессии, в то время как классификация часто требует сопоставления с реальными вероятностями [21].

Применение CVNNs для получения одиночных изображений с высоким разрешением. Интеграция CVNNs в технологию одиночных изображений с высоким разрешением SISR (Single-Image Super-Resolution) продемонстрировала значительное повышение производительности благодаря встроенной обработке данных сложного характера в частотной области или от таких датчиков, как MRI (Magnetic resonance imaging) и радар. Основное преимущество заключается в их способности одновременно моделировать амплитудную и фазовую информацию, что приводит к улучшенному сохранению деталей, резкости и точности цветопередачи по сравнению с их действительностнозначными аналогами.

Общей для всех приложений является адаптация успешных реальных архитектур к сложной предметной области. Например, сверточная нейронная сеть со сверхвысоким разрешением (SRCNN) была модифицирована в комплексную SRCNN (C-SRCNN), путем замены всех операций комплекснозначными свертками и активациями, такими как комплексный ReLU (CReLU), который применяет ReLU отдельно к действительной и мнимой частям [22]. Этот подход, использующий комплексную среднеквадратичную ошибку CMSE (Complex Root Mean Square Error), позволил получить более высокие оценки PSNR (Peak Signal-to-Noise Ratio) и SSIM (Structural similarity index) при меньшем количестве параметров.

Аналогичным образом, для MPT-реконструкции были адаптированы развернутые сети и U-сети с использованием сложных сверток. Были оценены различные сложные функции активации, и сеть CReLU часто оказывалась более эффективной. Эти сети превосходили действительностнозначные модели в таких показателях, как NRMSE (Normalized Root Mean Square Error) и PSNR, а также обеспечивали хорошую фазовую реконструкцию, что имеет решающее значение для клинических применений, таких как визуализация кровотока [23].

При получении радиолокационных и гиперспектральных изображений CVNNs используют сложные данные, поступающие от датчиков. Трехмерная комплекснозначная сеть CNN (Convolutional Neural Network) для гиперспектральной SISR использовала быстрое преобразование Фурье FFT (Fast Fourier Transform) для перемещения данных в комплексную область на основе слоев, подобных ReLU (iReLU), для достижения высокой пространственной и спектральной точности [24]. При получении радиолокационных изображений сеть CV-CNN продемонстрировала более низкий RMSE (Root Mean Square Error), но более высокое время обработки, чем аналог с действительными значениями, благодаря встроенной обработке сложных эхо-сигналов и использованию активации “Abs” [25].

Таким образом, CVNN расширяют возможности SISR для различных типов данных — от естественных изображений до медицинских данных и данных дистанционного зондирования — за счет обеспечения более эффективной и физически согласованной структуры для обработки сложной информации, что приводит к повышению качества реконструкции и эффективности.

Потенциальное влияние использования CVNN в VFI. Интеграция CVNNs в VFI открывает значительные перспективы для улучшения темпорального синтеза при обработке видео. CVNN, которые оперируют комплексными числами для получения информации об амплитуде и фазе, обладают преимуществами по сравнению с аналогами, использующими действительные значения, включая более информативное представление с меньшим количеством параметров, улучшенное обобщение и естественное согласование с операциями в частотной области, такими как преобразования Фурье [17, 26]. В VFI, где фазовые сдвиги кодируют едва заметные движения и временные зависимости, CVNN могут смягчить такие проблемы, как перекрытия, изменения освещения и нелинейные траектории, путем прямого моделирования движения на основе фазы, как это продемонстрировано в фазо-ориентированных методах [27].

Например, комплекснозначный стробируемый автокодировщик продемонстрировал высокую производительность при прогнозировании видеокадров, позволяя эффективно изучать трансляционные преобразования для более плавной генерации промежуточных кадров [28]. Это может быть распространено на приложения реального времени, сокращая вычислительные затраты и одновременно улучшая качество восприятия в таких областях, как медицинская визуализация и голографическая реконструкция видео [26]. Кроме того, способность CVNN обрабатывать коррелированные реальные и воображаемые компоненты может повысить надежность в зашумленных или размытых последовательностях, потенциально превосходя потоковые модели или модели на основе ядра в таких показателях, как PSNR и SSIM [27, 28]. В целом CVNNs могут революционизировать VFI, создав архитектуру с эффективными параметрами, в которой приоритет отдается фазовой когерентности, что открывает путь для улучшения качества видео с высокой точностью в новых технологиях, таких как виртуальная реальность и автономные системы.

Заключение. Комплекснозначные нейронные сети CVNNs представляют собой многообещающий инструмент для интерполяции видеокадров. Их основное преимущество заключается в том, что они автоматически обрабатывают информацию о фазе и амплитуде, предлагая надежную, эффективную с точки зрения параметров структуру для моделирования движения и решения таких проблем, как вариации освещения и помехи. Это может привести к высокому качеству интерполяции при плавном движении и повышенной точности восприятия.

Основные ограничения CVNNs включают сложность архитектуры, обусловленную ограничениями в сложно дифференцируемых функциях активации, и меньшую поддержку в основных системах глубокого обучения.

Будущая работа должна быть сосредоточена на разработке стабильных архитектур, специфичных для CVNN, а также стратегий обучения для VFI.

Литература

- [1] Seo W., Oh J., Kim M. Bim-vfi: directional motion field-guided frame interpolation for video with non-uniform motion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025, pp. 7244–7253.
- [2] Shen T., Li D., Gao Z., Tian L., Barsoum E. Ladder: an efficient framework for video frame interpolation. *Arxiv preprint, arXiv:2404.11108*, 2024.
- [3] Ha T., Lee S., Kim J. Motion compensated frame interpolation by new block-based motion estimation algorithm. *IEEE Transactions on Consumer Electronics*, 2004, vol. 50, № 2, pp. 752–759.
- [4] Niklaus S., Mai L., Liu F. Video frame interpolation via adaptive convolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 670–679.
- [5] Zhu X. Deformable convnets v2: more deformable, better results. X. Zhu, H. Hu, S. Lin, J. Dai. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.
- [6] Niklaus S., Liu F. Softmax splatting for video frame interpolation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5437–5446.
- [7] Jaderberg M., Simonyan K., Zisserman A. et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 2015, vol. 28, pp. 2017–2025.
- [8] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, vol. 43, № 3, pp. 933–948.
- [9] Meyer S., Wang O., Zimmer H., Grosse M., Sorkine-Hornung A. Phase-based frame interpolation for video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1410–1418.
- [10] Lu L., Wu R., Lin H., Lu J., Jia J. Video frame interpolation with transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3532–3542.
- [11] Zhang G., Liu C., Cui Y., Zhao X., Ma K., Wang L. Vfimbamba: video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 107225–107248.
- [12] Voleti V., Jolicoeur-Martineau A., Pal C. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 23371–23385.
- [13] Huang Z., Zhang T., Heng W., Shi B., Zhou S. Real-time intermediate flow estimation for video frame interpolation. *Proceedings of the European Conference on Computer Vision*, 2022, pp. 624–642.
- [14] Lee C.Y., Hasegawa H., Gao S.C. Complex-valued neural networks: a comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 2022, vol. 9, № 8, pp. 1406–1426.
- [15] Benvenuto N., Piazza F. On the complex backpropagation algorithm. *IEEE Transactions on Signal Processing*, 1992, vol. 40, № 4, pp. 967–969.
- [16] Mandic D.P. Complex valued recurrent neural networks for non-circular complex signals. *International Joint Conference on Neural Networks*, 2009, pp. 1987–1992.
- [17] Trabelsi C. et al. Deep complex networks. *CoRR, arXiv:1705.09792*, 2018.
<https://doi.org/10.48550/arXiv.1705.09792>
- [18] Virtue P., Yu S.X., Lustig M. Better than real: complex-valued neural nets for MRI fingerprinting. *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3953–3957.
- [19] Brandwood D.H. A complex gradient operator and its application in adaptive array theory. *IEE Proceedings H — Microwaves, Optics and Antennas*, 1983, vol. 130, pp. 11–16.
- [20] Aizenberg N.N., Aizenberg I.N., Krivosheev G.A. Multi-valued and universal binary neurons: mathematical model, learning, networks, application to image processing and pattern recognition.

- tion. *Proceedings of 13th International Conference on Pattern Recognition*, 1996, vol. 4, pp. 185–189.
- [21] Sarroff A.M. *Complex Neural Networks for Audio*. Tech. Rep. TR2018-859, Dartmouth College, Computer Science, Hanover, NH, 2018.
- [22] Zaher H., Jafar A., Alsahwa B. Image Super-Resolution Using Complex-Valued Deep Convolutional Neural Network. *Syrian Journal for Science and Innovation*, 2024, pp. 45–52.
- [23] Cole E. et al. Analysis of deep complex-valued convolutional neural networks for MRI reconstruction and phase-focused applications. *Magnetic Resonance in Medicine*, 2021, vol. 86, no. 2, pp. 1093–1109.
- [24] Sun Z., Xu X., Pan Z. SAR ATR using complex-valued CNN. *Proceedings of the 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 2020, pp. 125–128.
- [25] Aburaed N., Alkhatib M.Q., Marshall S., Zabalza J., Ahmad H.A. Complex-valued neural network for hyperspectral single image super resolution. *Proceedings of SPIE 12338, Hyperspectral Imaging and Applications II*, 2023, art. no. 123380.
- [26] Hirose A. A survey of complex-valued neural networks. *Arxiv:2101.12249*, 2021. <https://doi.org/10.48550/arXiv.2101.12249>
- [27] Meyer S. et al. PhaseNet for video frame interpolation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 498–507.
- [28] Fohlmeister M., Kirchner H. Complex valued gated auto-encoder for video frame prediction. *Arxiv:1903.03336*, 2019. <https://doi.org/10.48550/arXiv.1903.03336>

Overview of video frame interpolation methods

Ali Salman

slmanalii90@gmail.com

Karpenko Anatoly Pavlovich

apkarpenko@mail.ru

Spasenov Alexey Yurievich

a.spasenov@mail.ru

BMSTU, Moscow, Russia

This article explores the potential of Complex-Valued Neural Networks (CVNNs) for Video Frame Interpolation (VFI). It argues that CVNNs' inherent ability to process phase and amplitude information offers a robust, parameter-efficient framework for superior motion modeling and handling of challenges like occlusions. While acknowledging limitations in architectural design, the analysis concludes that CVNNs represent a promising direction for achieving higher-fidelity slow-motion generation and frame-rate conversion in emerging video technologies.

Keywords: video frame interpolation, complex-valued neural networks, deep learning, occlusions