Uma Abordagem a *DeepFake* via Algoritmos de Aprendizagem Profunda

Gustavo S. Rodrigues[1], **Carlos A. Silva**[1]

[1]Instituto Federal de Minas Gerais – Sabará (IFMG-Sabará)
CEP 34590-390 – Sabará – MG – Brasil

gustavosr_13@outlook.com, carlos.silva@ifmg.edu.br

Abstract. *The advancement of artificial intelligence has popularized and made actions to create and manipulate images and videos accessible. As a consequence, deepfakes have been widely used in the service of misinformation and the dissemination of fake news. Since deepfakes are very important in this new information age, this article aims to contribute to the understanding of this artificial intelligence technique, analyzing three relevant deep learning algorithms for the generation of deepfakes. The results show that Deepfacelab spends almost twice the processing time compared to FaceSwap and First Order Motion, however it presents a better quality of the generated results.*

Resumo. *Com o rápido avanço da inteligência artificial, ações de criação e manipulação de imagens e vídeos tem se tornado cada vez mais comum e acessível às pessoas. Como consequência, as deepfakes tem sido bastante utilizadas à serviço da desinformação e à disseminação de fake news. Considerando as deepfakes como um marco nessa nova era da informação, este artigo propõe a análise de três importantes algoritmos de aprendizagem profunda para a geração de deepfakes. Os resultados mostram que o Deepfacelab despende quase o dobro do tempo de processamento em relação ao FaceSwap e ao First Order Motion, porém apresenta uma melhor qualidade dos resultados gerados.*

1. Introdução

Deepfake, termo que vem de *Deep Learning* - "aprendizagem profunda" e *fake* - "falso", são conteúdos sintetizados por inteligência artificial que sobrepõe imagens de vídeo de uma pessoa-fonte em um vídeo de uma pessoa-alvo, manipulando falas e ações da pessoa-alvo por meio do vídeo da pessoa-fonte [Nguyen et al. 2019]. Em geral, estes métodos necessitam de um grande volume de dados de imagens e vídeos, desta forma, é comum figuras públicas e celebridades serem alvos de *deepfakes*, em virtude de sua exposição midiática. O primeiro vídeo *deepfake* surgiu em 2017, onde o rosto de uma celebridade foi trocada pelo rosto de um ator pornô [Yu et al. 2021]. Na literatura existem diversos casos destes tipos de manipulações visuais em cenários políticos [Appel and Prietzel 2022], de redes sociais [Fagni et al. 2021], de entretenimento [Usukhbayar and Homer 2020] entre outros.

Com a popularização da internet, tem sido cada vez mais comum pessoas buscarem informações por redes sociais, como Facebook, e plataformas de compartilhamento de vídeos, como Youtube [Anderson 2018]. Infelizmente, alinhado a esse crescimento de acesso à informação, tem se destacado a geração e disseminação de notícias falsas,

as quais podem causar consequências impactantes ao indivíduo e à sociedade. Há uma tendência de que as *deepfakes* sejam usadas cada vez mais em pornografia de vingança, *bullying*, vídeo falso de provas em tribunais, sabotagem política, propaganda terrorista, chantagem, manipulação de mercado e *fake news* [Maras and Alexandrou 2019].

Pesquisas que tratam da criação e detecção de *deepfakes* têm sido foco de interesse da comunidade científica nos últimos anos. É comum encontrar trabalhos na literatura focando em abordagens de reencenação, ou seja, mudar expressão, boca, pose, olhar ou corpo, ou abordagens de substituição, como substituir o rosto de um alvo [Mirsky and Lee 2021]. Alguns autores buscam categorizar os métodos de criação e detecção de *deepfakes*, como [Tolosana et al. 2020] que se baseia na síntese facial, troca de identidade e expressões e manipulação de atributos. A discussão sobre os impactos destes métodos na sociedade é bastante ampla, sendo cerne de pesquisas como de [Karnouskos 2020] que investiga as *deepfakes* por perspectivas multiangulares, incluindo mídia, sociedade, leis e regulamentações.

Neste trabalho buscamos ampliar a compreensão a respeito dos métodos de criação de *deepfakes* analisando a metodologia de funcionamento de três algoritmos, mais especificamente *Deepfacelab*, *Faceswap* e *First Order Motion*. Desse modo, para alcançar o objetivo central desse artigo, o texto encontra-se organizado da seguinte forma: Na seção 1 é feita a introdução da temática abordada no trabalho. Na seção 2, são apresentado relevantes trabalhos da literatura relacionados ao tema de pesquisa. Na seção 3, a metodologia empregada é descrita. Na seção 4 são apresentados os algoritmos de aprendizagem profunda pelos quais é realizada uma abordagem às *deepfakes*, e na seção 5 constam as análises e discussões a respeito dos resultados obtidos. E por fim, na seção 6 são apresentadas as conclusões finais e a expectativa dos trabalhos futuros.

2. Revisão bibliográfica

O artigo [Korshunov and Marcel 2018] foi um dos primeiros trabalhos acadêmicos a apresentar uma versão de algoritmo *open source* de *deepfake*, baseado em redes adversárias generativas (GANs, do inglês *Generative. Adversarial Networks*) para realizar a troca do rosto de uma pessoa por outro em um vídeo. Os autores reforçam a importância de como a quantidade de dados, o treinamento, e o ajuste fino dos parâmetros, interfere diretamente na qualidade dos vídeos gerados.

A transposição do movimento de duas pessoas em uma performance de dança foi o alvo da aplicação de *deepfake* em [Chan et al. 2019]. A execução com sucesso desta transposição, veio ratificar o grande potencial desta tecnologia, vislumbrando a ampliação de atuação desta ferramenta em outras áreas como animações e cinema.

Com o intuito de popularizar a tecnologia e demonstrar sua potencialidade, [Siarohin et al. 2019] desenvolveu um modelo intuitivo para o público leigo. Foi disponibilizada um rede neural treinada para o reconhecimento de imagens. O algoritmo desenvolvido permitia a manipulação de imagens para criação de vídeos ou *gifs* animados de modo bastante convincente.

Neste trabalho [Naruniec et al. 2020] os autores conduziram uma comparação de algoritmos de *deepfake* no qual ficou demonstrado a importância do treinamento progressivo na troca de rostos em alta resolução. Foi evidenciado que o uso de procedimentos

de estabilização de *landmarks* melhora os efeitos de tremor irreais e outras instabilidades temporais que podem ocorrer ao operar no domínio de alta resolução.

É fundamental discutir sobre o impacto que as *deepfakes* podem causar na sociedade. Essa discussão foi tema central de [Nguyen et al. 2019], no qual os autores buscaram apresentar alguns métodos para detecção de *deepfake*, promovendo uma ampla reflexão sobre desafios, tendências potenciais e direções futuras nessa área. Da mesma forma, [Vaccari and Chadwick 2020] destacou o impacto das *deepfakes* sobre a confiança nas informações das propagandas veiculadas aos tradicionais meios de comunicação, e no discurso público, criando um ambiente de constante desconfiança e desinformação. Os autores afirmam que a inexistência de leis que restrinjam o uso indevido na rede mundial de computadores relacionado ao mal uso da inteligência artificial, podem provocar significativos danos socioeconômicos no futuro.

3. Metodologia

O presente trabalho propõe um pesquisa exploratória a respeito do tema central *deepfake*, considerando três relevantes algoritmos de criação da tecnologia em estudo, mais especificamente os algoritmos *Deepfacelab* [Perov et al. 2020], *Faceswap*[1] e *First Order Motion* [Siarohin et al. 2019]. A escolha destes três algoritmos se deu após a análise de alguns parâmetros como, baixa complexidade de implementação, exigindo um menor nível de conhecimento técnico para sua utilização e configuração, sendo esses algoritmos amplamente utilizado por criadores de conteúdo audiovisual, acadêmicos e empresas, além de possuírem uma boa documentação *online* para auxiliar no manuseio e aprendizado dos mesmos. Outro ponto é a escalabilidade do sistema ao *hardware* disponível, tendo em vista que alguns algoritmos exigem especificações de *hardware* mais avançadas, tornando complexa sua utilização. Inicialmente foi realizado um levantamento bibliográfico em bases de dados científicas como *Web of Science*, *IEEE Xplore*, Google Acadêmico, *arXiv* entre outros, além dos inúmeros repositórios do GitHub, sobretudo referendado por [McCosker 2022], o qual aplica análise de conteúdo qualitativo contextual para explorar os repositórios mais populares do GitHub e contas do YouTube que ensinam "como fazer *deepfake*".

Construída a base teórica e compreendendo o estado-da-arte, o próximo passo foi estabelecer uma base única de dados para compor a instância a ser submetida aos três algoritmos. A base em questão foi extraída de [Rössler et al. 2018] a qual contém cerca de meio milhão de imagens editadas (de mais de 1.000 vídeos). Posteriormente foram implementados os algoritmos, e realizadas as simulações computacionais. A arquitetura utilizada foi de um processador Intel i7 10700F, com placa de vídeo Zotac RTX 3080 12GB, além de memória ram de 16Gb DDR4 a 3200Mhz. Desta forma, com intuito de aferir sobre os resultados e informação relevante a respeito dos métodos de *deep learning*, a seção 4 descreve os algoritmos empregados neste trabalho.

4. Desenvolvimento

Compreender o funcionamento dos métodos de geração de *deepfakes* é de suma importância para a promoção de uma discussão mais ampla a respeito do tema. Para isso,

[1] https://faceswap.dev

são apresentados os três agoritmos utilizados neste desenvolvimento nas subseções posteriores.

4.1. Deepfacelab

O *Deepfacelab* [Perov et al. 2020] é um sistema *deepfake* de código aberto criado para troca de rosto de alta qualidade. No ambiente do Github ele possui mais de 3.000 *forks* (novo repositório que compartilha configurações de código e visibilidade com o repositório original) e 14.000 *stars* (pode ser usado como métrica de popularidade). O sistema é amigável quanto a usabilidade e proporciona um equilíbrio entre velocidade e facilidade de uso. Atualmente duas diferentes arquiteturas e três tipos de modelos pré-treinados para a geração de *deepfakes* estão disponíveis no sistema. As arquiteturas suportadas são a DF (*DeepFake*) e a LIAE (*Lightly Improved Auto Encoder*), cujas principais diferenças estão relacionadas com alguns parâmetros como, interpretação do rosto com ou sem transformação, precisão e fidelidade aos dados de origem. A arquitetura DF funciona melhor quando a origem e o destino têm faces e cores semelhantes. Por outro lado, na LIAE a interpretação do rosto é mais tolerante com alguma transformação, adaptando-se mais ao rosto de destino, assim como as formas e cores diferentes. Partindo destas arquiteturas, temos os modelos Quick96, SAEHD e o AMP. Para fins deste trabalho, utilizamos o modelo SAEHD, pois com ele é possivel selecionar quaisquer uma das arquiteturas, sendo assim, o mais versátil dos três modelos. O sistema usa uma fase de conversão que consiste em um 'codificador' e um 'decodificador de destino' com uma camada 'interna' entre eles. A abordagem proposta LIAE é exibida na Figura 1.

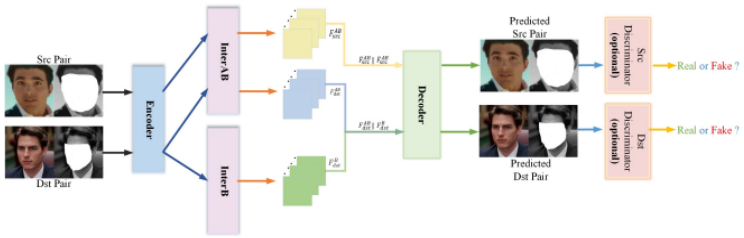

Figura 1. Arquitetura do *Deepfacelab* [Perov et al. 2020].

Para a extração de características na imagem fonte, o algoritmo utiliza-se de referência facial baseado em mapa de calor 2DFAN [Bulat and Tzimiropoulos 2017], que é usado enquanto a segmentação facial é obtida por uma rede de segmentação facial de granulação fina, **TernausNet** [Iglovikov and Shvets 2018].

4.2. Faceswap

Faceswap [Deepfakes 2017] é uma tecnologia *deepfake* que promove a troca de dois rostos utilizando uma rede neural profunda. Se o rosto de uma pessoa for trocado por um rosto vindo de um arquivo de código aberto, o rosto gerado torna-se irreconhecível ao original. De acordo com [Zhu et al. 2020], o rosto gerado preserva as informações faciais originais, como expressões e cor da pele, o que também pode ser aplicável a vídeos.

Um exemplo do funcionando do *Faceswap* é descrito em [Zhu et al. 2020] e ilustrado pela Figura 2.

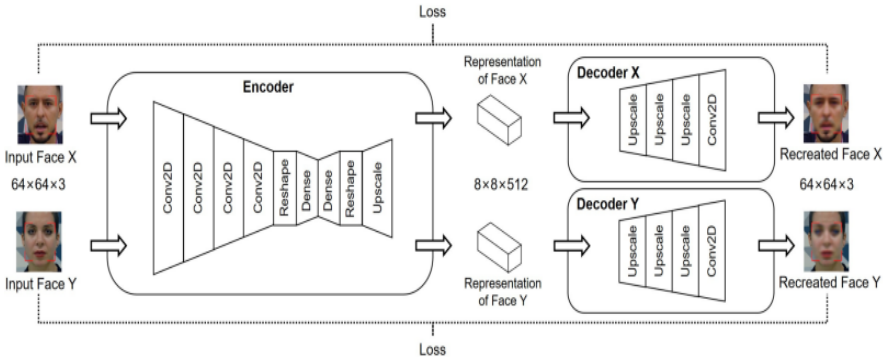

Figura 2. *Faceswap* [Zhu et al. 2020].

As entradas X e Y são exemplo de faces de código aberto. A face de entrada é processada por um codificador e torna-se um vetor de representação. O vetor é processado por um decodificador correspondente e se torna uma face recriada. A perda por retropropagação é calculada pela diferença entre faces de entrada e faces recriadas. Durante a troca, o modelo muda o decodificador para gerar imagens trocadas.

4.3. First Order Motion

Diferentes áreas de interesse como produção de filmes, fotografias e *e-commerce* podem usufruir de diversas aplicações provenientes da geração de vídeos por meio da animação de objetos em imagens estáticas. A animação de imagens é a sintetização de vídeos combinando a aparência extraída de uma imagem de origem com padrões de movimento derivados de um vídeo de condução. Problemas que se enquadram nessa classificação, são comumente tratados na literatura por prioridades na representação do objeto e por técnicas de computação gráfica assumindo conhecimento sobre o modelo do objeto específico para animar. A proposta do *First Order Motion* é usar modelos generativos profundos utilizando um conjunto de pontos-chave auto-aprendidos juntamente com transformações afins locais (justificativa do nome "primeira ordem") para modelar movimentos complexos, além de introduzir um gerador com reconhecimento de oclusão, que adota uma máscara de oclusão estimada automaticamente para indicar partes do objeto que não são visíveis na imagem de origem e que deve ser inferido do contexto [Siarohin et al. 2019].

A Figura 3 ilustra o funcionamento do *First Order Motion*.

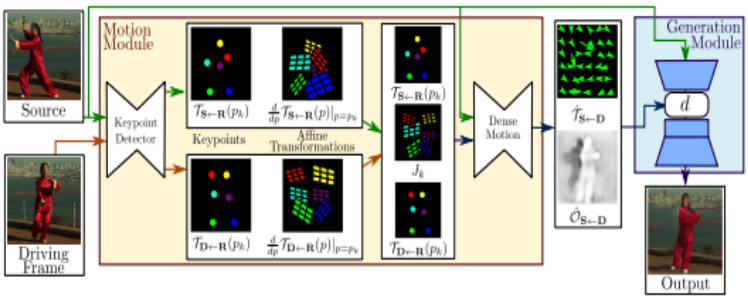

Figura 3. *First Order Motion* [Siarohin et al. 2019].

De acordo com [Siarohin et al. 2019], o método assume uma imagem S (origem)

e um *frame* D de um vídeo de condução como entradas. O detector de ponto-chave não supervisionado extrai o movimento de primeira ordem consistindo em pontos-chave esparsos e transformações afins locais em relação ao quadro de referência R. A densa rede de movimento usa a representação de movimento para gerar o denso fluxo ótico $\tau_S \leftarrow D$ de D para S e o mapa de oclusão $\mathcal{O}_S \leftarrow D$. A imagem de origem e as saídas da rede de movimento denso são usados pelo gerador para renderizar a imagem de destino.

5. Discussão dos resultados

Para além do extenso referêncial bibliográfico que constitui este trabalho, buscou-se compreender o funcionamento dos principais algoritmos de criação de *deepfakes* por meio de simulação computacional de uma base de dados da literatura.

A base de dados utilizada para realização dos teste foi a Faceforensic++[2]. Esta base de dados consiste em uma sequências de 1.000 vídeos originais que foram manipulados com quatro métodos automatizados de manipulação facial: *Deepfakes*, *Face2Face*, *Faceswap* e *NeuralTextures*. Uma parcela desses dados, equivalente a 100Gb de vídeos, foram separados para serem utilizados para os testes dos algoritmos utilizados neste trabalho.

A Tabela 5 apresenta o tempo despendido por cada algoritmo, considerando o treinamento acrescido da execução dos mesmos, a quantidade máxima de iterações, e os requisitos mínimos de *hardware* necessários para cada algoritmo.

Algoritmos	**Tamanho da base**	**Modelo**	**Tempo(s)**	**Nº de iterações**	**Arquitetura mínima**
DeepFaceLab	100Gb	SAEHD	$8,28 \times 10^4$	100k	GPU CUDA 3.5, $\geq$ 16Gb ram, CPU $\geq$ 8 núcleos, 8 Gb de VRAM
First Order Motion Model	100Gb	-	$4,02 \times 10^4$	100k	GPU CUDA
Faceswap	100Gb	-	$4,84 \times 10^4$	100k	GPU CUDA 3.5, 8 Gb de VRAM

Entre os resultados observados, percebe-se uma significativa diferença entre os modelos dos algoritmos, no que diz respeito ao tempo para executar uma mesma tarefa utilizando uma mesma base de dados. O *Deepfacelab*, utilizando o modelo SAEHD, despendeu quase um dia (23 horas) de processamento, enquanto os algoritmos *First Order Motion* e *Faceswap* gastaram em torno de aproximadamente 11 horas e 13,5 horas, respectivamente.

Neste trabalho, o *Deepfacelab* utilizando o modelo SAEHD, que apesar de ser um modelo que demanda mais recursos de *hardware* e necessita de um tempo maior de treinamento e execução, é capaz de oferecere resultados melhores com maior nitidez nas imagens geradas no final do processo.

Após a realização dos testes computacionais com os três algoritmos, o *Deepfacelab*, que apesar de exceder consideravelmente o tempo de treinamento e execução quando comparado aos outros dois algoritmos, apresentou a *deepfake* com qualidade visual maior, menos serrilhados, ou distorções na movimentação de partes chave nos vídeos resultantes como a boca e olhos, como pode ser observado nas Figuras 4 e 5.

[2] `https://github.com/ondyari/FaceForensics/blob/master/dataset/README.md`

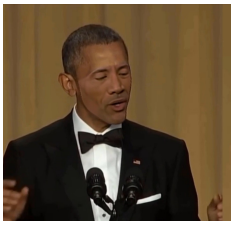

Figura 4. *Faceswap*.

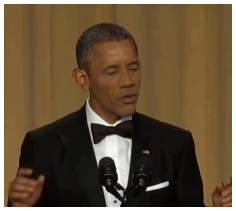

Figura 5. *Deepfacelab*.

6. Conclusão

Deepfakes são um conjunto de algoritmos com base na inteligência artificial que permite a criação de vídeos falsos, manipulados de maneira convincente para enganar os espectadores. Embora possam ser usados para fins de entretenimento, como em filmes e vídeos virais, também têm o potencial de serem usados para espalhar desinformação, difamar pessoas e causar danos graves. É importante que a sociedade esteja ciente dos riscos associados às *deepfakes* e trabalhe em conjunto para desenvolver soluções a fim de detectar e combater sua disseminação.

Neste trabalho foi apresentado uma análise exploratória sobre essa tecnologia abordando e comparando os algoritmos *Deepfacelab*, *Faceswap* e *First Order Motion*, a fim de verificar suas capacidades com o intuito de compreender melhor o funcionamento e apresentar uma discussão importante a respeito das *deepfakes*. Em essência, podemos afirmar que alguns modelos, embora sejam mais exigentes em termos de *hardware* e tempo computacional, são capazes de oferecer um melhor resultado final. Como trabalho futuro pretende-se estabelecer uma métrica de desempenho a respeito da qualidade dos vídeos/imagens gerados pelos métodos de aprendizagem profunda.

Referências

Anderson, K. E. (2018). Getting acquainted with social networks and apps: combating fake news on social media. *Library Hi Tech News*, 35(3):1–6.

Appel, M. and Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4):zmac008.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030.

Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. (2019). Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942.

Deepfakes (2017). Faceswap. https://github.com/deepfakes/faceswap. Acessado: 03-04-2023.

Fagni, T., Falchi, F., Gambini, M., Martella, A., and Tesconi, M. (2021). Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Iglovikov, V. I. and Shvets, A. A. (2018). Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *ArXiv*, abs/1801.05746.

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 1(3):138–147.

Korshunov, P. and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Maras, M.-H. and Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, 23(3):255–262.

McCosker, A. (2022). Making sense of deepfakes: Socializing ai and building data literacy on github and youtube. *New Media & Society*, page 14614448221093943.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41.

Naruniec, J., Helminger, L., Schroers, C., and Weber, R. M. (2020). High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., and Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 1:2.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. (2020). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

Usukhbayar, B. and Homer, S. (2020). Deepfake videos: The future of entertainment. http://dx.doi.org/10.13140/RG.2.2.28924.62085.

Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1):2056305120903408.

Yu, P., Xia, Z., Fei, J., and Lu, Y. (2021). A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624.

Zhu, B., Fang, H., Sui, Y., and Li, L. (2020). Deepfakes for medical video deidentification: Privacy protection and diagnostic information preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–420.