Transcrição Automática de Harmonias Funcionais

Fernando Tonucci de Cerqueira Oliveira[1], **Flavio Vinicius Diniz de Figueiredo**[1]

[1]Instituto de Ciências Exatas – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

fernandotonucci@dcc.ufmg.br, flavio@dcc.ufmg.br

Abstract. *In Western Music Tradition, chords and harmony are crucial elements which set the emotional tone and structure of a song, and, as such, understanding the full harmonical scope used by an artist is essential in the study of any musical piece. In the literature, however, works often focus solely on identifiyng the chords played at each time, thus losing valueable information on the function the chords serve in the given context. In this work, we attempt to classify chords in a music by their so-called harmonical function, which better shows how chords interact with each other to create the piece's harmony.*

Resumo. *Na tradição musical do ocidente, acordes e harmonia são elementos cruciais na definição do tom emocional e estrutura de qualquer canção, e, assim, compreender o escopo harmônico completo usado por um artista é essencial no estudo de sua obra. Na literatura, no entanto, trabalhos frequentemente focam apenas na identificação dos acordes tocados a cada momento, perdendo, portanto, informações valiosas sobre a função que cada acorde cumpre em um dado contexto. Neste trabalho, nós tentamos classificar acordes em um música pela sua Função Harmônica, que melhor representa como eles interagem entre si para a criação da harmonial da peça.*

1. Introdução

No âmbito de recuperação de informação musical (do inglês *Music Information Retrieval* - MIR), a tarefa de reconhecimento de acordes é uma das mais complexas e exploradas ao longo dos anos [Müller 2015, Pauwels et al. 2019]. Um acorde pode ser definido como um conjunto qualquer de semitons (ou notas) tocadas ao mesmo tempo, mas, para soar bem aos ouvidos, eles, no geral, seguem formas pré-definidas: temos uma nota dominante acompanhada de outras que a complementam, como ilustrado na Figura 1. A tarefa de reconhecimento de acordes, então, consiste em sair de uma representação musical mais crua (tipicamente a forma de onda, ou o espectrograma/cromagrama - ver Figura 2) para uma conjunto de tuplas (c_i, t_i^s, t_i^f), onde o acorde c_i é soado entre os tempos t_i^s e t_i^f da música. Porém, um problema menos abordado no reconhecimento de acordes consiste identificação de sua função harmônica, descrita a seguir.

Definimos harmonia como uma combinação de diferentes notas tocadas simultaneamente, de forma que elas soem coesas na percepção do ouvinte. Ao ouvido humano, tal coesão entre duas notas distintas ocorre quando suas frequências são harmônicas entre si, de forma que a junção de suas ondas crie um novo padrão, como ilustrado na Figura 3. Na tradição musical ocidental, os acordes são a peça básica de uma harmonia, sendo a análise harmônica, então, o estudo da suas interações, progressões e construções. Em

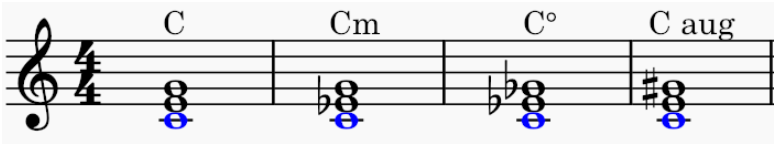

Figura 1. Diferentes variações de um acorde de dó. Em azul, as notas base do acorde.

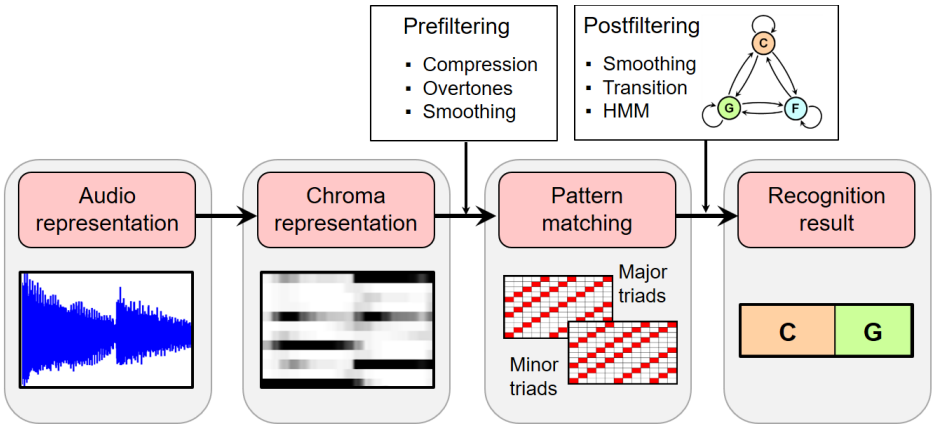

Figura 2. Pipeline clássica do reconhecimento de acordes. Imagem extraída de [Müller 2015]

contextos distintos, entretanto, o mesmo acorde pode ter diferentes funções dentro de uma harmonia, de forma que para o melhor entendimento da peça é necessário, também, identificar as suas respectivas funções harmônicas, isto é, o papel que eles exercem na construção harmônica da música.

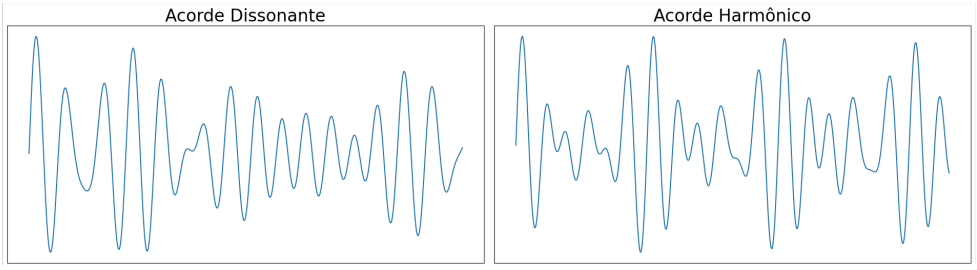

Figura 3. Na esquerda, um acorde dissonante, e na direita, um harmônico. Note como, no acorde harmônico, forma-se um padrão repetido, o que não ocorre no dissonante.

Dessa forma, podemos definir a identificação de harmonias funcionais como sendo, além do reconhecimento do acorde, a análise de sua tônica, modulação, qualidade, inversão e alterações. Esta grande abrangência traz consigo também uma nova complexidade à tarefa. Em particular (1) existe mais ambiguidade, dado que o mesmo acorde pode exercer diferentes funções; (2) o conjunto de classes para prever é bem maior (na casa dos milhares); (3) tal conjunto de classes é altamente desbalanceado, como mostrado na Figura 4. Nela, é perceptível que a grande prevalência dos acordes maiores, menores e maiores com sétima ocorre em detrimento das demais classes, que estão pouco presentes nos dados.

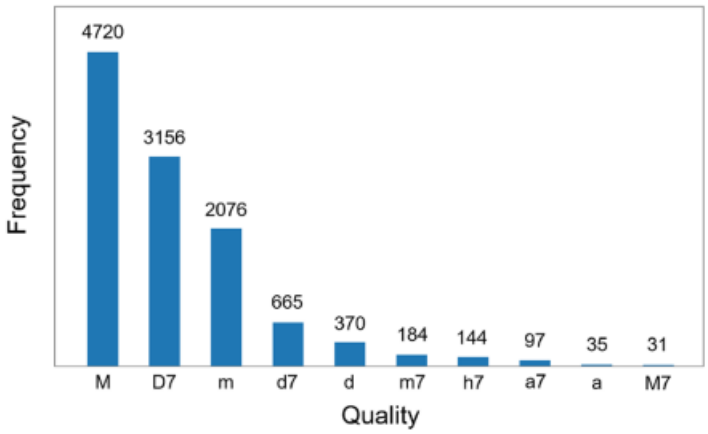

Figura 4. Frequência de diferentes qualidade de acordes no dataset Beethoven Piano Sonata Functional Harmony (BPS-FH) [Chen et al. 2018]. Note como a maioria dos acordes está pouco presente nos dados, formando uma cauda longa na distribuição das classes. Imagem extraída de [Chen and Su 2021].

Neste trabalho, portanto, focamos principalmente nas dificuldades trazidas pelos problemas (2) e (3). Por isso, propomos a utilização da focal loss [Lin et al. 2017], uma função de perda criada especialmente para classificações altamente desbalanceadas, como forma de minimizar as dificuldades presentes nos dados.

O restante do artigo será organizado da seguinte forma: na Seção 2 será apresentado um panorama geral atual da área de Identificação de Funções Harmônicas; na Seção 3 será apresentado o pipeline proposto, detalhando os aspectos técnicos da implementação do algoritmo; na Seção 4 será realizada uma avaliação do método proposto, bem como uma discussão acerca dos resultados; por fim, na Seção 5 será feito um resumo dos resultados obtidos, e serão discutidos alguns dos principais desafios encontrados pelos modelos, bem como possíveis passos para solucioná-los.

2. Trabalhos Relacionados

O emprego de métodos de aprendizado profundo para a identificação de harmonias funcionais é ainda um avanço recente na literatura. Em [Chen et al. 2018], os autores inicialmente usam uma redes neural recorrente LSTM com atenção bi-direcional em uma abordagem baseada em multi-task learning (MTL) em que a rede aprende a classificar sub-tarefas do problema. A classificação de harmonias funcionais, portanto, é dividida na identificação individual dos graus, da qualidade e da inversão do acorde. Com isso, embora a acurácia para cada sub-tarefa seja alta, elas, ao serem combinadas, geram uma previsão ainda insatisfatória da função harmônica.

Em [Micchi et al. 2020], os autores propõe uma rede dividida em dois segmentos interconectados. No primeiro, é utilizada uma arquitetura convolucional que aprende a reconhecer o contexto local da entrada, que é então passado para uma segunda parte recorrente que modela o contexto global. Após uma análise de seus resultados, os autores citam, ainda, as principais causas que levam a rede a fazer uma classificação equivocada. Primeiramente, os autores notam que, em grande parte dos casos, a rede tem

dificuldades de segmentar os acordes, isto é, o modelo é pouco preciso em identificar o momento em que o acorde muda. Nesses casos, a rede tem uma tendência de exacerbar a frequência de trocas. Outro fator crucial para o mau desempenho das redes é sua baixa acurácia na rotulação de funções raras, com a solução proposta tendo uma clara inibição na identificação de rótulos pouco comuns.

Em [Chen and Su 2021], os autores constroem em cima de seus trabalhos anteriores e propõe uma nova rede baseada em transformers, o Harmony Transformer V2 (HT). O HT, além de arquitetura mais moderna, tem duas melhorias claras em relação aos trabalhos de [Micchi et al. 2020]: o uso de uma abordagem MTL em que, além dos acordes, a rede também prevê o momento em que eles mudam, consequentemente melhorando a sua segmentação; e a consolidação da previsão das funções harmônicas em uma única saída da rede. Essa consolidação, no entanto, embora evite uma prejudicial independência entre as sub-tarefas, aumenta o vocabulário de saída da rede, o que pode intensificar o seu mau desempenho na identificação de classes raras.

Em todos os casos, as redes usam a entropia cruzada como a função de custo, de maneira que o modelo aprende a prever classes dominantes em detrimento das menos frequentes. Com o objetivo de solucionar esse problema, [Lin et al. 2017] propões o acréscimo de um fator modular no entropia cruzada, de forma que a rede seja incentivada, portanto, a focar em exemplos mais difíceis, melhor aprendendo, consequentemente, a modelar funções harmônicas menos presentes nos dados.

3. Metodologia

O entendimento de harmonias em uma música é uma tarefa altamente dependente do contexto local e global em que as notas estão inseridas. Dessa forma, a arquitetura escolhida para a identificação de harmonias funcionais deve ser capaz de modelar essa inter-dependência dos dados. Nesse sentido, os transfomers se destacam como uma arquitetura promissora, por sua alta capacidade de modelar, principalmente, dependências longas.

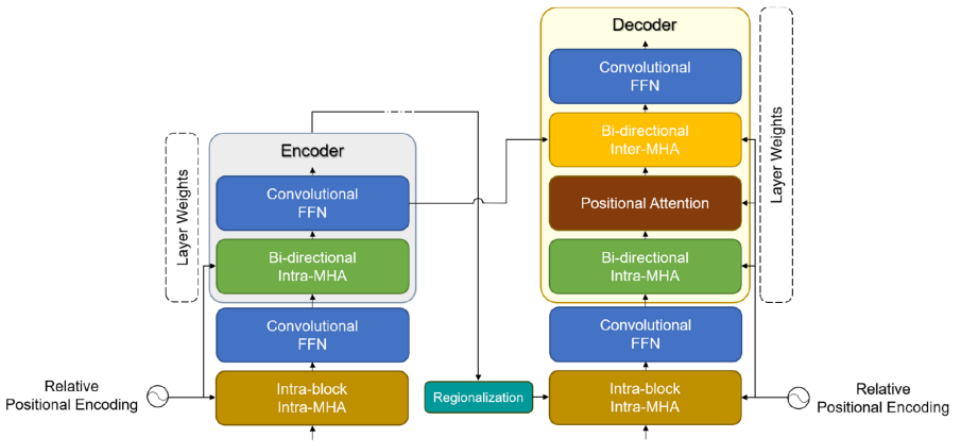

Figura 5. Arquitetura do Harmony Transformer V2. Imagem extraída de [Chen and Su 2021].

Nesse trabalho, portanto, o Harmony Transformer v2(HTv2) [Chen and Su 2021] foi referência para a construção do modelo proposto, que faz um processo end-2-end para a identificação de harmonias funcionais, gerando diretamente a classificação total da função harmônica. A arquitetura da rede, vista na Figura 5, é baseada em um encoder-decoder, que primeiro transforma o dado de entrada pra uma representação em um espaço latente, que passa, então, por um processo de regionalização antes de ser decodificada para a saída esperada.

É importante também ressaltar que, em contrapartida com a maior parte da literatura que parte de dados de áudio ou espectrograma para a realização da classificação, o HTv2 usa dados simbólicos. Essa escolha foi feita com o objetivo de atacar um buraco na literatura, que, dada a maior dificuldade na obtenção de dados simbólicos rotulados, raramente os usam. Tais dados, no entanto, são de fácil acesso, principalmente, na indústria musical moderna, que frequentemente criam peças partindo dos chamados piano rolls. Os piano rolls, que servem também de entrada para a rede, codificam as notas soadas em cada momento da música, gerando uma representação altamente compreensível e facilmente produzível, sendo, portanto, muito utilizados em diversas situações. Na Figura 6 podemos um exemplo dessa representação.

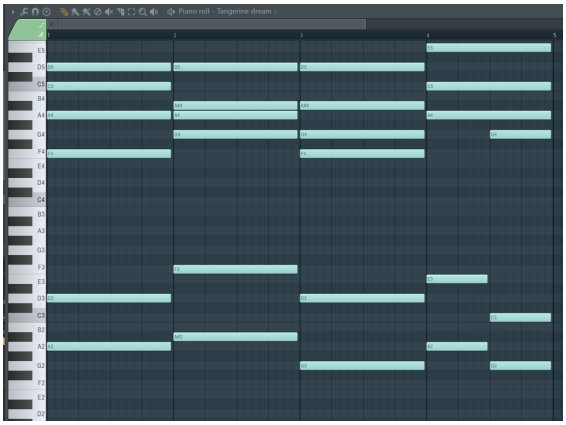

Figura 6. Exemplo de um piano roll

Ademais, o MTL foi utilizado para auxiliar a rede no aprendizado e segmentação dos acordes. Assim, além de prever a função harmônica em cada momento da música, a rede também gera como saída a chave atual e o momento de mudança dos acordes, tarefas essas que auxiliam o aprendizado de aspectos cruciais para a classificação funcional da harmonia. Quanto às chaves, essas são também parte da previsão final do modelo, visto que as funções harmônicas em si são independente da chave da música, de forma que, apenas sabendo a função de um acorde, não podemos determiná-lo sem saber o tom da música.

A saída final da rede, portanto, é composta por três previsões, sendo a mudança de acorde uma segmentação binária, a previsão de chaves uma classificação de 42 classes (21 tônicas multiplicadas por 2 qualidades (maior e menor)), e a classificação de funções harmônicas uma classificação de 5040 classes. Essas 5040 funções harmônicas partem de 9 graus primários para o acorde (temos 9 notas em uma chave), 12 graus secundários (qualquer um dos 12 semitons da música ocidental é válido) 10 qualidades (maior, menor, aumentado, diminuto, maior com sétima, menor com sétima, sétima dominante, sétima

diminuta, sétima meio-diminuta e sexta aumentada) e 4 inversões (alterações em qual nota é o baixo do acorde).

Com o intuito de melhor lidar com o alto desbalanceamento do problema em questão, utilizamos, em vez da entropia cruzada do HTv2 a focal loss. Essa função de perda multiplica a entropia cruzada por um fator modular baseada na confiança da saída p_t da rede, de forma que, exemplos com menor confiança, isto é, aqueles em que a rede têm mais dificuldade, são mais relevantes no treino.

$$FL = -\alpha_c(1 - p_t)^{\gamma} log(p_t) \tag{1}$$

Na focal loss, temos dois hiper-parâmetros a serem definidos: γ, que define o quanto a rede focará nos exemplos mais difíceis (note que, com γ = 0, a focal loss é equivalente à entropia cruzada), e α, um parâmetro opcional de rebalanceamento das classes. Note que deve ser definido um α para cada classe.

4. Experimentos

O HTv2 foi reimplementado em pytorch e treinado separadamente por 15 horas em uma GPU P100 utilizando duas funções de perda diferentes: a entropia cruzada, que serve de baseline para os experimentos, e a focal loss. Em ambos os casos, foi utilizada uma taxa de aprendizado de 0.0001, e o otimizador escolhido foi o ADAM. Para a focal loss, foi utilizada um gamma de 4, e α foi utilizada apenas na predição da mudança de acordes, em que foi definido $\alpha = 0.07$ para exemplos negativos, e $\alpha = 0.93$ para positivos.

4.1. Dataset

Ambos modelos foram treinados e avaliados em um conjunto de 32 sonatas de Bethoven no piano, rotuladas em [Chen et al. 2018] para a criação do dataset BPS-FH. Cada sonata foi dividida em pequenos intervalos de 32 tempos subdivididos em 4 semi-colcheias, para a criação de matrizes 88x128 (88 notas no piano x 128 semi-colcheias) que modelam um piano roll. Cada sonata, ainda, foi movida 9 vezes de chave, de forma a aumentar a diversidade e quantidade dos dados do dataset, totalizando mais de 200 000 piano rolls rotulados diferentes. Das 32 sonatas, 8 delas foram selecionadas para o conjunto de treino, em que não foi aplicado o mesmo processo de data augmentation.

4.2. Resultados

Tabela 1. Acurácia dos modelos na previsão das tarefas finais.

	Chave	Função Harmônica	Mudança de Acordes
Focal Loss	0.710	0.267	0.883
Entropia Cruzada	0.718	0.3061	0.883

A Tabela 1 mostra a acurácia nas predições da chave, função harmônica e segmentação das mudanças de acorde dos dois modelos no dataset de teste. É perceptível que a focal loss não é uma função de perda efetiva para o problema de Identificação da Harmonia Funcional, visto que, sua acurácia nesta tarefa foi significativamente mais baixa que a da entropia cruzada.

Indo mais a fundo nas predições, o modelo treinado com a focal loss previu a classe mais dominante do problema (a primeira tônica, com prevalência de 15% nos dados) em cerca de 40% das vezes, contra apenas 22% no treinado com a entropia cruzada. Esse fato indica que, neste problema, a focal loss agiu de maneira contrária ao esperado, incentivando o modelo a prever mais frequentemente as classes dominantes. Essa discrepância ocorreu, nós observamos, devido a dois grandes fatores: o elevado número de classes, quando comparado ao tamanho do dataset, e a utilização do multi-task learning.

Primeiro, as predições da rede foram, em todos os momentos do treino, bastante incertas. Dessa forma, mesmo em exemplos teoricamente mais fáceis, a probabilidade atribuída pela rede a cada classe se mantinha baixa. Analisando a Equação (1), percebemos que, neste caso, o fator modulador da focal loss permaneceria elevado em todos os casos, de forma que a rede não foi capaz de focar seu treinamento nos exemplos de classes menos frequentes nos dados.

Segundo, embora tipicamente o alto desbalanceamento serviria como um empecilho para a entropia cruzada tradicional, no cenário de MTL estudado, as diferentes tarefas contribuíram para prevenir a grande concentração das predições em torno de alguns poucos rótulos. Treinando a rede apenas para a segmentação de acordes, por exemplo, o HTv2 convergiu para nunca prever a mudança de acordes, o que não ocorreu quando ele foi treinado com a focal loss. No treino completo com o MTL, no entanto, após atingir, inicialmente, uma previsão composta apenas por 0s quando treinado com a entropia cruzada, a rede, guiada pelas demais tarefas, voltou a identificar mudanças após algumas épocas.

É importante ressaltar, ainda, que as 15 horas de treino não foram suficientes para o modelo convergir. Supomos, então, que após uma maior convergência, quando a rede tivesse previsões um pouco mais certeiras, o desempenho da focal loss pudesse melhorar, já que neste caso a rede já seria capaz de distinguir exemplos fáceis dos difíceis.

Por fim, observamos também, ao longo do treino, que, embora a focal loss não tenha se demonstrado adequada para a tarefa principal, ela aparentou melhor na segmentação de acordes, com a rede convergindo mais rapidamente para previsões adequadas e, ainda, mantendo sempre uma boa proporção de resultados positivos e negativos.

5. Conclusão

Neste trabalho, abordamos a tarefa de Identificação de Harmonias Funcionais em músicas partindo de um representação simbólica. O foco principal do estudo foi em analisar os impactos do alto desbalanceamento dos dados referentes ao problema, que estão amplamente concentrado em uma pequena parte das mais de 5000 funções harmônicas distintas. Por isso, testamos o uso da focal loss em vez da entropia cruzada, uma função de perda projetada para atenuar, justamente, os problemas de desbalanceamento. Após experimentos, percebemos que a grande dificuldade da tarefa servia de empecilho para a focal loss, que foi incapaz de focar em exemplos teoricamente mais difíceis dada a baixa confiança da rede em todos os casos. Reparamos, por fim, que para tarefas vizinhas, como a segmentação de acordes, a focal loss tem um bom potencial, que deve ser explorado em trabalhos futuros.

Por isso, o próximo foco da pesquisa será em combinar a entropia cruzada com a focal loss, usando a última no treinamento apenas da segmentação de acordes. Com isso,

esperamos, a rede melhorará seu desempenho na segmentação de acordes, e manterá a boa convergência apresentada pela entropia cruzada para os demais casos. Outra alternativa seria iniciar o treinamento com a entropia cruzada, valendo-se de sua convergência mais veloz, e fazer um fine-tuning com a focal loss, de forma que a rede pudesse, já mais confiante em suas previsões, aprender a melhor modelar as classes raras. Ademais, outras funções de perda focadas na atenuações dos supracitados problemas de desbalanceamento devem também ser exploradas. Dentre essas, ressaltamos a equalization loss [Tan et al. 2020] e a LDAM loss [Cao et al. 2019], que trazem uma abordagem distinta da focal loss que independe da confiança das previsões da rede e, por consequência, podem acelerar a convergência do modelo.

Referências

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Chen, T.-P. and Su, L. (2021). Attend to chords: Improving harmonic analysis of symbolic music using transformer-based models. *Transactions of the International Society for Music Information Retrieval*, 4(1).

Chen, T.-P., Su, L., et al. (2018). Functional harmony recognition of symbolic music data with multi-task recurrent neural networks. In *ISMIR*, pages 90–97.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Micchi, G., Gotham, M., and Giraud, M. (2020). Not all roads lead to rome: Pitch representation and model architecture for automatic harmonic analysis. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 3(1):42–54.

Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer.

Pauwels, J., O'Hanlon, K., Gómez, E., Sandler, M., et al. (2019). 20 years of automatic chord recognition from audio.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., and Yan, J. (2020). Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671.