INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS CORRENTE PIAUÍ
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

GUILHERME SANTO COSTA

CLASSIFICAÇÃO DE TEXTURAS COM VGG

CORRENTE

2025

1 INTRODUÇÃO

A capacidade de compreender e classificar texturas é um pilar fundamental na visão computacional. Ela impulsiona avanços em áreas diversas como inspeção de materiais, diagnóstico médico, robótica e análise de imagens de satélite. As texturas fornecem informações cruciais sobre as superfícies e estruturas dos objetos, complementando, e muitas vezes superando, características como cor e forma em diversas tarefas de reconhecimento.

Apesar de sua importância, a classificação de texturas não é uma tarefa fácil. Isso se deve, em grande parte, à dificuldade de definir e extrair características complexas e robustas das imagens.

Nesse contexto, este projeto tem como objetivo implementar e treinar um modelo VGG-16 customizado com fine-tuning, focal loss e visualização de ativações, utilizando o dataset DTD para a classificação de texturas, o focal loss para lidar com desbalanceamento entre classes, juntamente com métricas como acurácia, precisão, revocação e AUC para avaliação robusta. Os resultados obtidos demonstram a efetividade da arquitetura proposta na distinção de texturas visuais complexas.

2 TECNOLOGIAS UTILIZADAS

2.1 Describable Textures Dataset

O Describable Textures Dataset (DTD) é um banco de dados de texturas desenvolvido pelo Visual Geometry Group da Universidade de Oxford. É composto por 5640 imagens organizadas em 47 categorias baseadas em descrições semânticas inspiradas na percepção humana, como "riscado", "trançado" ou "pontilhado". Cada categoria possui 120 imagens com tamanhos entre 300×300 e 640×640, sendo que pelo menos 90% da área da imagem representa o atributo descrito.

Um diferencial do DTD é que suas categorias são subjetivas e anotadas manualmente, representando atributos perceptuais e não objetos concretos. Isso eleva o nível de complexidade do desafio de classificação, pois exige que os modelos de inteligência artificial reconheçam padrões estruturais e semânticos sutis, indo além do simples reconhecimento visual.

O dataset fornece 10 divisões (splits) pré-definidas para treinamento, validação e teste, promovendo comparações padronizadas entre métodos. Seu uso é comum em pesquisas que

buscam conectar linguagem e percepção visual, como reconhecimento de atributos, aprendizado com atenção e classificação baseada em descritores texturais.

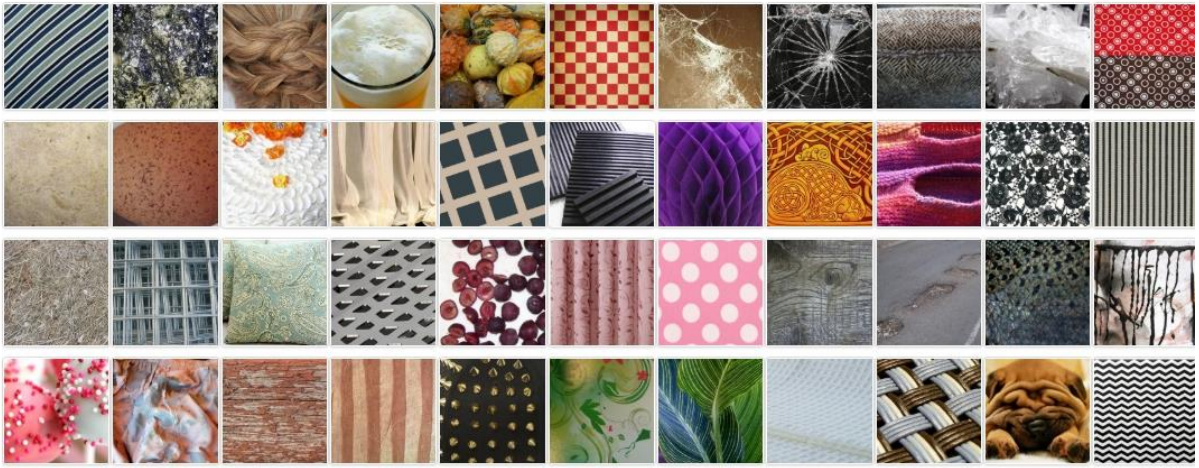

Figura 1: Exemplos de imagens do Describable Textures Dataset

2.2 Modelo VGG16

O modelo utilizado neste projeto é baseado na arquitetura VGG16, uma rede neural convolucional profunda desenvolvida pelo Visual Geometry Group da Universidade de Oxford. Este modelo foi originalmente projetado para a competição ImageNet em 2014 e é conhecido por sua simplicidade e eficácia na extração de características visuais.

Arquitetura da VGG16

A VGG16 é composta por 16 camadas com pesos treináveis, sendo 13 camadas convolucionais e 3 camadas totalmente conectadas (fully connected). Sua estrutura é organizada em 5 blocos principais, cada um contendo 2 ou 3 camadas convolucionais com filtros 3x3 e stride 1, seguidos por uma camada de MaxPooling 2x2, que reduz pela metade as dimensões espaciais da saída anterior.

A arquitetura enfatiza a profundidade e a uniformidade das operações, utilizando exclusivamente convoluções pequenas (3x3) e pooling fixo (2x2), o que simplifica o design e melhora a capacidade de generalização. Todas as camadas convolucionais utilizam a função de ativação ReLU, promovendo uma aprendizagem mais rápida e reduzindo problemas como o gradiente desvanecido.

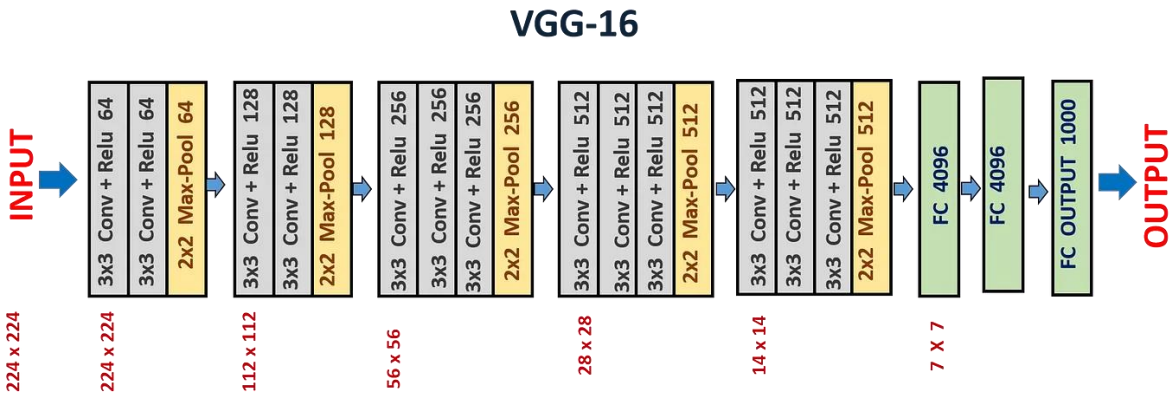

Figura 2: Modelo da arquitetura VGG16

Na versão original, após os blocos convolucionais, a rede é composta por duas camadas densas com 4096 neurônios cada, seguidas por uma camada densa final com 1000 neurônios e ativação softmax, voltada para a classificação de imagens do ImageNet.

Para o presente trabalho, foi utilizada a VGG16 como extratora de características (feature extractor), aproveitando os pesos pré-treinados no ImageNet. As 15 primeiras camadas foram congeladas, permitindo que apenas as camadas superiores fossem ajustadas durante o treinamento. Adicionalmente, a arquitetura foi adaptada com novas camadas densas, normalização por batch, dropout e regularização L1_L2 para melhor adequação à tarefa de classificação das 47 categorias do Describable Textures Dataset (DTD).

Essa abordagem permite o aproveitamento do conhecimento prévio da rede em reconhecer padrões visuais gerais, enquanto ajusta suas camadas finais para lidar com as especificidades e nuances do conjunto de texturas descritivas.

3 METODOLOGIA

3.1 Pré-processamento dos dados

Antes de usar as imagens do DTD no modelo VGG16, primeiro, todas as imagens, que tinham tamanhos variados, foram redimensionadas para 224x224 pixels. Mantendo as cores (três canais RGB) porque é o que o VGG16 precisa, depois, os valores dos pixels foram ajustados para ficarem entre 0 e 1 (dividindo por 255), isso ajuda o treinamento a ser mais rápido e evita problemas matemáticos.

Para que o modelo aprendesse melhor e não ficasse "viciado" só nas imagens originais — e para lidar com a diferença de quantidade de imagens em algumas das 47 categorias —, foi

usada uma técnica de aumento de dados. Basicamente, novas versões de cada imagem aplicando pequenas transformações aleatórias enquanto o modelo estava treinando. Isso incluiu girar a imagem, mover um pouco para os lados, dar zoom, mudar o brilho ou a cor, e até espelhar. O importante é que a textura principal da imagem nunca foi alterada. Esse truque faz com que o modelo veja muitas variações da mesma textura, tornando-o mais inteligente e capaz de reconhecer as texturas mesmo com pequenas diferenças.

Por último, as categorias das texturas foram convertidas para um formato que o modelo entende (chamado one-hot encoding). Os dados foram então separados em três grupos: um para o treinamento do modelo, um para validar como ele estava aprendendo, e outro para testar o desempenho final, garantindo que a avaliação fosse justa com imagens que o modelo nunca tinha visto.

3.2 Arquitetura do Modelo e Estratégia de Treinamento

A arquitetura proposta neste trabalho baseia-se na rede convolucional VGG16, amplamente reconhecida por sua simplicidade e desempenho consistente em tarefas de visão computacional. O modelo foi carregado com pesos pré-treinados no ImageNet, servindo como extrator de características. Para preservar o conhecimento previamente adquirido e evitar o sobreajuste, as primeiras 15 camadas da VGG16 foram congeladas durante a primeira fase de treinamento.

A arquitetura foi então estendida com um cabeçalho totalmente conectado, cuidadosamente projetado para permitir um refinamento específico da tarefa de classificação de texturas. A saída da VGG16 é conectada a uma camada de pooling global (GlobalAveragePooling2D), que reduz dimensionalmente o mapa de características mantendo sua expressividade. Na sequência, duas camadas densas (com 256 e 128 unidades, respectivamente) foram adicionadas, ambas com ativação ReLU, normalização em lote (Batch Normalization) e regularização L1/L2 para controle de complexidade. Para mitigar o risco de sobreajuste em um dataset com variação visual elevada, utilizou-se uma taxa elevada de Dropout (60%) após cada camada densa.

A saída final é composta por uma camada Dense com 47 neurônios (equivalente ao número de classes do DTD), ativada por função softmax, adequada à natureza multicategórica do problema.

O modelo foi treinado em duas fases: inicialmente, com as camadas convolucionais da base congeladas, permitindo que apenas as camadas superiores se adaptassem à tarefa de classificação de texturas. Na segunda fase, a técnica de fine-tuning foi aplicada, liberando

gradualmente algumas das camadas convolucionais superiores da VGG16 para aprendizado supervisionado, promovendo melhor especialização da extração de características para o domínio de texturas.

A função de perda utilizada foi a Focal Loss categórica, escolhida por sua robustez em cenários com classes desbalanceadas, pois atribui maior peso a exemplos difíceis ou menos representados. O algoritmo de otimização foi o Adam, configurado com uma taxa de aprendizado inicial moderada (1×10^{-4}), ajustada dinamicamente via callback ReduceLROnPlateau, que reduz a taxa sempre que o desempenho em validação estagna.

A seguir, a Tabela 1 resume a configuração dos principais hiperparâmetros utilizados no modelo proposto:

Hiperparâmetro	Valor
Arquitetura base	VGG16 (pré-treinada no ImageNet)
Tamanho das imagens	224 x 224 x 3
Otimizador	Adam
Taxa de aprendizado inicial	1×10^{-4}
Função de perda	Categoria Focal Loss
Métricas	Acuráci, AUC, Precisão, Revocação
Batch size	32
Épocas	50(Fase Inicial) + 30 (Fine-tuning)
Dropout	0,60
Regularização L1/L2	L1 = $1x10^{-5}$, L2 = $1x10^{-4}$
Camadas congeladas	Primeiras 15 camadas da VGG16
Estratégia de aumento de dados	Rotações, deslocamentos, zoom, brilho, flips

3 RESULTADOS E DISCUSSÕES

Após o treinamento do modelo VGG16 customizado, foram avaliadas diversas métricas de desempenho com o objetivo de compreender sua eficácia na tarefa de classificação de texturas do dataset DTD. As métricas analisadas incluem acurácia (accuracy), precisão (precision), revocação (recall), AUC (Área sob a Curva ROC) e perda (loss), tanto para os conjuntos de treinamento quanto de validação.

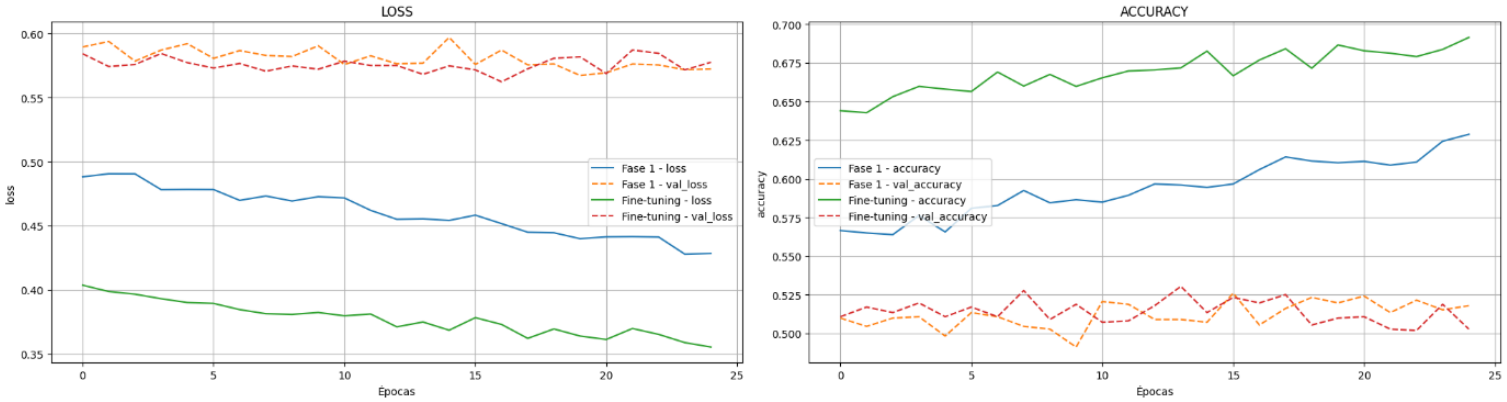

Figura 3: Gráficos de desempenho (Loss e Acurácia)

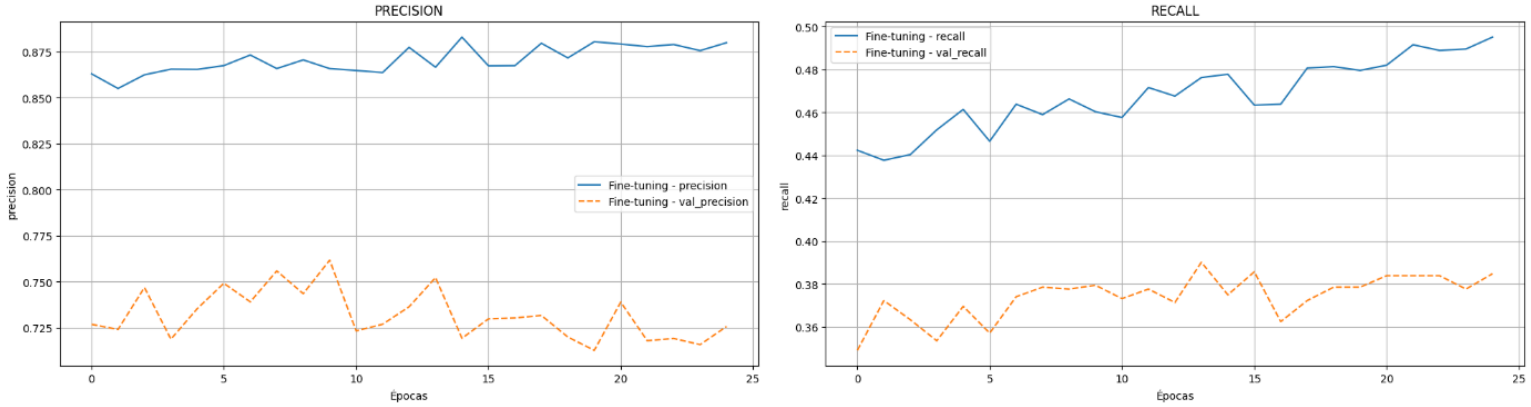

Figura 4: Gráfico de desempenho (Precision e Recall)

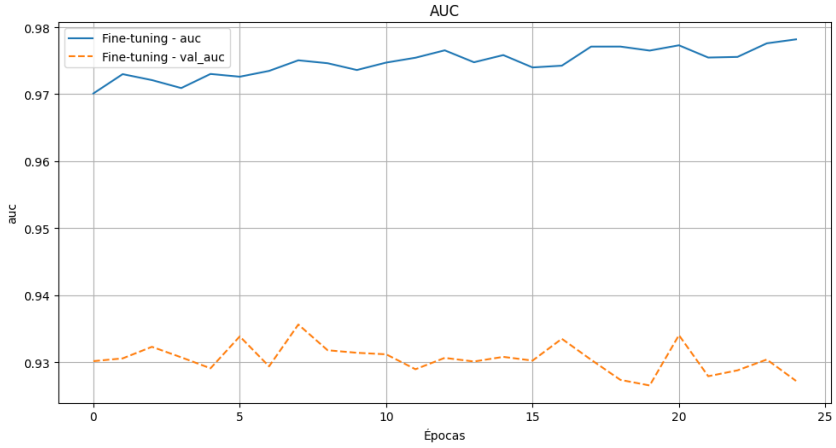

Figura 5: Gráfico de desempenho (AUC)

Os resultados revelam que, apesar do modelo atingir um valor elevado de AUC (97% em treino e 91% em validação), a acurácia e o recall apresentam desempenho mais modesto, principalmente no conjunto de validação. A alta AUC sugere que o modelo é capaz de distinguir bem entre as classes em termos de probabilidade, mas a baixa revocação (37% em validação) indica que ele ainda falha em recuperar corretamente todas as classes relevantes, o que pode estar relacionado ao desafio de lidar com categorias texturais altamente subjetivas do DTD.

A diferença significativa entre os desempenhos de treino e validação, principalmente nas métricas de precisão (87% vs. 73%) e recall (46% vs. 37%), aponta para indícios de overfitting, ou seja, o modelo se adapta bem aos dados vistos durante o treinamento, mas tem dificuldade em generalizar para dados novos. Esse comportamento é reforçado pela curva de perda (loss), que diminui no treino mas permanece alta na validação, sugerindo que o modelo está aprendendo padrões específicos demais em vez de características mais generalizáveis das texturas.

Tabela 1 – Médias das Métricas de Desempenho (Fine-tuning)

Métrica	Média (Treinamento)	Média (Validação)
Accuracy	59.00%	51.00%
Precision	87.00%	73.00%
Recall	46.00%	37.00%
AUC	97.00%	91.00%
Loss	43.00%	58.00%

Além disso, a matriz de confusão evidencia a complexidade da tarefa. Há confusão significativa entre diversas classes de texturas, como striped, banded, lined e grid, que compartilham características visuais semelhantes. Esse fator evidencia a dificuldade do modelo em distinguir atributos perceptuais subjetivos, mesmo utilizando uma arquitetura pré-treinada e adaptada.

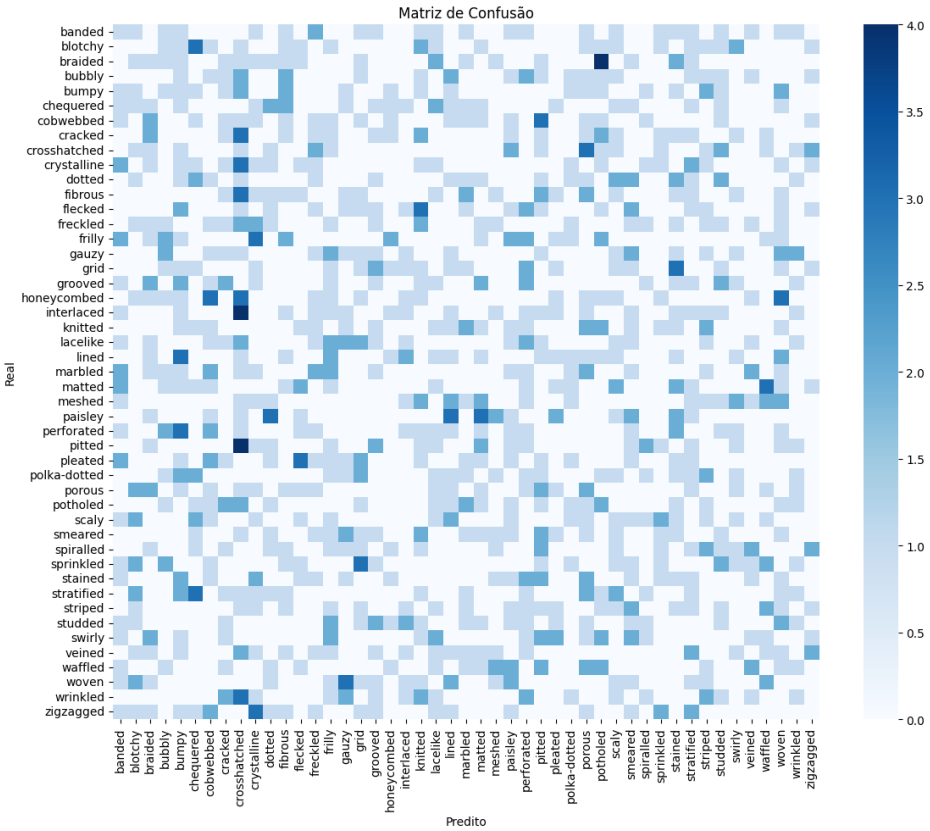

Figura 6: Matriz de confusão

Em resumo, os resultados demonstram que o modelo é promissor em termos de aprendizado, especialmente por sua elevada capacidade de separação (AUC), mas apresenta limitações em termos de generalização, revocação e discriminação fina entre classes semelhantes. Futuras abordagens poderiam considerar estratégias como o uso de mecanismos de atenção, aumento de dados direcionado, ou até mesmo modelos autoexplicáveis, com o objetivo de lidar melhor com a subjetividade e complexidade do domínio de texturas descritas por atributos perceptuais humanos.

5 CONCLUSÃO

Neste trabalho, foi investigada a aplicação de uma arquitetura VGG16 customizada com fine-tuning para a tarefa de classificação de texturas perceptuais no conjunto de dados Describable Textures Dataset (DTD). O modelo foi ajustado com técnicas como regularização L1_L2, Dropout e Batch Normalization, e obteve um desempenho expressivo em termos de separabilidade das classes, com uma AUC média de 91% no conjunto de validação. No entanto, a acurácia final de 51% revela desafios significativos na generalização do modelo para novos dados.

Esse desempenho pode ser explicado pelas características intrínsecas do próprio DTD. O dataset é composto por 47 classes de texturas naturais, que frequentemente apresentam grande variação visual interna e, ao mesmo tempo, alta similaridade entre diferentes classes. Texturas como striped, banded e lined, por exemplo, podem ter padrões bastante parecidos, dificultando a distinção mesmo para observadores humanos. Além disso, cada classe possui apenas 120 imagens, o que representa uma quantidade limitada de amostras para o treinamento de redes profundas, favorecendo o surgimento de overfitting.

Esses fatores tornam o DTD um benchmark notoriamente difícil, com valores típicos de acurácia entre 40% e 70% mesmo para arquiteturas robustas como VGG, ResNet e DenseNet. Portanto, os resultados obtidos com o modelo proposto estão dentro da expectativa para o cenário e reforçam a complexidade envolvida na tarefa.

Apesar das limitações, o experimento demonstrou que a VGG16 é capaz de aprender representações úteis para texturas, como evidenciado pelas ativações intermediárias e pela boa performance em métricas como precisão e AUC.

REFERÊNCIAS

OXFORD VISUAL GEOMETRY GROUP. Describable Textures Dataset (DTD). Disponível em: https://www.robots.ox.ac.uk/~vgg/data/dtd/ . Acesso em: 23 jun. 2025.

ROCHA, Anderson. Introdução ao reconhecimento de padrões em texturas. Instituto de Computação - UNICAMP. Disponível em: https://ic.unicamp.br/~rocha/msc/ipdi/texture_classification.pdf . Acesso em: 23 jun. 2025.

OLIVEIRA, Luciano S. de. Extração de características de textura. Departamento de Informática, Universidade Federal do Paraná. Disponível em: https://www.inf.ufpr.br/lesoliveira/padroes/haralick.pdf . Acesso em: 23 jun. 2025.

SOUSA, Felipe Silva de. Reconhecimento de texturas utilizando redes neurais convolucionais. Revista de Iniciação Científica - CPS, 2020. Disponível em: https://ric.cps.sp.gov.br/bitstream/123456789/3162/1/Reconhecimento_texturas.pdf . Acesso em: 23 jun. 2025.

BHOITE, Siddhesh. VGG Net — Architecture Explained. Medium, 2021. Disponível em: https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f . Acesso em: 23 jun. 2025.