Análise Comparativa de Modelos Visuais para Ambientes Internos com Variação Temporal e de Iluminação

William Azevedo Pessoa de Melo[1], **Alícia Caldeira da Silva**[1], **Carlos Victor de Araújo Lima**[1], **Alternei de Souza Brito**[1,2], **Felipe Gomes de Oliveira**[1,2]

[1]Instituto de Ciências Exatas e Tecnologia – Universidade Federal do Amazonas
Rua Nossa Senhora do Rosário, 3863 - Tiradentes - Itacoatiara, AM, Brasil

[2]Instituto de Computação – Universidade Federal do Amazonas
Av. General Rodrigo Octávio, 6200 - Coroado I - Manaus, AM, Brasil

william.melo, alicia.silva, carlos-victor.lima, alternei

felipeoliveira @ufam.edu.br

Abstract. *Visual classification models are essential in applications such as autonomous navigation and mobile robotics, but they still face challenges in indoor environments with lighting and temporal variations. This work compares the performance of DINOv2 feature extractor, a state-of-the-art self-supervised model, with supervised architectures such as ConvNeXt, EfficientNet, ResNet, and ViT. Using the KTH-IDOL2 dataset, we evaluated the models under different environmental conditions. Results show that DINOv2 consistently outperformed the others, achieving up to 98.02% accuracy. These findings highlight the robustness of self-supervised representations in the face of visual variability, positioning DINOv2 as a promising alternative for realistic indoor scene classification.*

Resumo. *Modelos de classificação visual são fundamentais em aplicações como navegação autônoma e robótica móvel, mas ainda enfrentam desafios em ambientes internos com variações de iluminação e mudanças temporais. Este trabalho compara o desempenho do extrator de características DINOv2, modelo auto-supervisionado de última geração, com arquiteturas supervisionadas como ConvNeXt, EfficientNet, ResNet e ViT. Utilizando o dataset KTH-IDOL2, avaliamos os modelos em diferentes condições ambientais. Os resultados mostram que o DINOv2 superou consistentemente os demais, alcançando até 98,02% de acurácia. Os achados destacam a robustez das representações auto-supervisionadas frente à variabilidade visual, posicionando o DINOv2 como uma alternativa promissora para classificação de ambientes em cenários realistas.*

1. Introdução

A classificação visual de ambientes constitui um componente central em sistemas inteligentes, com aplicações relevantes em robótica, realidade aumentada e navegação autônoma [Barros et al. 2021, Garg et al. 2021]. Esses sistemas requerem representações visuais robustas e generalizáveis para operarem com eficácia em ambientes com variações temporais e espaciais [Masone and Caputo 2021].

Entretanto, modelos supervisionados enfrentam desafios significativos quando expostos a alterações no ambiente, como variações de iluminação, deslocamento de objetos ou mudanças sazonais [Pronobis et al. 2010, Zaffar et al. 2020]. Tais variações frequentemente comprometem a acurácia dos modelos, especialmente quando são treinados com dados rotulados que não contemplam todos os contextos possíveis [Zhang et al. 2021].

Nesse cenário, métodos auto-supervisionados vêm ganhando destaque por sua capacidade de aprender representações visuais discriminativas sem a necessidade de rótulos. O *DINOv2* é um exemplo notável, utilizando uma arquitetura baseada em *Vision Transformers (ViT)* para aprender recursos robustos e transferíveis a partir de grandes volumes de dados não rotulados. Seus autores demonstram que o *DINOv2* supera abordagens auto-supervisionadas anteriores e atinge desempenho competitivo com modelos supervisionados em tarefas como classificação, segmentação e recuperação de instâncias [Oquab et al. 2024].

Paralelamente aos avanços em auto-supervisão, as arquiteturas supervisionadas continuam a evoluir significativamente. A *ConvNeXt*, por exemplo, representa uma atualização moderna das redes convolucionais clássicas, incorporando características inspiradas em *Transformers*, como *kernels* ampliados e camadas invertidas de *bottleneck*, alcançando desempenho competitivo em benchmarks como *ImageNet* e *COCO* [Liu et al. 2022]. De forma semelhante, a *EfficientNet* propõe um método sistemático de escalonamento para melhorar a relação entre desempenho e custo computacional [Tan and Le 2019].

Desse modo, este trabalho apresenta uma análise comparativa entre o modelo auto-supervisionado *DINOv2* e quatro arquiteturas supervisionadas amplamente utilizadas: *ConvNeXt*, *EfficientNet*, *ResNet* e *ViT*. A investigação é conduzida em um cenário realista, utilizando o dataset *KTH-IDOL2*, conhecido por suas variações temporais e de iluminação, que impõem desafios típicos de ambientes internos reais. Como principal contribuição, o estudo fornece uma avaliação sistemática da robustez e capacidade de generalização do *DINOv2* frente a modelos supervisionados, considerando condições visuais que simulam situações práticas enfrentadas em aplicações como robótica e navegação autônoma. Além disso, o trabalho explora cenários com separação entre sessões temporais e mudanças ambientais naturais, uma configuração pouco abordada em estudos anteriores sobre classificação de ambientes com redes profundas.

2. Objetivos

O presente trabalho tem como objetivo principal avaliar e comparar o desempenho do modelo auto-supervisionado *DINOv2* com arquiteturas supervisionadas amplamente utilizadas, como a *ConvNeXt, EfficientNet, ResNet* e *ViT*, na tarefa de classificação visual de ambientes internos, considerando cenários com variação temporal e condições de iluminação distintas, utilizando o dataset *KTH-IDOL2*.

Como objetivos específicos, temos:

- Aplicar o modelo *DINOv2* a dados de ambientes internos sob diferentes condições visuais, investigando sua capacidade de generalização frente a variações temporais e de iluminação;
- Comparar o desempenho do *DINOv2* com modelos supervisionados consolidados, utilizando métricas de acurácia em diferentes cenários de teste;

- Avaliar o impacto da variação de iluminação na estabilidade e precisão dos modelos;
- Investigar a robustez temporal dos modelos, analisando seu desempenho em sequências capturadas em momentos distintos ao longo do tempo;

3. Fundamentação Teórica

A classificação visual de ambientes internos é uma tarefa essencial em sistemas inteligentes, particularmente em aplicações como robótica móvel, navegação autônoma e localização sem mapa. Esses sistemas exigem modelos capazes de lidar com variações estruturais, temporais e visuais, frequentemente presentes em ambientes reais. Para enfrentar esses desafios, a literatura recente tem explorado diferentes estratégias baseadas em aprendizado profundo, com destaque para abordagens supervisionadas, arquiteturas baseadas em *Transformers*, métodos híbridos e, mais recentemente, técnicas de aprendizado auto-supervisionado [Garg et al. 2021, Zaffar et al. 2020].

3.1. Modelos Supervisionados

Modelos supervisionados têm sido amplamente utilizados em tarefas de classificação de cenas e ambientes, com ênfase em redes neurais convolucionais. Trabalhos como o de Zhou et al. [Zhou et al. 2014] introduziram o *dataset Places* e demonstraram que redes treinadas com grandes volumes de dados rotulados podem aprender representações visuais discriminativas. Arquiteturas como *ResNet* [He et al. 2016], EfficientNet [Tan and Le 2019] e ConvNeXt [Liu et al. 2022] destacam-se por sua eficiência e desempenho em benchmarks de classificação. No entanto, tais modelos são geralmente sensíveis a mudanças de domínio, como variações de iluminação e rearranjos no ambiente, o que limita sua robustez em cenários dinâmicos [Barros et al. 2021]].

3.2. Transformers e Arquiteturas Híbridas

A introdução do *Vision Transformer (ViT)* [Dosovitskiy et al. 2021] propôs uma mudança conceitual ao substituir convoluções por mecanismos de atenção, possibilitando o aprendizado de dependências globais desde as primeiras camadas da rede. Essa abordagem tem mostrado resultados promissores em diversas tarefas visuais. No contexto da classificação de ambientes, modelos como o *TransVPR* [Wang et al. 2022] aplicam *ViTs* com múltiplos níveis de atenção para reconhecer lugares internos de forma robusta. No entanto, *ViTs* puros geralmente demandam grandes volumes de dados e alto poder computacional. Como resposta, surgiram arquiteturas híbridas, como o *ConvNeXt*, que incorporam elementos inspirados em *Transformers* à estrutura das redes convolucionais tradicionais [Liu et al. 2022].

3.3. Aprendizado Auto-supervisionado e o DINOv2

Técnicas de aprendizado auto-supervisionado ganharam destaque nos últimos anos por dispensarem rótulos durante o treinamento, o que é vantajoso em contextos com alta variabilidade visual e baixo custo de anotação. O modelo *DINO* [Caron et al. 2021] introduziu um mecanismo de auto-destilação que explora diferentes visões de uma mesma imagem, promovendo a emergência de representações semânticas. Sua evolução, o *DINOv2*, aprimorou esse processo incorporando grandes conjuntos de dados curados, melhorias na arquitetura base e funções de perda mais robustas. O *DINOv2* demonstrou desempenho competitivo com modelos supervisionados em tarefas como classificação de cenas, segmentação e correspondência de instâncias [Oquab et al. 2024].

3.4. Classificação de Ambientes Internos

A classificação de ambientes internos é um campo específico da visão computacional que lida com desafios como variação de iluminação, reorganização de objetos e mudanças temporais. O *dataset KTH-IDOL*, proposto por [Luo et al. 2006], foi desenvolvido justamente para avaliar algoritmos sob essas condições, utilizando imagens coletadas por robôs em diferentes salas e condições ambientais (ensolarado, nublado, noturno). Abordagens multimodais também têm sido exploradas, como a proposta por [Anwer et al. 2019], que combina informações de RGB, profundidade e textura (LBP) em redes convolucionais para melhorar a acurácia da classificação em cenas internas.

Apesar dos avanços, a maioria das abordagens supervisionadas ainda não avalia sistematicamente a capacidade dos modelos de generalizar sob condições visuais não vistas, como aquelas causadas por variações temporais ou mudanças de iluminação. Trabalhos recentes com *Transformers* ou *CNNs* geralmente testam os modelos em condições similares às de treinamento, o que limita a avaliação da robustez dos métodos em contextos reais [Zaffar et al. 2020, Masone and Caputo 2021, Garg et al. 2021].

3.5. Lacunas e Direcionamento deste Trabalho

Embora haja progresso significativo na arquitetura dos modelos de visão, ainda persistem lacunas importantes na avaliação de robustez frente a mudanças ambientais. Poucos estudos utilizam protocolos experimentais com separação temporal entre treino e teste ou troca explícita de condições de iluminação. Este trabalho busca preencher parte dessa lacuna ao comparar o desempenho do modelo auto-supervisionado *DINOv2* com arquiteturas supervisionadas consolidadas (*ConvNeXt, EfficientNet, ResNet e ViT*), em um cenário realista com o dataset *KTH-IDOL2*, focando na generalização visual sob condições variáveis típicas de ambientes internos não controlados.

4. Metodologia

Neste trabalho, a metodologia adotada se fundamenta integralmente no modelo auto-supervisionado *DINOv2*, que representa uma das abordagens mais recentes e robustas no campo do aprendizado profundo aplicado à visão computacional. A escolha por esse modelo como eixo central da investigação se deve à sua capacidade comprovada de aprender representações visuais generalistas, escaláveis e altamente discriminativas, mesmo na ausência de rótulos, o que o torna ideal para tarefas em cenários visuais dinâmicos e desafiadores, como ambientes internos com variações de iluminação e temporalidade [Oquab et al. 2024].

O *DINOv2* é baseado em *Vision Transformers* e opera segundo um mecanismo de auto-destilação sem rótulos, no qual duas redes neurais idênticas em arquitetura, denominadas aluno e professor, interagem durante o treinamento. A rede aluno é otimizada para prever as distribuições de saída da rede professor, que são calculadas a partir de diferentes vistas da mesma imagem. Os parâmetros da rede professor não são aprendidos diretamente, mas atualizados como uma média móvel exponencial (EMA) dos parâmetros da rede aluno, conforme introduzido por [Caron et al. 2021].

As previsões são geradas a partir da aplicação de uma *MLP (Multilayer Perceptron)* aos *tokens* de classe ou de *patch* produzidos pelo *ViT*, seguidas de uma normalização

com *softmax* em diferentes temperaturas. A principal função de perda utilizada é a entropia cruzada entre as distribuições da rede professor (pt) e da rede aluno (ps), expressa por:

$$\mathcal{L}_{\text{DINO}} = -\sum p_t \log p_s \tag{1}$$

O *DINOv2* aprofunda esse processo incorporando uma segunda função de perda baseada no método *iBOT*, que atua no nível dos *patches* mascarados. Com isso, o modelo é incentivado a prever regiões ocultas da imagem, promovendo uma compreensão espacial mais refinada. Para garantir estabilidade e eficiência no treinamento, a metodologia emprega o algoritmo *Sinkhorn-Knopp* como mecanismo de *centering* e o regularizador *KoLeo*, que maximiza a diversidade dos vetores latentes a partir de uma estimativa de entropia diferencial.

Além disso, o *DINOv2* utiliza técnicas computacionais de alto desempenho, como *FlashAttention*, *stochastic depth* otimizado e paralelismo com FSDP (*Fully Sharded Data Parallel*). Tais recursos possibilitam o treinamento eficiente mesmo com modelos de grande porte, como o *ViT-g/14* com mais de 1 bilhão de parâmetros. Em arquiteturas menores, como *ViT-S* e *ViT-L*, é empregada destilação a partir de um professor congelado, permitindo alcançar alta qualidade com menor custo computacional.

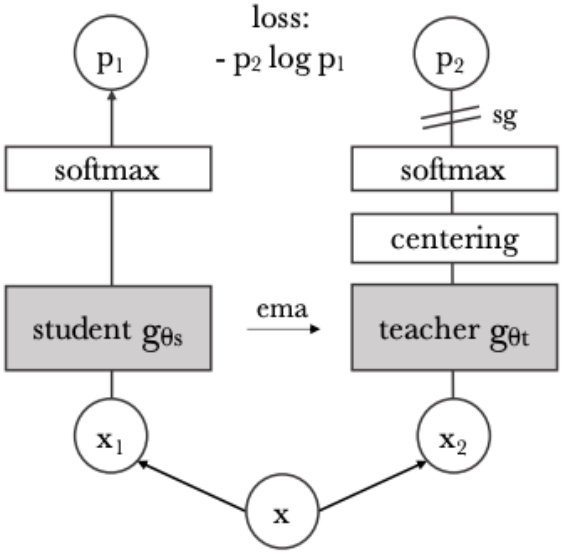


Figure 1. Arquitetura do DINOv2 adaptada de [Oquab et al. 2024].

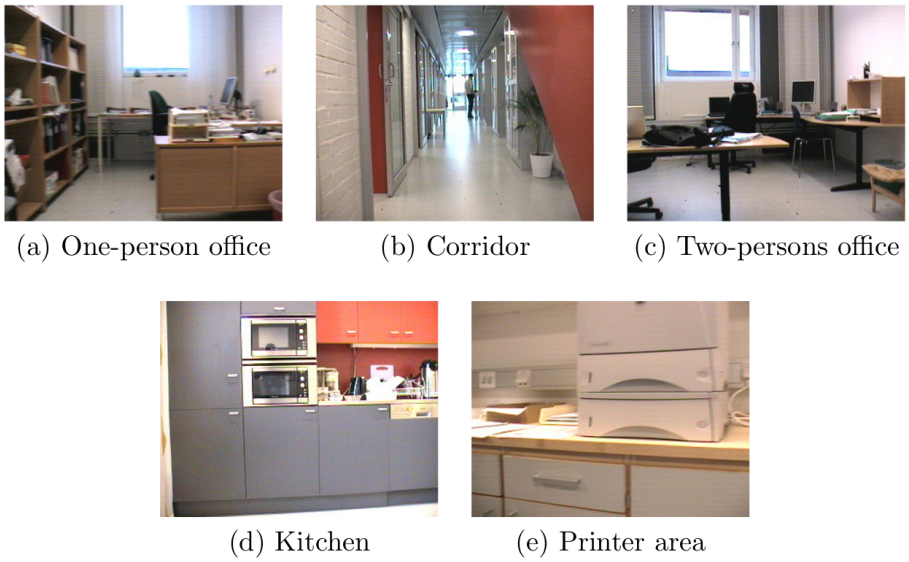
Essa metodologia visa, portanto, avaliar de forma sistemática o desempenho do *DINOv2* em um cenário realista e desafiador, explorando seu potencial para generalização visual em ambientes internos com variações naturais. Ao combinar um modelo de ponta em auto-supervisão com um conjunto de dados rico em diversidade temporal e luminosa, buscamos compreender até que ponto as representações aprendidas pelo *DINOv2* são capazes de sustentar um desempenho robusto frente à complexidade visual do mundo real. A seguir, apresentamos os resultados obtidos e a comparação com modelos supervisionados amplamente utilizados na literatura.

5. Resultados e Discussões

Nesta seção, apresentamos detalhes dos experimentos realizados para avaliar o método proposto.

5.1. Conjunto de Dados

O *KTH-IDOL2* (*Image Database for Indoor and Outdoor Localization*) [Luo et al. 2006] é um conjunto de dados desenvolvido com o objetivo de avaliar a robustez e adaptabilidade de algoritmos de reconhecimento visual de lugares em ambientes internos reais e dinâmicos. O *dataset* foi coletado no laboratório CVAP da KTH (*Royal Institute of Technology*), na Suécia, utilizando dois robôs móveis, *Minnie* e *Dumbo*.



(a) One-person office (b) Corridor (c) Two-persons office

(d) Kitchen (e) Printer area

Figure 2. Imagens apresentando o interior de cada ambiente.

O ambiente de coleta é composto por cinco salas com diferentes funcionalidades: corredor, área da impressora, cozinha e dois escritórios. As sequências de imagens foram registradas enquanto os robôs percorriam trajetórias semelhantes, com variações de ponto de vista devido ao controle manual. Cada imagem foi automaticamente rotulada com a posição e orientação do robô, além do cômodo em que foi capturada.

Um dos principais diferenciais do *KTH-IDOL2* é a ênfase na variação visual causada por fatores reais, sendo as imagens adquiridas sob três condições distintas: a) ensolarado (*sunny*), com forte presença de luz solar, sombras e reflexos; b) nublado (*cloudy*), com luz difusa e ausência de sombras marcantes; e c) noturno (*night*), com baixa iluminação ambiente e predominância de luz artificial.

Sobre as variações temporais e humanas, as coletas ocorreram ao longo de seis meses, permitindo capturar mudanças no ambiente como a presença e ausência de pessoas, alterações na posição de móveis e objetos e modificações na decoração e reorganização de salas. O banco de dados contém 24 sequências de imagens, sendo que cada sequência possui entre 800 e 1100 quadros capturados a 5 fps. Os dados incluem também odometria e varreduras a laser, úteis para estudos que integram percepção visual e localização.

Figure 3. Exemplo de imagens capturadas sob diferentes condições de iluminação.

5.2. Detalhes da Implementação

Para a execução dos experimentos da nossa abordagem, utilizamos a biblioteca *PyTorch* em um computador *Dell* com uma *CPU Intel® XeonT M Silver 4114* de 2,20 GHz, 128 GB de memória principal DDR4-2133 e uma *GPU NVIDIA® GeForce® RTX A4000* de 16 GB GDDR6.

A fase de treinamento do modelo *DINOv2* incluiu a otimização dos hiperparâmetros por meio de *Grid Search*, ajustando variáveis como *learning rate*, *batch size*, *epochs* e *dropout* para maximizar o desempenho do modelo. Para avaliar a capacidade de generalização do modelo, foi adotada uma estratégia de *transfer learning* onde o treinamento e a validação foram realizados com imagens de uma sequência específica, enquanto o teste foi conduzido com imagens de uma sequência distinta, respeitando a separação temporal entre os conjuntos.

5.3. Resultados

Apresentamos os resultados obtidos nos experGeneralização entre Plataformasimentos conduzidos com os modelos *DINOv2* (nossa abordagem), *ConvNeXt* [Liu et al. 2022], *EfficientNet* [Tan and Le 2019], *ResNet* [He et al. 2016] e *ViT* [Dosovitskiy et al. 2021] aplicados à tarefa de classificação de ambientes internos no *dataset KTH-IDOL2*. As análises foram organizadas em duas etapas complementares, com o objetivo de avaliar a robustez dos modelos frente a variações temporais e alterações nas condições de iluminação, ambas características naturais e frequentes em ambientes reais.

Na primeira parte, investigamos a capacidade dos modelos de generalizar ao longo do tempo, ou seja, diante de mudanças na disposição de objetos, presença de pessoas, modificações na decoração e outras transformações naturais que ocorrem entre capturas feitas em diferentes momentos. Na segunda parte, analisamos o impacto das condições de iluminação na acurácia dos modelos, considerando três cenários distintos: ensolarado,

nublado e noturno. A seguir, são apresentados os resultados obtidos em cada cenário experimental, acompanhados de análises comparativas entre os modelos e discussão dos principais resultados.

5.3.1. Avaliação Temporal

Nesta primeira análise, avaliamos a capacidade dos modelos de generalizar sob variações temporais em ambientes internos, considerando mudanças naturais que ocorrem com o passar do tempo, como alterações na disposição de objetos, movimentação de móveis, presença ocasional de pessoas e modificações na iluminação ambiente. Adotamos um protocolo experimental em que o treinamento e a validação foram realizados utilizando as sequências 1 e 2, enquanto o teste foi conduzido com as sequências 3 e 4, capturadas meses depois, no mesmo ambiente. Dessa forma, é possível avaliar a robustez dos modelos frente à evolução visual dos espaços, evitando o sobreajuste a padrões momentâneos.

Dumbo12	*sunny34*	*cloudy34*	*night34*
Modelo	**Acurácia (%)**	**Acurácia (%)**	**Acurácia (%)**
ViT	91.84	92.72	92.30
EfficientNetB0	93.95	91.02	94.22
Resnet152	94.56	92.62	94.01
ConvNeXt-base	95.64	95.36	95.82
DINOv2	**96.77**	**97.06**	**96.33**

Table 1. Resultados dos experimentos ao treinar e testar sob a mesma condição de iluminação, mas utilizando sequências diferentes, para a plataforma *Dumbo*.

Na Tabela 1, apresentamos os resultados obtidos na plataforma *Dumbo*, utilizando as sequências 1 e 2 para treinamento e validação, e as sequências 3 e 4 para teste, mantendo a mesma condição de iluminação entre os conjuntos. Essa configuração permite avaliar o impacto da variação temporal, isolando outros fatores como mudança de domínio visual ou iluminação.

Observa-se que todos os modelos supervisionados apresentam desempenho satisfatório, com acurácia acima de 91% em todos os cenários. No entanto, o modelo *DINOv2* superou todas as demais arquiteturas em todas as condições de iluminação avaliadas, alcançando 97,06% de acurácia no cenário nublado (*cloudy*) e mantendo desempenho elevado mesmo na condição ensolarada (96,77%, *sunny*) e noturna (96,33%, *night*).

A Tabela 2 apresenta os resultados obtidos na plataforma *Minnie*, também considerando treinamento e validação nas sequências 1 e 2, e teste nas sequências 3 e 4, com a mesma condição de iluminação entre os conjuntos. Embora os valores de acurácia sejam ligeiramente inferiores aos obtidos na plataforma *Dumbo*, o modelo *DINOv2* manteve-se como o de melhor desempenho em todas as condições.

O *DINOv2* atingiu 93,75% de acurácia sob condição ensolarada (*sunny*), 90,80% em nublado (*cloudy*) e 92,93% à noite (*night*), superando os modelos supervisionados em todos os casos. Os demais modelos apresentaram maior variação de desempenho, com quedas mais acentuadas especialmente em ambientes nublados, como observado no *EfficientNetB0* (83,68%) e no *ViT* (86,20%).

Minnie	*sunny34*	*cloudy34*	*night34*
Modelo	**Acurácia (%)**	**Acurácia (%)**	**Acurácia (%)**
EfficientNetB0	85.03	83.68	91.92
ConvNeXt-base	91.23	89.41	89.12
Resnet152	91.91	86.94	87.66
ViT	92.49	86.20	89.90
DINOv2	**93.75**	**90.80**	**92.93**

Table 2. Resultados dos experimentos ao treinar e testar sob a mesma condição de iluminação, mas utilizando sequências diferentes, para a plataforma *Minnie*.

Esses resultados reforçam o comportamento consistente do *DINOv2* frente às mudanças temporais, mesmo em uma plataforma com câmera em altura diferente e possíveis variações de perspectiva, como é o caso da *Minnie*. A robustez apresentada pelo modelo auto-supervisionado o destaca como uma solução promissora para aplicações de classificação visual em ambientes reais, sujeitos a modificações ao longo do tempo.

5.3.2. Avaliação por Condições de Iluminação

Nesta etapa, investigamos a robustez dos modelos frente a diferentes condições de iluminação, um fator crítico para tarefas de classificação visual em ambientes reais. A variação na iluminação afeta diretamente a aparência das cenas, com alterações em contraste, sombras, reflexos e intensidade luminosa, características que podem comprometer a estabilidade dos modelos supervisionados tradicionais.

Adotamos um protocolo em que os modelos foram treinados e validados em uma determinada condição de iluminação (por exemplo, ensolarado) e testados em outra distinta (por exemplo, nublado ou noturno), sempre utilizando sequências diferentes para garantir separação temporal. Essa abordagem permite avaliar até que ponto cada modelo consegue generalizar visualmente para cenários com iluminação diversa, sem exposição prévia a essas variações.

A Tabela 3 apresenta os resultados obtidos na plataforma *Dumbo*, considerando o cenário de variação entre condições de iluminação. Neste experimento, os modelos foram treinados com dados de duas condições de iluminação (por exemplo, *sunny* e *cloudy*) e testados com dados de uma terceira condição não vista durante o treinamento (*night*). Essa configuração permite avaliar diretamente a capacidade de generalização visual dos modelos a condições de iluminação desconhecidas.

Observa-se que, embora todos os modelos supervisionados tenham apresentado desempenho razoável nas condições ensolarada e nublada, houve uma queda significativa quando testados em ambiente noturno, com destaque para o *ViT*, que obteve apenas 81,24% de acurácia nesse cenário. O modelo *DINOv2*, por outro lado, demonstrou excelente estabilidade entre as condições, atingindo 98,21% na condição ensolarada (*sunny*), 96,94% em nublado (*cloudy*) e 93,81% à noite (*night*), a maior entre todos os modelos em todos os casos.

Esses resultados reforçam a robustez do *DINOv2* a variações de iluminação, evidenciando sua capacidade de aprender representações visuais menos sensíveis a artefatos

Dumbo	*sunny3*	*cloudy3*	*night3*
Modelo	**Acurácia (%)**	**Acurácia (%)**	**Acurácia (%)**
ViT	88.74	87.54	81.24
EfficientNetB0	91.47	94.97	86.65
Resnet152	94.84	95.41	88.10
ConvNeXt-base	95.26	95.63	92.75
DINOv2	**98.21**	**96.94**	**93.81**

Table 3. Resultados dos experimentos ao treinar com as sequências 1 e 2 de diferentes condições de iluminação e teste com a sequência 3 condição restante, para a plataforma *Dumbo*.

luminosos, como sombras e variações de intensidade. Essa característica é especialmente relevante para aplicações em ambientes não controlados, onde mudanças nas condições de iluminação ocorrem de forma imprevisível.

A Tabela 4 apresenta os resultados obtidos na plataforma *Minnie*, seguindo o mesmo protocolo de avaliação por condição de iluminação: os modelos foram treinados com duas das três condições disponíveis e testados com a condição restante, nunca vista durante o treinamento. Assim como na plataforma *Dumbo*, os resultados evidenciam uma diferença de desempenho entre os modelos supervisionados e o modelo autosupervisionado *DINOv2*.

Minnie	*sunny3*	*cloudy3*	*night3*
Modelo	**Acurácia (%)**	**Acurácia (%)**	**Acurácia (%)**
ConvNeXt-base	80.18	92.51	90.11
EfficientNetB0	81.85	86.91	89.78
ViT	84.81	87.81	82.83
Resnet152	85.01	92.95	87.83
DINOv2	**92.70**	**94.85**	**93.59**

Table 4. Resultados dos experimentos ao treinar com as sequências 1 e 2 de diferentes condições de iluminação e teste com a sequência 3 condição restante, para a plataforma *Minnie*.

Embora alguns modelos supervisionados tenham apresentado resultados competitivos em condições específicas — como a *ResNet152* em nublado (92,95%, *cloudy*), a maioria demonstrou maior sensibilidade às mudanças de iluminação, com quedas expressivas no cenário ensolarado. O *ConvNeXt-base*, por exemplo, obteve apenas 80,18% nesse cenário, enquanto o *ViT* teve desempenho ainda inferior à noite (82,83%, *night*).

O modelo *DINOv2* novamente se destacou, com os melhores resultados em todas as condições: 92,70% (*sunny*), 94,85% (*cloudy*) e 93,59% (*night*). A consistência apresentada pelo *DINOv2*, mesmo diante de uma plataforma com características visuais distintas, como a posição da câmera da *Minnie*, reforça sua capacidade de abstração visual e generalização frente a variações luminosas.

Os resultados apresentados nesta seção demonstram de forma consistente que o modelo *DINOv2* supera as arquiteturas supervisionadas em cenários com variação de iluminação, tanto na plataforma *Dumbo* quanto na *Minnie*. Enquanto os modelos su-

pervisionados mostraram maior oscilação de desempenho, especialmente em condições mais extremas como ambientes ensolarados e noturnos, o *DINOv2* manteve alta acurácia e estabilidade entre diferentes cenários, mesmo quando exposto a condições de teste não vistas durante o treinamento.

Esses achados indicam que as representações aprendidas de forma auto-supervisionada pelo *DINOv2* são menos sensíveis a artefatos visuais induzidos pela iluminação, como sombras, reflexos e baixa luminosidade. Tal robustez é especialmente desejável em aplicações do mundo real, como robótica móvel e sistemas autônomos, nos quais a variabilidade das condições visuais é inevitável.

Portanto, a avaliação por condição de iluminação reforça as evidências de que o *DINOv2* não apenas atinge melhor desempenho em termos de acurácia, mas também oferece maior generalização e adaptabilidade visual, consolidando-se como uma alternativa promissora para tarefas de classificação de ambientes em cenários não controlados.

6. Considerações Finais

Este trabalho apresentou uma análise comparativa entre o modelo auto-supervisionado *DINOv2* e arquiteturas supervisionadas amplamente utilizadas, como *ConvNeXt*, *EfficientNet*, *ResNet* e *ViT*, aplicadas à tarefa de classificação de ambientes internos utilizando o *dataset KTH-IDOL2*. Os experimentos foram conduzidos com foco em dois desafios comuns em ambientes reais: variação temporal e alterações nas condições de iluminação.

Os resultados mostraram que, embora os modelos supervisionados apresentem desempenho competitivo em cenários controlados, o *DINOv2* obteve acurácia superior e mais estável em todas as condições avaliadas. Em cenários com mudanças temporais, que envolvem modificações no ambiente ao longo do tempo, e sob diferentes condições de iluminação (ensolarado, nublado e noturno), o *DINOv2* demonstrou maior robustez e capacidade de generalização, alcançando acurácia de até 98,21%.

Esses achados reforçam o potencial do aprendizado auto-supervisionado como uma abordagem eficaz para tarefas de classificação visual em ambientes dinâmicos e não controlados, onde o acesso a grandes volumes de dados rotulados nem sempre é viável. O desempenho consistente do *DINOv2* diante de diferentes fontes de variação visual indica que esse tipo de modelo pode ser especialmente útil em aplicações como robótica móvel, navegação autônoma e sistemas inteligentes de percepção visual.

Como trabalho futuro, pretende-se explorar a combinação de representações auto-supervisionadas com técnicas de agregação espacial, além da avaliação em outros *datasets* e cenários de domínio cruzado, ampliando o escopo da análise para contextos ainda mais desafiadores.

References

Anwer, R. M., Khan, F. S., Laaksonen, J., and Zaki, N. (2019). Multi-stream convolutional networks for indoor scene recognition. In *Computer Analysis of Images and Patterns: 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Proceedings, Part I 18*, pages 196–208. Springer.

Barros, T., Pereira, R., Garrote, L., Premebida, C., and Nunes, U. J. (2021). Place recognition survey: An update on deep learning approaches. *CoRR*, abs/2106.10458.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Garg, S., Fischer, T., and Milford, M. (2021). Where is your place, visual place recognition? *CoRR*, abs/2103.06443.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, J., Pronobis, A., Caputo, B., and Jensfelt, P. (2006). The KTH-IDOL2 Database. Technical Report CVAP304, KTH Royal Institute of Technology, CVAP/CAS, Stockholm, Sweden.

Masone, C. and Caputo, B. (2021). A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547.

Oquab, M., Darcet, T., Moutakanni, T., Ramé, A., Taylor, L., Misra, I., and Caron, M. (2024). Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, published online.

Pronobis, A., Jie, L., and Caputo, B. (2010). The more you learn, the less you store: Memory-controlled incremental svm for visual place recognition. *Image and Vision Computing*, 28(7):1080–1097.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR.

Wang, R., Shen, Y., Zuo, W., Zhou, S., and Zheng, N. (2022). Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657.

Zaffar, M., Ehsan, S., Milford, M., Flynn, D., and McDonald-Maier, K. D. (2020). Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *CoRR*, abs/2005.08135.

Zhang, X., Wang, L., and Su, Y. (2021). Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 113:107760.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.