

UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO- ICOMP

PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Fusão Multimodal via Atenção Cruzada Bidirecional para Mapeamento Semântico de Ambientes Internos

Alternei de Souza Brito

Manaus - AM

Janeiro de 2026

Alternei de Souza Brito

Fusão Multimodal via Atenção Cruzada Bidirecional para Mapeamento Semântico de Ambientes Internos

Dissertação apresentada ao Programa de Pós-graduação em Informática, Instituto de Computação, da Universidade Federal do Amazonas, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador

Dr. Felipe Gomes de Oliveira

Universidade Federal do Amazonas - UFAM

Instituto de Computação- ICOMP

Manaus - AM

Janeiro de 2026

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

B862f Brito, Alternei de Souza
Fusão Multimodal via Atenção Cruzada Bidirecional para Mapeamento Semântico de Ambientes Internos / Alternei de Souza Brito. - 2026.
117 f. : il., color. ; 31 cm.

Orientador(a): Felipe Gomes de Oliveira.
Dissertação (mestrado) - Universidade Federal do Amazonas, Programa de Pós-Graduação em Informática, Manaus, 2026.

1. mapeamento semântico. 2. aprendizado auto-supervisionado. 3. fusão multi- modal. 4. atenção cruzada. 5. robótica móvel. I. Oliveira, Felipe Gomes de. II. Universidade Federal do Amazonas. Programa de Pós-Graduação em Informática. III. Título

---

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"FUSÃO MULTIMODAL VIA ATENÇÃO CRUZADA BIDIRECIONAL PARA MAPEAMENTO SEMÂNTICO DE AMBIENTES INTERNOS"

ALTERNEI DE SOUZA BRITO

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Felipe Gomes de Oliveira - **Presidente**

Prof. Dr. José Luiz de Souza Pio - **Membro Interno**

Prof. Dr. Anselmo Rafael Cukla - **Membro Externo**

Manaus, 28 de janeiro de 2026.

Documento assinado eletronicamente por **Anselmo Rafael Cukla**, **Usuário Externo**, em 02/02/2026, às 15:50, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Jose Luiz de Souza Pio**, **Professor do Magistério Superior**, em 02/02/2026, às 20:12, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **Felipe Gomes de Oliveira**, **Professor do Magistério Superior**, em 03/02/2026, às 11:57, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2983115** e o código CRC **BD82F14F**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

À minha mãe, que com coragem e amor assumiu todos os papéis, sendo alicerce, direção e abrigo.

Agradecimentos

Agradeço, primeiramente a Deus, por me conceder saúde, força, sabedoria e perseverança ao longo desta caminhada, tornando possível a superação dos desafios enfrentados durante o desenvolvimento deste trabalho.

Agradeço ao meu orientador, Professor Felipe Gomes de Oliveira, pela orientação, disponibilidade, confiança e valiosas contribuições ao longo de todo o processo de pesquisa, fundamentais para meu amadurecimento acadêmico e científico.

À minha família, expresso minha profunda gratidão pelo apoio incondicional e incentivo constantes. Em especial, agradeço à minha esposa, Lisania Brito, pelo companheirismo, paciência e apoio diário; aos meus filhos que foram fonte permanente de motivação e inspiração; à minha mãe, Ana Brito, pelo amor, incentivo e ensinamentos que sustentaram minha trajetória.

Agradeço ao Programa de Pós-Graduação em Informática (PPGI) e ao Instituto de Computação (IComp), da Universidade Federal do Amazonas, pelo suporte institucional favorável à realização desta pesquisa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD, como também pelo Projeto de P&D Inteligência Artificial Avançada para Sistemas Móveis Neuromórficos – Mob4Ai.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

"O Senhor é a minha força e o meu escudo; nele o meu coração confia, e dele recebo ajuda."

Salmos 28:7

Fusão Multimodal via Atenção Cruzada Bidirecional para Mapeamento Semântico de Ambientes Internos

Autor: Alternei de Souza Brito

Orientador: Dr. Felipe Gomes de Oliveira

Resumo

O mapeamento semântico é fundamental para robôs móveis, pois fornece uma representação espacial associada a rótulos semânticos que favorecem navegação e tomada de decisão contextualizada. Esta dissertação aborda o mapeamento semântico de ambientes internos tratando-o como classificação supervisionada de cenas a partir de imagens *RGB* e profundidade monocular estimada. Propõe-se uma arquitetura multimodal baseada em representações auto-supervisionadas e fusão por Atenção Cruzada Bidirecional entre aparência e geometria. A abordagem é avaliada nos *datasets KTH-IDOL2* e *COLD* sob protocolos de generalização temporal, variação e interpolação de iluminação, e transferência entre domínios. Os resultados mostram desempenho elevado e ganhos consistentes da fusão multimodal em cenários desafiadores, além de maior coerência espacial nos mapas semânticos 2D gerados a partir das predições. Essa estratégia contribui para aumentar a confiabilidade da percepção em robôs móveis em ambientes internos reais, favorecendo navegação e tomada de decisão sob mudanças fotométricas e variações de domínio, sem exigir sensores adicionais além de uma câmera *RGB*.

Palavras-chave: mapeamento semântico; aprendizado auto-supervisionado; fusão multimodal; atenção cruzada; robótica móvel.

Fusão Multimodal via Atenção Cruzada Bidirecional para Mapeamento Semântico de Ambientes Internos

Autor: Alternei de Souza Brito

Orientador: Dr. Felipe Gomes de Oliveira

Abstract

Semantic mapping is fundamental for mobile robots, as it provides a spatial representation associated with semantic labels that support navigation and context-aware decision-making. This dissertation addresses indoor semantic mapping by treating it as supervised scene classification from *RGB* images and estimated monocular depth. We propose a multimodal architecture based on self-supervised representations and Bidirectional Cross-Attention fusion between appearance and geometry. The approach is evaluated on *KTH-IDOL2* and *COLD* under protocols of temporal generalization, illumination variation and interpolation, and cross-domain transfer. Results show high performance and consistent gains from multimodal fusion in challenging scenarios, as well as greater spatial coherence in the 2D semantic maps generated from the predictions. This strategy increases the reliability of perception for mobile robots in real indoor environments, supporting navigation and decision-making under photometric changes and domain shifts, without requiring additional sensors beyond an *RGB* camera.

Keywords: semantic mapping; self-supervised learning; multimodal fusion; cross-attention.

LISTA DE ILUSTRAÇÕES

Figura 1 – Mapa semântico conceitual *2D*: o robô percorre o ambiente, classifica cenas ao longo da trajetória e projeta as predições no plano para compor o mapa semântico. 22
Figura 2 – Exemplo ilustrativo de mapa semântico em ambiente interno apresentado por (SüNDERHAUF et al., 2016). 29
Figura 3 – Arquitetura baseada em *CNN* para a classificação de cenas, proposta por (WANG et al., 2019) 30
Figura 4 – Resultados do *Depth Anything v2* comparados a outros modelos (YANG et al., 2024) 31
Figura 5 – Arquitetura de fusão de dados *RGB* e profundidade proposta por (WU et al., 2024b) 32
Figura 6 – Visão geral da arquitetura proposta *MFISC* + Mapeamento Semântico *2D*. A partir de pares *RGB* e profundidade (estimada), extraem-se representações com *DINOv2*; projeções e objetivos auxiliares favorecem alinhamento e complementaridade entre modalidades; a fusão ocorre via Atenção Cruzada Bidirecional; e as predições de cena são projetadas na trajetória do robô para compor mapas semânticos 2D. . 46
Figura 7 – Exemplo de par multimodal gerado a partir do *dataset COLD*. O mapa de profundidade foi estimado pelo modelo *Depth Anything v2*, normalizado e adaptado para compatibilidade com o *backbone DINOv2*. 47

Figura 8 – Mapa semântico 2D gerado para a sequência *Freiburg-night34* do conjunto *COLD* (*Ground-Truth*). Cada ponto da trajetória é colorido de acordo com a classe de cena associada. 53
Figura 9 – Exemplo de imagens dos ambientes retiradas de uma das sequências apresentando o interior das salas, para o conjunto *KTH-IDOL2* (LUO et al., 2006). 55
Figura 10 – Exemplo de imagens capturadas sob diferentes condições de ilumina-ção, para o conjunto *KTH-IDOL2* (LUO et al., 2006). 55
Figura 11 – Exemplos de imagens apresentando os interiores de algumas salas em cada um dos três laboratórios, para o conjunto *COLD* (PRONOBIS; CAPUTO, 2009). 56
Figura 12 – Resultados do Esquema 1 no *KTH-IDOL2* (plataforma *Dumbo*): acurá-cia (%) do *MFISC* em generalização temporal, treinando nas sequên-cias *1–2* e testando em *3–4* sob *sunny34*, *night34* e *cloudy34*. 65
Figura 13 – Resultados do Esquema 1 no *KTH-IDOL2* (plataforma *Minnie*): acurá-cia (%) do *MFISC* em generalização temporal, treinando nas sequên-cias *1–2* e testando em *3–4* sob *sunny34*, *night34* e *cloudy34*. 66
Figura 14 – Esquema 1 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* para cada condição (*sunny3*, *night3*, *cloudy3*). 67
Figura 15 – Esquema 1 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* por condição. 68
Figura 16 – Esquema 1 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* por condição. . 69
Figura 17 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *sunny34* e *cloudy34*. . 70
Figura 18 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *night34* e *cloudy34*. . 71
Figura 19 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *sunny3* e *cloudy3*. 72

Figura 20 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*. 73
Figura 21 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*. . . . 74
Figura 22 – Esquema 3 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny34* (treino *cloudy1+night2*); teste em *night34* (treino *cloudy1+sunny2*); teste em *cloudy34* (treino *sunny1+night2*). 76
Figura 23 – Esquema 3 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny34* (treino *cloudy1+night2*); teste em *night34* (treino *cloudy1+sunny2*); teste em *cloudy34* (treino *sunny1+night2*). 77
Figura 24 – Esquema 3 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*). 78
Figura 25 – Esquema 3 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*). 78
Figura 26 – Esquema 3 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*). 79
Figura 27 – Esquema 4 no *KTH-IDOL2* (transferência *Dumbo* → *Minnie*). Acurácia do *MFISC* variando apenas o *backbone*. 81
Figura 28 – Esquema 4 no *KTH-IDOL2* (transferência *Minnie* → *Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*. 82
Figura 29 – Esquema 4 no *COLD* (transferência *Freiburg* → *Ljubljana*). Treino em *Freiburg-sunny12* e teste em *Ljubljana-cloudy3* e *Ljubljana-night3*. . . . 83

Figura 30 – Esquema 4 no *COLD* (transferência *Ljubljana* → *Saarbrücken*). Treino em *Ljubljana-cloudy12* e teste em *Saarbrücken-night3* e *Saarbrücken-sunny3*. 84
Figura 31 – Esquema 4 no *COLD* (transferência *Saarbrücken* → *Freiburg*). Treino em *Saarbrücken-night12* e teste em *Freiburg-cloudy3* e *Freiburg-sunny3*. 84
Figura 32 – Mapas semânticos *2D* na sequência *Minnie/night34* (*KTH-IDOL2*, Esquema 2). Comparação entre (a) *ground-truth*, (b) predição com *MFISC+ConvNeXt* e (c) predição com *MFISC+DINOv2*. As siglas no quadro de cores correspondem às classes do *dataset*. 92
Figura 33 – Mapas semânticos *2D* na sequência *Freiburg/night3* (*COLD*, Esquema 2). Comparação entre (a) *ground-truth*, (b) predição com *MFISC+ConvNeXt* e (c) predição com *MFISC+DINOv2*. As siglas no quadro de cores correspondem às classes do *dataset* 94
Figura 34 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *sunny34* e *night34*. . 105
Figura 35 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *night34* e *cloudy34*. . 105
Figura 36 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *sunny34* e *night34*. . 106
Figura 37 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy34* e *sunny34*. . 106
Figura 38 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*. 107
Figura 39 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*. 107
Figura 40 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*. 108
Figura 41 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy3* e *sunny3*. 108

Figura 42 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*. . . . 109
Figura 43 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy3* e *sunny3*. . . . 109
Figura 44 – Esquema 4 no *COLD* (*Freiburg* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-cloudy12* e teste em *Ljubljana-night3* e *Ljubljana-sunny3*. . . . 110
Figura 45 – Esquema 4 no *COLD* (*Freiburg* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-night12* e teste em *Ljubljana-cloudy3* e *Ljubljana-sunny3*. . . . 111
Figura 46 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-cloudy12* e teste em *Saarbrücken-night3* e *Saarbrücken-sunny3*. . . . 111
Figura 47 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-night12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-sunny3*. . . . 112
Figura 48 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-sunny12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-night3*. . . . 112
Figura 49 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-cloudy12* e teste em *Freiburg-night3* e *Freiburg-sunny3*. . . . 113
Figura 50 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-night12* e teste em *Freiburg-cloudy3* e *Freiburg-sunny3*. . . . 113
Figura 51 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-sunny12* e teste em *Freiburg-cloudy3* e *Freiburg-night3*. . . . 114

Figura 52 – Esquema 4 no *COLD* (*Ljubljana* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-night12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-sunny3*. 114
Figura 53 – Esquema 4 no *COLD* (*Ljubljana* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-sunny12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-night3*. 115
Figura 54 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-cloudy12* e teste em *Ljubljana-night3* e *Ljubljana-sunny3*. 115
Figura 55 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-night12* e teste em *Ljubljana-cloudy3* e *Ljubljana-sunny3*. 116
Figura 56 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-sunny12* e teste em *Ljubljana-cloudy3* e *Ljubljana-night3*. 116
Figura 57 – Esquema 4 no *COLD* (*Saarbrücken* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-cloudy12* e teste em *Freiburg-night3* e *Freiburg-sunny3*. 117
Figura 58 – Esquema 4 no *COLD* (*Saarbrücken* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-sunny12* e teste em *Freiburg-cloudy3* e *Freiburg-night3*. 117

LISTA DE TABELAS

Tabela 1 – Síntese comparativa dos trabalhos relacionados e posicionamento da abordagem proposta. 43
Tabela 2 – Cenários experimentais e classes semânticas por conjunto de dados (e por local no *COLD*). 57
Tabela 3 – Resumo do protocolo experimental para avaliação da abordagem proposta. 61
Tabela 4 – Estudo de ablação no *KTH-IDOL2* (*Minnie*) sob o Esquema 2. Em todos os casos, o *backbone* é o *DINOv2*; variam-se apenas os componentes do *MFISC*. 86
Tabela 5 – Estudo de ablação no *COLD* (*Ljubljana*) sob o Esquema 2. Em todos os casos, o *backbone* é o *DINOv2*; variam-se apenas os componentes do *MFISC*. 88
Tabela 6 – Comparação com *baselines* no *KTH-IDOL2* (*Minnie*) sob o mesmo cenário do ablation (Esquema 2). 89
Tabela 7 – Comparação com *baselines* no *COLD* (*Ljubljana*) sob o mesmo cenário do ablation (Esquema 2). 90

LISTA DE ABREVIATURAS E SIGLAS

CBAM *Convolutional Block Attention Module*

CNN *Convolutional Neural Network*

DINOv2 *Distillation with No Labels v2*

GNNs *Graph Neural Networks*

LiDAR *Light Detection and Ranging*

MLP *Multi Layer Perceptron*

RGB *Red, Green, Blue*

RGB-D *Red, Green, Blue, Depth*

ROS *Robot Operating System*

SLAM *Simultaneous Localization and Mapping*

SUMÁRIO

1 INTRODUÇÃO 21
1.1 Contextualização 21
1.2 Definição do Problema e Justificativa 23
1.3 Objetivos 24
1.4 Contribuições 25
1.5 Organização do Trabalho 27

2 FUNDAMENTOS 28
2.1 Mapeamento Semântico em Robótica Móvel 28
2.2 Classificação de Cenas e Representações Visuais 29
2.3 Estimativa de Profundidade Monocular 31
2.4 Fusão de Representações Visuais 32
2.5 Classificação Supervisionada Multimodal 33

3 TRABALHOS RELACIONADOS 34
3.1 Mapeamento Semântico em Ambientes Internos 34
3.2 Reconhecimento Visual de Lugares e Classificação de Cenas 36
3.3 Fusão Multimodal 38
3.4 Eficiência Computacional para Robôs Móveis 40
3.5 Síntese Crítica dos Trabalhos Relacionados 41

4 METODOLOGIA 44
4.1 Arquitetura da Abordagem Proposta 45
4.1.1 Geração dos Mapas de Profundidade 46
4.1.2 Extração de Características com o *DINOv2* 47
4.1.3 Projeções Específicas por Modalidade 48

4.1.4 Alinhamento e Regularização 49
4.1.4.1 Alinhamento Contrastivo 49
4.1.4.2 Regularização por Ortogonalidade 50
4.1.5 Mecanismo de Fusão por Atenção Cruzada Bidirecional 51
4.1.6 A Etapa de Classificação Final 51
4.1.7 Geração dos Mapas Semânticos 2D 52
4.2 Conjuntos de Dados 53
4.2.1 *KTH-IDOL2* (*Image Database for rObot Localization*) 54
4.2.2 *COLD* (*COsy Localization Database*) 54
4.2.3 Pré-processamento dos Dados 56
4.3 Configuração Experimental 58
4.3.1 Ambiente Computacional 58
4.3.2 Parâmetros de Otimização e Treinamento 58
4.3.3 Estratégias de Regularização e Critérios de Parada 59
4.4 Protocolo e Delineamento Experimental 59
4.4.1 Estrutura da Avaliação e Modelos 60
4.4.2 Cenários de Teste, Divisão de Dados e Métricas 61

5 RESULTADOS E ANÁLISE DOS EXPERIMENTOS 62
5.1 Esquema 1: Generalização Temporal na Mesma Condição de Iluminação 64
5.1.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 1 65
5.1.2 Resultados no *Dataset COLD* para o Esquema 1 67
5.2 Esquema 2: Generalização para Condições de Iluminação Não Vistas 69
5.2.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 2 70
5.2.2 Resultados no *Dataset COLD* para o Esquema 2 72
5.3 Esquema 3: Interpolação entre Iluminações 74
5.3.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 3 75
5.3.2 Resultados no *Dataset COLD* para o Esquema 3 77

5.4 Esquema 4: Generalização entre Domínios (Locais e Plataformas) . . . 80
5.4.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 4 . . . 81
5.4.2 Resultados no *Dataset COLD* para o Esquema 4 . . . 82
5.5 Estudo de Ablação . . . 85
5.5.1 Estudo de Ablação no *KTH-IDOL2* (*Minnie*) . . . 86
5.5.2 Estudo de Ablação no *COLD* (*Ljubljana*) . . . 87
5.5.3 Síntese e implicações . . . 88
5.6 Comparação com *Baselines* da Literatura . . . 89
5.6.1 Resultados no *KTH-IDOL2* (*Minnie*) . . . 89
5.6.2 Resultados no *COLD* (*Ljubljana*) . . . 90
5.7 Mapas Semânticos . . . 90
5.7.1 Construção dos mapas e convenções de visualização . . . 91
5.7.2 Resultados qualitativos em *KTH-IDOL2* (*Minnie/night34*) . . . 91
5.7.3 Resultados qualitativos em *COLD* (*Freiburg/night3*) . . . 92
5.7.4 Síntese . . . 93

6 CONSIDERAÇÕES FINAIS . . . 95
6.1 Conclusões . . . 95
6.2 Trabalhos Futuros . . . 97

Referências . . . 99

APÊNDICE A RESULTADOS COMPLEMENTARES . . . 104
A.1 Esquema 2: Generalização para Condições de Iluminação Não Vistas . . . 104
A.1.1 KTH-IDOL2 . . . 104
A.1.2 COLD . . . 104
A.2 Esquema 4: Generalização entre Domínios . . . 110
A.2.1 COLD: Transferência entre Locais . . . 110

1

INTRODUÇÃO

O mapeamento semântico em ambientes internos é fundamental para que robôs móveis naveguem e tomem decisões com segurança, pois permite associar significado às regiões observadas ao longo da exploração. Neste trabalho, formulamos o mapeamento semântico como um processo em duas etapas: (i) categorização semântica de cenas a partir de informação visual e geométrica (imagens *RGB* e profundidade) e (ii) projeção dessas predições sobre a trajetória do robô para compor mapas semânticos *2D* mais robustos e consistentes.

1.1 Contextualização

A compreensão semântica de ambientes internos é um requisito para que robôs móveis naveguem e tomem decisões de forma segura, eficiente e inteligente. No contexto de mapeamento semântico, o robô precisa associar significado às regiões observadas e consolidar essas informações ao longo da trajetória, formando um mapa interpretável do ambiente (BALASKA et al., 2020). Uma forma prática de realizar isso é tratar a atribuição de categorias semânticas como um problema de classificação de cenas, vinculando cada predição à pose do robô para compor o mapa.

Apesar dos avanços em *CNNs* e modelos modernos de representação visual, ainda é desafiador manter desempenho sob variações de iluminação e mudanças entre domínios, como diferentes estruturas, rotas e condições ambientais, além de integrar de forma coerente múltiplas fontes de informação com custo computacional compatível

com robôs (CABRERA et al., 2024b; PELTOMÄKI, 2023). Trabalhos recentes indicam que a fusão multimodal pode enriquecer a representação e melhorar a robustez, sobretudo ao combinar *RGB* e profundidade (LI; WANG, 2024; SILVA et al., 2023; PAPAPETROS et al., 2023). Neste trabalho, consideramos especificamente *RGB* e mapas de profundidade derivados das próprias imagens, evitando sensores adicionais.

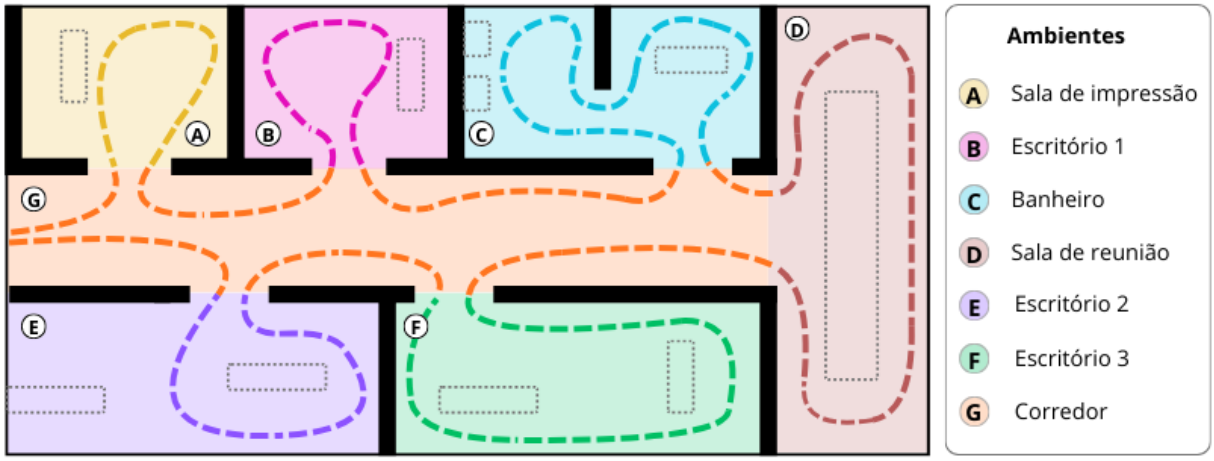


Figura 1 – Mapa semântico conceitual *2D*: o robô percorre o ambiente, classifica cenas ao longo da trajetória e projeta as predições no plano para compor o mapa semântico.

Para enfrentar esses desafios, propomos um arcabouço multimodal que integra *RGB* e profundidade para categorização semântica de cenas internas e posterior composição de mapas semânticos *2D* ao longo da trajetória do robô. A profundidade é estimada a partir das próprias imagens, garantindo aplicabilidade em cenários com apenas câmera *RGB*. O núcleo do método explora representações auto-supervisionadas e introduz mecanismos para (i) adaptar e alinhar as modalidades em um espaço latente compartilhado, (ii) preservar complementaridade entre aparência e geometria, e (iii) promover interação dinâmica entre as modalidades por Atenção Cruzada Bidirecional.

A eficácia do método é avaliada nos *datasets KTH-IDOL2* e *COLD*, considerando cenários desafiadores de variação de iluminação e generalização entre ambientes, além de estudo de ablação e análise qualitativa dos mapas semânticos gerados.

1.2 Definição do Problema e Justificativa

A construção de mapas semânticos *2D* robustos e informativos a partir da percepção visual de robôs móveis em ambientes internos permanece um desafio central em robótica móvel. Em cenários reais, variações de iluminação, mudanças estruturais e dinâmica do ambiente degradam a consistência das predições ao longo da trajetória, comprometendo a coerência espacial do mapa gerado (ALFARO et al., 2024; PELTOMÄKI, 2023).

Embora métodos supervisionados e multimodais tenham alcançado avanços revelantes (SOUSA; BASSANI, 2022; CHEN et al., 2024), ainda existem limitações importantes: (i) dificuldade de generalizar para condições fora da distribuição de treinamento (novos ambientes e configurações), (ii) integração multimodal nem sempre é estável, podendo ocorrer colapso ou redundância entre modalidades, e (iii) restrições de eficiência que dificultam a adoção em plataformas embarcadas com orçamento limitado de computação e energia (YANG et al., 2023a; TAN et al., 2024).

Adotamos uma perspectiva prática para mitigar essas limitações ao investigar mapeamento semântico *2D* baseado em classificação de cenas ao longo da trajetória, usando uma câmera *RGB* como sensor primário e profundidade monocular estimada como pista geométrica complementar. Essa escolha reduz o acoplamento sensorial e facilita a implantação em robôs móveis sem sensores dedicados, melhorando portabilidade e custo. Em contrapartida, a profundidade estimada pode ser ruidosa e pouco discriminativa quando usada isoladamente, exigindo um mecanismo de fusão capaz de explorar informação útil sem amplificar inconsistências entre modalidades.

O problema desta dissertação pode ser resumido como: como integrar, de forma eficiente e robusta, pistas complementares de aparência e geometria para classificar cenas internas sob variações ambientais e projetar essas predições na trajetória do robô, produzindo mapas semânticos *2D* mais consistentes? Exploramos a fusão de imagens *RGB* e profundidade monocular estimada, investigando mecanismos de representação e fusão que promovam simultaneamente alinhamento entre modalidades e preservação de complementaridade. Avaliamos uma arquitetura multimodal com Atenção Cruzada Bidirecional e objetivos auxiliares voltados a reduzir a sensibilidade a mudanças fotométricas e de domínio.

1.3 Objetivos

O objetivo geral desta dissertação é desenvolver e validar uma abordagem de mapeamento semântico *2D* para ambientes internos baseada em duas etapas: (i) categorização semântica de cenas a partir de percepção visual multimodal e (ii) projeção das predições ao longo da trajetória do robô para compor mapas semânticos *2D*. A abordagem explora representações visuais auto-supervisionadas e integração *RGB*–profundidade, visando aumentar robustez e generalização sob variações ambientais, sem exigir sensores dedicados de profundidade.

Como objetivos específicos, destacam-se:

- Caracterizar o mapeamento semântico *2D* como categorização de cenas associada à trajetória, definindo critérios de avaliação para desempenho preditivo e consistência espacial do mapa;
- Analisar o potencial de representações auto-supervisionadas para classificação de cenas internas sob variações de iluminação e mudança de domínio, identificando configurações mais estáveis nesse contexto;
- Demonstrar o impacto da integração de uma pista geométrica derivada de profundidade monocular estimada na robustez da classificação, analisando seu efeito em comparação a variantes monomodais e a fusões simplificadas;
- Comprovar que uma estratégia de fusão multimodal com interação entre aparência e geometria, acompanhada de mecanismos de estabilização, produz ganhos consistentes em cenários de maior deslocamento de domínio;
- Avaliar quantitativamente a abordagem em conjuntos de dados públicos e sob protocolos controlados de generalização, incluindo estudo de ablação para isolar a contribuição dos principais componentes e comparação com a literatura;
- Evidenciar, por análise qualitativa, como as diferenças de desempenho se refletem na coerência espacial dos mapas semânticos *2D* gerados, comparando *ground truth* e variantes representativas do método.

1.4 Contribuições

As contribuições desta dissertação concentram-se na proposição e na validação experimental de um arcabouço multimodal para mapeamento semântico *2D* em ambientes internos, formulado sob a perspectiva de classificação de cenas associada à trajetória. A abordagem explora representações visuais auto-supervisionadas e mecanismos de fusão *RGB*–profundidade para aumentar robustez e generalização sob variações de iluminação e mudança de domínio.

De forma resumida, as principais contribuições são:

- Proposição de uma arquitetura que combina representações auto-supervisionadas e fusão por Atenção Cruzada Bidirecional para classificação de cenas, permitindo interação e refinamento mútuo entre aparência e geometria antes da decisão de classe.

- Demonstração de que mapas de profundidade estimados a partir de uma única câmera podem ser incorporados ao *pipeline* multimodal, ampliando o uso de pistas geométricas sem exigir sensores dedicados.

- Incorporação de projeções específicas por modalidade, alinhamento contrastivo entre pares *RGB*–profundidade e regularização por ortogonalidade, com evidências de que essas estratégias promovem consistência *cross-modal* sem induzir redundância ou colapso modal.

- Definição e aplicação de esquemas de avaliação que cobrem generalização temporal, iluminação não vista, interpolação de iluminação e transferência entre domínios, validados nos *datasets* públicos *KTH-IDOL2* e *COLD* sob múltiplas condições fotométricas e cenários de mudança de plataforma e local.

- Evidência quantitativa e qualitativa do impacto no mapeamento estudo de ablação que quantifica a contribuição relativa das modalidades e dos principais componentes do método, complementado por análise qualitativa por mapas semânticos *2D* derivados das predições, indicando ganhos de coerência espacial e redução de rótulos espúrios em condições desafiadoras.

Durante o desenvolvimento deste trabalho, partes da metodologia proposta foram apresentados em artigos que se encontram publicados em conferências nacionais e internacionais.

- Artigo apresentado e publicado: Alternei de Souza Brito, Emanuelle de Souza Gil and Felipe Gomes de Oliveira, *"An Appearance-Based Approach for Indoor Environment Semantic Classification for Autonomous Navigation"*, *2025 Brazilian Symposium on Robotics (SBR)*, Vitória, Brasil, 2025, pp. 255–260. Este trabalho apresenta uma abordagem baseada em aparência para classificação semântica de ambientes internos visando navegação autônoma, servindo como base conceitual e experimental para os avanços propostos nesta dissertação. *doi: 10.1109/SBR/WRE66973.2025.11249563*.

- Artigo apresentado e publicado: José Marcos Cabrera Neto, Alternei de Souza Brito, Vandermi João da Silva, Paulo L. J. Drews-J, Douglas Macharet, João Marcos Cavalcanti, José Luiz de Souza Pio e Felipe Gomes de Oliveira, *"This Place Looks Familiar: Vision-based Semantic Mapping of Indoor Scenes"*, *International Conference on Advanced Robotics (ICAR 2025)*. Este artigo apresenta uma versão preliminar da abordagem de mapeamento semântico aqui desenvolvida, com foco na classificação de cenas em ambientes internos utilizando informações visuais para apoiar a navegação robótica. *doi.org/10.1109/ICAR65334.2025.11338653*

- Artigo submetido: Alternei de Souza Brito e Felipe Gomes de Oliveira, *"Self-Supervised Multimodal Fusion via Cross-Attention and Contrastive Alignment for Indoor Scene Categorization"*, submetido à *International Conference on Robotics and Automation (ICRA 2026)*. Este trabalho apresenta a arquitetura desta dissertação, incorporando mecanismos adicionais de alinhamento contrastivo entre modalidades e regularização das representações, alinhando-se às direções de trabalhos futuros discutidas neste manuscrito.

- Artigo submetido: Alternei S. Brito, Paulo V. K. Borges, Paulo L. J. Drews-Jr e Felipe G. Oliveira, *"SRD-Fusion: Self-Supervised RGB–Depth Fusion for Indoor Scene Categorization"*, submetido à *International Conference on Pattern Recognition (ICPR) 2026*. Este trabalho aprofunda e estende a arquitetura desta dissertação, apresen-

tando alinhamento contrastivo intermodal e regularização das representações, adicionando a fusão bidirecional hierárquica e mecanismo de ponderação adaptativo dependente da amostra, visando aumentar a robustez sob variações de iluminação e transferência entre localidades.

1.5 Organização do Trabalho

Esta dissertação está organizada em cinco capítulos, além desta introdução, de modo a apresentar o embasamento teórico, o panorama do estado da arte, a medotologia proposta e a avaliação experimental.

Capítulo 2 — Fundamentação Teórica apresenta os conceitos que sustentam este trabalho, incluindo mapeamento semântico e categorização de cenas, estimativa de profundidade monocular, fusão multimodal, aprendizado auto-supervisionado e classificação supervisionada.

Capítulo 3 — Trabalhos Relacionados revisa e discute abordagens recentes para mapeamento semântico e reconhecimento de cenas em ambientes internos, com ênfase em robustez a variações de iluminação e mudança de domínio, fusão de modalidades visuais e aplicabilidade em robótica móvel.

Capítulo 4 — Abordagem Proposta descreve a metodologia desenvolvida, incluindo a extração de representações, a geração de profundidade monocular, mecanismos de adaptação e alinhamento entre modalidades, fusão por Atenção Cruzada Bidirecional e a etapa de classificação. O capítulo também detalha o procedimento de associação das predições às poses do robô para compor mapas semânticos *2D*.

Capítulo 5 — Avaliação Experimental apresenta os *datasets*, protocolos experimentais, métricas e resultados obtidos, incluindo um estudo de ablação, além da análise da qualidade dos mapas semânticos *2D* gerados.

Por fim, o Capítulo 6 — Considerações Finais sintetiza as conclusões, discute as contribuições e limitações, e aponta direções para trabalhos futuros, incluindo extensões da arquitetura e aprimoramentos no processo de mapeamento semântico.

2

FUNDAMENTOS

Este capítulo apresenta os fundamentos teóricos e técnicos que embasam esta dissertação. São discutidos conceitos de mapeamento semântico em robótica móvel, categorização de cenas, estimativa de profundidade monocular a partir de imagens *RGB*, fusão multimodal de representações e estratégias de treinamento para integração consistente entre modalidades. Esses tópicos fornecem o suporte necessário para a abordagem proposta, que utiliza apenas imagens *RGB* e mapas de profundidade estimados e combinando pistas de aparência e geometria para melhorar a robustez do mapeamento semântico em ambientes internos.

2.1 Mapeamento Semântico em Robótica Móvel

O mapeamento é uma tarefa fundamental em robótica móvel, permitindo que agentes autônomos construam uma representação espacial do ambiente à medida que exploram o espaço físico. Inicialmente centrado em mapas puramente geométricos, como mapas métricos ou topológicos, o foco evoluiu para incluir informações de alto nível, como categorias de objetos, funções de áreas e relações espaciais. Essa evolução culminou no conceito de mapa semântico, que associa significado ao espaço mapeado, facilitando a tomada de decisão contextualizada por parte do robô (GALINDO et al., 2005; CADENA et al., 2016).

Os mapas semânticos podem assumir diferentes formas. Em abordagens métrico–semânticas, informações semânticas são sobrepostas a reconstruções espaciais

detalhadas, como nuvens de pontos ou grades de ocupação. Já em mapas topológico–semânticos, os ambientes são representados como grafos em que nós correspondem a locais com atributos semânticos, e as arestas representam conexões navegáveis entre eles (PRONOBIS; JENSFELT, 2012). Esses mapas são particularmente úteis em ambientes internos, onde categorias como corredor, cozinha ou sala de reunião são importantes para a execução de tarefas orientadas a contexto.

Neste trabalho, adotamos uma formulação baseada em categorização de cenas: cada posição do robô é associada a um rótulo semântico, como corredor, cozinha ou laboratório, e essas predições são projetadas sobre a trajetória para compor um mapa semântico *2D*.



Figura 2 – Exemplo ilustrativo de mapa semântico em ambiente interno apresentado por (SüNDERHAUF et al., 2016).

2.2 Classificação de Cenas e Representações Visuais

A classificação de cenas refere-se à tarefa de atribuir um rótulo semântico a uma imagem ou parte do ambiente com base em sua aparência visual. Diferentemente da segmentação, que rotula cada *pixel* ou região individualmente, a classificação de cenas opera sobre unidades maiores, como quadros inteiros, regiões agrupadas ou nós de um grafo. Em robótica móvel, essa tarefa é essencial para identificar funcionalmente locais como banheiro, escritório ou laboratório, permitindo que o sistema tome decisões adequadas

ao contexto (ZHOU et al., 2014).

Historicamente, essa tarefa foi baseada em descritores manuais, como *SIFT*, *HOG* e *Gist* (OLIVA; TORRALBA, 2001). Com o avanço das redes neurais convolucionais (*CNNs*), passaram a ser empregadas representações visuais aprendidas automaticamente, capazes de capturar padrões visuais e semânticos em múltiplas escalas. Modelos como *AlexNet*, *VGG16*, *ResNet* e *EfficientNet* tornaram-se amplamente utilizados para extrair *embeddings* visuais, que podem alimentar classificadores supervisionados ou ser comparados diretamente por similaridade, aplicando-se a tarefas como reconhecimento ou recuperação de lugares.

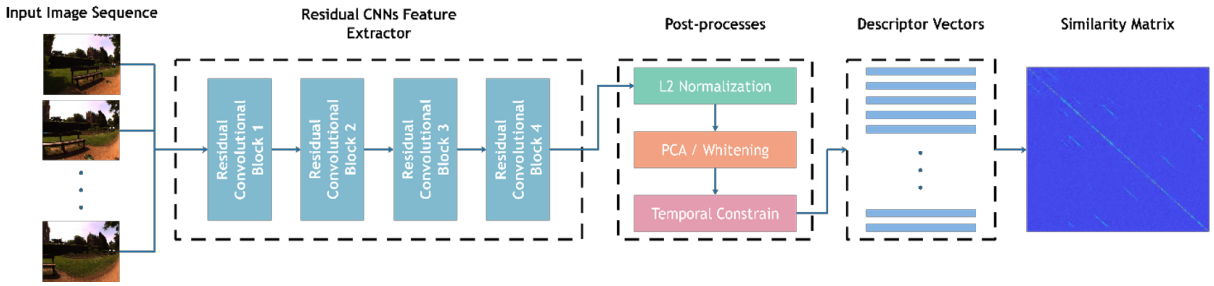


Figura 3 – Arquitetura baseada em *CNN* para a classificação de cenas, proposta por (WANG et al., 2019)

Mais recentemente, arquiteturas baseadas em *Transformers*, como *Vision Transformer* (DOSOVITSKIY et al., 2021) e *Swin Transformer v2* (LIU et al., 2022a), e modelos pré-treinados (*BEit* (BAO et al., 2022) e *DEit* (TOUVRON et al., 2021)) em larga escala têm se destacado por produzir representações mais transferíveis e robustas a mudanças de domínio. Em particular, abordagens auto-supervisionadas permitem aprender descritores visuais com menor dependência de rótulos, favorecendo adaptação a variações de iluminação e mudanças estruturais típicas de ambientes internos. Essa tendência motiva o uso de *backbones* auto-supervisionados como base para a extração de representações na abordagem proposta.

Em cenários de mapeamento semântico baseado em cenas, como o considerado nesta dissertação, essas predições de classe são associadas à pose do robô fornecida, por exemplo, por odometria ou sensores de localização, permitindo construir mapas em que cada célula ou pose é rotulada com uma categoria semântica.

2.3 Estimativa de Profundidade Monocular

A estimativa de profundidade a partir de uma única imagem (*monocular depth estimation*) é uma alternativa viável ao uso de sensores físicos como *LiDAR* ou câmeras *RGB-D*. Essa técnica emprega redes neurais profundas treinadas para prever a estrutura tridimensional de uma cena a partir da informação bidimensional de uma imagem *RGB*. Modelos como *Monodepth2* (GODARD et al., 2019), *DPT* (*Dense Prediction Transformer*) (RANFTL; BOCHKOVSKIY; KOLTUN, 2021) e, mais recentemente, o *Depth Anything v2* (YANG et al., 2024) demonstraram alta capacidade de inferir profundidade a partir de imagens *RGB*, utilizando grande volume de dados supervisionados e auto-supervisionados.

O *Depth Anything v2*, em particular, destaca-se pela capacidade de generalização *zero-shot* e compatibilidade com diversas cenas do mundo real, sendo treinado com uma ampla variedade de domínios visuais. Essas estimativas permitem reconstruir mapas parciais do ambiente e apoiar a projeção de elementos detectados em representações *3D*. A principal vantagem dessas abordagens é eliminar a necessidade de sensores físicos adicionais, como *LiDAR* e câmeras *RGB-D*), simplificando a configuração sensorial do robô, ao custo de uma etapa adicional de inferência de profundidade.

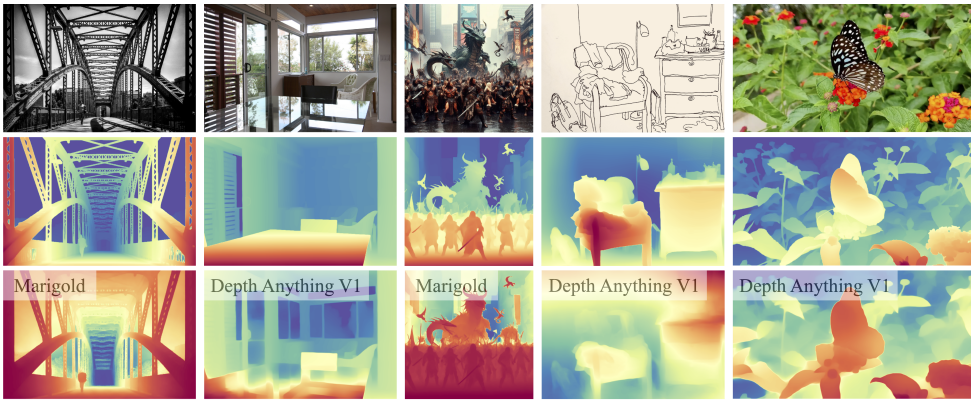


Figura 4 – Resultados do *Depth Anything v2* comparados a outros modelos (YANG et al., 2024)

No contexto desta dissertação, os mapas de profundidade estimados a partir das imagens *RGB* são utilizados como uma segunda modalidade de entrada, complementando a aparência visual com pistas geométricas para a categorização de cenas e, consequentemente, para o mapeamento semântico.

2.4 Fusão de Representações Visuais

A fusão de dados em visão computacional pode ocorrer em diferentes níveis: *early fusion* (na entrada, concatenando *RGB* com profundidade), *middle fusion* (em camadas intermediárias da rede) ou *late fusion* (na saída, por exemplo, combinando resultados de múltiplos módulos). Em contexto de mapeamento semântico, a fusão de representações derivadas de imagens *RGB*, como *embeddings* visuais, mapas de profundidade estimados e outros tipos de dados, permite enriquecer a descrição da cena sem aumentar a complexidade sensorial do sistema.

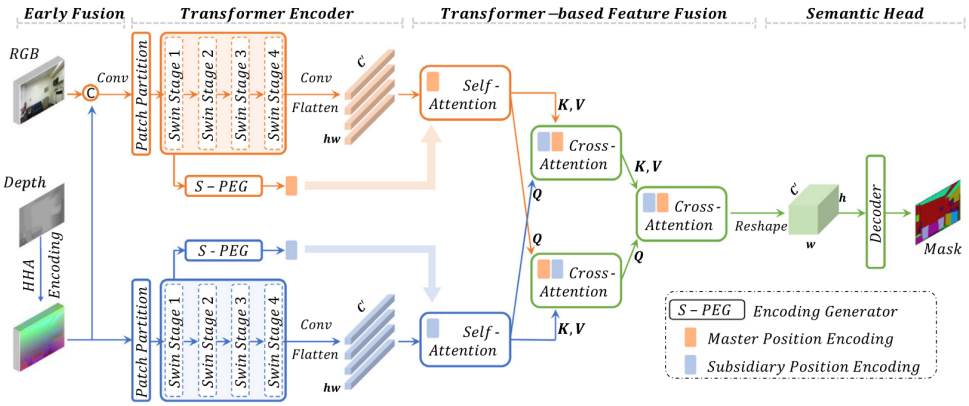


Figura 5 – Arquitetura de fusão de dados *RGB* e profundidade proposta por (WU et al., 2024b)

Essa fusão pode ser feita por concatenação de vetores, por atenção cruzada entre representações ou por *pipelines* modulares que integram diferentes redes especializadas. Trabalhos recentes também exploram estratégias baseadas em *Graph Neural Networks* (*GNNs*s) ou atenção multivariada para fundir pistas semânticas diversas, gerando mapas mais informativos, mesmo em ambientes com ambiguidade visual (QI et al., 2020; WU et al., 2024b).

Neste trabalho, a fusão entre *RGB* e profundidade é realizada em nível intermediário (*middle fusion*) por meio de Atenção Cruzada Bidirecional, permitindo que cada modalidade refine suas representações a partir da outra. Estratégias de adaptação e regularização são empregadas para estabilizar a integração multimodal, promovendo consistência entre modalidades sem perda de complementaridade. O objetivo é explorar de forma conjunta informações de textura/cor e geometria para produzir representações mais robustas para a classificação de cenas e para o mapeamento semântico.

2.5 Classificação Supervisionada Multimodal

A classificação supervisionada consiste em aprender uma função que mapeia variáveis de entrada para uma classe de saída a partir de exemplos rotulados. No contexto desta dissertação, as entradas são representações extraídas de modalidades complementares, como imagens *RGB* e profundidade estimada, incluindo descritores produzidos por um *backbone* auto-supervisionado e representações resultantes da fusão multimodal. Essas variáveis são agregadas e utilizadas por um classificador supervisionado para predizer a categoria semântica da cena observada.

Em aplicações robóticas, além de desempenho, busca-se robustez a variações do ambiente e estabilidade na inferência. Por essa razão, o classificador final é tipicamente combinado com estratégias de regularização e com representações visuais previamente aprendidas, visando melhorar generalização em cenários fora da distribuição de treinamento.

Em síntese, os conceitos apresentados neste capítulo fundamentam a abordagem proposta, que integra categorização de cenas, estimativa de profundidade monocular, fusão multimodal e projeção das predições ao longo da trajetória do robô para compor mapas semânticos *2D*. No próximo capítulo, são discutidos trabalhos relacionados que contextualizam o estado da arte e as lacunas abordadas por esta dissertação.

3

TRABALHOS RELACIONADOS

Este capítulo reúne trabalhos relacionados aos eixos desta dissertação, com foco em (i) mapeamento semântico tratado como categorização de cenas associada à trajetória, (ii) robustez a variações de iluminação e mudança de domínio, e (iii) integração multimodal com baixo acoplamento sensorial. Ao longo da revisão, destacamos as lacunas que motivam esta pesquisa e posicionamos como a abordagem proposta se diferencia das alternativas existentes.

3.1 Mapeamento Semântico em Ambientes Internos

O mapeamento semântico em ambientes internos consolidou-se como componente essencial para navegação inteligente de robôs móveis, permitindo não apenas a representação da geometria do ambiente, mas também a atribuição de significados funcionais às suas partes. Abordagens recentes exploram a integração entre percepção visual, aprendizado profundo e representações espaciais, com foco em sistemas capazes de operar em tempo real e adaptar-se a variações estruturais e ambientais.

Entre os trabalhos mais recentes, (TAO et al., 2020) propõem um sistema de *3D Semantic Visual SLAM* que combina imagens *RGB-D* com segmentação por *Mask Scoring R-CNN*, permitindo a construção de mapas com rotulagem semântica de objetos em ambientes internos complexos. A abordagem utiliza *ORB-SLAM2* como base para rastreamento e construção do mapa, aplicando a segmentação apenas em quadros-chave para reduzir a carga computacional. As informações semânticas são projetadas do *2D*

para o *3D*, associando os objetos detectados à nuvem de pontos reconstruída.

Com foco na organização estrutural dos dados semânticos, (RUAN; GUO; HUANG, 2022) apresentam um método de mapeamento baseado em *OctoMap* e na arquitetura *CBAM-PSPNet*, incorporando mecanismos de atenção espacial e de canal para refinar a segmentação. A proposta utiliza dados *RGB-D* para construir uma grade *Octree 3D* em tempo real, atualizada com rótulos semânticos propagados a partir da segmentação *2D*. A inclusão do módulo *CBAM* (*Convolutional Block Attention Module*) permite que o modelo foque em regiões informativas da imagem, melhorando a precisão da rotulagem com menor esforço computacional.

(XIN et al., 2023) propõem uma abordagem voltada à alta precisão da rotulagem semântica de cenas internas, empregando sensores *RGB-D* e estratégias de refinamento de bordas para melhorar a coerência espacial das classes. O sistema introduz um método de fusão baseado em pontos de contorno (*edge points*) para suavizar as transições entre rótulos vizinhos na nuvem de pontos, corrigindo ruídos típicos em mapas *3D* densos, especialmente em regiões com objetos parcialmente ocluídos ou mal iluminados.

De forma complementar, (SONG et al., 2022) combinam dados de câmera monocular e *LiDAR*, com foco na robustez e na interpretabilidade em tarefas de navegação robótica. A proposta é baseada em classificação de cenas com correlação temporal, utilizando uma arquitetura composta por *InceptionV3* para extração de características visuais e uma *LSTM* para capturar dependências entre quadros consecutivos. Essa combinação permite lidar melhor com transições ambíguas entre espaços adjacentes minimizando erros de classificação quando quadros são analisados isoladamente. A informação semântica é propagada sobre os mapas de ocupação por meio de estimação bayesiana, resultando em mapas probabilísticos de categorias de cena.

Soluções comerciais de robôs de limpeza popularizaram a noção de mapas do ambiente com informações de alto nível, incluindo delimitação de cômodos e, em alguns casos, rótulos semânticos utilizados para planejamento e interação com o usuário. Embora tais sistemas sejam proprietários e frequentemente combinem múltiplos sensores e heurísticas, eles reforçam a relevância prática de mapas enriquecidos semanticamente. Neste trabalho, investigamos esse problema sob uma perspectiva científica e reprodu-

tível, avaliando robustez e generalização na classificação de cenas e na consistência espacial de mapas semânticos *2D* gerados a partir das predições ao longo da trajetória.

Em contraste com esses trabalhos, que em geral dependem de sensores dedicados (*LiDAR*/*RGB-D*) ou segmentação densa para construir mapas semânticos *3D*, esta dissertação investiga um caminho complementar: mapeamento semântico *2D* baseado em categorização de cenas, utilizando apenas uma câmera *RGB* como sensor. A profundidade é obtida por estimativa monocular e integrada como modalidade complementar, permitindo explorar pistas geométricas sem aumentar a complexidade sensorial do robô. Além disso, o foco aqui não é apenas mapear, mas aumentar a consistência das predições ao longo da trajetória sob variações de iluminação e mudança de domínio, avaliando esse comportamento por protocolos controlados.

3.2 Reconhecimento Visual de Lugares e Classificação de Cenas

O reconhecimento visual de lugares e a classificação de cenas são fundamentais para robôs móveis em ambientes internos, pois permitem associar a percepção atual a informações semânticas previamente aprendidas. Nos últimos anos, a área avançou com o uso de redes neurais profundas, desde arquiteturas supervisionadas clássicas até métodos auto-supervisionados e baseados em atenção. Os trabalhos revisados abordam mudanças de iluminação, variações temporais e alta similaridade entre lugares distintos.

No campo supervisionado, (CABRERA et al., 2022) propõem uma arquitetura baseada em *AlexNet* adaptada para imagens omnidirecionais, tratando a localização como uma tarefa hierárquica: primeiro uma classificação grosseira da sala; em seguida, uma localização fina baseada em descritores visuais extraídos de camadas intermediárias. O modelo é otimizado por meio da busca bayesiana de hiperparâmetros e aumento de dados como rotação, brilho, saturação e ruído, alcançando bons resultados sob variações de iluminação.

(ALFARO et al., 2024) utilizam redes *Triplet* para diferenciar locais semantica-

mente distintos mesmo sob mudanças visuais sutis. A arquitetura é composta por três ramificações baseadas em *VGG-16*, treinadas com diferentes funções de perda como *Triplet Margin Loss* e *Circle Loss*. Os autores analisam o impacto do tamanho da janela, da distância espacial entre amostras e do *batch size* no desempenho da localização fina.

Em avaliação comparativa, (CABRERA et al., 2024a) comparam diferentes *CNNs* (*AlexNet*, *ResNet*, *ConvNeXt* e *EfficientNet*) para localização hierárquica baseada em imagens panorâmicas. Modelos mais profundos, como o *ConvNeXt*, obtêm melhor acurácia sob condições nubladas e noturnas, enquanto a robustez sob iluminação ensolarada depende fortemente das técnicas de contraste utilizadas no treinamento.

(PAPAPETROS et al., 2023) apresentam uma abordagem não supervisionada inovadora baseada em representações de sequência no domínio distância–espaço, na qual vetores de descritores com *NetVLAD* são transformados em representações normalizadas por *z-score* e comparadas por correlação cruzada. A técnica se mostra robusta à desordem temporal e a variações estruturais no ambiente, alcançando alto desempenho mesmo quando as sequências estão parcialmente embaralhadas ou o ambiente sofreu modificações.

Adicionalmente, (CABRERA et al., 2024b) avaliam o uso de redes siamesas para discriminação binária de cenas, explorando pares positivos e negativos extraídos de *datasets* reais. A pesquisa considera diferentes configurações de sensores, iluminação e tipos de ambiente, com resultados que demonstram a viabilidade dessas redes para tarefas de reidentificação de locais, especialmente em robôs dotados de câmeras omnidirecionais.

Mais recentemente, (KHAN; OTHMAN, 2024) introduzem uma arquitetura *dual-stream* para classificação de cenas internas voltada a aplicações em robótica móvel, explorando um fluxo para capturar contexto global e outro para enfatizar pistas locais. O método busca lidar com variabilidade intra-classe e similaridade inter-classe em ambientes internos, discutindo o impacto de cenas visualmente semelhantes (por exemplo, corredores e escritórios) e mudanças moderadas de aparência ao longo do tempo. Esse tipo de desenho é particularmente relevante como referência de *baseline* para cenários robóticos, pois persegue robustez sem recorrer explicitamente a uma segunda modalidade geométrica.

Embora esses trabalhos avancem a robustez do reconhecimento e da categorização sob diferentes condições, a maioria explora predominantemente uma única modalidade visual (*RGB*) e não investiga, de forma sistemática, mecanismos de integração que conciliem consistência entre modalidades e preservação de complementaridade. Esta dissertação se diferencia ao combinar representações auto-supervisionadas e profundidade monocular em um *pipeline* único de classificação multimodal, com mecanismos explícitos para estabilizar a fusão e melhorar generalização. Os ganhos são quantificados em cenários desafiadores de iluminação e mudança de domínio, comparando versões monomodais e multimodais sob o mesmo protocolo.

3.3 Fusão Multimodal

A fusão de diferentes representações derivadas de imagens *RGB* consolidou-se como uma estratégia eficaz para enriquecer a percepção sem depender de múltiplos sensores físicos. Em robótica, essa direção é particularmente atraente porque permite combinar pistas de aparência e geometria, incluindo profundidade estimada monocularmente, para tornar as predições mais robustas sob variações fotométricas e ambiguidades estruturais.

Na literatura recente, observa-se uma tendência de empregar mecanismos de atenção para integrar modalidades *RGB* e profundidade de forma *cross-modal*, explorando tanto consistência quanto complementaridade. Por exemplo, o *BiANet* (ZHANG et al., 2021) introduz uma rede de atenção bilateral para detecção de objetos salientes em *RGB-D*, modelando interações entre modalidades com ênfase em realçar regiões informativas. Embora o foco seja detecção saliente, o trabalho é relevante por explicitar como a atenção pode atuar como operador de fusão guiada, em que pistas geométricas ajudam a filtrar ruído visual e a refinar decisões baseadas em aparência.

De forma relacionada, o *CMANet* (PAN et al., 2025) propõe um mecanismo de atenção mútua *context-aware* para segmentação por referência, investigando como diferentes fontes de informação podem se condicionar mutuamente com base no contexto. O princípio de *mutual attention* é alinhado com o objetivo de estabilizar decisões ao

permitir que uma modalidade corrija ou reforce evidências da outra, evitando que a fusão se reduza a uma combinação ingênua de descritores.

Do ponto de vista de modelagem por Transformers, o *SCMT* (ZHAO et al., 2024) introduz um Transformer *structure-aware* para *depth completion*, explorando atenção cross-modal para transferir estrutura da modalidade *RGB* para a profundidade. Esse tipo de arquitetura reforça o papel de blocos *cross-attention* como mecanismo geral para alinhar indícios estruturais e aparência no espaço latente. Para esta dissertação, esse grupo de métodos é especialmente relevante por apoiar a ideia de que a interação *RGB*↔*depth* pode ser modelada explicitamente, ao invés de depender apenas de concatenação ou fusões tardias.

Além disso, abordagens recentes em detecção exploram fusão espacial *cross-modal* com mecanismos de atenção para combinar *RGB* e profundidade em nível de características. Trabalhos como o *YOLO-RGBDtea* (WU et al., 2024a) enfatizam que a profundidade pode fornecer *priors* geométricos úteis para localizar e discriminar alvos quando a aparência é degradada (por exemplo, em baixa iluminação), desde que exista um mecanismo de integração capaz de selecionar regiões e escalas relevantes.

No contexto de classificação de cenas, trabalhos recentes destacam que a profundidade pode fornecer pistas geométricas complementares, mas que fusões simplificadas tendem a ser sensíveis a ruídos do *depth* e podem não explorar adequadamente a interação entre modalidades. Por exemplo, (ZHOU; JIAN; LIU, 2025) propõem um desenho de rede para classificação *RGB-D* que enfatiza diferenças e contrastes entre características, argumentando que mecanismos de fusão mais estruturados são importantes para lidar com inconsistências e imprecisões do *depth* em cenários internos. De forma complementar, (WAQAS; AHMAD, 2024) combinam *Vision Transformers* com refinamento por *CRF* e módulos auxiliares para impor consistência contextual às predições, explorando informação geométrica e estrutural para reduzir ambiguidades em cenas internas.

Em síntese, esses trabalhos apontam um padrão comum: (i) a simples combinação de modalidades não é suficiente para explorar complementaridade; (ii) mecanismos de atenção (*bilateral*, *mutual* ou *cross-attention*) são frequentemente utilizados para promover interação informativa; e (iii) a geometria (profundidade) tende a atuar como

pista auxiliar para reduzir ambiguidades quando a aparência falha. Contudo, a maioria dessas abordagens foi desenvolvida para tarefas como detecção/segmentação ou para cenários com profundidade medida por sensores *RGB-D*, não investigando de forma sistemática a robustez da fusão sob protocolos controlados de generalização temporal, variação fotométrica e transferência de domínio em classificação de cenas.

Esta dissertação avança nesse ponto ao investigar mapeamento semântico *2D* tratado como classificação de cenas, integrando profundidade monocular estimada e *RGB* em um *pipeline* único. A fusão é conduzida por Atenção Cruzada Bidirecional, com módulos adicionais para estabilizar a integração e preservar complementaridade entre modalidades. Os ganhos são quantificados em cenários desafiadores de iluminação e mudança de domínio, comparando variantes monomodais, fusão simples e a arquitetura completa sob o mesmo protocolo experimental.

3.4 Eficiência Computacional para Robôs Móveis

A implementação de mapeamento semântico em robôs móveis impõe restrições de capacidade computacional, consumo energético, tempo de resposta e memória. Por isso, várias abordagens recentes desenvolvem arquiteturas leves, otimizadas para execução em dispositivos embarcados ou com suporte limitado de *hardware*, como *Raspberry Pi* e *NVIDIA Jetson*.

(LU et al., 2022) propõem segmentação semântica sensível à instância para *NVIDIA Jetson Xavier*, combinando um detector leve com um módulo de segmentação que aproveita mapas de profundidade estimados para gerar máscaras semânticas em tempo real. O sistema atinge mais de 30 *FPS*, mantendo bom equilíbrio entre desempenho e precisão, e demonstrando viabilidade para aplicação embarcada em robôs móveis que operam em ambientes internos com limitações de recursos computacionais.

(ALQOBALI et al., 2024) apresentam mapeamento semântico executado em tempo real sobre a *Raspberry Pi*, utilizando câmera e sensores *LiDAR*. Implementado no *ROS* e validado com simulações no *Gazebo*, o sistema alcança cerca de 4 *FPS*, considerado suficiente para aplicações de navegação em ambientes internos. A proposta destaca a

importância da modularidade e da leveza computacional para manter o desempenho em plataformas de baixo custo, reforçando a aplicabilidade do sistema em robôs reais.

(ZHOU et al., 2024) desenvolvem segmentação de cenas *RGB-D* com redes otimizadas, combinando *SegFormer* e *MobileNet-V1* com distilação de conhecimento. A arquitetura é projetada para plataformas com restrição de memória e processamento, utilizando módulos multifluxo leves e um *encoder–decoder* compacto. A proposta é avaliada em *benchmarks* padrão de segmentação *indoor* e apresenta resultados competitivos em termos de acurácia, com menor latência e uso reduzido de parâmetros, tornando-a adequada para robôs autônomos embarcados.

Esses trabalhos evidenciam que soluções robóticas devem equilibrar desempenho e custo computacional. Nesta dissertação, esse equilíbrio é buscado ao evitar sensores adicionais e ao explorar um *pipeline* modular, no qual a etapa de extração representacional é compartilhada entre modalidades e a fusão é realizada por um mecanismo de atenção com baixa sobrecarga adicional em relação ao *backbone*. Embora o foco principal seja robustez e generalização, a proposta também considera a viabilidade prática ao restringir a percepção a uma câmera *RGB* e módulos de integração compactos.

3.5 Síntese Crítica dos Trabalhos Relacionados

A literatura recente em mapeamento semântico para robótica móvel apresenta avanços importantes, sobretudo em abordagens baseadas em segmentação e *SLAM* semântico, bem como em métodos de reconhecimento e categorização de lugares mais robustos a variações ambientais. Trabalhos recentes em fusão multimodal têm consolidado o uso de mecanismos de atenção *cross-modal* como estratégia para modelar a interação entre aparência e geometria, ainda que frequentemente em tarefas-alvo distintas, como detecção e segmentação.

Apesar desses avanços, a revisão evidencia lacunas relevantes para cenários internos realistas em robótica móvel. Em particular, observa-se que: (i) muitos sistemas de mapeamento semântico dependem de sensores dedicados e segmentação densa para construir mapas *3D*, elevando custo e complexidade; (ii) quando a classificação de cenas

é adotada como forma de rotulagem semântica ao longo da trajetória, a integração multimodal nem sempre é investigada de forma sistemática sob variações fotométricas e transferência de domínio; e (iii) embora mecanismos de atenção *cross-modal* sejam recorrentes na literatura *RGB-D*, são menos frequentes estudos que combinam *backbones* auto-supervisionados modernos com componentes explícitos para promover simultaneamente alinhamento entre modalidades e preservação de complementaridade, especialmente quando a profundidade é obtida por estimativa monocular.

Diante dessas lacunas, esta dissertação se diferencia ao:

- adotar um cenário *RGB*-only, gerando profundidade por estimativa monocular para evitar sensores adicionais e manter portabilidade;
- tratar o mapeamento semântico como categorização de cenas associada à trajetória, produzindo mapas semânticos *2D* de forma direta e interpretável;
- utilizar representações auto-supervisionadas para favorecer transferência e robustez a mudanças fotométricas e de domínio;
- empregar fusão multimodal por Atenção Cruzada Bidirecional e componentes auxiliares para estabilizar a interação e preservar complementaridade;
- quantificar ganhos por protocolos controlados e por ablação, comparando variantes monomodais, fusão simples e a arquitetura completa sob as mesmas condições experimentais.

Em conjunto, a contribuição é sustentada por ganhos consistentes de desempenho e robustez em relação a baselines competitivos e variantes do próprio método, especialmente em cenários de iluminação não vista e transferência entre domínios, além de evidências qualitativas em mapas semânticos *2D* com maior coerência espacial.

A Tabela 1 resume o posicionamento desta dissertação em relação aos trabalhos revisados. Diferentemente dessas abordagens, nossa proposta combina profundidade monocular e representações auto-supervisionadas com componentes explícitos para estabilizar a fusão e interação dinâmica via Atenção Cruzada Bidirecional, avaliando robustez por protocolos controlados de iluminação e transferência de domínio em classificação de cenas e na geração de mapas semânticos *2D*.

Trabalho	Sensores	Paradigma	Saída	Multimodal	Diferencial / Observação
(TAO et al., 2020)	*RGB-D*	*SLAM* + seg.	Mapa sem. 3D	*RGB-D* + seg.	Seg. em *keyframes*
(RUAN; GUO; HUANG, 2022)	*RGB-D*	Seg. + map.	*OctoMap* 3D	*RGB-D* + atenção	*CBAM-PSPNet*
(XIN et al., 2023)	*RGB-D*	Seg. + refino	Mapa sem. 3D	*RGB-D*	Refino por bordas
(SONG et al., 2022)	*RGB* + *LiDAR*	Classif. temporal	Mapa ocup.	Visão+*LiDAR*	*LSTM* + *Bayes*
(CABRERA et al., 2022)	*RGB* (omni)	Superv. hierárq.	Sala + local	Monomodal	*Aug.* + *HPO*
(ALFARO et al., 2024)	*RGB*	Métrica (*Triplet*)	*Embeddings*	Monomodal	*Triplet/Circle loss*
(CABRERA et al., 2024a)	*RGB* (omni)	Superv. (*CNNs*)	Comparativo	Monomodal	*AlexNet–ConvNeXt*
(PAPAPETROS et al., 2023)	*RGB* (seq.)	Não superv.	*Matching*	Monomodal	*NetVLAD* + *xcorr*
(CABRERA et al., 2024b)	*RGB*	Siamese	*ReID* binária	Monomodal	Pares pos/neg
(KHAN; OTHMAN, 2024)	*RGB*	Classif. (*dual-stream*)	Classe de cena	Monomodal	Contexto global + pistas locais
(ZHOU; JIAN; LIU, 2025)	*RGB-D*	Classif. *RGB-D*	Classe de cena	*RGB+depth*	Fusão baseada em contrastes
(ZHANG et al., 2021)	*RGB-D*	Detecção saliente	Mapa/objetos salientes	Atenção bilateral	Fusão guiada por atenção (*SOD*)
(PAN et al., 2025)	*RGB* + texto	Seg. por referência	Máscaras	Atenção mútua	*Mutual attention* contextual (*RIS*)
(ZHAO et al., 2024)	*RGB* + prof. (parcial)	Completação de profundidade	Prof. densa	Transformer *cross-modal*	Guiada por *RGB* (*depth completion*)
(WU et al., 2024a)	*RGB-D*	Detecção	*Bounding boxes*	Atenção espacial *cross-modal*	Fusão *RGB*/prof. em detecção
(HOU; ZHAO; LIN, 2020)	*RGB-only*	Prof. + objetos	Mapa sem.	Prof. est. + det.	*FCN* + *Faster R-CNN*
(WAQAS; AHMAD, 2024)	*RGB-D*	Classif. + refino	Classe de cena	*RGB+depth*	ViT + consistência contextual (*CRF*)
(FU et al., 2023)	*RGB-only*	Objetos + descr.	Mapa 3D	Det. + descr.	Descr. por objetos
(SOUSA; BASSANI, 2022)	*RGB-only*	Topológico	Grafo sem.	*Embeddings*	Nós com *feats*
Esta dissertação	***RGB-only***	**Classificação multimodal**	**Mapa semântico *2D* (trajetória)**	***RGB* + prof. monocular (*Depth Anything V2*)**	***DINOv2* + projeções + alinhamento contrastivo + ortogonalidade + atenção cruzada bidirecional**

Tabela 1 – Síntese comparativa dos trabalhos relacionados e posicionamento da abordagem proposta.

4

METODOLOGIA

Este capítulo descreve a arquitetura e o protocolo experimental adotados para avaliar a hipótese central desta dissertação: a integração multimodal de imagens *RGB* e mapas de profundidade pode aumentar a robustez da categorização de cenas internas quando combinada com (i) representações auto-supervisionadas extraídas por *DINOv2*, (ii) mecanismos de adaptação e alinhamento entre modalidades e (iii) fusão por Atenção Cruzada Bidirecional. Investigamos estratégias que promovem consistência entre modalidades sem eliminar sua complementaridade, visando melhorar desempenho sob variações de iluminação e mudanças de domínio.

A metodologia também contempla a etapa de mapeamento semântico propriamente dita. As predições geradas pelos modelos são projetadas sobre o plano do ambiente, utilizando as coordenadas associadas a cada imagem nos *datasets KTH-IDOL2* e *COLD*, resultando em mapas semânticos *2D* que permitem uma análise visual da consistência espacial das classificações.

Este capítulo está organizado da seguinte forma: a Seção 4.1 detalha a arquitetura do modelo, incluindo a extração de característcias e o mecanismo de fusão; a Seção 4.2 apresenta os conjuntos de dados utilizados e o pré-processamento; e a Seção 4.3 descreve o protocolo experimental, a configuração de treinamento, os modelos de comparação e as métricas de avaliação.

4.1 Arquitetura da Abordagem Proposta

Nesta dissertação, propomos um *pipeline* de mapeamento semântico *2D* baseado em duas etapas: (i) categorização semântica de cenas a partir de entradas *RGB* e profundidade estimada e (ii) projeção dos rótulos ao longo da trajetória do robô para compor o mapa semântico.

Denominamos *MFISC* (*Multimodal Fusion for Indoor Scene Categorization*) o módulo de categorização utilizado na etapa (i), composto por projeções por modalidade, objetivos auxiliares de alinhamento e fusão por Atenção Cruzada Bidirecional. Por consistência com a nomenclatura adotada em publicações derivadas deste trabalho, passamos a nos referir ao módulo de categorização semântica como *MFISC* ao longo da dissertação.

A arquitetura ilustrada na Figura 6, opera em cinco estágios principais:

1. Geração da segunda modalidade: para cada imagem *RGB*, um mapa de profundidade é estimado por um modelo de profundidade monocular (*Depth Anything V2*), resultando em pares alinhados (I_{RGB}, I_{Depth});

2. Extração de características densas: ambas as modalidades são processadas por um *backbone* compartilhado (*DINOv2*), produzindo sequências de *tokens* que capturam informações globais e locais da cena;

3. Adaptação e alinhamento entre modalidades: projeções específicas por modalidade tornam as representações mais compatíveis antes da integração, enquanto objetivos auxiliares promovem consistência entre pares *RGB*–profundidade sem eliminar a complementaridade das modalidades;

4. Fusão multimodal por Atenção Cruzada Bidirecional: cada modalidade consulta a outra via atenção cruzada, com refinamento mútuo e controle de intensidade da fusão por um parâmetro aprendível, produzindo representações enriquecidas com pistas de aparência e geometria;

5. Classificação e mapeamento: os *tokens [CLS]* refinados são agregados e alimentam um cabeçalho de classificação supervisionada para predição da classe de cena. Em

seguida, as predições são associadas às poses do robô e projetadas no plano para gerar mapas semânticos *2D* ao longo da trajetória.

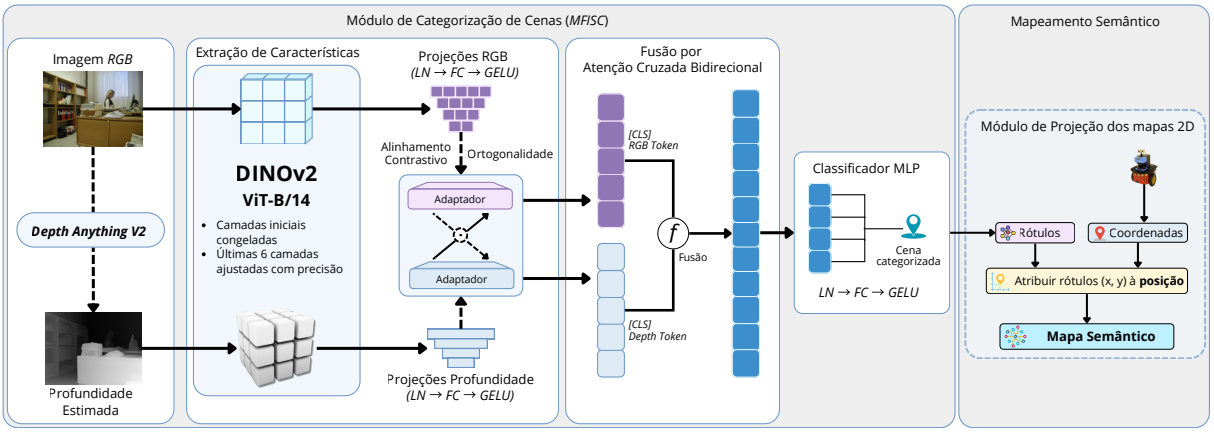


Figura 6 – Visão geral da arquitetura proposta *MFISC* + Mapeamento Semântico *2D*. A partir de pares *RGB* e profundidade (estimada), extraem-se representações com *DINOv2*; projeções e objetivos auxiliares favorecem alinhamento e complementaridade entre modalidades; a fusão ocorre via Atenção Cruzada Bidirecional; e as predições de cena são projetadas na trajetória do robô para compor mapas semânticos 2D.

4.1.1 Geração dos Mapas de Profundidade

Os conjuntos de dados utilizados neste trabalho não disponibilizam mapas de profundidade. Para viabilizar a fusão multimodal, adotamos o modelo *Depth Anything v2* (YANG et al., 2024), estado da arte em estimação monocular de profundidade. Esse modelo foi escolhido por sua robustez e capacidade de generalização, obtida a partir de pré-treinamento em larga escala.

Dada uma imagem *RGB* I_{RGB}, o modelo produz um mapa de profundidade I_{Depth} em escala de cinza. Para fins de armazenamento e reprodutibilidade, essas estimativas foram normalizadas e salvas em formato *PNG* de 16 *bits* (0–65535), preservando gradientes sutis de profundidade. No entanto, durante o treinamento e a inferência, os mapas salvos são carregados pelo *pipeline* de dados, convertidos para tensor e expandidos para três canais por replicação, de modo a atender ao formato de entrada do *DINOv2*. Em seguida, assim como as imagens *RGB*, essas entradas são normalizadas utilizando

as estatísticas do *ImageNet* ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), garantindo compatibilidade com o *backbone* pré-treinado e padronização do pré-processamento entre modalidades.

A Figura 7 ilustra amostras como resultado do processo descrito, apresentando uma imagem *RGB* do dataset e o respectivo mapa de profundidade estimado pelo modelo *Depth Anything v2*.

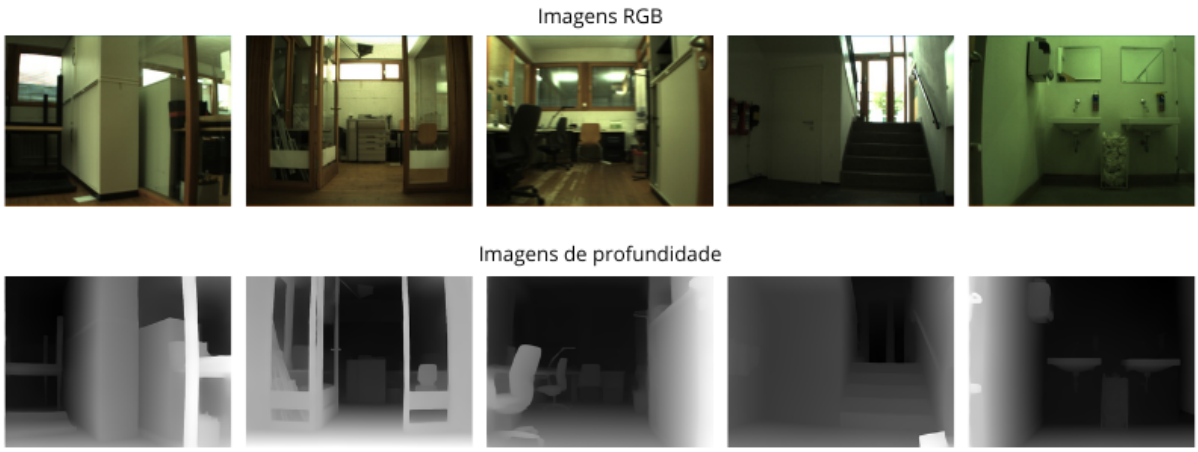


Figura 7 – Exemplo de par multimodal gerado a partir do *dataset COLD*. O mapa de profundidade foi estimado pelo modelo *Depth Anything v2*, normalizado e adaptado para compatibilidade com o *backbone DINOv2*.

4.1.2 Extração de Características com o *DINOv2*

Como extrator de características primário, utilizamos o *DINOv2* (OQUAB et al., 2024) (ViT-B/14) pré-treinado, compartilhado por ambas as modalidades. Nesse *backbone*, cada imagem é dividida em *patches* de 14×14 *pixels*, projetados em uma sequência de *tokens*. Um *token* especial *[CLS]* é adicionado para agregar informação global da cena.

O processo de extração pode ser representado funcionalmente. Dada uma imagem I, o *DINOv2* produz uma sequência $F \in \mathbb{R}^{(L+1)\times d}$, composta por L *tokens* de *patch* e um *token [CLS]*, com dimensão d.

$$F = \Phi_{\text{DINOv2}}(I) \in \mathbb{R}^{N\times d} \tag{4.1}$$

Este processo é aplicado em dois ramos paralelos:

- Processamento *RGB*: A imagem de entrada I_{RGB} é processada diretamente pelo *DINOv2* para gerar o mapa de características visuais F_{RGB}.

$$F_{RGB} = \Phi_{\text{DINOv2}}(I_{RGB}) \tag{4.2}$$

- Processamento de Profundidade: O mapa de profundidade gerado, de um único canal, é então pré-processado (replicado para três canais e normalizado) para gerar a entrada I'_{Depth} que é processada pelo mesmo *backbone*:

$$F_{Depth} = \Phi_{\text{DINOv2}}(I'_{Depth}) \tag{4.3}$$

A saída desta etapa consiste nos dois mapas de características, F_{RGB} e F_{Depth}, que retêm informação espacial detalhada e servem como entrada para o mecanismo de fusão.

4.1.3 Projeções Específicas por Modalidade

Embora *RGB* e profundidade sejam processadas pelo mesmo *backbone*, suas representações podem apresentar descompasso devido às diferenças de conteúdo (aparência vs. geometria). Para tornar as sequências de *tokens* mais compatíveis antes das etapas de alinhamento e fusão, empregamos projeções específicas por modalidade, atuando como adaptadores leves.

Sejam F_{RGB} e F_{Depth} as sequências extraídas pelo *DINOv2*. Aplicamos duas funções de projeção, f^{rgb}_{proj} e f^{dep}_{proj}, produzindo:

$$\hat{F}_{RGB} = f^{rgb}_{proj}(F_{RGB}), \quad \hat{F}_{Depth} = f^{dep}_{proj}(F_{Depth}), \tag{4.4}$$

onde cada f_{proj} é implementada como um bloco leve do tipo *LayerNorm–Linear–GELU–LayerNorm*, preservando a dimensionalidade dos *tokens*. Essas projeções reduzem incompatibilidades entre espaços de características e estabilizam as etapas subsequentes de integração multimodal.

4.1.4 Alinhamento e Regularização

Mesmo após a adaptação por projeções, as representações podem não estar alinhadas semanticamente. Para reforçar consistência entre pares correspondentes *RGB*–profundidade e, ao mesmo tempo, evitar colapso (representações redundantes entre modalidades), utilizamos objetivos auxiliares: alinhamento contrastivo e regularização por ortogonalidade.

4.1.4.1 Alinhamento Contrastivo

Para alinhar explicitamente as representações de *RGB* e profundidade referentes ao mesmo instante, adotamos uma perda contrastiva do tipo *InfoNCE*. Para um lote de pares correspondentes $(h^{(i)}_{RGB}, h^{(i)}_{Dep})$, maximiza-se a similaridade do par positivo i em relação aos demais pares do lote (negativos), com temperatura τ. Essa formulação é particularmente adequada ao nosso cenário, pois a correspondência *RGB–Depth* é conhecida por construção (pares gerados para cada quadro), permitindo impor alinhamento ponto-a-ponto no espaço latente.

Sejam $\hat{h}^{rgb}$ e $\hat{h}^{dep} \in \mathbb{R}^d$ os *embeddings* normalizados extraídos dos *tokens [CLS]* de $\hat{F}_{RGB}$ e $\hat{F}_{Depth}$. A similaridade cosseno é definida por:

$$\mathrm{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}. \tag{4.5}$$

Para um *batch* de tamanho N, definimos a probabilidade de a amostra i em *RGB* corresponder à amostra j em profundidade como:

$$p_{ij}^{rgb\to dep} = \frac{\exp(\mathrm{sim}(\hat{h}_i^{rgb}, \hat{h}_j^{dep})/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\hat{h}_i^{rgb}, \hat{h}_k^{dep})/\tau)}, \tag{4.6}$$

onde τ é um parâmetro de temperatura. A perda *InfoNCE* simétrica é:

$$\mathcal{L}_{cl} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log p_{ii}^{rgb\to dep} + \log p_{ii}^{dep\to rgb} \right]. \tag{4.7}$$

Optamos por *InfoNCE* por fornecer um sinal direto de correspondência entre modalidades e ser simples de integrar ao treinamento supervisionado. Em contraste, métodos como *Barlow Twins* e *VICReg* removem a necessidade de negativos via termos de redundância/variância e são mais comumente usados em *setups* simétricos

de duas visões do mesmo dado. Já abordagens do tipo *DINO style loss* dependem de auto-destilação e *teacher–student*, aumentando complexidade de implementação. No nosso caso, *InfoNCE* funciona como objetivo auxiliar para aproximar pares *RGB–Depth*, enquanto a regularização por ortogonalidade atua de forma complementar para evitar colapso e preservar informação não redundante entre modalidades.

4.1.4.2 Regularização por Ortogonalidade

Embora o alinhamento contrastivo aproxime as representações *RGB*–profundidade correspondentes, somente alinhar não garante que as modalidades preservem informação complementar. Esse aspecto é especialmente crítico neste trabalho, pois a profundidade é estimada a partir do próprio *RGB*, o que tende a induzir correlação entre as modalidades. Assim, introduzimos uma regularização por ortogonalidade aplicada às projeções específicas por modalidade, penalizando correlação excessiva entre os vetores latentes e incentivando que cada modalidade ocupe direções distintas no espaço de características.

Diferentemente de regularizações genéricas (por exemplo, *weight decay* ou *dropout*) ou de termos focados apenas em dispersão/variância, a ortogonalidade atua diretamente sobre a redundância *cross-modal*, favorecendo fusão mais informativa e evitando que a atenção cruzada converta a integração em duplicação de sinais.

Adotamos uma penalização de ortogonalidade entre matrizes de *embeddings* do *batch*. Seja $H^{rgb}, H^{dep} \in \mathbb{R}^{N \times d}$ a concatenação dos vetores *[CLS]* projetados de cada modalidade no *batch*. Definimos:

$$\mathcal{L}_{\perp} = \left\| \frac{{H^{rgb}}^{\top} H^{dep}}{\|H^{rgb}\|_F \, \|H^{dep}\|_F} \right\|_F^2 . \tag{4.8}$$

Esse termo desencoraja alinhamento trivial (colapso) e incentiva que *RGB* e profundidade contribuam com informações complementares.

4.1.5 Mecanismo de Fusão por Atenção Cruzada Bidirecional

Após a adaptação por projeções e a supervisão por objetivos auxiliares, as sequências $\hat{F}_{RGB}$ e $\hat{F}_{Depth}$ são integradas por Atenção Cruzada Bidirecional. Nesse mecanismo, cada modalidade atua alternadamente como consulta (*Query*) e como fonte (*Key/Value*) da outra, permitindo refinamento mútuo.

A atenção por produto escalar com escala é definida por:

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V. \tag{4.9}$$

Utilizamos uma versão multi-cabeças (*MHA*) com conexões residuais e um fator de modulação aprendível que controla a intensidade da fusão. Formalmente:

$$\tilde{F}_{RGB} = \hat{F}_{RGB} + \sigma(g_{rgb})\,\mathrm{MHA}(Q = \hat{F}_{RGB}, K = \hat{F}_{Depth}, V = \hat{F}_{Depth}), \tag{4.10}$$

$$\tilde{F}_{Depth} = \hat{F}_{Depth} + \sigma(g_{dep})\,\mathrm{MHA}(Q = \hat{F}_{Depth}, K = \hat{F}_{RGB}, V = \hat{F}_{RGB}), \tag{4.11}$$

onde $\sigma(\cdot)$ é a sigmoide e g_{rgb}, g_{dep} são parâmetros aprendíveis. Intuitivamente, *RGB* incorpora pistas geométricas provenientes da profundidade, enquanto a profundidade incorpora informações semânticas de aparência, produzindo representações mais robustas para a etapa de classificação.

4.1.6 A Etapa de Classificação Final

A partir das sequências refinadas, extraímos os *tokens [CLS]* de cada modalidade, $\tilde{c}_{RGB}$ e $\tilde{c}_{Depth}$, e os concatenamos:

$$f = [\tilde{c}_{RGB} \,\|\, \tilde{c}_{Depth}]. \tag{4.12}$$

O vetor f é normalizado e alimenta um classificador *MLP* que produz os *logits* das C classes:

$$\hat{y} = f_{\mathrm{MLP}}(\mathrm{LN}(f)). \tag{4.13}$$

O treinamento supervisionado utiliza entropia cruzada categórica:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i). \tag{4.14}$$

A função objetivo total combina o termo supervisionado com os objetivos auxiliares de alinhamento e regularização:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{\perp}\mathcal{L}_{\perp}, \tag{4.15}$$

onde λ_{cl} e $\lambda_{\perp}$ controlam a contribuição do alinhamento contrastivo e da regularização por ortogonalidade, respectivamente.

4.1.7 Geração dos Mapas Semânticos 2D

Embora a saída primária da arquitetura proposta seja a classe semântica atribuída a cada imagem, o objetivo final deste trabalho é construir mapas semânticos que associem essas predições à trajetória do robô no ambiente. Para isso, exploramos o fato de que os *datasets KTH-IDOL2* e *COLD* fornecem, para cada quadro, informações de posição no plano (coordenadas *2D*) ou dados que permitem reconstruir a trajetória.

A partir do modelo treinado, cada imagem I_t de uma sequência gera uma predição de classe $\hat{y}_t$. Em paralelo, obtemos a posição correspondente (x_t, y_t) do robô no instante t, conforme disponibilizado pelo *dataset*. O mapa semântico *2D* é então construído associando, a cada ponto da trajetória, a classe predita pelo modelo:

$$\mathcal{M} = \{(x_t, y_t, \hat{y}_t) \mid t = 1, \ldots, T\}.$$

Na prática, esse procedimento é implementado por meio de um *pipeline* simples em *Python*: (i) leitura das predições de classe para cada quadro da sequência; (ii) alinhamento dessas predições com as coordenadas (x_t, y_t); (iii) mapeamento de cada classe para uma cor fixa; e (iv) plotagem da trajetória em um plano *2D*, colorindo cada amostra de acordo com a categoria semântica atribuída. O resultado é um mapa semântico *2D*, em que regiões do ambiente aparecem destacadas de acordo com o tipo de lugar (por exemplo, corredores, escritórios, salas de reunião, laboratórios).

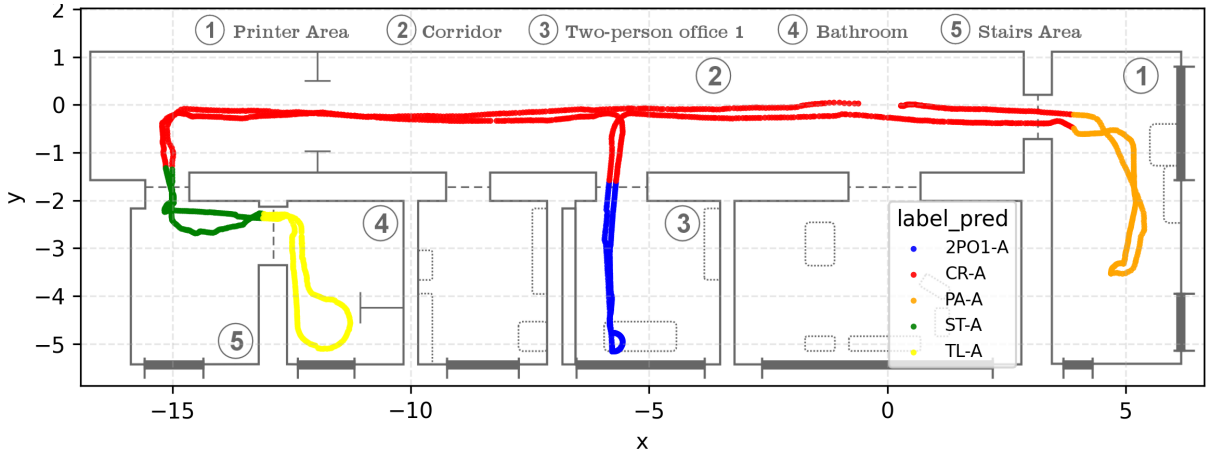


Figura 8 – Mapa semântico 2D gerado para a sequência *Freiburg-night34* do conjunto *COLD* (*Ground-Truth*). Cada ponto da trajetória é colorido de acordo com a classe de cena associada.

Para permitir uma análise qualitativa mais clara do impacto da arquitetura proposta no mapeamento semântico, foram selecionadas sequências representativas em cada *dataset*. Em cada sequência, são gerados três mapas: (i) o *ground truth*, obtido a partir dos rótulos originais; (ii) o mapa produzido pelo melhor *backbone* supervisionado de referência integrado à arquitetura; e (iii) o mapa gerado pela arquitetura proposta com *DINOv2* e Atenção Cruzada Bidirecional.

A Figura 8 ilustra exemplo desses mapas semânticos para o ambiente *Freiburg-night* do *dataset COLD*. Essa visualização permite comparar não apenas a acurácia global, mas também a consistência espacial das classificações ao longo da trajetória, evidenciando em que medida a fusão multimodal reduz erros localmente concentrados como confusões sistemáticas em regiões de baixa iluminação e melhora a coerência do mapa como um todo.

4.2 Conjuntos de Dados

Foram utilizados os conjuntos de dados *KTH-IDOL2* (LUO et al., 2006) e *COLD* (PRONOBIS; CAPUTO, 2009), amplamente reconhecidos na literatura de mapeamento semântico. A escolha se deve à variabilidade visual significativa, incluindo diferentes condições de iluminação, perspectivas de câmera e *layouts*, o que permite avaliar a robustez e a capacidade de generalização da abordagem. Além disso, as informações de posição

disponibilizadas pelos *datasets* são utilizadas para projetar as predições de classe sobre o plano do ambiente, permitindo a construção dos mapas semânticos *2D* discutidos na Seção 4.1.7. Além disso, ambos os conjuntos disponibilizam dados de navegação, como as poses, odometria e, em parte, localização baseada em *laser*, os quais são utilizados para reconstruir a trajetória *2D* e projetar as predições de classe no plano do ambiente.

4.2.1 *KTH-IDOL2* (*Image Database for rObot Localization*)

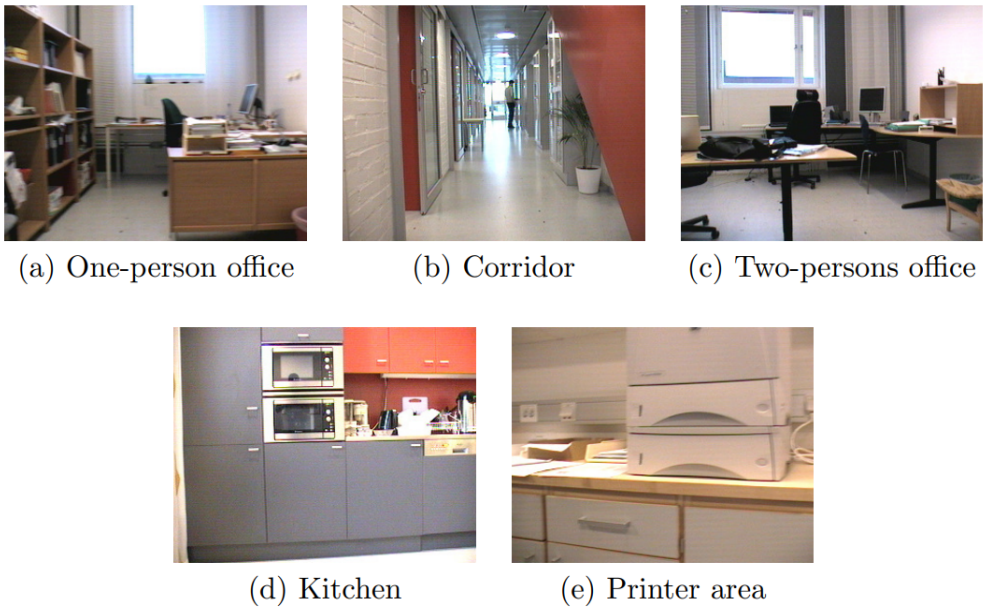
O conjunto de dados *KTH-IDOL2* contém 24 sequências de imagens capturadas em um ambiente interno composto por cinco classes de ambientes. As sequências foram adquiridas por duas plataformas móveis distintas, *PeopleBot Minnie* e *PowerBot Dumbo*, cada uma com configuração física diferente, como altura da câmera em relação ao solo, o que induz mudanças relevantes de ponto de vista entre plataformas.

As imagens foram capturadas por uma câmera perspectiva enquanto os robôs eram conduzidos manualmente, e o conjunto inclui, para cada sequência, dados associados de odometria e varreduras a *laser*, bem como a pose estimada durante a aquisição por um método de localizaçãor. Essas informações permitem recuperar a trajetória do robô e associar, a cada *frame*, uma pose *2D* no plano do ambiente. Embora o *dataset* original também forneça varreduras a *laser* e dados de odometria, este trabalho utiliza exclusivamente as imagens *RGB*.

As aquisições foram realizadas ao longo de seis meses, sob três condições de iluminação: nublado (*cloudy*), noturno (*night*) e ensolarado (*sunny*). As imagens possuem uma resolução nativa de 320x240 *pixels*. As figuras 9 e 10 ilustram a diversidade de ambientes e condições de iluminação presentes.

4.2.2 *COLD* (*COsy Localization Database*)

O conjunto de dados *COLD* expande o escopo experimental para três locais distintos: as universidades de *Ljubljana* (Eslovênia), *Freiburg* (Alemanha) e *Saarbrücken* (Alemanha). O *dataset* foi coletado em ambientes de laboratório, seguindo as mesmas condições de



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

(d) Kitchen (e) Printer area

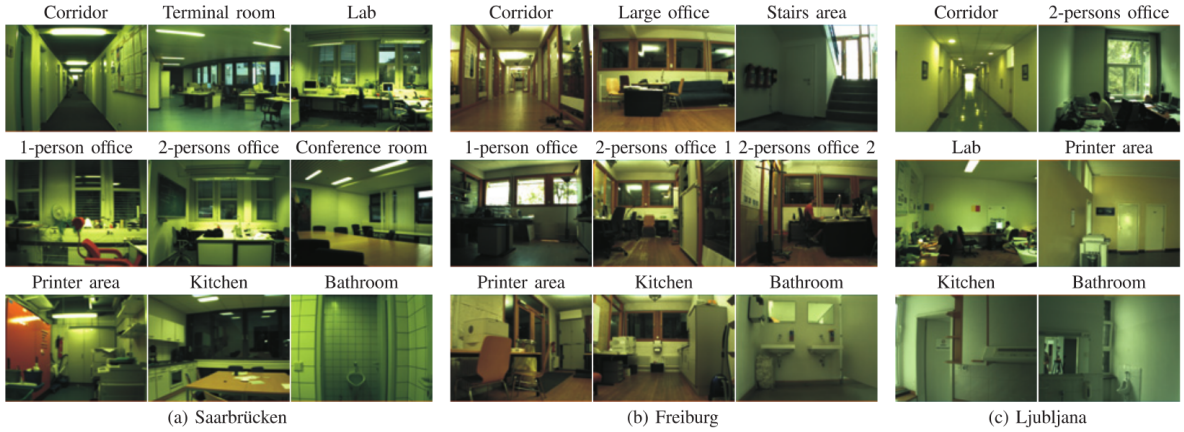
Figura 9 – Exemplo de imagens dos ambientes retiradas de uma das sequências apresentando o interior das salas, para o conjunto *KTH-IDOL2* (LUO et al., 2006).



Figura 10 – Exemplo de imagens capturadas sob diferentes condições de iluminação, para o conjunto *KTH-IDOL2* (LUO et al., 2006).

iluminação do *KTH-IDOL2*. As imagens, com resolução original de 640x480 *pixels*, são organizadas por classes semânticas, permitindo uma análise robusta da generalização dos modelos entre domínios visualmente distintos. As poses de cada imagem são fornecidas juntamente com o conjunto, sendo obtidas por localização baseada em *laser* quando

disponível; no caso de *Ljubljana*, a anotação de pose foi realizada a partir da odometria com correções manuais. A Figura 11 ilustra exemplos das imagens capturadas.



(a) Saarbrücken (b) Freiburg (c) Ljubljana

Figura 11 – Exemplos de imagens apresentando os interiores de algumas salas em cada um dos três laboratórios, para o conjunto *COLD* (PRONOBIS; CAPUTO, 2009).

Para construir os mapas semânticos, extraímos as poses disponibilizadas pelo *dataset* e as alinhamos temporalmente aos *frames*. Em particular, utilizamos o registro de pose fornecido para o instante de aquisição de cada imagem, obtido por localização baseada em *laser* quando disponível, e, quando necessário, odometria como fonte auxiliar. A trajetória é então definida pela sequência ordenada de poses, e cada predição $\hat{y}_t$ é associada ao ponto (x_t, y_t) correspondente, permitindo a projeção dos rótulos ao longo do percurso do robô no plano *2D*.

A Tabela 2 lista as classes semânticas consideradas em cada conjunto de dados. No *COLD*, o conjunto de classes varia entre locais, refletindo diferenças funcionais e estruturais dos laboratórios.

4.2.3 Pré-processamento dos Dados

Todas as amostras foram organizadas como pares alinhados (I_{RGB}, I_{Depth}) e redimensionadas para 224×224 *pixels*, compatível com o *backbone DINOv2*. Para facilitar a leitura e padronizar o carregamento, as imagens foram reorganizadas em pastas por classe. No *COLD*, cujos rótulos são fornecidos em arquivo texto, as anotações foram associadas às imagens e padronizadas no mesmo formato de organização adotado no *KTH-IDOL2*.

Dataset	**Local**	**Classes**
KTH-IDOL2	*Dumbo/Minnie*	*One-person Office*, *Corridor*, *Printer Area*, *Two-person Office*, *Kitchen*
COLD	*Freiburg*	*Two-person Office*, *Corridor*, *Printer Area*, *Stair Area*, *Bathroom*
	Ljubljana	*Two-person Office*, *Corridor*, *Printer Area*, *Laboratory*, *Bathroom*
	Saarbrücken	*One-person Office*, *Corridor*, *Printer Area*, *Kitchen*, *Bathroom*

Tabela 2 – Cenários experimentais e classes semânticas por conjunto de dados (e por local no *COLD*).

Durante o treinamento, aplicamos transformações geométricas pareadas, garantindo consistência espacial entre as modalidades. Especificamente, utilizamos: (i) espelhamento horizontal com probabilidade 0.5; (ii) espelhamento vertical com probabilidade 0.2; e (iii) rotação aleatória no intervalo $[-15°, 15°]$. Após essas operações, as imagens foram redimensionadas para 224×224 *pixels*, utilizando interpolação bicúbica para *RGB* e interpolação por vizinho mais próximo para profundidade, preservando descontinuidades e contornos estruturais do mapa.

Para simular variações fotométricas realistas apenas no domínio *RGB*, aplicamos *Color Jitter* com parâmetros $(0.3, 0.3, 0.3, 0.1)$ (brilho, contraste, saturação e matiz). Para os mapas de profundidade, nenhuma transformação fotométrica foi aplicada, evitando introduzir perturbações artificiais no domínio de intensidade da profundidade.

Em seguida, as imagens foram convertidas para tensores. Os mapas de profundidade, originalmente de um canal, foram expandidos para três canais por replicação para atender à dimensionalidade de entrada do *DINOv2*. Por fim, ambas as modalidades foram normalizadas com as estatísticas do *ImageNet* ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), mantendo o mesmo padrão de escala de entrada do *backbone*.

No processo de validação e teste, aplicamos apenas o redimensionamento para 224×224, a conversão para tensor, a replicação do canal da profundidade e a normalização, removendo a aleatoriedade introduzida pelo aumento de dados.

4.3 Configuração Experimental

Esta seção descreve os detalhes técnicos adotados para garantir reprodutibilidade, incluindo o ambiente computacional, hiperparâmetros de treinamento e estratégias de regularização empregadas para melhorar generalização e permitir comparação justa entre *baselines* e a abordagem proposta.

4.3.1 Ambiente Computacional

Todos os experimentos foram conduzidos em uma estação de trabalho *Dell* equipada com processador *Intel® Xeon™ Silver* 4114 (2.20 GHz), 128 GB de memória DDR4 e uma GPU *NVIDIA® RTX* A4000 com 16 GB de memória GDDR6. As implementações foram realizadas em *Python* utilizando os *frameworks PyTorch* (versão 2.6.0) e *CUDA* (versão 12.3.r12.3).

4.3.2 Parâmetros de Otimização e Treinamento

Os modelos foram treinados por 20 épocas, com *batch size* 16 para os experimentos multimodais. A otimização utilizou *AdamW*, com taxa de aprendizado inicial de 1×10^{-4}. Para estabilizar o ajuste fino do *DINOv2*, aplicou-se uma taxa de aprendizado $10\times$ menor no *backbone* (1×10^{-5}) e *weight decay* distinto: 0.01 no *backbone* e 0.05 nos módulos adicionados (projeções, atenção e classificador). A taxa de aprendizado foi ajustada com *CosineAnnealingLR* com $T_{\max} = 20$.

O treinamento empregou precisão mista (AMP) com *autocast* e *GradScaler*, utilizando *bfloat16* quando suportado, caso contrário *float16*. A função de perda supervisionada foi entropia cruzada com *label smoothing* de 0.1.

Nos modelos com alinhamento multimodal, a função objetivo combinou entropia cruzada com alinhamento contrastivo (*InfoNCE*) e regularização por ortogonalidade, com temperatura $\tau = 0.1$, $\lambda_{cl} = 0.10$ e $\lambda_{\perp} = 10^{-4}$.

4.3.3 Estratégias de Regularização e Critérios de Parada

Para mitigar sobreajuste e melhorar generalização, adotaram-se as seguintes estratégias:

- Regularização L2 (*Weight Decay*): foi aplicada regularização L2 diferencial. Para os pesos do *backbone* pré-treinado, utilizou-se *weight decay* de 0.01. Para o classificador (*MLP*), treinado do zero, utilizou-se 0.05, promovendo regularização mais intensa.
- Parada antecipada (*Early Stopping*): a acurácia no conjunto de validação foi monitorada ao final de cada época. O treinamento era interrompido caso não houvesse melhora por 5 épocas consecutivas, e o modelo com melhor desempenho em validação era selecionado para teste.

Além do *weight decay* diferencial, utilizou-se *gradient clipping* com norma máxima 1.0 e *dropout* no classificador (0.3), bem como *dropout* nos blocos de atenção/*MLP* (0.1). Ao final de cada época, a acurácia em validação foi monitorada e o modelo com melhor desempenho foi salvo e utilizado na etapa de teste. Esta configuração experimental foi mantida consistente em todos os testes, garantindo comparação justa entre a abordagem proposta e os modelos *baseline*.

4.4 Protocolo e Delineamento Experimental

Para avaliar rigorosamente a arquitetura proposta e validar as escolhas de seus componentes fundamentais, foi definido um protocolo experimental multifásico. O objetivo é responder às seguintes questões de pesquisa: (Q1) qual *backbone* maximiza o desempenho quando integrado ao *pipeline* multimodal completo, mantendo fixos o mecanismo de fusão e o cabeçalho de classificação? (Q2) em que medida o *MFISC* mantém robustez sob diferentes formas de deslocamento de domínio — variação temporal, mudanças de iluminação e transferência entre domínios? e (Q3) a abordagem multimodal proposta, com alinhamento entre modalidades e fusão por Atenção Cruzada Bidirecional, supera variantes monomodais e estratégias de fusão simplificadas, conforme verificado no estudo de ablação?

4.4.1 Estrutura da Avaliação e Modelos

A avaliação foi estruturada em dois estágios para isolar e quantificar a contribuição dos principais componentes da abordagem proposta.

Estágio 1: Validação comparativa de *backbones* no contexto do MFISC. No Estágio 1, buscamos identificar quais arquiteturas de *backbone* são mais eficazes quando acopladas ao *pipeline* multimodal completo, mantendo fixos o mecanismo de fusão e o cabeçalho de classificação. Para isso, o *DINOv2* foi comparado com onze arquiteturas representativas, incluindo modelos baseados em *Transformers*: *ViT* (DOSOVITSKIY et al., 2021), *BeiT* (BAO et al., 2022), *DeiT* (TOUVRON et al., 2021), *Swin Transformer V2* (LIU et al., 2022a); modelos convolucionais: *EfficientNet* (TAN; LE, 2019), *ConvNeXt* (LIU et al., 2022b), *DenseNet* (HUANG et al., 2018), *ResNet* (HE et al., 2015), *Inception-v3* (SZEGEDY et al., 2015)); arquiteturas híbridas: *MLP-Mixer* (TOLSTIKHIN et al., 2021), *ConvFormer* (GU et al., 2022)); e o modelo auto-supervisionado *CoMAE* (YANG et al., 2023b). Todos os modelos foram inicializados com pesos do *ImageNet*, quando aplicável e ajustados (*fine-tuned*) sob condições controladas para garantir uma comparação justa.

Estágio 2: Estudo de ablação da arquitetura multimodal. Com o *DINOv2* estabelecido como o *backbone* mais robusto no contexto do *MFISC*, o Estágio 3 quantifica o ganho incremental dos componentes da arquitetura multimodal. O desempenho do modelo final é comparado com variantes que removem ou adicionam módulos específicos, permitindo analisar a contribuição de (i) projeções por modalidade, (ii) alinhamento contrastivo, (iii) regularização por ortogonalidade e (iv) fusão por Atenção Cruzada Bidirecional. Além disso, comparamos diretamente a configuração multimodal com variantes monomodais, a fim de mensurar o benefício da integração entre aparência e geometria.

Além da análise comparativa de *backbones* e do estudo de ablação da arquitetura proposta, também realizamos uma comparação com *baselines* recentes da literatura de classificação *RGB-D*, executados sob o mesmo cenário experimental do estudo de ablação, de modo a assegurar uma comparação direta e justa.

4.4.2 Cenários de Teste, Divisão de Dados e Métricas

A avaliação dos modelos em cada estágio foi conduzida sobre quatro cenários de teste, projetados para avaliar sistematicamente a robustez frente a desafios centrais na robótica móvel, conforme resume a Tabela 3.

Esquema	Objetivo de Avaliação	Conjunto de Treinamento	Conjunto(s) de Teste
1. Generalização Temporal	Avaliar robustez a mudanças temporais na cena.	KTH: *cloudy* 1 + 2 COLD: *cloudy* 1 + 2	KTH: *cloudy* 3 + 4 COLD: *cloudy* 3
2. Variação de Iluminação	Medir capacidade de generalizar para iluminação não vista.	KTH: *cloudy* 1 + 2 COLD: *cloudy* 1 + 2	KTH: *night* 3 + 4; *sunny* 3 + 4 COLD: *night* 3; *sunny* 3
3. Interpolação de Iluminação	Testar se o aprendizado em condições extremas ajuda na generalização.	KTH: *cloudy* 1 + *night* 2 COLD: *cloudy* 1 + *night* 2	KTH: *sunny* 3 COLD: *sunny* 3
4. Generalização entre Domínios	Avaliar robustez a mudança de local/plataforma.	KTH: *Dumbo*, *cloudy* 1 + 2 COLD: Freiburg, *cloudy* 1 + 2	KTH: *Minnie*, *cloudy* 3 + 4 COLD: Ljubljana, *night* 3

Tabela 3 – Resumo do protocolo experimental para avaliação da abordagem proposta.

Além dos cenários de generalização temporal e variação e interpolação de iluminação, incluímos um quarto esquema voltado à generalização entre domínios. Nesse cenário, os modelos são treinados em sequências de um local sob uma condição de iluminação fixa e avaliados em um local/plataforma diferente sob a condição correspondente. No *COLD*, os diferentes laboratórios (*Freiburg*, *Ljubljana* e *Saarbrücken*) são tratados como domínios distintos. No *KTH-IDOL2*, as plataformas *Dumbo* e *Minnie* são utilizadas como domínios análogos a locais, permitindo avaliar robustez a mudanças de plataforma e ponto de vista.

Para cada experimento, o conjunto de dados de treinamento foi particionado em 80% para treino efetivo e 20% para validação, garantindo que a etapa de ajuste de hiperparâmetros não tivesse acesso ao conjunto de teste. O conjunto de teste foi mantido completamente isolado durante todo o processo e corresponde a uma das sequências específicas de cada condição de iluminação, sendo utilizado exclusivamente para a avaliação final do modelo. A métrica primária de desempenho foi a Acurácia (Equação 4.16).

$$\text{Acurácia} = \frac{N_{\text{corretos}}}{N_{\text{total}}} \tag{4.16}$$

5

Resultados e Análise dos Experimentos

Este capítulo apresenta e discute os resultados empíricos obtidos a partir do protocolo experimental descrito no capítulo anterior. O objetivo central da avaliação é verificar a hipótese de que a combinação de um *backbone* auto-supervisionado com uma arquitetura multimodal baseada em fusão por Atenção Cruzada Bidirecional, complementada por objetivos auxiliares de alinhamento e preservação de complementaridade, resulta em um modelo de classificação de cenas mais robusto e preciso para aplicações em robótica móvel.

Para organizar a análise e manter consistência com a nomenclatura adotada em publicações derivadas deste trabalho, denominamos *MFISC* (*Multimodal Fusion for Indoor Scene Categorization*) o módulo de categorização de cenas empregado na etapa de percepção, responsável por produzir rótulos semânticos a partir de entradas *RGB* e de mapas de profundidade estimados. Os resultados quantitativos reportam o desempenho do *MFISC*, enquanto a etapa de mapeamento semântico é discutida por meio da projeção dos rótulos preditos na trajetória do robô, resultando em mapas semânticos *2D* utilizados na análise qualitativa.

Ao longo dos experimentos, avaliamos o desempenho do *MFISC* em sua configuração multimodal completa, variando apenas o *backbone* de extração de características (*ConvNeXt*, *Swin Transformer V2*, *ViT*, entre outros). Para garantir uma comparação justa, todos os *backbones* são ajustados sob as mesmas condições e analisados segundo

o mesmo protocolo de teste. Dessa forma, as diferenças observadas refletem exclusivamente o impacto do extrator de características no contexto de um mesmo *pipeline*, permitindo identificar a configuração que maximiza desempenho e robustez, com destaque para o papel do *DINOv2* como extrator de representações transferíveis em condições desafiadoras.

O capítulo também incorpora um estudo de ablação que quantifica a contribuição dos principais componentes do *MFISC*: (i) entradas monomodais (*RGB-only* e profundidade-*only*); (ii) fusão por Atenção Cruzada Bidirecional; e (iii) módulos adicionais de adaptação e alinhamento multimodal. Essa análise evidencia em que medida cada componente contribui para os ganhos observados em robustez e acurácia.

No *dataset KTH-IDOL2*, avaliamos as plataformas *Dumbo* e *Minnie*, tratando-as como domínios distintos para investigar robustez a mudanças de plataforma e ponto de vista. No *COLD*, analisamos três ambientes de aquisição — *Freiburg*, *Saarbrücken* e *Ljubljana* —, o que permite avaliar generalização entre locais com diferentes configurações e estatísticas de aparência. Ambos os *datasets* incluem três condições de iluminação — ensolarado (*sunny*), e nublado (*night*) e noturno (*cloudy*) — permitindo explorar de forma controlada o impacto de variações fotométricas.

A apresentação dos resultados segue os quatro esquemas experimentais definidos na metodologia, organizados para refletir diferentes dimensões de generalização e robustez:

- Esquema 1 — Generalização Temporal na Mesma Condição de Iluminação: avalia a consistência do modelo ao longo do tempo sob iluminação constante, medindo estabilidade intra-domínio.

- Esquema 2 — Generalização para Condições de Iluminação Não Vistas: investiga a capacidade de generalizar para domínios de iluminação não observados durante o treinamento.

- Esquema 3 — Interpolação entre Condições de Iluminação: examina se o aprendizado em condições extremas melhora a generalização para uma condição intermediária.

- Esquema 4 — Generalização entre Domínios (Locais/Plataformas): avalia a robustez do modelo ao transferir conhecimento entre locais distintos no *COLD* e entre plataformas no *KTH-IDOL2*, mantendo a condição de iluminação correspondente.

Em cada esquema, apresentamos inicialmente os resultados quantitativos de desempenho e discutimos o comportamento frente às variações temporais, fotométricas e de domínio. Em seguida, realizamos o estudo de ablação em cenários representativos, detalhando como cada componente contribui para o desempenho final. Complementarmente, uma análise qualitativa dos mapas semânticos *2D* derivados das predições de classificação ilustra como as diferenças de desempenho se refletem na consistência espacial dos rótulos ao longo das trajetórias dos robôs.

5.1 Esquema 1: Generalização Temporal na Mesma Condição de Iluminação

O Esquema 1 avalia a generalização temporal do modelo sob iluminação constante e no mesmo domínio. Nesse cenário, o modelo é treinado nas sequências *1–2* e testado em sequências registradas em momentos distintos (*3–4*), mantendo a mesma condição de iluminação. As sequências foram registradas em instantes distintos e incorporam mudanças naturais do ambiente, além de variações de pose, enquadramento e exposição. Esse cenário caracteriza um *shift* intra-domínio típico de robótica móvel, no qual o objetivo trata da estabilidade temporal das predições para evitar fragmentação e "ruído semântico" no mapa *2D*.

Reportamos os resultados do *MFISC* em sua configuração completa, variando exclusivamente o *backbone* de extração de características para isolar seu efeito no desempenho temporal. A discussão dos resultados enfatiza: (i) o desempenho relativo dos *backbones*, com destaque para a consistência do *DINOv2*; (ii) a influência da condição de iluminação, mesmo em um cenário temporal; e (iii) o impacto do ponto de vista e das características da plataforma na generalização. O Esquema 1 mede a confiabilidade temporal do classificador como pré-requisito para mapas semânticos *2D* consistentes.

5.1.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 1

No *dataset KTH-IDOL2*, em todos os experimentos deste esquema, o modelo é treinado nas sequências *1–2* (por exemplo, *cloudy12*, *night12* e *sunny12*) e testado nas sequências *3–4* (por exemplo, *cloudy34*, *night34* e *sunny34*) da condição correspondente. Os resultados são apresentados por meio de gráficos de barras horizontais, facilitando a comparação direta entre configurações.

As Figuras 12 e 13 reportam as acurácias do *MFISC* em sua configuração multimodal completa, variando exclusivamente o *backbone* de extração de características. Assim, cada barra corresponde ao desempenho do *MFISC+backbone*, mantendo fixos os módulos de projeção por modalidade, alinhamento contrastivo, regularização por ortogonalidade e fusão por Atenção Cruzada Bidirecional.

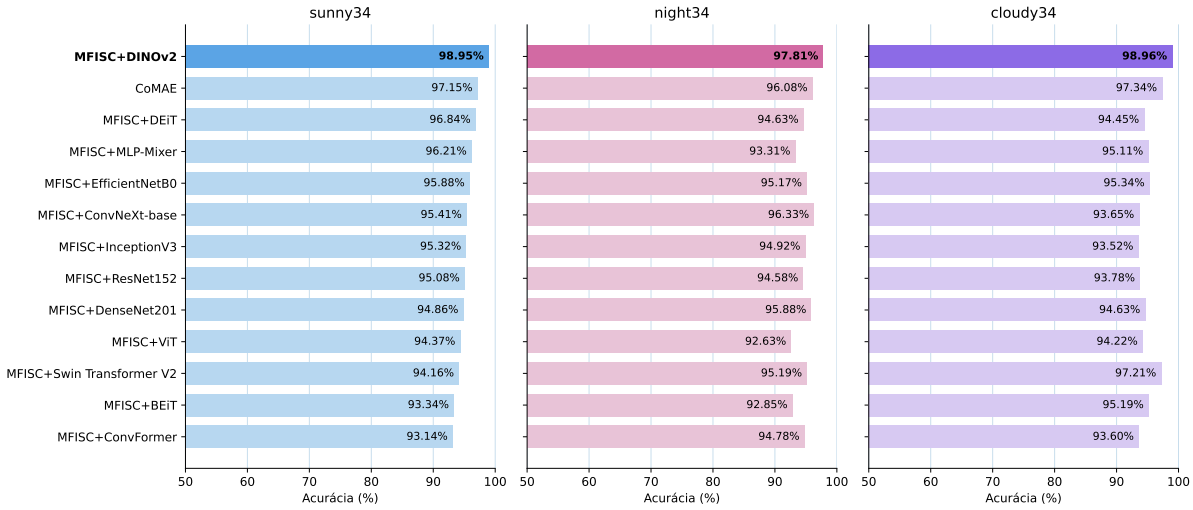


Figura 12 – Resultados do Esquema 1 no *KTH-IDOL2* (plataforma *Dumbo*): acurácia (%) do *MFISC* em generalização temporal, treinando nas sequências *1–2* e testando em *3–4* sob *sunny34*, *night34* e *cloudy34*.

Na plataforma *Dumbo* (Figura 12), observa-se desempenho elevado e estável em todas as condições fotométricas. O *MFISC+DINOv2* atinge as melhores acurácias do esquema, com 98.96% na condição nublada (*cloudy*), 98.95% em ensolarada (*sunny*) e 97.81% na condição noturna (*night*). Como referência auto-supervisionada adicional, o *MFISC+CoMAE* também apresenta resultados altos, porém consistentemente inferiores ao *DINOv2* (97.34% em *cloudy*, 97.15% em *sunny* e 96.08% em *night*), indicando que o

DINOv2 oferece representações mais favoráveis para maximizar o ganho do *pipeline* multimodal neste cenário temporal. *Backbones* supervisionados como *DEiT*, *ConvNeXt* e *Swin Transformer V2* permanecem competitivos, mas sem superar o *DINOv2* nas três condições.

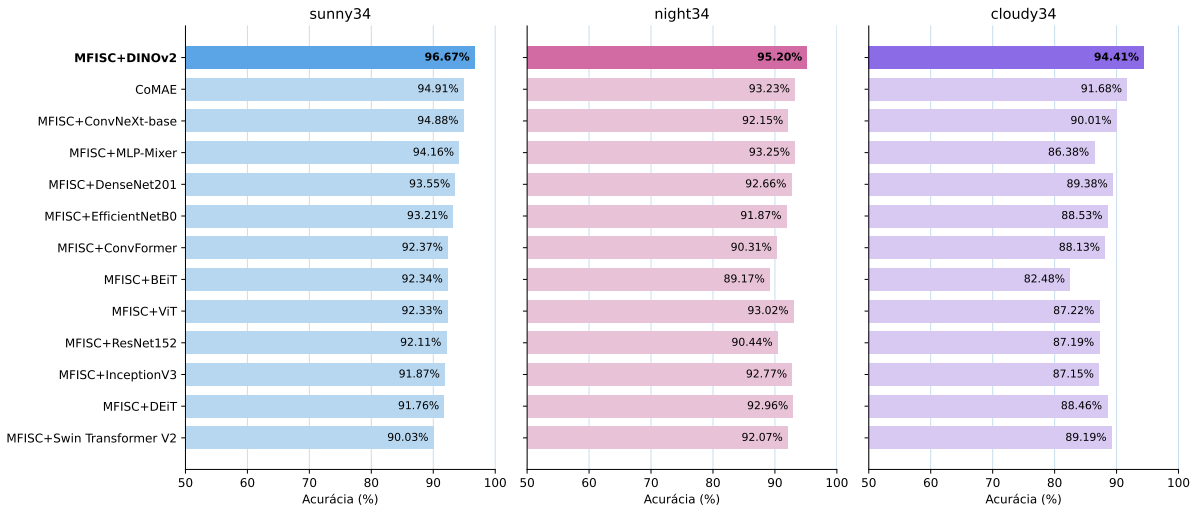


Figura 13 – Resultados do Esquema 1 no *KTH-IDOL2* (plataforma *Minnie*): acurácia (%) do *MFISC* em generalização temporal, treinando nas sequências *1–2* e testando em *3–4* sob *sunny34*, *night34* e *cloudy34*.

Na plataforma *Minnie* (Figura 13), o *MFISC+DINOv2* apresenta a melhor performance, alcançando 96.67% na condição ensolarada e mantendo acurácias elevadas nas condições noturna (95.20%) e nublada (94.41%). O *MFISC+CoMAE* aparece como alternativa auto-supervisionada forte, mas com queda mais pronunciada em relação ao *DINOv2*, especialmente na condição nublada (94.91% em *sunny*, 93.23% em *night* e 91.68% em *cloudy*). Observa-se ainda maior dispersão entre *backbones* em comparação à plataforma *Dumbo*, sugerindo que diferenças de ponto de vista e dinâmica de navegação podem amplificar as discrepâncias entre extratores mesmo em um cenário temporal.

Em síntese, o Esquema 1 evidencia que o *MFISC* apresenta alta estabilidade temporal no *KTH-IDOL2* e que a escolha do *DINOv2* como *backbone* maximiza o desempenho de forma consistente nas duas plataformas. A comparação com o *CoMAE* sugere que, embora outros extratores auto-supervisionados também sejam competitivos, o *DINOv2* fornece representações mais eficazes para explorar o potencial da fusão multimodal proposta.

5.1.2 Resultados no *Dataset COLD* para o Esquema 1

No *COLD*, em cada ambiente, o modelo é treinado nas sequências *1–2* (por exemplo, *sunny12*, *night12* e *cloudy12*) e testado na sequência *3* da condição correspondente (*sunny3*, *night3* e *cloudy3*). Para facilitar a comparação entre configurações, os resultados são apresentados em gráficos de barras horizontais.

As Figuras 14–16 reportam as acurácias do *MFISC* em sua configuração multimodal completa, variando exclusivamente o *backbone* de extração de características. Assim, cada barra corresponde ao desempenho do *MFISC+backbone*, mantendo fixos os módulos de projeção por modalidade, alinhamento contrastivo, regularização por ortogonalidade e fusão por Atenção Cruzada Bidirecional.

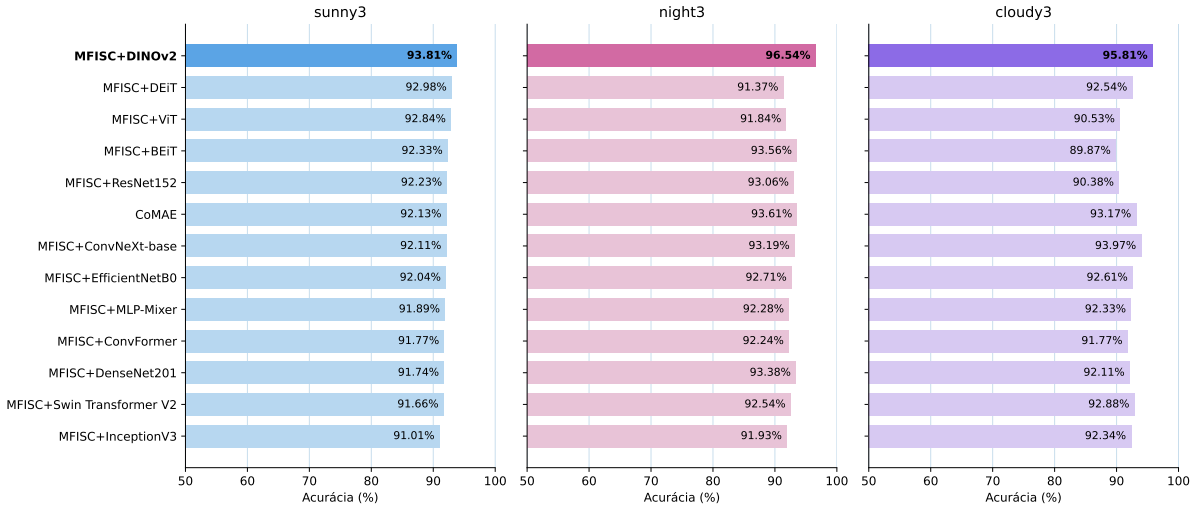


Figura 14 – Esquema 1 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* para cada condição (*sunny3*, *night3*, *cloudy3*).

No ambiente *Freiburg* (Figura 14), o *MFISC+DINOv2* obtém o melhor desempenho em todas as condições, atingindo 93.81% na condição ensolarada (*sunny3*), 96.54% em noturno (*night3*) e 95.81% em nublado (*cloudy3*). Na condição nublada, observa-se aproximação de um *backbone* supervisionado forte, *ConvNeXt-base* (93.97%), mas em *night3* a vantagem do *DINOv2* permanece mais evidente, sugerindo maior estabilidade temporal quando a aparência torna-se menos informativa.

Em *Ljubljana* (Figura 15), o *MFISC+DINOv2* novamente apresenta as maio-

res acurácias e mantém consistência entre as três condições: 97.21% (*sunny*), 96.59% (*night*) e 96.74% (*cloudy*). O *CoMAE* aparece como o segundo melhor extrator auto-supervisionado na comparação, alcançando 94.38% em *sunny*, 94.25% em *night* e 93.69% em *cloudy*, mas ainda com margem relevante em relação ao *DINOv2*. Nota-se maior dispersão entre *backbones* supervisionados, sobretudo na condição noturna, onde há queda acentuada para *ConvNeXt-base* (84.16%), indicando sensibilidade temporal mesmo sob iluminação constante; em contraste, o *DINOv2* preserva alta estabilidade, reforçando sua adequação para cenários com maior complexidade visual.

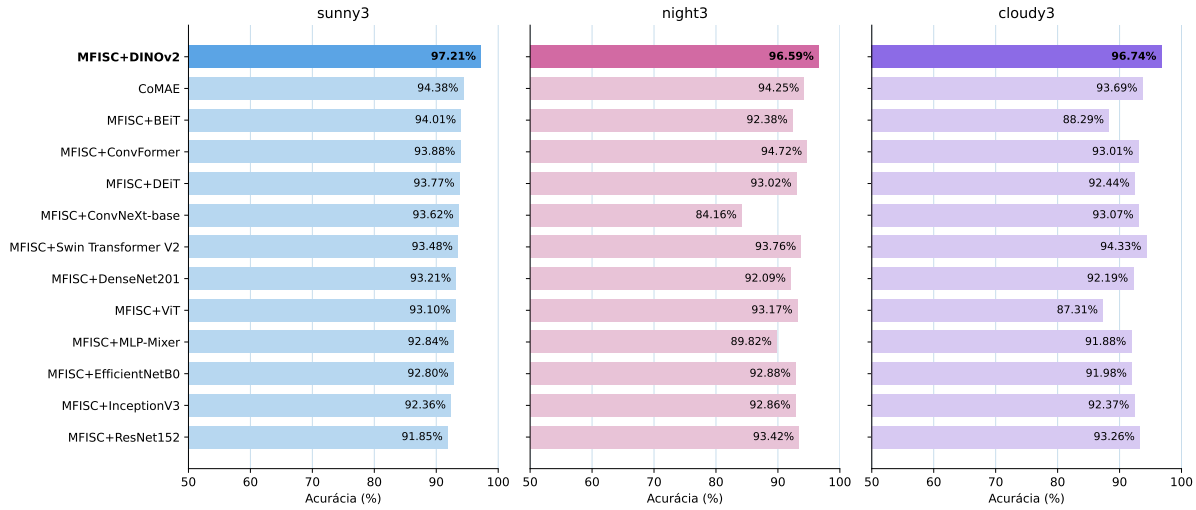


Figura 15 – Esquema 1 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* por condição.

Para *Saarbrücken* (Figura 16), o *MFISC+DINOv2* lidera em todas as condições, com 96.59% em ensolarado (*sunny*), 95.01% em noturno (*night*) e 97.06% em nublado (*cloudy*). Um aspecto relevante é que, na condição nublada, *ConvNeXt-base* (96.52%) e *ConvFormer* (96.31%) se aproximam do topo, sugerindo que, nesse domínio, o deslocamento temporal é relativamente pequeno e múltiplos extratores conseguem sustentar alto desempenho quando acoplados ao *MFISC*.

Em síntese, os resultados do *COLD* no Esquema 1 confirmam a alta estabilidade temporal do *MFISC* e indicam que o *DINOv2* é o *backbone* que maximiza o desempenho de forma mais consistente nos três ambientes e condições. Além disso, a inclusão do *CoMAE* como alternativa auto-supervisionada reforça a tendência observada: embora

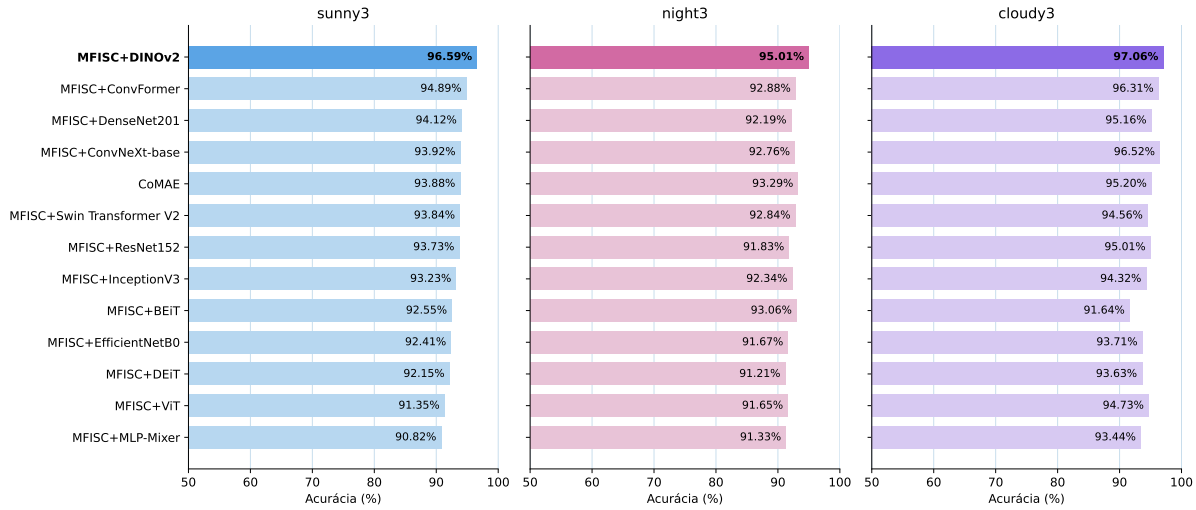


Figura 16 – Esquema 1 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *seq. 1–2* e teste em *seq. 3* por condição.

extratores *SSL* recentes já sejam competitivos, o *DINOv2* mantém vantagem sistemática quando integrado ao *pipeline* multimodal, especialmente em cenários mais desafiadores (*night3*).

5.2 Esquema 2: Generalização para Condições de Iluminação Não Vistas

O Esquema 2 avalia a capacidade de generalizar para condições de iluminação não vistas durante o treinamento. Aqui o modelo é treinado em uma única condição e avaliado diretamente em condições distintas, sem qualquer ajuste adicional. Esse delineamento induz um deslocamento fotométrico controlado, representando cenários em que o robô precisa manter desempenho diante de mudanças bruscas de iluminação ao longo da navegação, mantendo fixos o domínio e a categoria de cenas.

Apresentamos os resultados do *MFISC* em sua configuração multimodal completa, mantendo fixos os módulos auxiliares e a fusão por Atenção Cruzada Bidirecional. Variamos apenas (i) a condição de iluminação utilizada no treinamento e (ii) o *backbone* de extração de características, de modo a isolar o efeito do extrator sob deslocamento

fotométrico. Para manter a apresentação concisa, reportamos nesta seção um subconjunto representativo das configurações. Os resultados complementares para todas as combinações avaliadas estão disponíveis no Apêndice A.

5.2.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 2

Neste conjunto, utilizamos a configuração de treinamento com as sequências 1 e 2 da condição *night* e testes nas sequências 3 e 4 das condições *sunny* e *cloudy*, para a plataforma *Dumbo*. Para *Minnie*, o treinamento foi realizado na condição *sunny* e treinado nas demais condições.

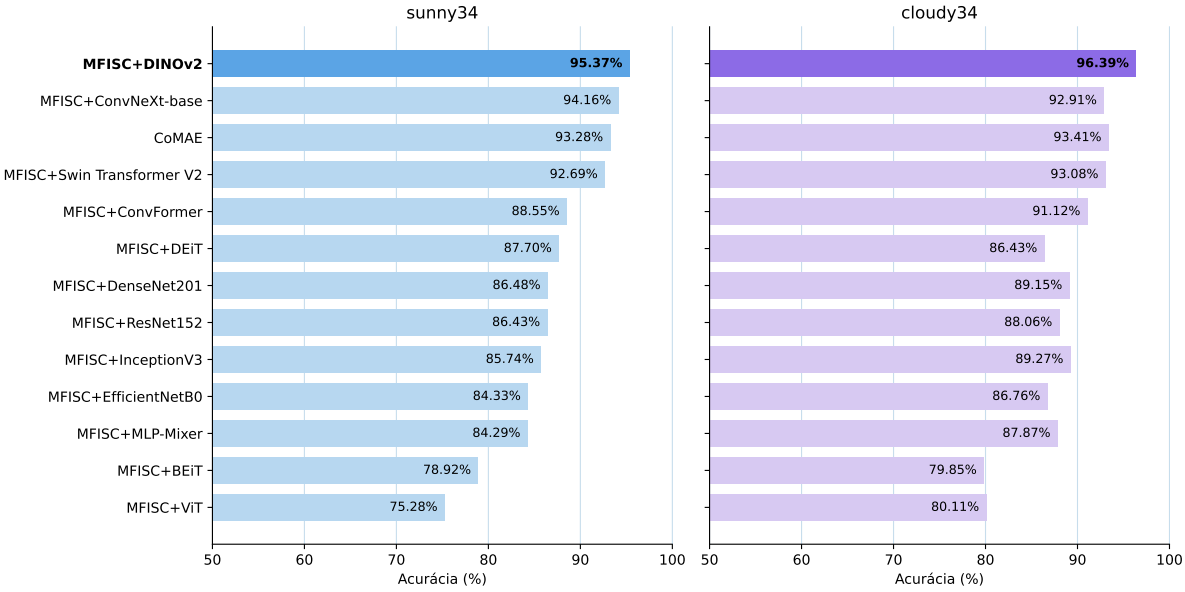


Figura 17 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *sunny34* e *cloudy34*.

A Figura 17 apresenta os resultados ao treinar em noturno (*night*) e testar em ensolarado (*sunny*) e nublado (*cloudy*). Nessa transferência, o *MFISC+DINOv2* atinge as maiores acurácias em ambos os testes, com 95.37% em *sunny* e 96.39% em *cloudy*. A *ConvNeXt-base* se aproxima em *sunny* (94.16%), enquanto em *cloudy* os melhores supervisionados permanecem abaixo do *DINOv2*, como *Swin Transformer V2* (93.08%) e *ConvNeXt-base* (92.91%). Esses resultados indicam que, mesmo sob deslocamento foto-

métrico severo, o *MFISC* preserva desempenho elevado quando acoplado ao *DINOv2*.

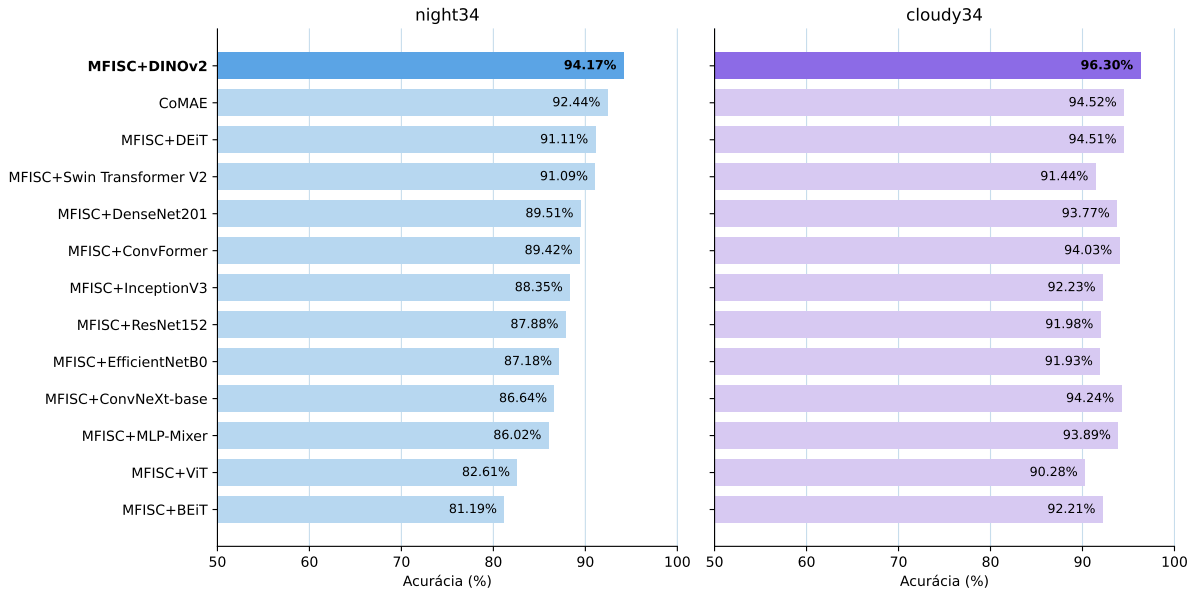


Figura 18 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *night34* e *cloudy34*.

A Figura 18 reporta o treinamento na condição ensolarada e testes em noturno e nublado. Novamente, o *MFISC+DINOv2* lidera, alcançando 94.17% em *night* e 96.30% em *cloudy*. Observa-se maior dispersão entre *backbones* na condição noturna, com quedas mais acentuadas em modelos como *BEiT* (81.19%) e *ViT* (82.61%), enquanto alternativas como *DeiT* (91.11%) e *Swin Transformer V2* (91.09%) permanecem competitivas, porém abaixo do *DINOv2*. Vale notar que o *CoMAE* surge como uma alternativa autosupervisionada competitiva, atingindo 92.44% em *night* e 94.52% em *cloudy*, ainda que abaixo do *DINOv2*.

Em síntese, o Esquema 2 impõe um deslocamento fotométrico mais severo que o Esquema 1 e evidencia que o *MFISC* acoplado ao *DINOv2* oferece a configuração mais consistente para lidar com iluminação não vista nas duas plataformas avaliadas.

5.2.2 Resultados no *Dataset COLD* para o Esquema 2

No conjunto *COLD* reportamos três configurações representativas, cada uma em um local diferente (*Freiburg*, *Ljubljana* e *Saarbrücken*). O treinamento foi realizado nas sequências 1 e 2 na condição noturna e testes na sequência 3 das condições ensolarada e nublada, para Freiburg. Para os demais ambientes, adotamos o mesmo protocolo, alterando apenas a condição de treinamento e o local considerado.

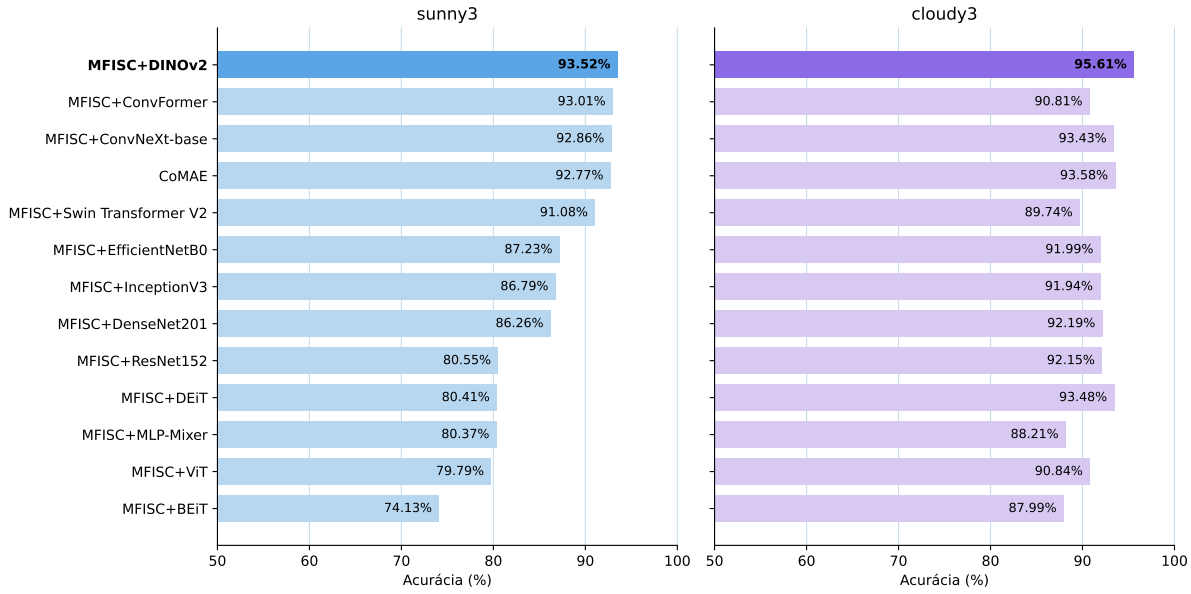


Figura 19 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *sunny3* e *cloudy3*.

A Figura 19 apresenta os resultados ao treinar em noturno (*night*) e testar em ensolarado (*sunny*) e nublado (*cloudy*), no ambiente *Freiburg*. De forma geral, observa-se que o deslocamento noturno→ensolarado tende a ser mais severo do que noturno→nublado, refletindo a diferença fotométrica mais acentuada entre noite e ensolarado. O *MFISC* acoplado ao *DINOv2* tipicamente sustenta as maiores acurácias nos dois testes, alcançando 93.52% na condição ensolarada e 95.61% em nublado, o que indica maior estabilidade de representação sob mudanças de iluminação.

A Figura 20 reporta o treinamento na condição ensolarada e testes nas condições nublado e noturno. Aqui, o deslocamento ensolarado→noturno costuma induzir queda mais pronunciada para diversos *backbones*, pois reduz fortemente contraste e visibilidade

de estruturas. Ainda assim, o *MFISC* com *DINOv2* tende a manter vantagem consistente, obtendo 97.56% de acurácia para a condição nublado e 95.81% para noturno. Esses resultados sugerem que as representações auto-supervisionadas capturam invariâncias úteis mesmo quando textura e cor se tornam menos confiáveis.

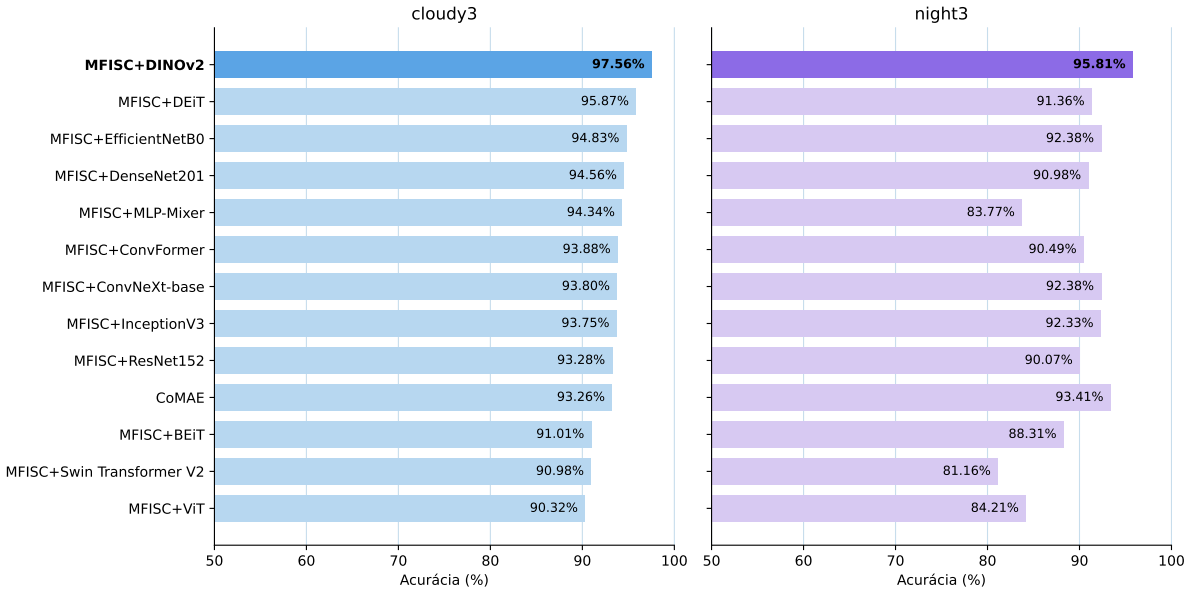


Figura 20 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*.

A Figura 21 apresenta a configuração de treino em nublado (*cloudy*) e testes em noturno (*night*) e ensolarado (*sunny*). Esse caso é interessante por usar nublado como condição intermediária, frequentemente servindo como um ponto de partida menos extremo do que noturno e ensolarado. A tendência é apresentar uma transferência mais equilibrada para ambos os testes, embora a direção nublado→noturno ainda imponha desafios por perda de informação fotométrica. Novamente, o *MFISC++DINOv2* costuma apresentar o melhor compromisso entre desempenho e estabilidade entre os dois domínios de testem, alcançando 96.42% em noturno e 95.37% em ensolarado.

Em síntese, o Esquema 2 no *COLD* evidencia que a mudança de iluminação introduz um deslocamento de domínio mais severo do que a variação temporal do Esquema 1, e reforça que o *MFISC* acoplado ao *DINOv2* é a configuração mais consistente para generalização fotométrica nos três ambientes avaliados. Esses resultados complementam a análise do *KTH-IDOL2*, mostrando que a robustez do *MFISC* sob

iluminação não vista se mantém também em locais com diferentes configurações e estatísticas de aparência.

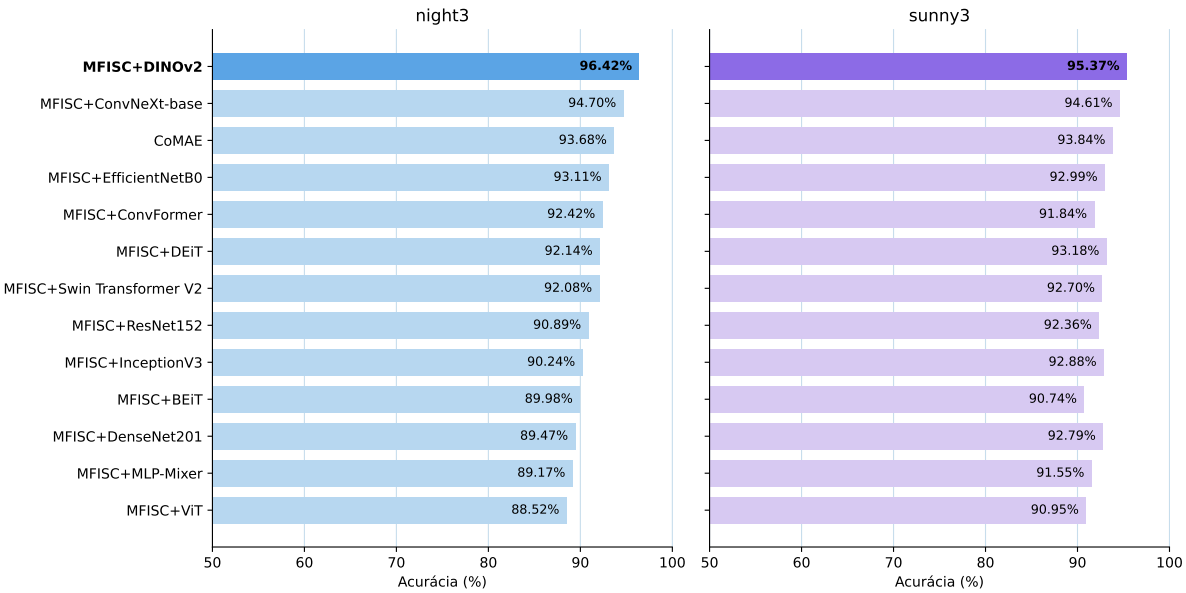


Figura 21 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*.

5.3 Esquema 3: Interpolação entre Iluminações

O Esquema 3 investiga se a maior diversidade de iluminação no treinamento melhora a generalização para uma condição não utilizada no aprendizado. Para isso, o *MFISC* é treinado com duas condições de iluminação e avaliado diretamente na terceira condição restante. Esse delineamento permite analisar a habilidade do modelo em interpolar conhecimento entre domínios fotométricos distintos e verificar se a exposição conjunta a variações complementares de luz favorece representações mais invariantes.

Reportamos os resultados (acurácia %) do *MFISC* em sua configuração multimodal completa, mantendo fixos os módulos auxliares e a fusão por Atenção Cruzada Bidirecional. Variamos apenas o *backbone* de extração de características, de modo a isolar seu efeito sob o cenário de interpolação fotométrica.

A avaliação do Esquema 3 é composta por três configurações de treino e teste,

cobrindo todas as combinações de interpolação entre as condições ensolarado (*sunny*), nublado (*cloudy*) e noturno (*night*). Em cada configuração, o treinamento utiliza as sequências *1–2* das duas condições selecionadas, e o teste é realizado na sequência *3* da condição restante:

- **Teste em *sunny3*:** treino com *cloudy1* + *night2*.
- **Teste em *cloudy3*:** treino com *sunny1* + *night2*.
- **Teste em *night3*:** treino com *cloudy1* + *sunny2*.

5.3.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 3

No *KTH-IDOL2*, o *MFISC* é treinado com duas condições e testado na terceira condição restante. Como essa configuração não fica explícita nos gráficos, reforçamos abaixo o protocolo adotado (válido para *Dumbo* e *Minnie*):

- Teste em *sunny34*: treino em *cloudy1* + *night2*.
- Teste em *night34*: treino em *cloudy1* + *sunny2*.
- Teste em *cloudy34*: treino em *sunny1* + *night2*.

As Figuras 22 e 23 apresentam as acurácias do *MFISC* variando apenas o *backbone* de extração. Assim, cada barra corresponde ao desempenho do *MFISC+backbone* sob o cenário de interpolação fotométrica.

Conforme os resulados apresentados na Figura 22 (*Dumbo*), o *MFISC+DINOv2* lidera nos três testes, alcançando 98.11% em ensolarado (*sunny*), 95.88% em noturno (*night*) e 98.74% em nublado (*cloudy*). Observa-se que o teste em noturno tende a ser o mais desafiador, com maior dispersão entre *backbones* (por exemplo, *BEiT*: 84.39%; *ViT*: 87.72%), enquanto os testes em ensolarado e nublado concentram acurácias mais altas e próximas do teto, como visto no *CoMAE*. Ainda assim, o *DINOv2* mantém a melhor estabilidade, indicando que a exposição conjunta a duas iluminações no treino favorece representações mais invariantes quando acopladas ao *MFISC*.

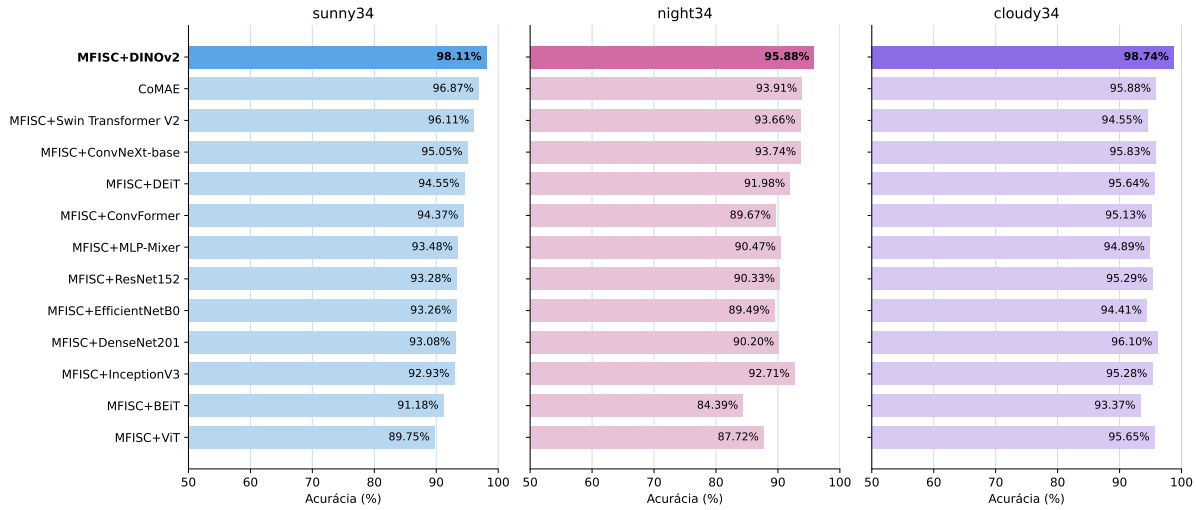


Figura 22 – Esquema 3 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny34* (treino *cloudy1+night2*); teste em *night34* (treino *cloudy1+sunny2*); teste em *cloudy34* (treino *sunny1+night2*).

Na Figura 23 (*Minnie*), o *MFISC+DINOv2* novamente apresenta os melhores resultados: 94.10% em ensolarado (*sunny*), 95.50% em noturno (*night*) e 96.18% em nublado (*cloudy*). Um aspecto notável é que, em *sunny*, a diferença entre *backbones* se amplia (por exemplo, *BEiT*: 84.77%; *ViT*: 85.50%), sugerindo que, para *Minnie*, a condição ensolarada pode induzir maior variabilidade de aparência e exposição. Já em noturno e nublado, alguns *backbones* supervisionados se aproximam (por exemplo, *ConvNeXt-base*: 93.42% em *night*), mas o *DINOv2* preserva a liderança. O modelo auto-supervisionado *CoMAE* aparece como backbone competetitivo, superando métodos supervisionados na condição ensolarada.

Os resultados do Esquema 3 no *KTH-IDOL2* mostram que treinar o *MFISC* com duas iluminações tende a sustentar acurácias elevadas ao testar na terceira condição, especialmente quando o *DINOv2* é utilizado como extrator. Enquanto o Esquema 2 avalia a transferência a partir de uma única iluminação, o Esquema 3 evidencia que a diversidade fotométrica no treino pode reduzir o impacto do deslocamento de domínio e melhorar a estabilidade do modelo.

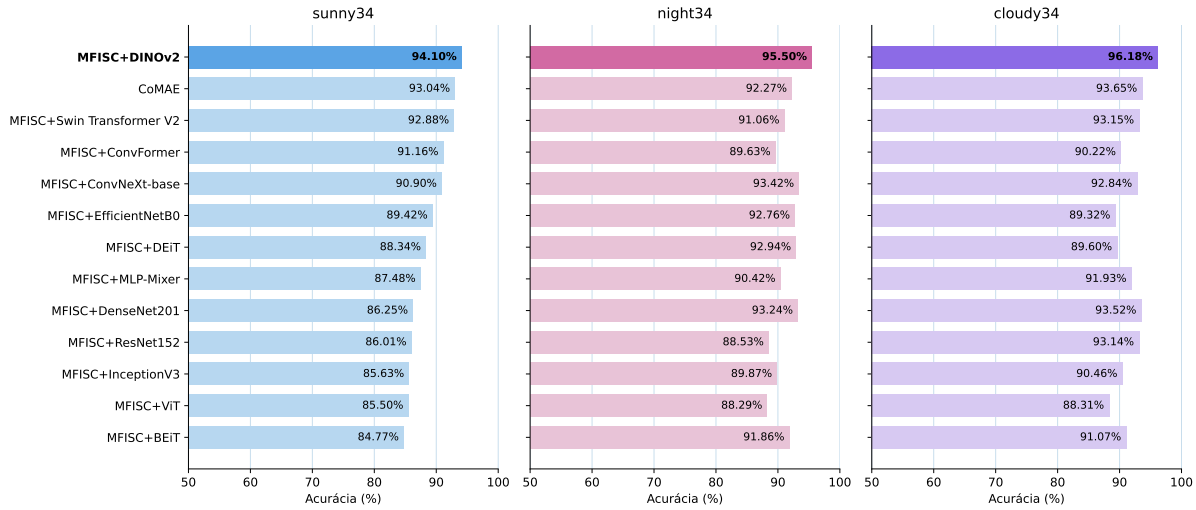


Figura 23 – Esquema 3 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny34* (treino *cloudy1+night2*); teste em *night34* (treino *cloudy1+sunny2*); teste em *cloudy34* (treino *sunny1+night2*).

5.3.2 Resultados no *Dataset COLD* para o Esquema 3

No *COLD*, o Esquema 3 avalia a capacidade de interpolação entre condições de iluminação no mesmo local, verificando se treinar o *MFISC* com duas iluminações favorece a generalização para a terceira condição restante. Reforçamos abaixo o protocolo adotado, aplicado separadamente em cada ambiente (*Freiburg*, *Ljubljana* e *Saarbrücken*):

- **Teste em *sunny3*:** treino em *cloudy1* + *night2*.
- **Teste em *night3*:** treino em *cloudy1* + *sunny2*.
- **Teste em *cloudy3*:** treino em *sunny1* + *night2*.

As Figuras 24, 25 e 26 reportam as acurácias do *MFISC*, variando apenas o *backbone* de extração. Assim, cada barra representa o desempenho do *MFISC+backbone* no cenário de interpolação fotométrica.

Em *Freiburg* (Figura 24), o *MFISC+DINOv2* apresenta o melhor desempenho nos três testes, com 94.85% em ensolarado (*sunny*), 94.24% em noturno (*night*) e 96.67% em nublado (*cloudy*). Observa-se que o teste em condição noturna tende a aumentar a

dispersão entre *backbones*, indicando maior sensibilidade a variações fotométricas mais severas, enquanto nublado concentra as maiores acurácias.

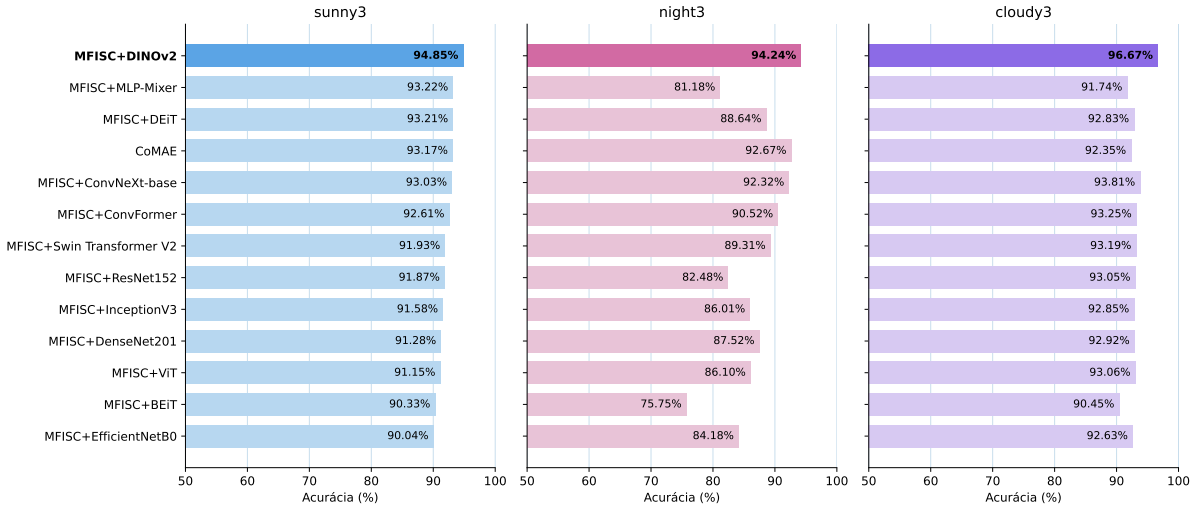


Figura 24 – Esquema 3 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*).

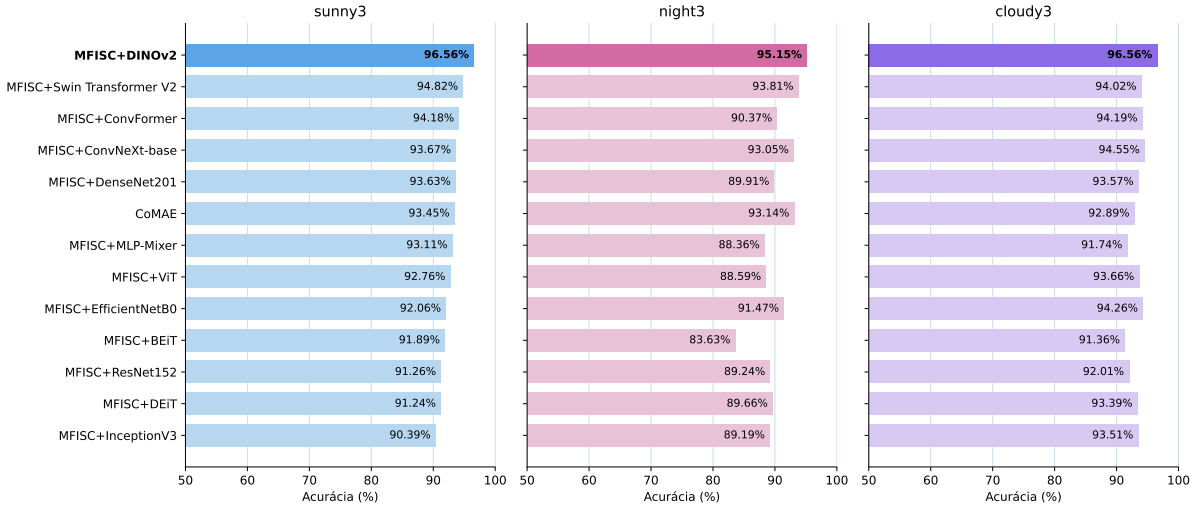


Figura 25 – Esquema 3 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*).

Em *Ljubljana* (Figura 25), o *MFISC+DINOv2* novamente lidera, atingindo 96.56% em ensolarado (*sunny3*), 95.15% em noturno (*night3*) e 96.56% em nublado (*cloudy3*). O

cenário mantém acurácias elevadas mesmo sob interpolação, sugerindo que a exposição conjunta a duas iluminações no treino contribui para reduzir a queda de desempenho no domínio de teste restante.

Em *Saarbrücken* (Figura 26), o *MFISC+DINOv2* também obtém os melhores resultados, com 96.30% em ensolarado, 95.57% em noturno e 95.39% em nublado. Nota-se ainda que alguns *backbones* podem apresentar quedas acentuadas em condições específicas (casos isolados de baixa acurácia), reforçando que o cenário de interpolação é útil para expor fragilidades de generalização fotométrica que não aparecem em avaliações intra-iluminação.

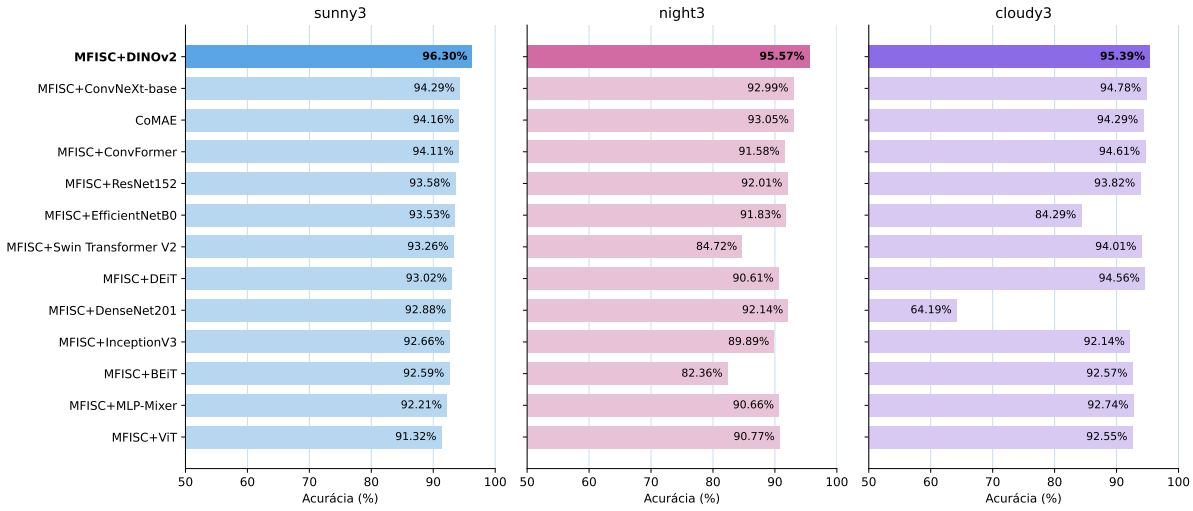


Figura 26 – Esquema 3 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*. Protocolo de interpolação: teste em *sunny3* (treino *cloudy1+night2*); teste em *night3* (treino *cloudy1+sunny2*); teste em *cloudy3* (treino *sunny1+night2*).

Os resultados do Esquema 3 no *COLD* indicam que treinar o *MFISC* com duas condições de iluminação sustenta acurácias elevadas ao testar na terceira condição, com desempenho consistentemente superior quando o *DINOv2* é utilizado como extrator. Esse comportamento complementa o Esquema 2 que mede generalização a partir de uma iluminação, o Esquema 3 evidencia o ganho de robustez quando o treinamento incorpora diversidade fotométrica.

5.4 Esquema 4: Generalização entre Domínios (Locais e Plataformas)

O Esquema 4 avalia a robustez do método sob mudança de domínio, quando o modelo é treinado em um domínio fonte e testado em um domínio alvo distinto. No *KTH-IDOL2*, essa transferência ocorre entre plataformas (*Dumbo* $\rightarrow$ *Minnie* ou vice-versa), capturando variações de ponto de vista, altura da câmera e dinâmica de navegação. No *COLD*, a transferência ocorre entre locais (*Freiburg* $\rightarrow$ *Ljubljana*/*Saarbrücken*), refletindo mudanças de *layout*, materiais, iluminação e distribuição de classes. Esse é o cenário mais crítico para mapeamento semântico, pois erros sistemáticos de domínio tendem a se propagar ao longo da trajetória, degradando a coerência espacial do mapa *2D*.

Nesse arranjo, o *MFISC* é treinado nas sequências 1 e 2 de uma condição de iluminação em um domínio-fonte (por exemplo, *Freiburg-sunny12*) e avaliado em um domínio-alvo diferente, mantendo o modelo congelado e sem ajustes adicionais. Para ampliar o desafio, o teste é realizado nas demais condições de iluminação do domínio-alvo (por exemplo, *Ljubljana-night3* e *Ljubljana-cloudy3*), de modo que o experimento combine simultaneamente mudança espacial e mudança fotométrica.

Reportamos a acurácia (%) do *MFISC* em sua configuração multimodal completa, mantendo fixos os módulos auxiliares e a fusão por Atenção Cruzada Bidirecional. Variamos apenas o *backbone* de extração de características, permitindo isolar seu efeito sob um cenário de transferência mais severo. Assim, as diferenças observadas refletem o quanto cada extrator contribui para a robustez do *pipeline* frente ao deslocamento de domínio, requisito essencial para aplicações reais de robótica móvel, nas quais mudanças de plataforma e ambientes são frequentes.

Apresentamos os resultados separadamente para *KTH-IDOL2*, tratando *Dumbo* e *Minnie* como domínios distintos, e *COLD*, avaliando transferência entre *Freiburg*, *Ljubljana* e *Saarbrücken*. Para manter a apresentação concisa, reportamos nesta seção um subconjunto representativo das configurações. Os resultados completos para todas as combinações avaliadas estão disponíveis no Apêndice A.

5.4.1 Resultados no *Dataset KTH-IDOL2* para o Esquema 4

No *KTH-IDOL2*, o *MFISC* é treinado nas sequências *1 e 2* de uma plataforma (domínio-fonte) e testado nas sequências *3 e 4* da outra plataforma (domínio-alvo), mantendo fixa a condição de iluminação correspondente. Reportamos os resultados em três condições, sempre utilizando o *MFISC* multimodal completo e variando apenas o *backbone* de extração. As figuras a seguir apresentam os resultados para as plataformas *Dumbo* e *Minnie*, respectivamente.

A Figura 27 apresenta os resultados de transferência *Dumbo* → *Minnie*. O *MFISC + DINOv2* obtém as melhores acurácias nas três condições, alcançando 91.12% em ensolarado (*sunny34*), 92.04% em noturno (*night34*) e 91.34% em nublado (*cloudy34*). Observa-se que, mesmo sob mudança de plataforma, o desempenho permanece acima de 90% nas três iluminações, sugerindo maior estabilidade das representações extraídas quando acopladas ao mecanismo de fusão do *MFISC*. O *MFISC+CoMAE* surge com acurácias competitivas nas três condições de iluminação, superando os modelos supervisionados, mas ainda abaixo do *MFISC+DINOv2*.

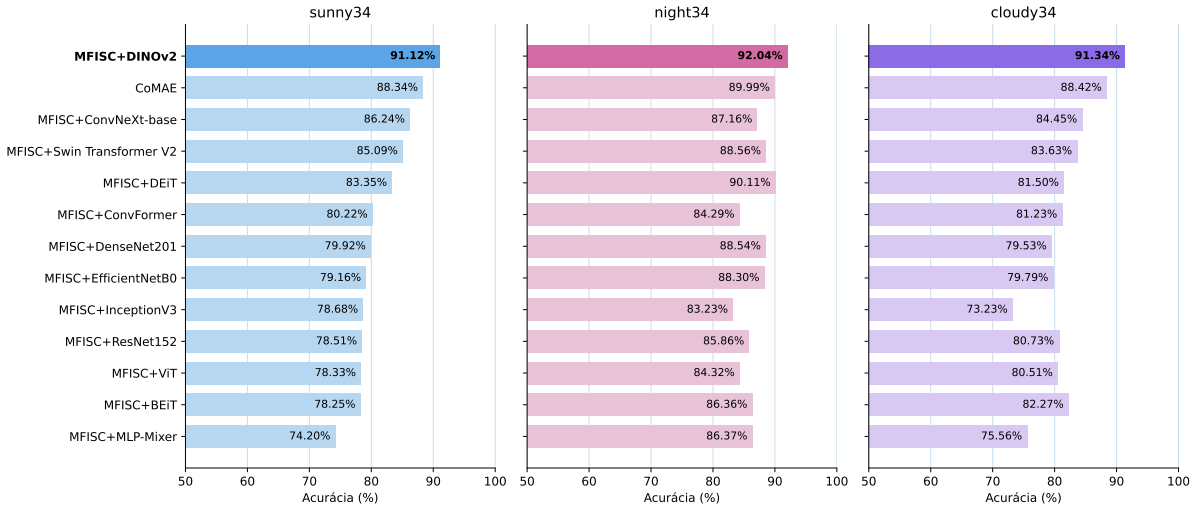


Figura 27 – Esquema 4 no *KTH-IDOL2* (transferência *Dumbo* → *Minnie*). Acurácia do *MFISC* variando apenas o *backbone*.

A Figura 28 reporta a transferência inversa, *Minnie* → *Dumbo*. Novamente, o *MFISC+DINOv2* lidera em todas as condições, com 87.16% em ensolrado (*sunny34*), 87.59% em noturno (*night34*) e 89.62% em nublado (*cloudy34*). Nota-se uma queda

moderada nas acurácias, indicando que a direção de transferência pode influenciar a severidade do deslocamento de domínio, como diferenças de ponto de vista, dinâmica de captura e distribuição de amostras por classe entre as plataformas.

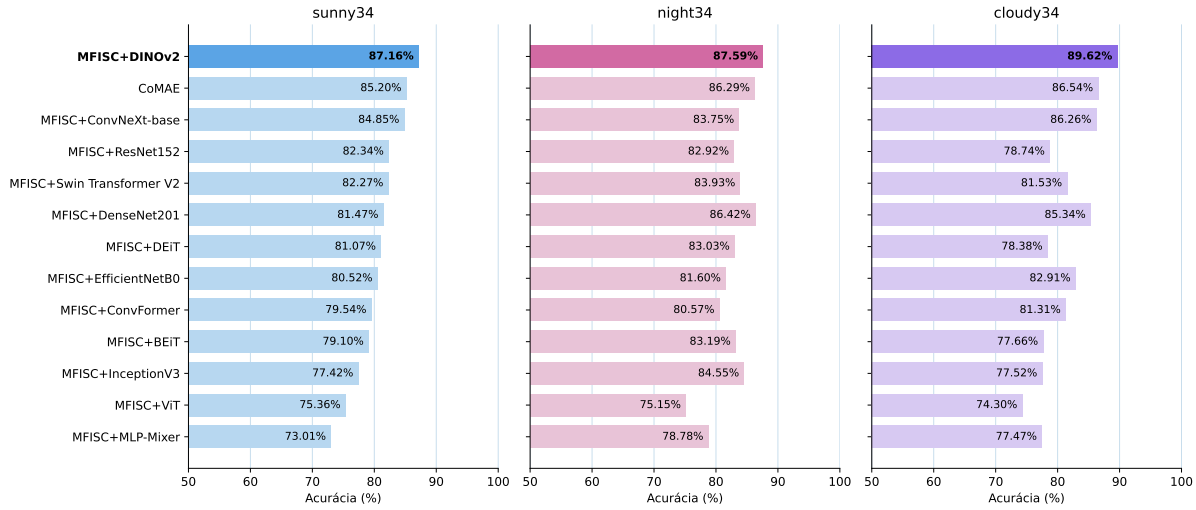


Figura 28 – Esquema 4 no *KTH-IDOL2* (transferência *Minnie* → *Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*.

Em síntese, os resultados do Esquema 4 confirmam que o cenário de mudança de plataforma é mais desafiador do que a generalização temporal (Esquema 1), mas o *MFISC* mantém desempenho competitivo, especialmente quando combinado ao *DINOv2* como extrator de características. Esses achados reforçam que a robustez do *pipeline* depende não apenas da fusão multimodal, mas também da capacidade do extrator em fornecer representações transferíveis sob deslocamento de domínio.

5.4.2 Resultados no *Dataset COLD* para o Esquema 4

No *COLD*, o Esquema 4 avalia a transferência entre locais, caracterizando um cenário de deslocamento de domínio mais severo do que os esquemas intra-local. Selecionamos um experimento representativo por local-fonte. Em todos os casos, treinamos o *MFISC* nas sequências 1 e 2 de uma única condição de iluminação no local-fonte e testamos no local-alvo em duas condições restantes (sequência 3). As Figuras 29, 30 e 31 reportam as acurácias do *MFISC*, variando apenas o *backbone*.

A Figura 29 mostra os resultados a partir do treinamento em *Freiburg-sunny12* e teste em *Ljubljana-cloudy3* e *Ljubljana-night3*. O *MFISC+DINOv2* obtém as melhores acurácias, com 63.10% na condição nublado (*cloudy3*) e 62.17% em noturno (*night3*). O *MFISC+CoMAE* surge com resultados competitivos, mas ainda abaixo da arquitetura completa do *MFISC*. Nota-se que os melhores *backbones* supervisionados ficam próximos, indicando que o cenário ainda preserva alguma transferibilidade entre os ambientes.

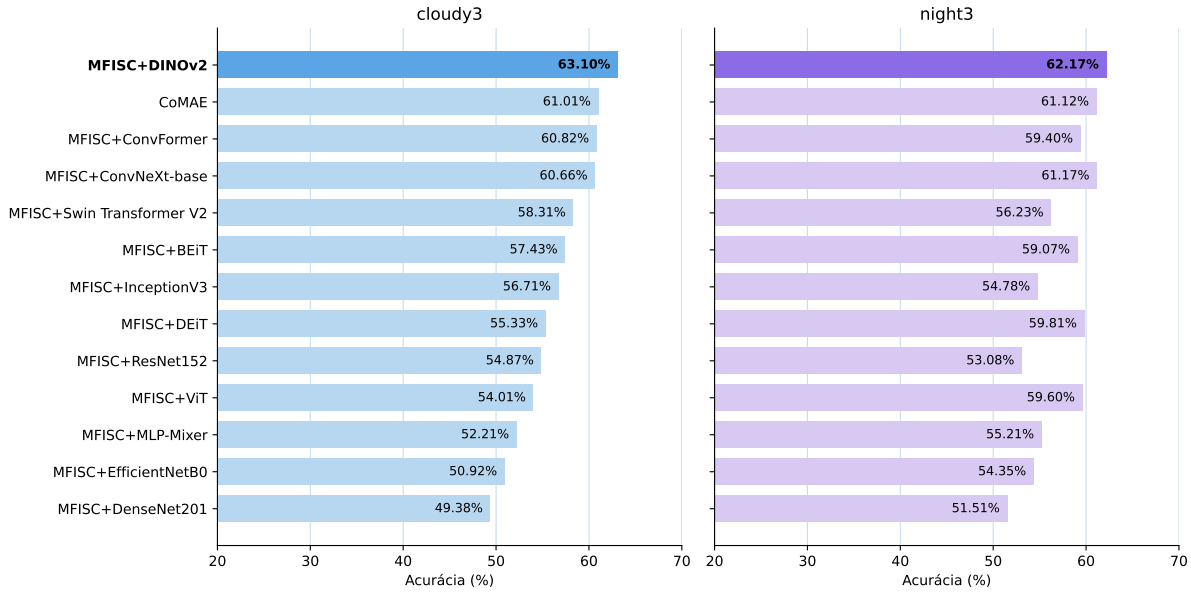


Figura 29 – Esquema 4 no *COLD* (transferência *Freiburg* → *Ljubljana*). Treino em *Freiburg-sunny12* e teste em *Ljubljana-cloudy3* e *Ljubljana-night3*.

Na Figura 30, treinamos em *Ljubljana-cloudy12* e testamos em *Saarbrücken-night3* e *Saarbrücken-sunny3*. Este é o caso mais desafiador entre os três recortes apresentados: o *MFISC+DINOv2* lidera com 50.62% na condição noturno (*night3*) e 46.84% em ensolarado (*sunny3*), evidenciando o impacto do deslocamento de domínio combinado à mudança fotométrica no local-alvo. Novamente o *MFISC+CoMAE* apresenta acurácias próximas ao topo, mas não superaa arquitetura completa do *MIFSC+DINOv2*.

Por fim, a Figura 31 apresenta o treino em *Saarbrücken-night12* e teste em *Freiburg-cloudy3* e *Freiburg-sunny3*. O *MFISC+DINOv2* novamente alcança o melhor desempenho, com 56.58% em nublado (*cloudy3*) e 57.64% ensolarado (*sunny3*), mantendo uma vantagem consistente sobre os demais *backbones*. O *MFISC+CoMAE* se revela um extrator de características com desempenho superior ao modelos supervisionados.

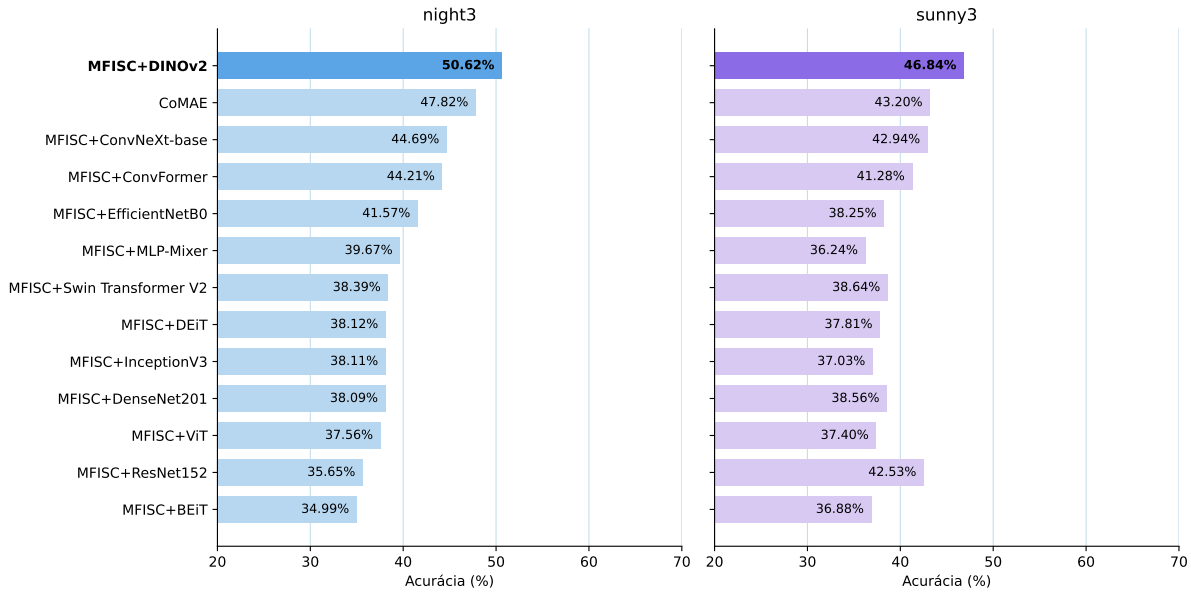


Figura 30 – Esquema 4 no *COLD* (transferência *Ljubljana* → *Saarbrücken*). Treino em *Ljubljana-cloudy12* e teste em *Saarbrücken-night3* e *Saarbrücken-sunny3*.

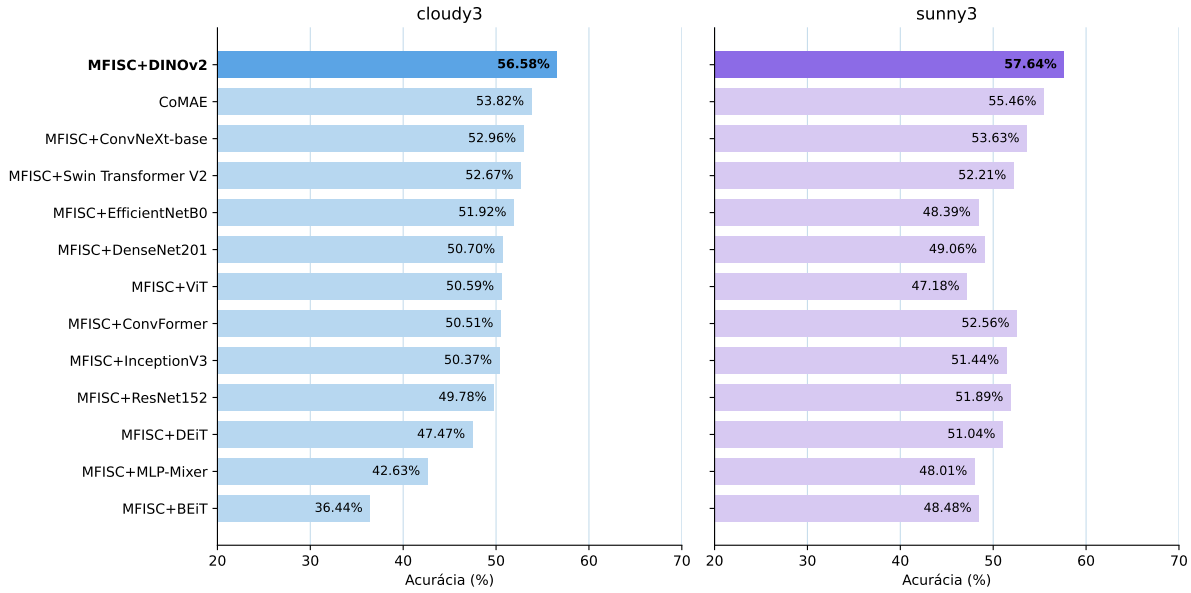


Figura 31 – Esquema 4 no *COLD* (transferência *Saarbrücken* → *Freiburg*). Treino em *Saarbrücken-night12* e teste em *Freiburg-cloudy3* e *Freiburg-sunny3*.

Em síntese, os resultados de transferência entre locais confirmam que o Esquema 4 impõe um desafio substancial, reduzindo as acurácias absolutas em comparação com os esquemas intra-local. Ainda assim, o *MFISC+DINOv2* mantém liderança

consistente nos três recortes avaliados, reforçando o papel de representações auto-supervisionadas como base mais transferível quando acopladas ao *pipeline* de fusão do *MFISC*.

5.5 Estudo de Ablação

O estudo de ablação quantifica, de forma controlada, a contribuição dos principais componentes do *pipeline* proposto para a tarefa de mapeamento semântico tratado como classificação de cenas. Nesta etapa, buscamos responder a três questões: (i) qual é o ganho da fusão multimodal em relação às variantes monomodais quando utilizamos o mesmo extrator; (ii) qual é o impacto de módulos adicionais projetados para tornar a fusão mais estável e transferível (projeções por modalidade, alinhamento contrastivo e regularização por ortogonalidade); e (iii) em que medida tais componentes contribuem especificamente em cenários de mudança fotométrica severa, onde o deslocamento de domínio é mais pronunciado.

Para garantir uma comparação justa, todas as variantes desta seção mantêm fixos: (i) o *backbone* (*DINOv2*), (ii) o protocolo de treinamento e avaliação e (iii) o cabeçalho de classificação. Variamos apenas os componentes do *MFISC* ativados em cada configuração. Os cenários escolhidos seguem o Esquema 2 (generalização para iluminação não vista), pois esse protocolo impõe um descompasso entre treino e teste que é representativo de aplicações reais, como transições para baixa iluminação.

De forma resumida, avaliamos as seguintes variantes:

- *RGB-only:* classificação usando apenas a modalidade *RGB*.
- *Depth-only:* classificação usando apenas profundidade monocular.
- *Fusão simples:* concatenação das duas modalidades.
- *C-Att (RGB+Depth):* fusão por Atenção Cruzada Bidirecional.
- *C-Att + Proj:* C-Att acrescida de projeções específicas por modalidade.

- *C-Att + Proj + CL + Ortho* (*MFISC* completo): adiciona regularização por ortogonalidade para preservar complementaridade.

Foram selecionados dois contextos representativos: (i) plataforma *Minnie* no *KTH-IDOL2* em testes com destino em condição noturna (*night*); e (ii) ambiente *Ljubljana* no *COLD* com teste em noturno. Em ambos os casos, o objetivo é avaliar como cada componente contribui para sustentar o desempenho sob mudança fotométrica severa.

5.5.1 Estudo de Ablação no *KTH-IDOL2* (*Minnie*)

A Tabela 4 apresenta os resultados do estudo de ablação no *KTH-IDOL2*, considerando a plataforma *Minnie* sob o Esquema 2. Em ambos os cenários, o modelo é treinado em uma condição (*sunny* ou *cloudy*) e avaliado em *night*, caracterizando uma transferência fotométrica severa.

Como *baseline* monomodal, *RGB-only* já atinge acurácias elevadas (89.12% e 91.98%), refletindo a robustez das representações do *DINOv2* sob mudança de iluminação. Por outro lado, *Depth-only* apresenta desempenho substancialmente inferior (52.10% e 51.71%), indicando que a profundidade monocular, isoladamente, é menos discriminativa para categorias semânticas em ambientes internos, especialmente quando há similaridade estrutural entre classes.

Tabela 4 – Estudo de ablação no *KTH-IDOL2* (*Minnie*) sob o Esquema 2. Em todos os casos, o *backbone* é o *DINOv2*; variam-se apenas os componentes do *MFISC*.

Configuração (*DINOv2*)	**Acurácia (%) — Esquema 2 / *Minnie***	
	Treino: sunny12, Teste: night34	*Treino: cloudy12, Teste: night34*
RGB-only	89.12	91.98
Depth-only	52.10	51.71
Fusão simples	89.46	91.80
C-Att (RGB + Depth)	92.49	92.82
C-Att + Proj	93.16	93.54
C-Att + Proj + CL + Ortho (MFISC)	**94.17**	**94.74**

Para verificar se a simples combinação das modalidades já é suficiente, avaliamos também uma *fusão simples* baseada na concatenação direta das representações ($RGB \,\|\, Depth$) antes do classificador. Essa variante produz ganhos marginais em um cenário (89.12% $\rightarrow$ 89.46%) e leve queda no outro (91.98% $\rightarrow$ 91.80%), sugerindo que a

concatenação, por si só, não garante integração consistente e pode amplificar ruídos da profundidade estimada dependendo da condição de treinamento.

Em contraste, a fusão por Atenção Cruzada (*C-Att*) obtém ganhos claros sobre *RGB-only* em ambos os casos (92.49% e 92.82%), evidenciando que a profundidade atua como pista geométrica complementar quando integrada por um mecanismo de interação entre modalidades, reduzindo ambiguidades fotométricas e estabilizando as predições.

Ao introduzir projeções específicas por modalidade (*C-Att + Proj*), o desempenho aumenta e se torna mais estável (93.16% e 93.54%), indicando que a adaptação das representações antes da fusão reduz incompatibilidades entre os espaços latentes de *RGB* e profundidade. Em seguida, a inclusão do alinhamento contrastivo (*CL*) reforça a consistência entre pares correspondentes *RGB*–profundidade, enquanto a regularização por ortogonalidade (*Ortho*) atua para preservar complementaridade e mitigar colapso de representações.

A variante completa (*C-Att + Proj + CL + Ortho*, referida como *MFISC*) apresenta o melhor desempenho global (94.17% e 94.74%). Em particular, no cenário *Treino: sunny12* $\rightarrow$ *Teste: night34*, o ganho absoluto em relação a *RGB-only* é de 5.05 pontos percentuais (89.12% $\rightarrow$ 94.17%), indicando que o ganho não decorre apenas da atenção cruzada, mas do conjunto de mecanismos que equilibram alinhamento e complementaridade.

5.5.2 Estudo de Ablação no *COLD* (*Ljubljana*)

A Tabela 5 apresenta o estudo de ablação no *COLD* para o ambiente *Ljubljana*, sob o Esquema 2 (treino em *sunny12* ou *cloudy12* e teste em *night3*). O *RGB-only* constitui um *baseline* forte, com 92.34% (*sunny12* $\rightarrow$ *night3*) e 91.90% (*cloudy12* $\rightarrow$ *night3*), enquanto *Depth-only* apresenta desempenho inferior (73.59% e 72.78%).

Para verificar se a combinação das modalidades é suficiente, avaliamos também a *fusão simples* por concatenação. Essa variante obtém ganho marginal em *sunny12* $\rightarrow$ *night3* (92.34% $\rightarrow$ 92.57%), mas degrada em *cloudy12* $\rightarrow$ *night3* (91.90% $\rightarrow$ 91.06%), indicando que a concatenação direta não produz integração consistente e pode amplificar

Tabela 5 – Estudo de ablação no *COLD* (*Ljubljana*) sob o Esquema 2. Em todos os casos, o *backbone* é o *DINOv2*; variam-se apenas os componentes do *MFISC*.

Configuração (*DINOv2*)	**Acurácia (%) — Esquema 2 / *Ljubljana***	
	Treino: sunny12, Teste: night3	*Treino: cloudy12, Teste: night3*
RGB-only	92.34	91.90
Depth-only	73.59	72.78
Fusão simples	92.57	91.06
C-Att (RGB + Depth)	93.30	92.25
C-Att + Proj	94.23	93.11
C-Att + Proj + CL + Ortho (MFISC)	**95.81**	**93.99**

ruídos da profundidade estimada dependendo da condição de treinamento.

Em contraste, a fusão por Atenção Cruzada (*C-Att*) melhora os resultados em ambos os cenários (93.30% e 92.25%). A inclusão de projeções por modalidade (*C-Att + Proj*) eleva o desempenho para 94.23% e 93.11%, sugerindo melhor compatibilização entre os espaços latentes. Por fim, a variante completa (*C-Att + Proj + CL + Ortho*, *MFISC*) atinge 95.81% e 93.99%, correspondendo a um ganho máximo de 3.47 pontos percentuais sobre *RGB-only* no cenário *sunny12* $\rightarrow$ *night3*.

Esse padrão reforça que *RGB* concentra a maior parcela do poder discriminativo, mas a profundidade monocular fornece informação geométrica complementar que, quando integrada de forma estruturada (atenção + alinhamento + complementaridade), contribui para reduzir erros localizados em regiões ambíguas e estabilizar a decisão do classificador.

5.5.3 Síntese e implicações

Em síntese, o estudo de ablação indica que (i) *RGB-only* com *DINOv2* é um *baseline* forte sob mudança de iluminação; (ii) *Depth-only* é menos discriminativo quando utilizado isoladamente, mas fornece uma pista geométrica complementar; e (iii) a combinação *Proj + CL + Ortho + C-Att* é a que apresenta desempenho mais consistente, pois favorece alinhamento *cross-modal* e preservação de complementaridade, resultando no melhor desempenho global. Esses achados sustentam o uso de uma fusão multimodal estruturada como base para gerar mapas semânticos *2D* mais coerentes, discutidos na análise qualitativa.

5.6 Comparação com *Baselines* da Literatura

Além do estudo de ablação, conduzimos uma comparação direta com abordagens recentes da literatura voltadas à classificação de cenas com *RGB* e profundidade. O objetivo é posicionar o *MFISC* frente a *baselines* competitivos sob um protocolo idêntico ao utilizado no estudo ablação, assegurando uma comparação justa e reprodutível.

Para garantir equidade, todos os métodos foram avaliados sob o mesmo cenário do estudo de ablação do Esquema 2, mantendo (i) as mesmas divisões de treino/teste, (ii) a mesma métrica (acurácia) e (iii) as mesmas modalidades de entrada (*RGB* e profundidade monocular estimada). Adicionalmente, todos os modelos foram inicializados com pesos pré-treinados, seguindo a prática comum na literatura e reduzindo vieses decorrentes de treinamento do zero em conjuntos relativamente menores.

5.6.1 Resultados no *KTH-IDOL2* (*Minnie*)

A Tabela 6 apresenta os resultados no *KTH-IDOL2* (plataforma *Minnie*) sob o mesmo cenário do estudo de ablação (Esquema 2). Entre os *baselines*, o melhor desempenho é obtido pelo *ISC-RGB-D* (WAQAS; AHMAD, 2024), com 90.08% (treino: *sunny12* $\rightarrow$ teste: *night34*) e 89.93% (treino: *cloudy12* $\rightarrow$ teste: *night34*). Ainda assim, o *MFISC* supera consistentemente os métodos comparados, alcançando 94.17% e 94.74%, o que representa ganhos de 4.09 e 4.81 pontos percentuais, respectivamente, em relação ao melhor *baseline* de cada configuração. Esses resultados reforçam que o ganho não decorre apenas do uso de *RGB-D*, mas do desenho de fusão estruturada adotado pelo *MFISC*.

Tabela 6 – Comparação com *baselines* no *KTH-IDOL2* (*Minnie*) sob o mesmo cenário do ablation (Esquema 2).

Método	**Acurácia (%) — Esquema 2 / *Minnie***	
	Treino: sunny12, Teste: night34	*Treino: cloudy12, Teste: night34*
Resnet101-RNN (CAGLAYAN et al., 2022)	85.14	86.13
FCDENet (ZHOU; JIAN; LIU, 2025)	88.52	87.22
MAPE-ViT (AHMED et al., 2025)	87.10	88.09
ISC-RGB-D (WAQAS; AHMAD, 2024)	90.08	89.93
MFISC	**94.17**	**94.74**

5.6.2 Resultados no *COLD* (*Ljubljana*)

A Tabela 7 reporta os resultados no *COLD* (ambiente *Ljubljana*) sob o mesmo cenário do estudo de ablação (Esquema 2). Novamente, o *ISC-RGB-D* (WAQAS; AHMAD, 2024) apresenta o melhor desempenho entre os *baselines*, com 91.79% e 90.73%. O *MFISC* mantém vantagem consistente, alcançando 95.81% e 93.99%, superando o melhor *baseline* em 4.02 e 3.26 pontos percentuais, respectivamente. Esse comportamento sugere que a estratégia de integração do *MFISC* é mais robusta ao deslocamento fotométrico característico do Esquema 2, mesmo quando comparada a arquiteturas recentes inicializadas com pré-treinamento.

Tabela 7 – Comparação com *baselines* no *COLD* (*Ljubljana*) sob o mesmo cenário do ablation (Esquema 2).

Método	**Acurácia (%) — Esquema 2 / *Ljubljana***	
	Treino: sunny12, Teste: night3	*Treino: cloudy12, Teste: night3*
Resnet101-RNN (CAGLAYAN et al., 2022)	87.32	86.98
FCDENet (ZHOU; JIAN; LIU, 2025)	90.17	89.46
MAPE-ViT (AHMED et al., 2025)	89.93	89.02
ISC-RGB-D (WAQAS; AHMAD, 2024)	91.79	90.73
MFISC	**95.81**	**93.99**

Em síntese, a comparação com *baselines* recentes evidencia que os ganhos do *MFISC* não se devem apenas à disponibilidade de profundidade, mas ao modo como aparência e geometria são integradas. Em particular, a vantagem consistente sobre métodos competitivos sugere que a fusão multimodal estruturada é mais adequada para cenários robóticos internos sujeitos a mudanças fotométricas e variações de domínio, mantendo desempenho elevado sem exigir sensores adicionais além de uma câmera *RGB*.

5.7 Mapas Semânticos

Embora os experimentos anteriores tratem o mapeamento semântico como um problema de classificação de cenas, é fundamental verificar como essas predições se manifestam no espaço físico percorrido pelo robô. Nesta seção, avaliamos qualitativamente a consistência espacial das predições projetando as saídas do classificador ao longo das

trajetórias *2D* fornecidas pelos *datasets*. Dessa forma, os resultados quantitativos são complementados por visualizações que evidenciam a estabilidade dos rótulos ao longo do percurso, aspecto diretamente relacionado à qualidade do mapa semântico final.

5.7.1 Construção dos mapas e convenções de visualização

A construção dos mapas segue um procedimento simples e reprodutível. Para cada quadro t de uma sequência, associamos a posição (x_t, y_t) (fornecida pelo *dataset* ou derivada da odometria) ao rótulo correspondente: o rótulo verdadeiro (y_t) para o *ground-truth* ou o rótulo predito ($\hat{y}_t$) para cada modelo. Em seguida, a trajetória é plotada no plano *2D* e cada amostra é colorida conforme sua classe. Para garantir comparabilidade direta, utilizamos a mesma paleta de cores em todos os mapas (incluindo *ground-truth* e predições).

Como o objetivo desta seção é isolar o impacto do *backbone* dentro do mesmo *pipeline*, os mapas de predição apresentados referem-se sempre ao *MFISC* (módulo de fusão e classificação) acoplado a dois *backbones*: um supervisionado competitivo (*ConvNeXt*) e o auto-supervisionado (*DINOv2*). Assim, as diferenças visuais observadas refletem principalmente a qualidade das representações extraídas pelo *backbone* quando integradas ao *MFISC*.

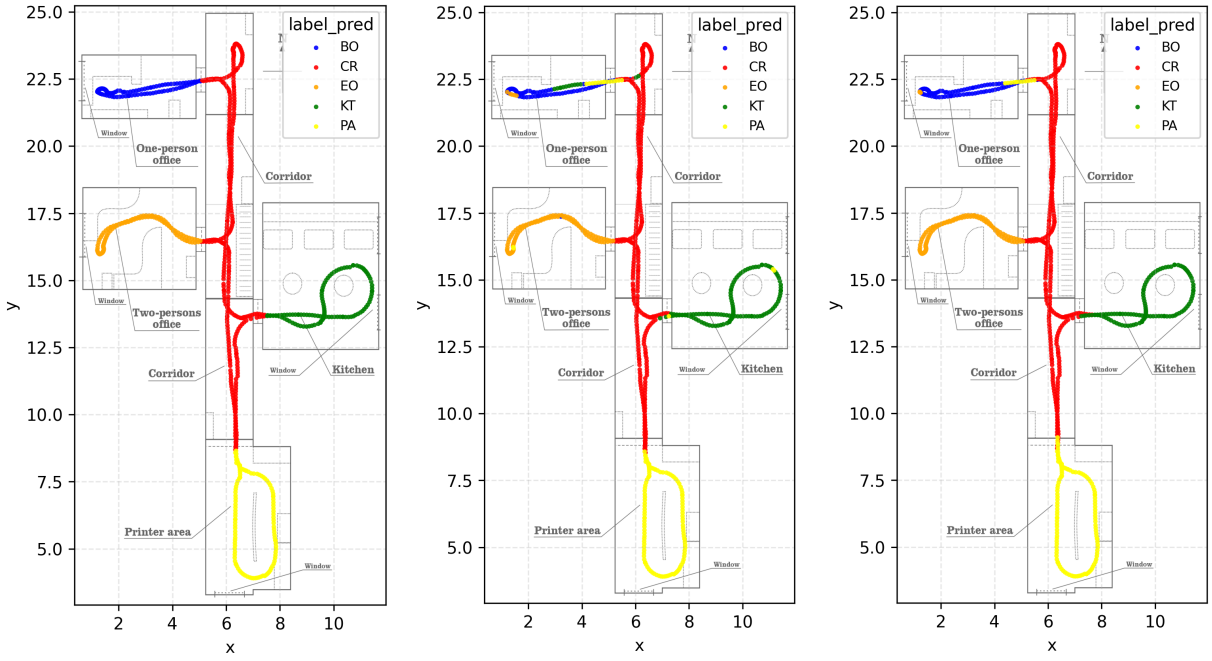
Para comparar o comportamento espacial em um contexto desafiador, selecionamos dois cenários representativos sob o protocolo do Esquema 2 (generalização para iluminação não vista), ambos envolvendo teste em condição *night*:(i) *KTH-IDOL2/Minnie-night34* e (ii) *COLD/Freiburg-night3*. Em cada cenário, apresentamos três mapas: *ground-truth*, *MFISC+ConvNeXt* e *MFISC+DINOv2*.

5.7.2 Resultados qualitativos em *KTH-IDOL2* (*Minnie/night34*)

A Figura 32 compara o mapa de *ground-truth* com as predições obtidas por *MFISC + ConvNeXt* e *MFISC + DINOv2* na sequência *Minnie-night34*. Visualmente, observa-se que *MFISC + DINOv2* tende a produzir trajetórias com menor alternância de rótulos ao

longo de trechos contínuos e menor presença de predições espúrias (pontos isolados com cor divergente), aproximando-se mais dos segmentos semanticamente coerentes vistos no *ground-truth*.

Em contraste, *MFISC + ConvNeXt* apresenta maior incidência de vazamento de classes em regiões de transição (por exemplo, mudanças de sala/corredor), o que se manifesta como pequenos "pontos fora do padrão" ao longo de trechos que deveriam permanecer semanticamente estáveis. Esse comportamento é consistente com o cenário do Esquema 2, no qual a mudança fotométrica amplia ambiguidades de aparência e induz oscilações locais de predição.



(a) *Ground-truth* (b) MFISC + *ConvNeXt* (c) MFISC + *DINOv2*

Figura 32 – Mapas semânticos *2D* na sequência *Minnie/night34* (*KTH-IDOL2*, Esquema 2). Comparação entre (a) *ground-truth*, (b) predição com *MFISC+ConvNeXt* e (c) predição com *MFISC+DINOv2*. As siglas no quadro de cores correspondem às classes do *dataset*.

5.7.3 Resultados qualitativos em *COLD* (*Freiburg/night3*)

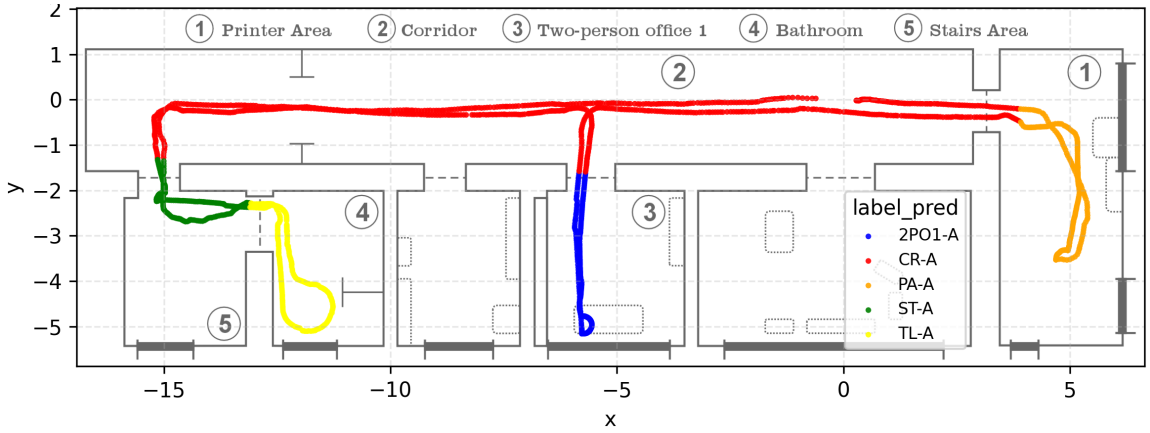
A Figura 33 apresenta a mesma comparação para a sequência *Freiburg/night3* do *COLD*. Novamente, *MFISC + DINOv2* produz um traçado com maior coerência por região, reduzindo trocas abruptas de rótulo ao longo de trechos extensos (por exemplo, segmentos

de corredor) e preservando melhor a estrutura esperada do *ground-truth*.

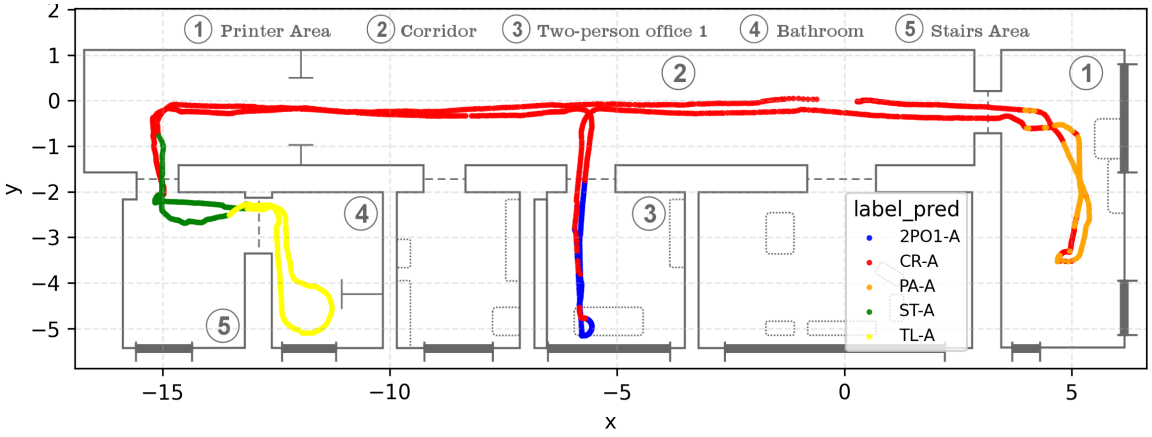
Já *MFISC+ConvNeXt* tende a introduzir inconsistências mais visíveis em partes da trajetória onde a iluminação noturna altera significativamente textura e contraste, o que se traduz em pequenas porções com rótulos incompatíveis com o contexto espacial local. Em termos práticos, essas inconsistências resultam em mapas com maior ruído e fronteiras semânticas menos nítidas, o que pode comprometer a leitura do mapa para tarefas de navegação e tomada de decisão.

5.7.4 Síntese

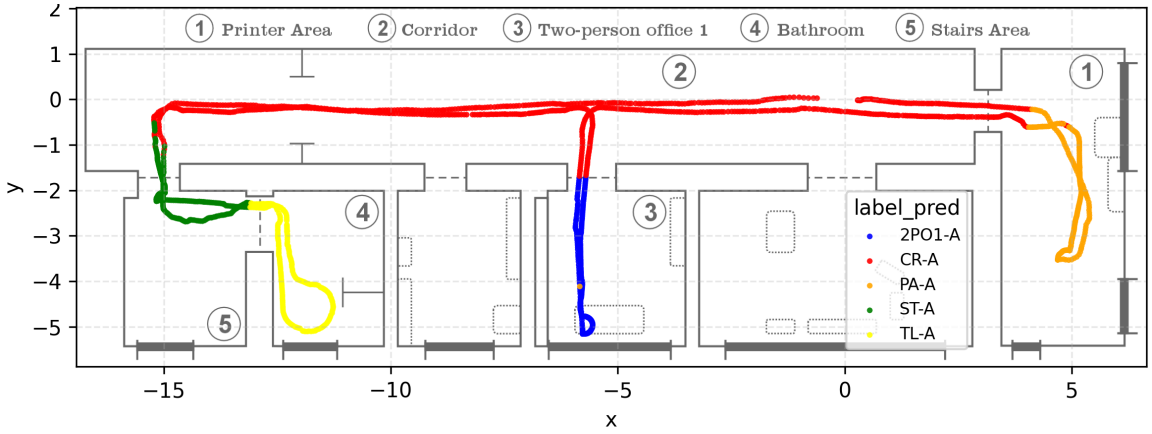
Em ambos os *datasets*, os mapas mostram que *MFISC+DINOv2* tende a gerar predições mais estáveis ao longo da trajetória, com menor alternância local de rótulos e maior aderência às regiões semânticas do *ground-truth*. Essa evidência qualitativa é consistente com os ganhos observados nas métricas quantitativas e reforça que melhorias de acurácia no cenário de iluminação não vista (Esquema 2) se traduzem, na prática, em mapas semânticos 2D mais coerentes e interpretáveis.



(a) *Ground-truth*



(b) C-Att com *ConvNeXt*



(c) C-Att com *DINOv2*

Figura 33 – Mapas semânticos *2D* na sequência *Freiburg/night3* (*COLD*, Esquema 2). Comparação entre (a) *ground-truth*, (b) predição com *MFISC+ConvNeXt* e (c) predição com *MFISC+DINOv2*. As siglas no quadro de cores correspondem às classes do *dataset*

.

6

CONSIDERAÇÕES FINAIS

6.1 Conclusões

A construção de mapas semânticos em ambientes internos é um componente essencial para a navegação autônoma de robôs móveis, pois permite associar significado às regiões do ambiente e sustentar decisões mais seguras e contextualizadas. Nesta dissertação, o mapeamento semântico foi abordado sob a perspectiva de classificação supervisionada de cenas: cada quadro observado ao longo da trajetória recebe um rótulo semântico e essas predições são posteriormente projetadas no plano do ambiente, resultando em mapas semânticos *2D* alinhados às poses do robô.

Investigamos como estratégias de fusão multimodal podem tornar essa classificação mais robusta em condições realistas de operação. Para isso, foi desenvolvida uma arquitetura multimodal que combina representações auto-supervisionadas extraídas pelo *DINOv2* com um mecanismo de Atenção Cruzada Bidirecional entre duas modalidades derivadas de uma única câmera: a imagem *RGB* e um mapa de profundidade monocular estimado. A abordagem integra (i) a geração dos mapas de profundidade, (ii) a extração de características por modalidade, (iii) a fusão por atenção cruzada e módulos auxiliares de alinhamento e (iv) a classificação supervisionada final, produzindo rótulos de cena diretamente utilizáveis na etapa de mapeamento.

A avaliação experimental foi conduzida nos conjuntos *KTH-IDOL2* e *COLD*, contemplando diferentes plataformas/locais e variações fotométricas. Considerando os quatro esquemas de teste definidos, os resultados evidenciam que a arquitetura multimodal proposta mantém desempenho elevado e consistente em cenários de deslo-

camento temporal, mudanças bruscas de iluminação e transferências entre plataformas e ambientes. De forma sistemática, o *DINOv2* mostrou-se o extrator de características mais eficaz quando acoplado ao *pipeline* de fusão, superando *backbones* supervisionados representativos, como *ConvNeXt* e *Swin Transformer V2*, e sustentando maior estabilidade, em especial nas transferências envolvendo a condição noturna, que se apresentou como a mais desafiadora ao longo dos protocolos.

O estudo de ablação permitiu isolar a contribuição das modalidades e do mecanismo de fusão. Observou-se que o canal *RGB* isolado já constitui um *baseline* forte sob as representações auto-supervisionadas do *DINOv2*, enquanto a profundidade monocular, quando utilizada de forma independente, apresenta desempenho inferior, refletindo limitações e ruídos inerentes ao processo de estimação. Ainda assim, ao ser integrada ao *RGB* via Atenção Cruzada e seus módulos auxiliares, a profundidade atua como pista geométrica complementar e produz ganhos consistentes em cenários adversos. Por exemplo, no *KTH-IDOL2* (*Minnie*), no protocolo *Treino: sunny12* $\rightarrow$ *Teste: night34*, a fusão eleva a acurácia de 89.12% (*RGB-only*) para 94.17% (*MFISC*), um aumento de 5.05 pontos percentuais. Esse comportamento reforça que a fusão não adiciona apenas redundância: ela explora complementaridade entre aparência e estrutura, reduzindo ambiguidades fotométricas e estabilizando as predições.

Para posicionar a proposta frente à literatura, comparamos o *MFISC* a *baselines* recentes executados no mesmo protocolo do estudo de ablação (Esquema 2), caracterizado por forte deslocamento fotométrico com teste em *night*. Em ambos os *datasets* avaliados nesse cenário (*KTH-IDOL2/Minnie* e *COLD/Ljubljana*), o *MFISC* apresentou vantagem consistente, indicando que a combinação de extração auto-supervisionada e fusão por atenção cruzada com mecanismos de alinhamento e complementaridade melhora a robustez em condições de iluminação não vista.

Por fim, a análise qualitativa dos mapas semânticos *2D* complementou os achados quantitativos ao evidenciar como as diferenças de desempenho se manifestam no espaço. Em sequências desafiadoras do Esquema 2, os mapas gerados pela arquitetura multimodal (*MFISC+DINOv2*) apresentaram maior coerência espacial, com menor incidência de rótulos espúrios e menor concentração de erros em regiões de baixa ilumi-

nação ou alta ambiguidade visual, quando comparados ao *ground-truth* e a variantes com *backbones* supervisionados.

Quanto aos objetivos específicos, este trabalho (i) caracterizou o mapeamento semântico *2D* como categorização de cenas associada à trajetória e definiu critérios para avaliar desempenho e consistência espacial; (ii) verificou o potencial de representações auto-supervisionadas para robustez sob variações fotométricas e de domínio; (iii) demonstrou que a profundidade monocular estimada atua como pista geométrica complementar quando integrada à aparência; (iv) comprovou ganhos consistentes com uma estratégia de fusão multimodal estruturada; e (v) evidenciou, por resultados quantitativos e análise qualitativa, o impacto dessas escolhas na coerência dos mapas semânticos *2D*.

Em conjunto, os resultados indicam que a combinação de representações auto-supervisionadas e fusão multimodal por Atenção Cruzada Bidirecional, com os módulos auxiliares de projeção por modalidade e objetivos de alinhamento e complementaridade no espaço latente, constitui uma base sólida para sistemas leves de percepção voltados ao mapeamento semântico em robótica móvel, aproximando o desempenho do sistema das exigências de operação em ambientes internos reais.

6.2 Trabalhos Futuros

Como desdobramento natural desta dissertação, há frentes promissoras a serem exploradas, tanto no nível de modelagem quanto na ampliação do escopo experimental. Do ponto de vista arquitetural, uma direção relevante consiste em enriquecer as representações multimodais incorporando pistas semânticas de alto nível provenientes de detecção de objetos. Em particular, informações como classes detectadas, confiança, *bounding boxes* e descritores regionais podem ser integradas ao *pipeline* para (i) fornecer contexto explícito sobre a composição da cena, (ii) reduzir ambiguidades em locais visualmente semelhantes e (iii) favorecer maior consistência espacial durante a projeção das predições na trajetória do robô. Essa integração pode ocorrer por meio de concatenação de descritores globais e regionais, por mecanismos de atenção, como atenção cruzada

entre *tokens* de cena e *tokens* de objetos, ou ainda por um módulo de fusão hierárquica que combine pistas de cena e de instância.

No escopo experimental, considera-se também a definição de um esquema adicional voltado à avaliação entre plataformas e ambientes distintos, com transferências diretas entre domínios. Exemplos incluem treinar o modelo em um laboratório (*KTH-IDOL2* e avaliar em outro no *COLD*. Esse tipo de protocolo complementa os esquemas já explorados, permitindo investigar de forma mais direta a robustez do sistema frente a mudanças de locais, ponto de vista e estatísticas de aparência, aproximando a avaliação das condições de operação encontradas em robôs móveis em ambientes internos reais.

REFERÊNCIAS

AHMED, M. W. et al. Mape-vit: multimodal scene understanding with novel wavelet-augmented vision transformer. *PeerJ Computer Science*, PeerJ Inc., v. 11, p. e2796, 2025. 89, 90

ALFARO, M. et al. Triplet neural networks for the visual localization of mobile robots. In: INSTICC. *Proceedings of the 21st International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO*. [S.l.]: SciTePress, 2024. p. 125–132. ISBN 978-989-758-717-7. 23, 36, 43

ALQOBALI, R. et al. A real-time semantic map production system for indoor robot navigation. *Sensors*, MDPI, v. 24, n. 20, p. 6691, 2024. 40

BALASKA, V. et al. Unsupervised semantic clustering and localization for mobile robotics tasks. *Robotics and Autonomous Systems*, Elsevier, v. 131, p. 103567, 2020. 21

BAO, H. et al. *BEiT: BERT Pre-Training of Image Transformers*. 2022. Disponível em: <https://arxiv.org/abs/2106.08254>. 30, 60

CABRERA, J. J. et al. Training, optimization and validation of a cnn for room retrieval and description of omnidirectional images. *SN Computer Science*, Springer, v. 3, n. 4, p. 271, 2022. 36, 43

CABRERA, J. J. et al. An evaluation of cnn models and data augmentation techniques in hierarchical localization of mobile robots. *Evolving Systems*, Springer, v. 15, n. 6, p. 1991–2003, 2024. 37, 43

CABRERA, J. J. et al. An experimental evaluation of siamese neural networks for robot localization using omnidirectional imaging in indoor environments. *Artificial Intelligence Review*, Springer, v. 57, n. 8, p. 198, 2024. 22, 37, 43

CADENA, C. et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, v. 32, n. 6, p. 1309–1332, 2016. 28

CAGLAYAN, A. et al. When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition. *Computer Vision and Image Understanding*, Elsevier, v. 217, p. 103373, 2022. 89, 90

CHEN, Y. et al. An improved dense-to-sparse cross-modal fusion network for 3d object detection in rgb-d images. *Multimedia Tools and Applications*, Springer, v. 83, n. 4, p. 12159–12184, 2024. 23

DOSOVITSKIY, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. Disponível em: <https://arxiv.org/abs/2010.11929>. 30, 60

FU, T. et al. An rgb-d semantic map building and global localization method. In: IEEE. *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. [S.l.], 2023. p. 979–984. 43

GALINDO, C. et al. Multi-hierarchical semantic maps for mobile robotics. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.: s.n.], 2005. p. 2278–2283. 28

GODARD, C. et al. *Digging Into Self-Supervised Monocular Depth Estimation*. 2019. Disponível em: <https://arxiv.org/abs/1806.01260>. 31

GU, P. et al. *ConvFormer: Combining CNN and Transformer for Medical Image Segmentation*. 2022. Disponível em: <https://arxiv.org/abs/2211.08564>. 60

HE, K. et al. *Deep Residual Learning for Image Recognition*. 2015. Disponível em: <https://arxiv.org/abs/1512.03385>. 60

HOU, C.; ZHAO, X.; LIN, Y. Depth estimation and object detection for monocular semantic slam using deep convolutional network. In: IEEE. *2020 IEEE 20th international conference on software quality, reliability and security companion (QRS-C)*. [S.l.], 2020. p. 256–263. 43

HUANG, G. et al. *Densely Connected Convolutional Networks*. 2018. Disponível em: <https://arxiv.org/abs/1608.06993>. 60

KHAN, S. D.; OTHMAN, K. M. Indoor scene classification through dual-stream deep learning: a framework for improved scene understanding in robotics. *Computers*, MDPI, v. 13, n. 5, p. 121, 2024. 37, 43

LI, Z.; WANG, C. A deep learning-based robot analysis model for semantic context capturing by using predictive models in public management. *Journal of Global Information Management (JGIM)*, IGI Global Scientific Publishing, v. 32, n. 1, p. 1–23, 2024. 22

LIU, Z. et al. *Swin Transformer V2: Scaling Up Capacity and Resolution*. 2022. Disponível em: <https://arxiv.org/abs/2111.09883>. 30, 60

LIU, Z. et al. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. 60

LU, J. et al. Real-time instance-aware segmentation and semantic mapping on edge devices. *IEEE Transactions on Instrumentation and Measurement*, IEEE, v. 72, p. 1–9, 2022. 40

LUO, J. et al. *The KTH-IDOL2 Database*. Stockholm, Sweden, 2006. Disponível em: <http://www.pronobis.pro/publications/luo2006idol2>. 11, 53, 55

OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, Springer, v. 42, p. 145–175, 2001. 30

OQUAB, M. et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. Disponível em: <https://arxiv.org/abs/2304.07193>. 47

PAN, X. et al. Cmanet: Context-aware mutual attention network for referring image segmentation. *IEEE Transactions on Multimedia*, IEEE, 2025. 38, 43

PAPAPETROS, I. T. et al. Visual place recognition in changing environments with sequence representations on the distance-space domain. *Machines*, MDPI, v. 11, n. 5, p. 558, 2023. 22, 37, 43

PELTOMÄKI, J. *LiDAR Place Recognition with Image Retrieval.* Tese (Doutorado) — University of Tampere, Finland, 2023. 22, 23

PRONOBIS, A.; CAPUTO, B. COLD: COsy Localization Database. *International Journal of Robotics Research (IJRR)*, v. 28, n. 5, p. 588–594, maio 2009. Disponível em: <http://www.pronobis.pro/publications/pronobis2009ijrr>. 11, 53, 56

PRONOBIS, A.; JENSFELT, P. Large-scale semantic mapping and reasoning with heterogeneous modalities. In: *2012 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2012. p. 3515–3522. 29

QI, C. R. et al. *ImVoteNet: Boosting 3D Object Detection in Point Clouds with Image Votes*. 2020. Disponível em: <https://arxiv.org/abs/2001.10692>. 32

RANFTL, R.; BOCHKOVSKIY, A.; KOLTUN, V. *Vision Transformers for Dense Prediction*. 2021. Disponível em: <https://arxiv.org/abs/2103.13413>. 31

RUAN, X.; GUO, P.; HUANG, J. A semantic octomap mapping method based on cbam-pspnet. *Journal of web engineering*, River Publishers, v. 21, n. 3, p. 879–910, 2022. 35, 43

SILVA, I. et al. Industrial environment multi-sensor dataset for vehicle indoor tracking with wi-fi, inertial and odometry data. *Data*, v. 8, n. 10, 2023. ISSN 2306-5729. Disponível em: <https://www.mdpi.com/2306-5729/8/10/157>. 22

SONG, X. et al. A semantic mapping system based on scene classification for indoor mobile robots. In: IEEE. *2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT)*. [S.l.], 2022. p. 1–6. 35, 43

SOUSA, Y. C.; BASSANI, H. F. Topological semantic mapping by consolidation of deep visual features. *IEEE Robotics and Automation Letters*, IEEE, v. 7, n. 2, p. 4110–4117, 2022. 23, 43

SZEGEDY, C. et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. Disponível em: <https://arxiv.org/abs/1512.00567>. 60

SüNDERHAUF, N. et al. Place categorization and semantic mapping on a mobile robot. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. [S.l.: s.n.], 2016. p. 5729–5736. 10, 29

TAN, C. J. et al. 3d semantic mapping based on rgb-d camera and lidar sensor in beach environment. In: IEEE. *2024 1st International Conference on Robotics, Engineering, Science, and Technology (RESTCON)*. [S.l.], 2024. p. 21–26. 23

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *Proceedings of the 36th International Conference on Machine Learning*. [S.l.], 2019. p. 6105–6114. 60

TAO, C. et al. 3d semantic vslam of indoor environment based on mask scoring rcnn. *Discrete Dynamics in Nature and Society*, Wiley Online Library, v. 2020, n. 1, p. 5916205, 2020. 34, 43

TOLSTIKHIN, I. et al. *MLP-Mixer: An all-MLP Architecture for Vision*. 2021. Disponível em: <https://arxiv.org/abs/2105.01601>. 60

TOUVRON, H. et al. *Training data-efficient image transformers and distillation through attention*. 2021. Disponível em: <https://arxiv.org/abs/2012.12877>. 30, 60

WANG, S. et al. Compressed holistic convolutional neural network-based descriptors for scene recognition. In: *2019 4th International Conference on Robotics and Automation Engineering (ICRAE)*. [S.l.: s.n.], 2019. p. 135–139. 10, 30

WAQAS, A.; AHMAD, J. Indoor scene classification using rgb-d data: A vision transformer and conditional random field approach. In: *ICIC*. [S.l.: s.n.], 2024. 39, 43, 89, 90

WU, Y. et al. An improved yolov7 network using rgb-d multi-modal feature fusion for tea shoots detection. *Computers and Electronics in Agriculture*, Elsevier, v. 216, p. 108541, 2024. 39, 43

WU, Z. et al. Transformer fusion for indoor rgb-d semantic segmentation. *Computer Vision and Image Understanding*, v. 249, p. 104174, 2024. ISSN 1077-3142. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1077314224002558>. 10, 32

XIN, J. et al. An improved high precision 3d semantic mapping of indoor scenes from rgb-d images. *CMES-Computer Modeling in Engineering & Sciences*, v. 137, n. 3, 2023. 35, 43

YANG, G. et al. Real-time dense 3d semantic mapping using rgb-d camera. In: IEEE. *2023 42nd Chinese Control Conference (CCC)*. [S.l.], 2023. p. 4419–4424. 23

YANG, J. et al. Comae: Single model hybrid pre-training on small-scale rgb-d datasets. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2023. v. 37, n. 3, p. 3145–3154. 60

YANG, L. et al. *Depth Anything V2*. 2024. Disponível em: <https://arxiv.org/abs/2406.09414>. 10, 31, 46

ZHANG, Z. et al. Bilateral attention network for rgb-d salient object detection. *IEEE transactions on image processing*, IEEE, v. 30, p. 1949–1961, 2021. 38, 43

ZHAO, L. et al. Structure-aware cross-modal transformer for depth completion. *IEEE Transactions on Image Processing*, IEEE, v. 33, p. 1016–1031, 2024. 39, 43

ZHOU, B. et al. Learning deep features for scene recognition using places database. In: . Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 487–495. 30

ZHOU, W.; JIAN, B.; LIU, Y. Feature contrast difference and enhanced network for rgb-d indoor scene classification in internet of things. *IEEE Internet of Things Journal*, IEEE, 2025. 39, 43, 89, 90

ZHOU, W. et al. An efficient rgb-d indoor scene-parsing solution via lightweight multiflow intersection and knowledge distillation. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, 2024. 41

A

Resultados Complementares

Este apêndice reúne os resultados completos dos Esquemas 2 e 4 que não foram incluídos no corpo do texto por limitação de espaço. Em todos os gráficos, a arquitetura *MFISC* é mantida fixa e variamos apenas o *backbone* de extração de características.

A.1 Esquema 2: Generalização para Condições de Iluminação Não Vistas

Apresentamos os resultados complementares do Esquema 2 que não foram incluídos no capítulo principal. Em todos os gráficos, o *framework MFISC* é mantido fixo e variamos apenas o *backbone* de extração de características. As configurações seguem o padrão treino em sequências 1–2 de uma condição de iluminação e teste em sequências posteriores (*KTH-IDOL2*: 3–4; *COLD*: 3) sob condições de iluminação distintas.

A.1.1 KTH-IDOL2

A seguir (Figuras 34 a 37), reportamos os resultados por plataforma (*Dumbo* e *Minnie*) considerando duas configurações representativas de treino e teste para o Esquema 2.

A.1.2 COLD

Apresentamos os resultados (Figuras 38 a 43) do Esquema 2 por local (*Freiburg*, *Ljubljana* e *Saarbrücken*).

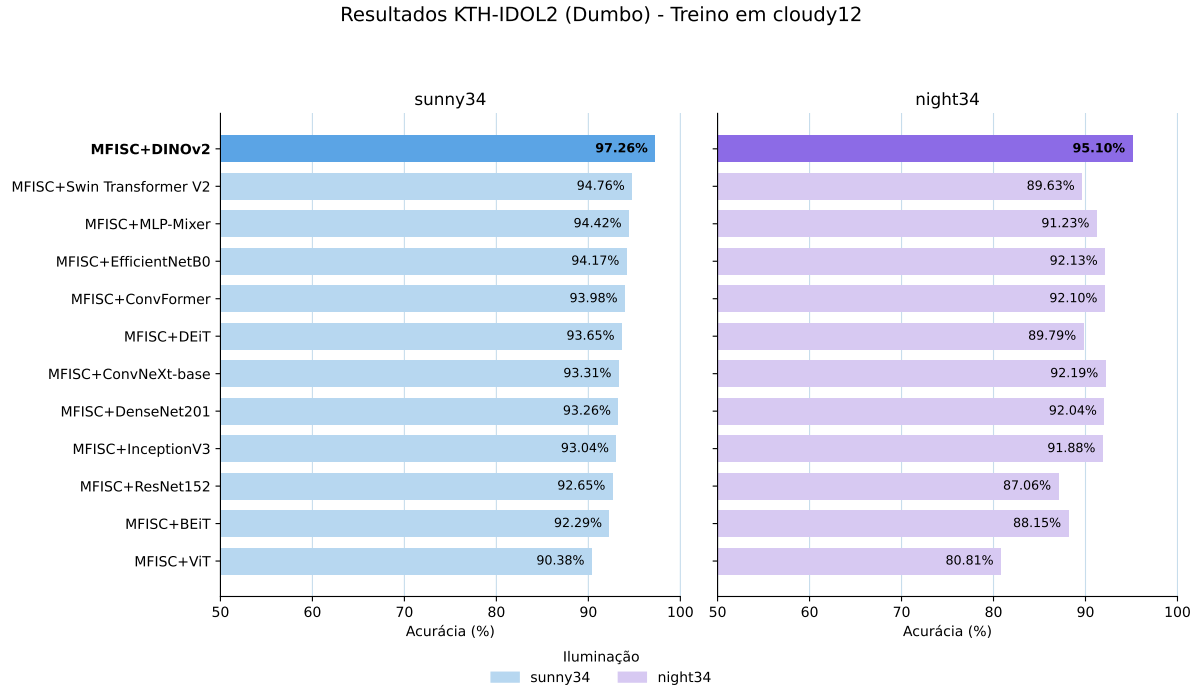


Figura 34 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *sunny34* e *night34*.

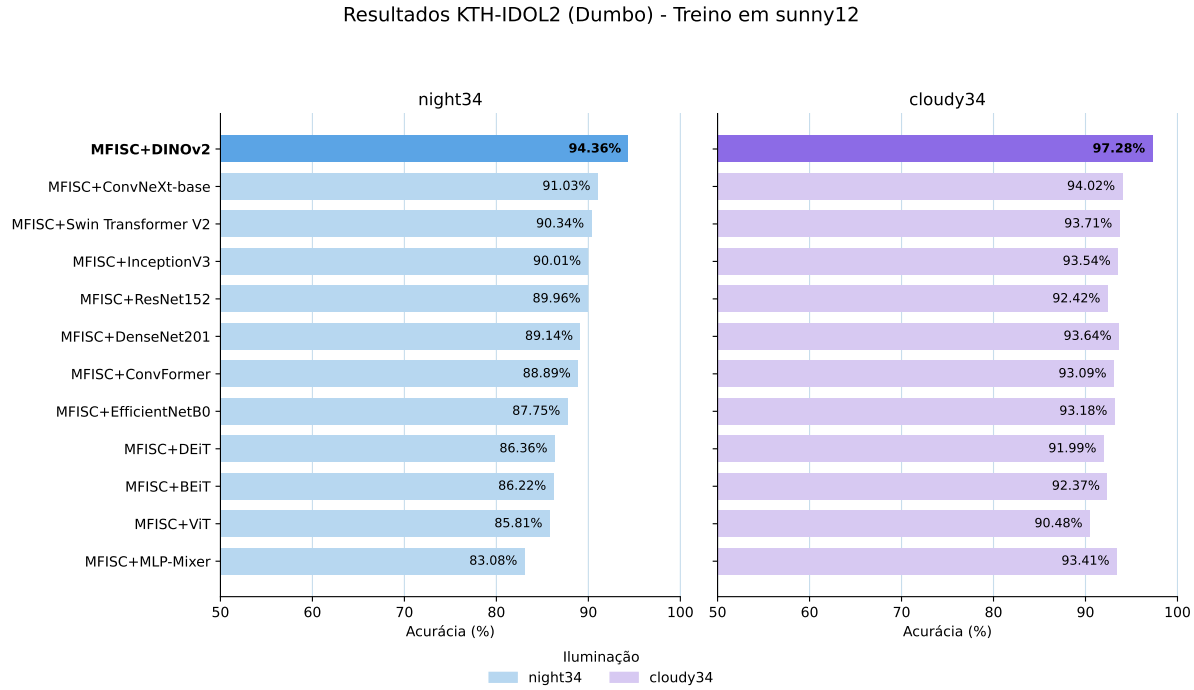


Figura 35 – Esquema 2 no *KTH-IDOL2* (*Dumbo*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *night34* e *cloudy34*.

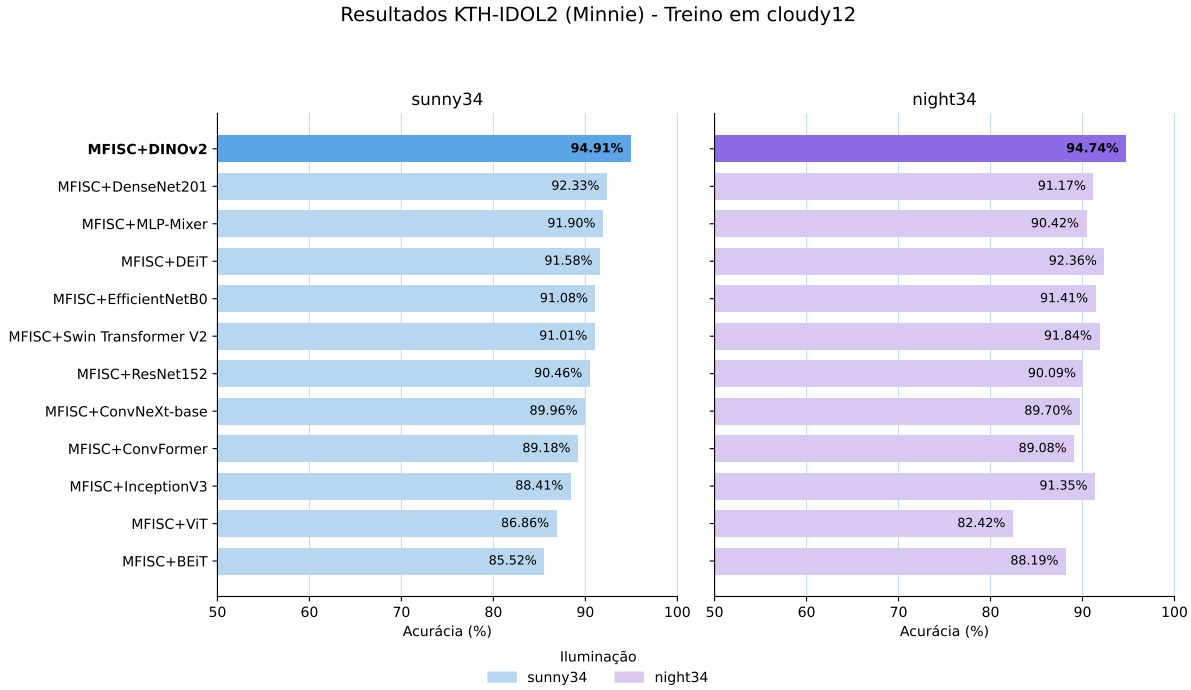


Figura 36 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *sunny34* e *night34*.

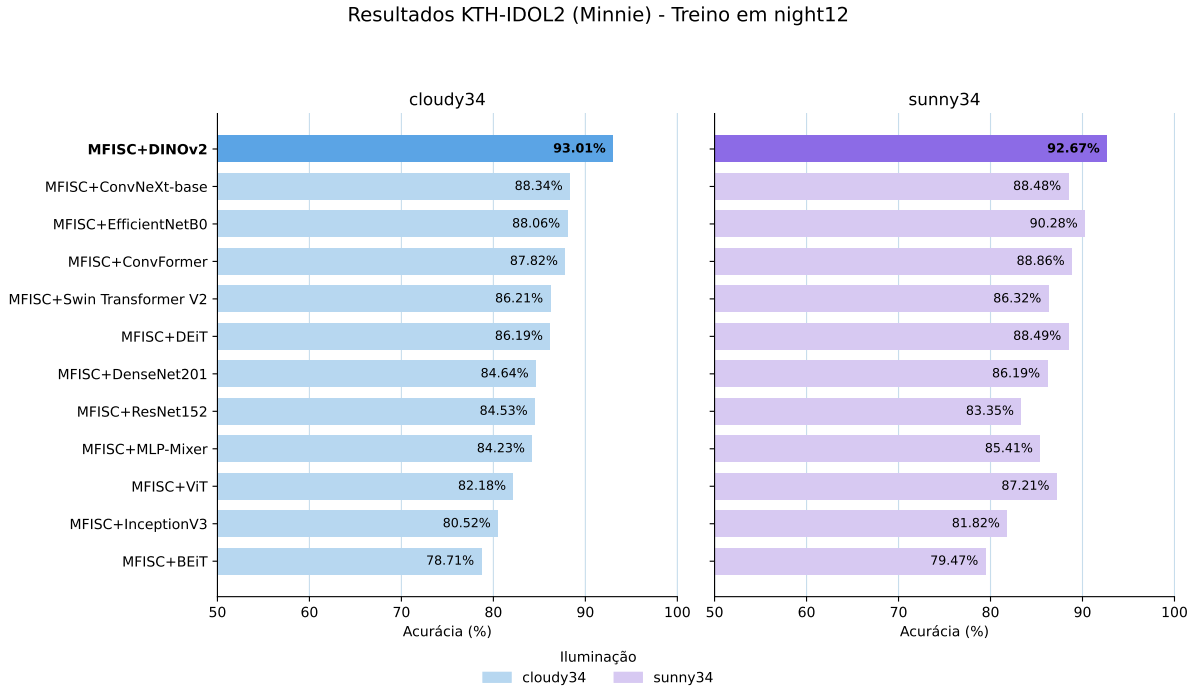


Figura 37 – Esquema 2 no *KTH-IDOL2* (*Minnie*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy34* e *sunny34*.

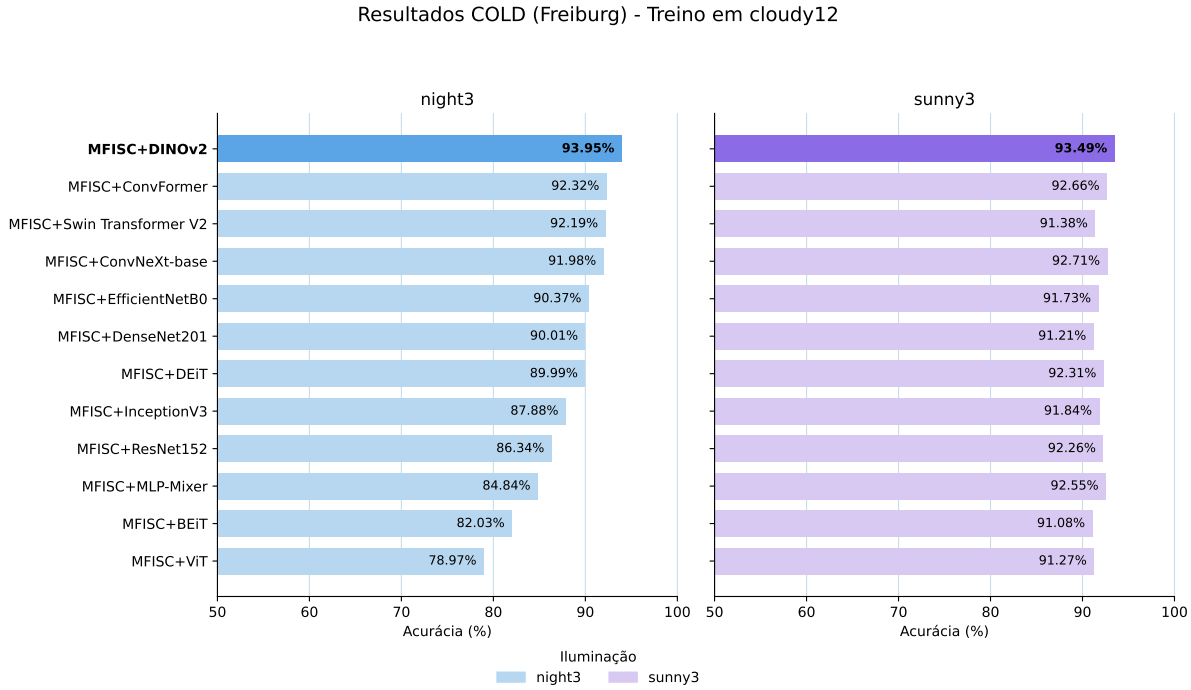


Figura 38 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*.

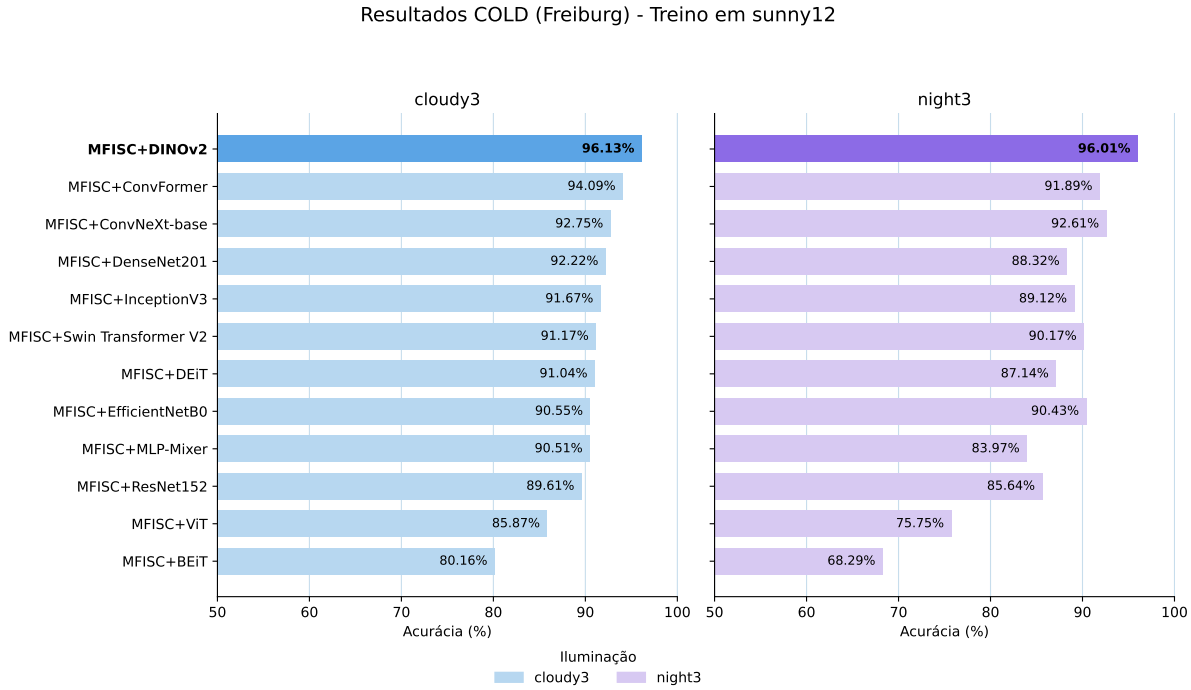


Figura 39 – Esquema 2 no *COLD* (*Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*.

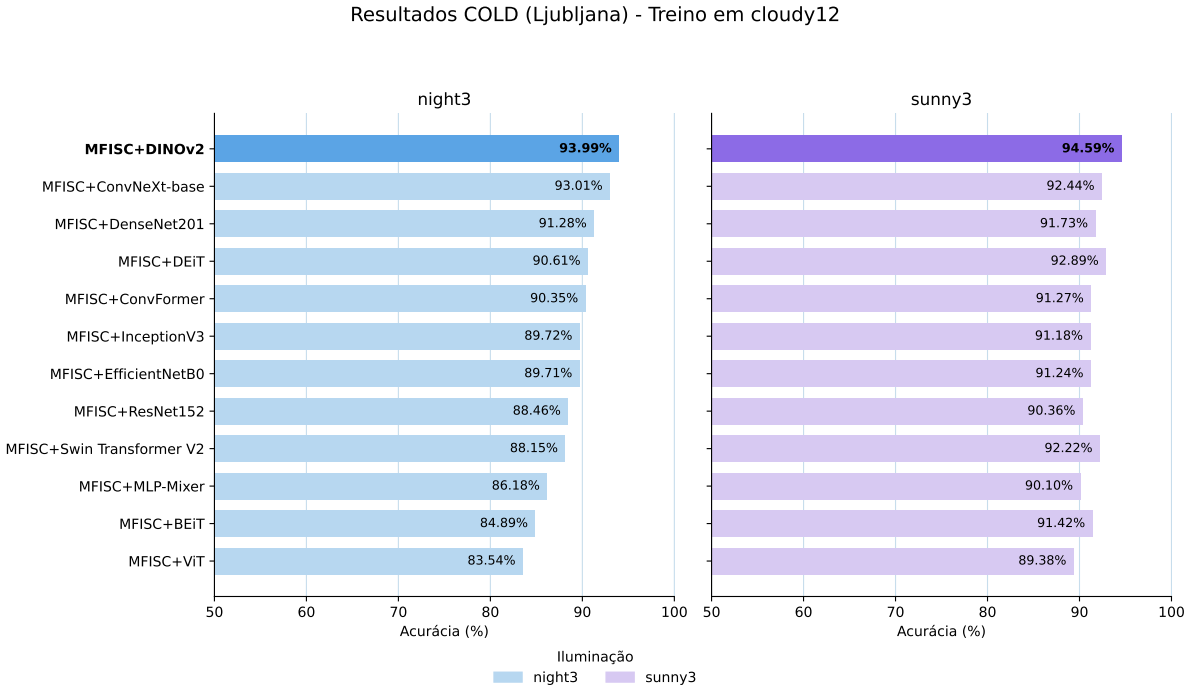


Figura 40 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *cloudy12* e teste em *night3* e *sunny3*.

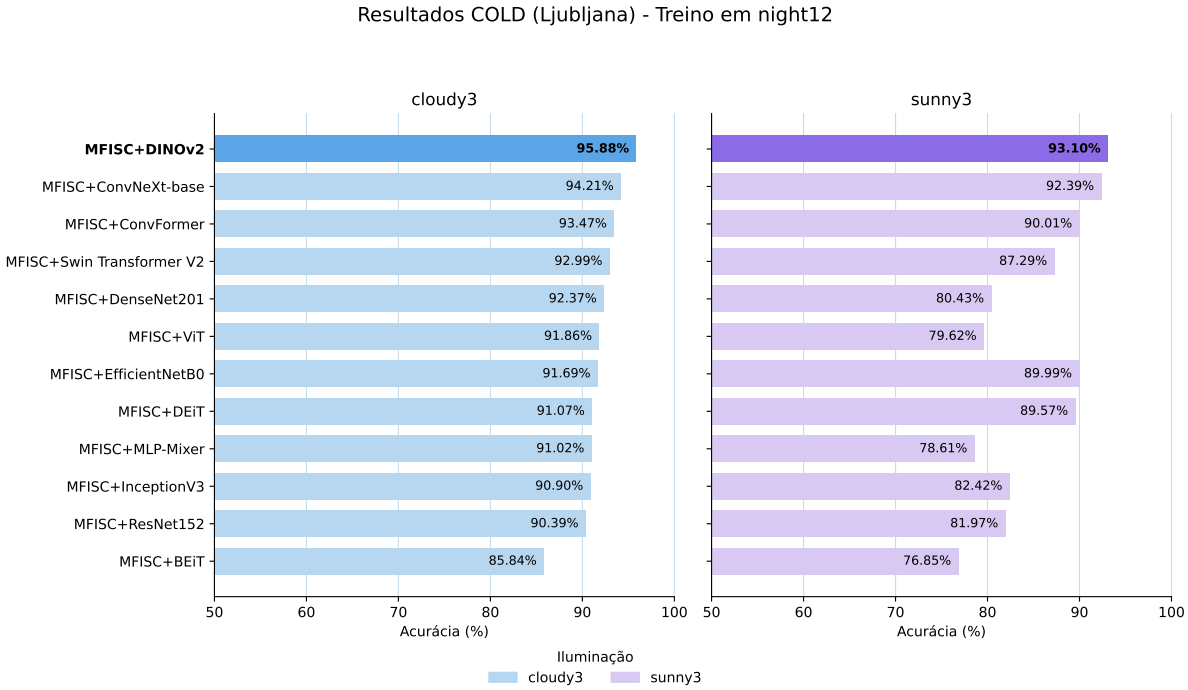


Figura 41 – Esquema 2 no *COLD* (*Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy3* e *sunny3*.

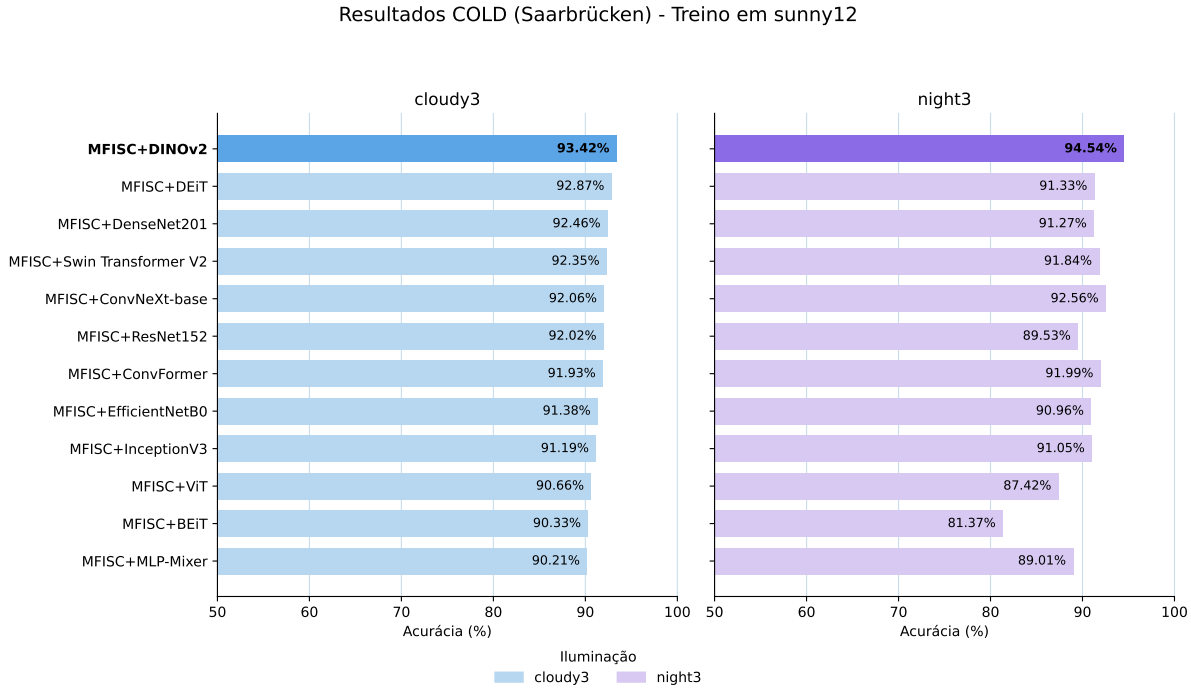


Figura 42 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *sunny12* e teste em *cloudy3* e *night3*.

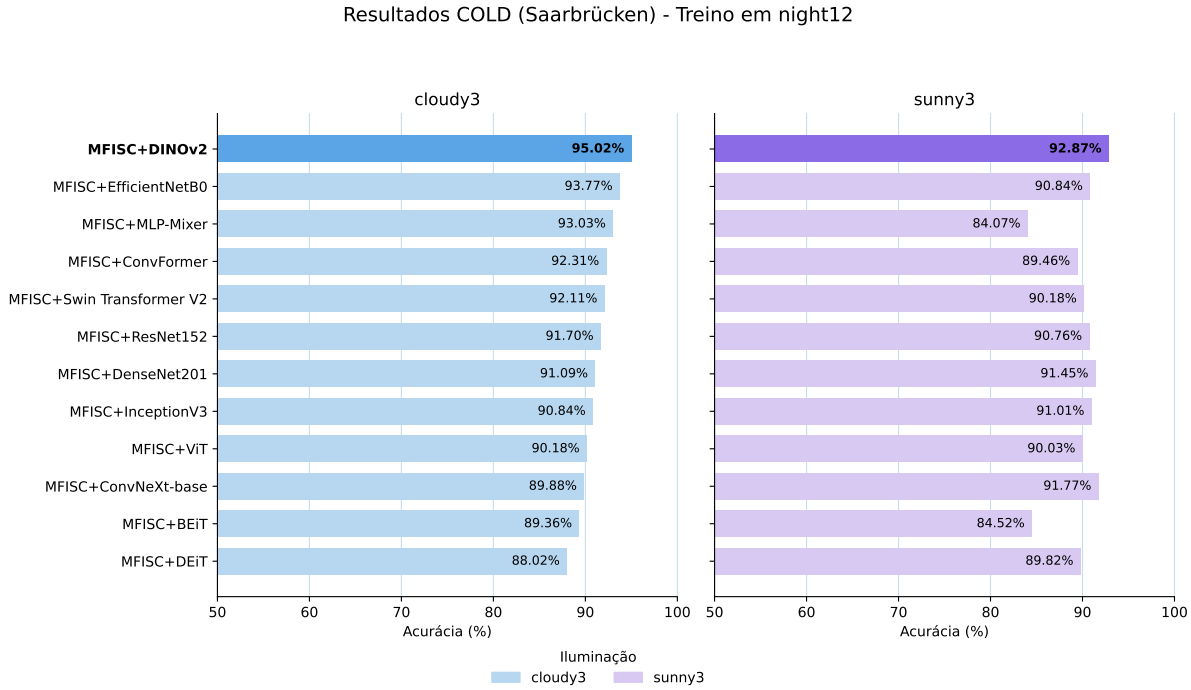


Figura 43 – Esquema 2 no *COLD* (*Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *night12* e teste em *cloudy3* e *sunny3*.

A.2 Esquema 4: Generalização entre Domínios

Este apêndice reúne os resultados complementares do Esquema 4, no qual avaliamos a capacidade de transferência do *framework MFISC* entre domínios distintos. No *COLD* o domínio corresponde ao local de aquisição (*Freiburg, Ljubljana e Saarbrücken*). Em todos os gráficos, o *MFISC* é mantido fixo e variamos apenas o *backbone* de extração de características. As configurações seguem o padrão: treino nas sequências 1–2 de um domínio sob uma condição de iluminação e teste no domínio-alvo sob as condições de iluminação restantes.

A.2.1 COLD: Transferência entre Locais

Reportamos resultados de transferência entre locais no *COLD* (Figuras 44 a 58), combinando o treinamento em um domínio (por exemplo, *Freiburg* e testando nos domínios restantes (por exemplo, *Ljubljana e Saarbrücken*).

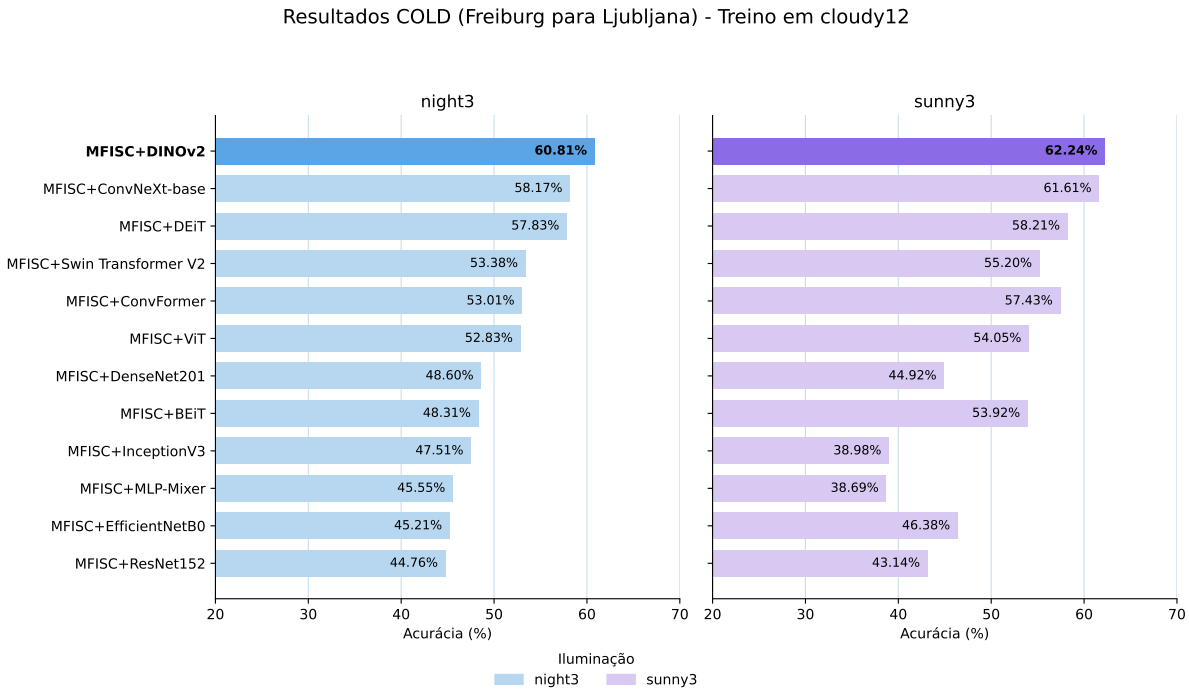


Figura 44 – Esquema 4 no *COLD* (*Freiburg* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-cloudy12* e teste em *Ljubljana-night3* e *Ljubljana-sunny3*.

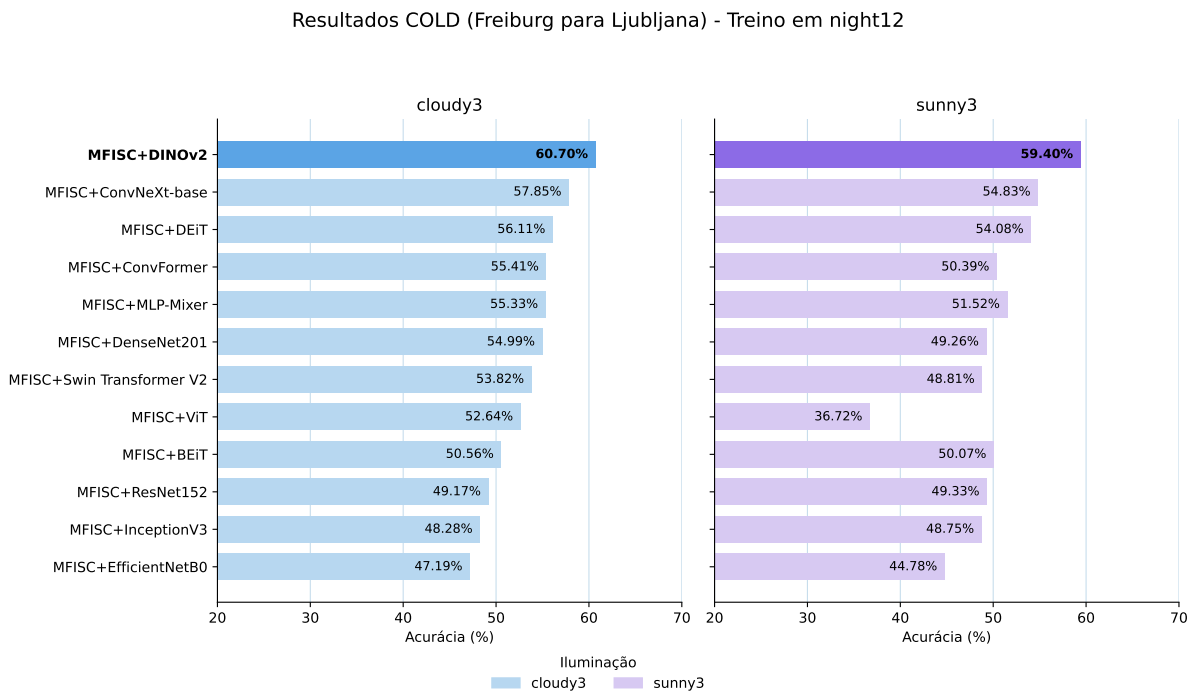


Figura 45 – Esquema 4 no *COLD* (*Freiburg* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-night12* e teste em *Ljubljana-cloudy3* e *Ljubljana-sunny3*.

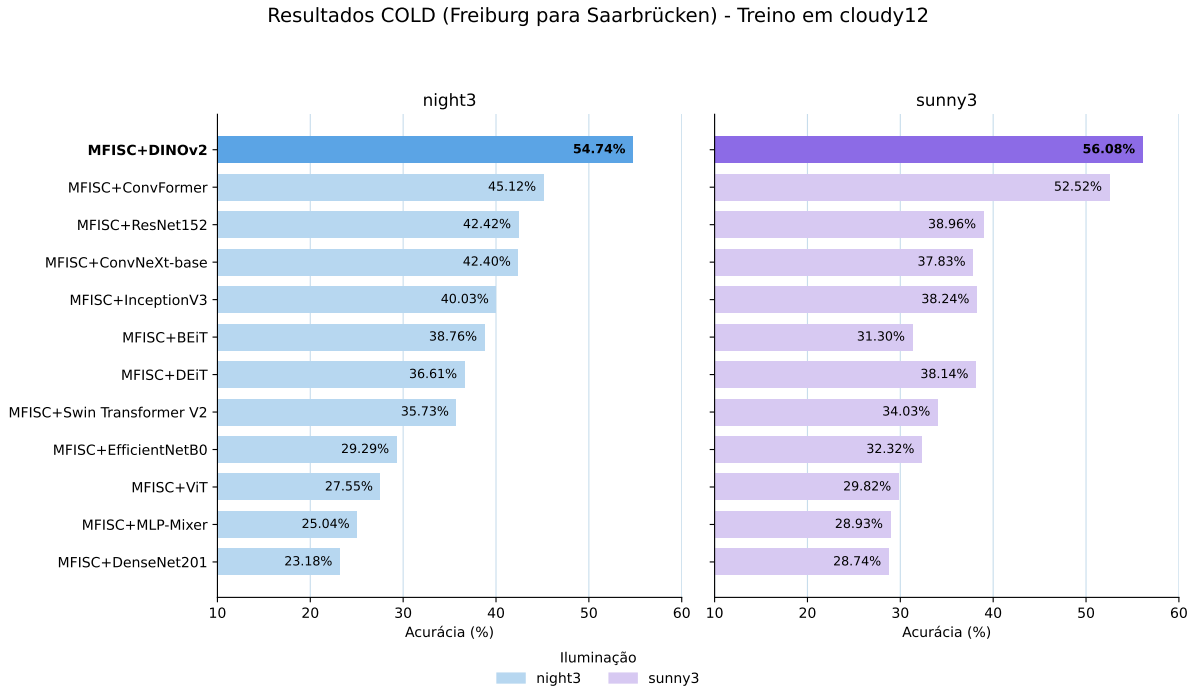


Figura 46 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-cloudy12* e teste em *Saarbrücken-night3* e *Saarbrücken-sunny3*.

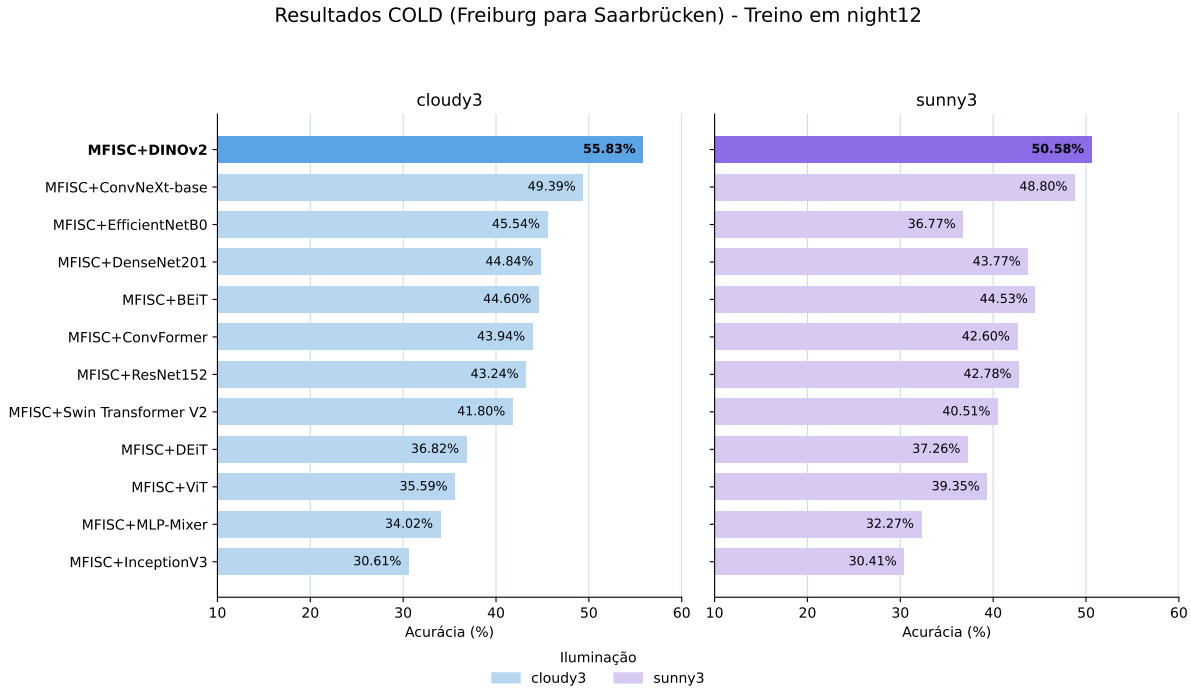


Figura 47 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-night12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-sunny3*.

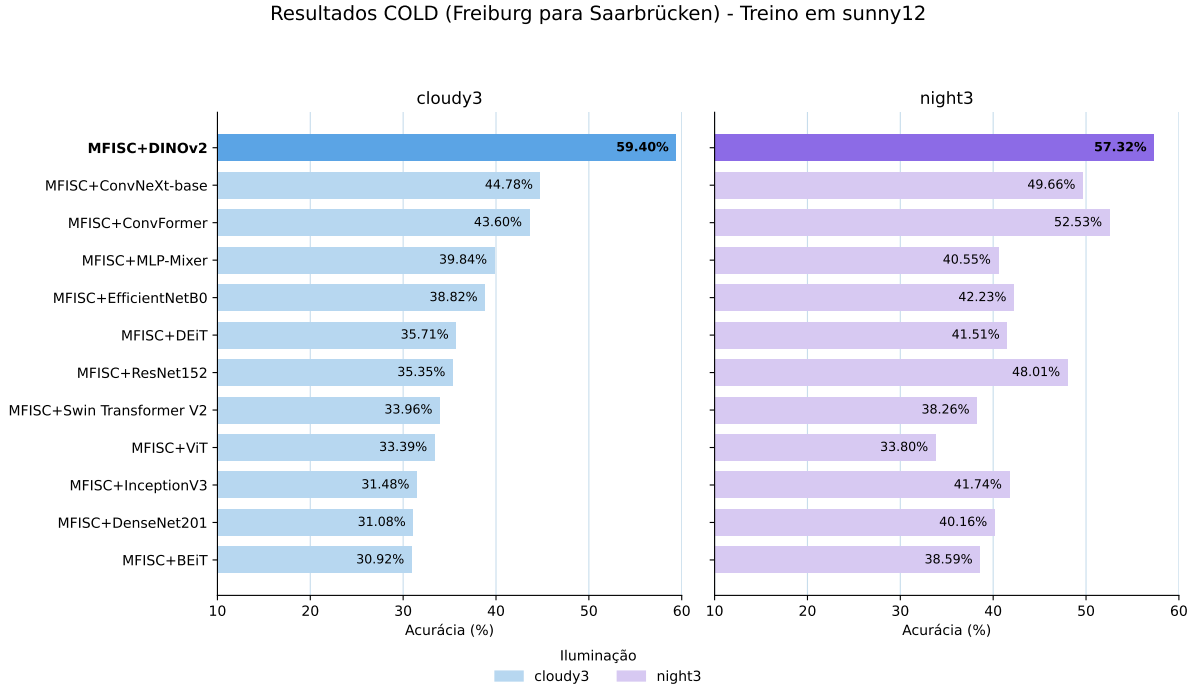


Figura 48 – Esquema 4 no *COLD* (*Freiburg* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Freiburg-sunny12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-night3*.

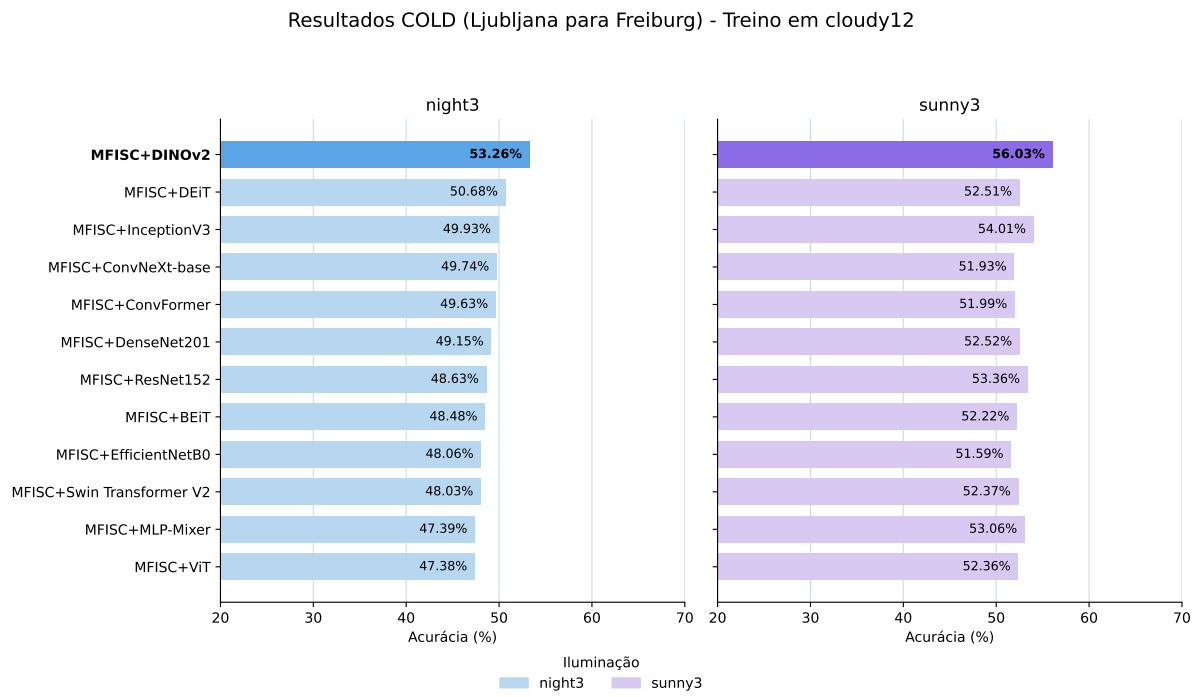


Figura 49 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-cloudy12* e teste em *Freiburg-night3* e *Freiburg-sunny3*.

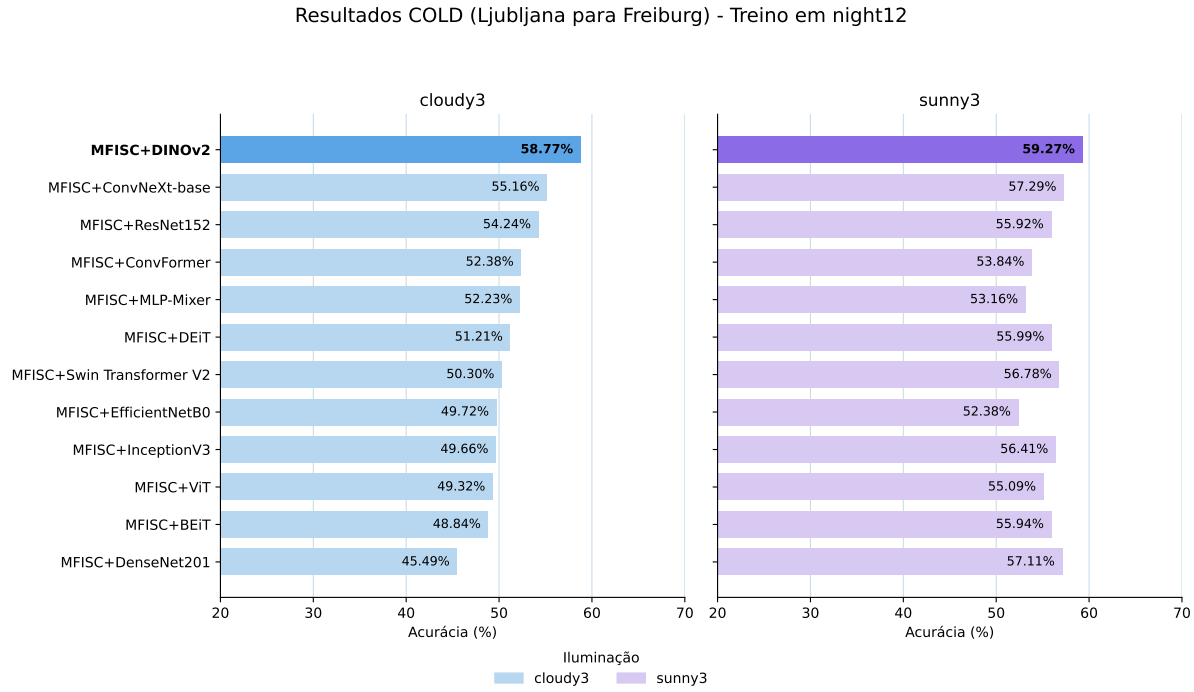


Figura 50 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-night12* e teste em *Freiburg-cloudy3* e *Freiburg-sunny3*.

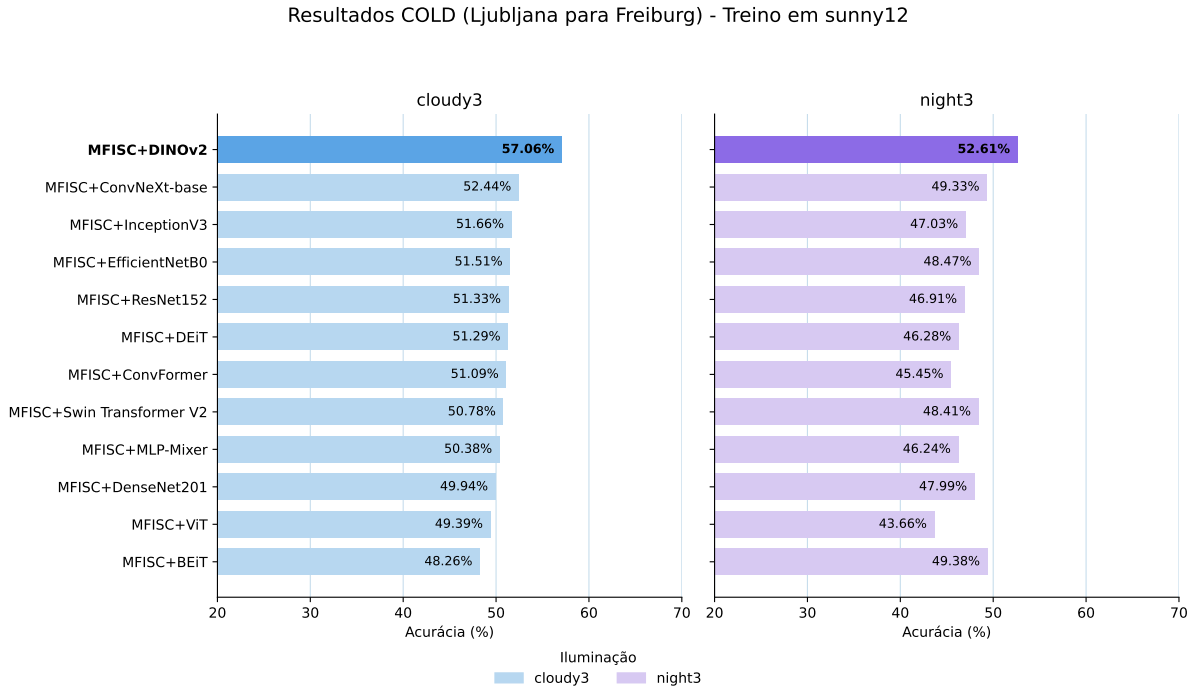


Figura 51 – Esquema 4 no *COLD* (*Ljubljana* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-sunny12* e teste em *Freiburg-cloudy3* e *Freiburg-night3*.

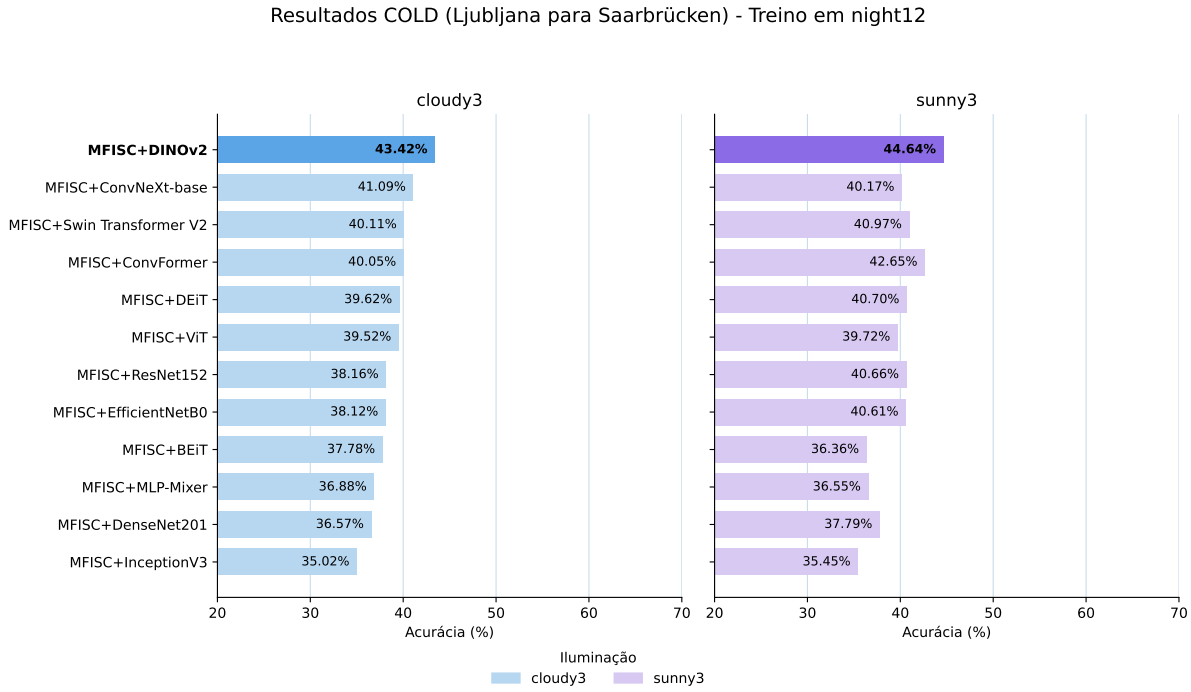


Figura 52 – Esquema 4 no *COLD* (*Ljubljana* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-night12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-sunny3*.

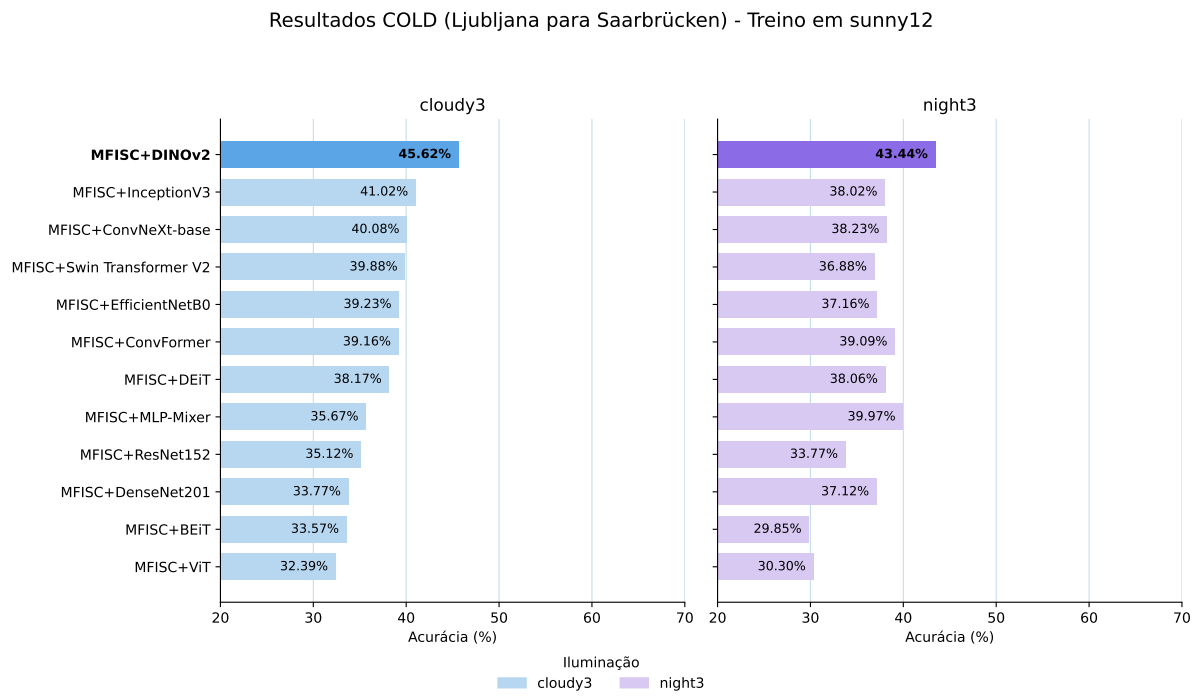


Figura 53 – Esquema 4 no *COLD* (*Ljubljana* → *Saarbrücken*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Ljubljana-sunny12* e teste em *Saarbrücken-cloudy3* e *Saarbrücken-night3*.

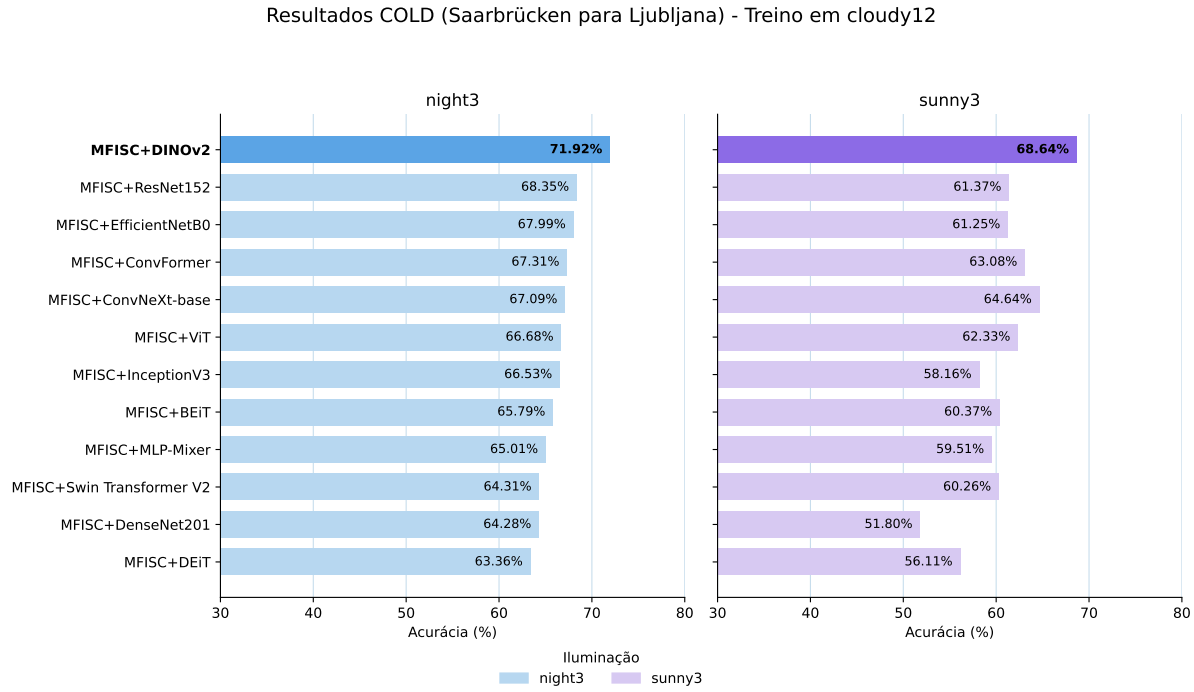


Figura 54 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-cloudy12* e teste em *Ljubljana-night3* e *Ljubljana-sunny3*.

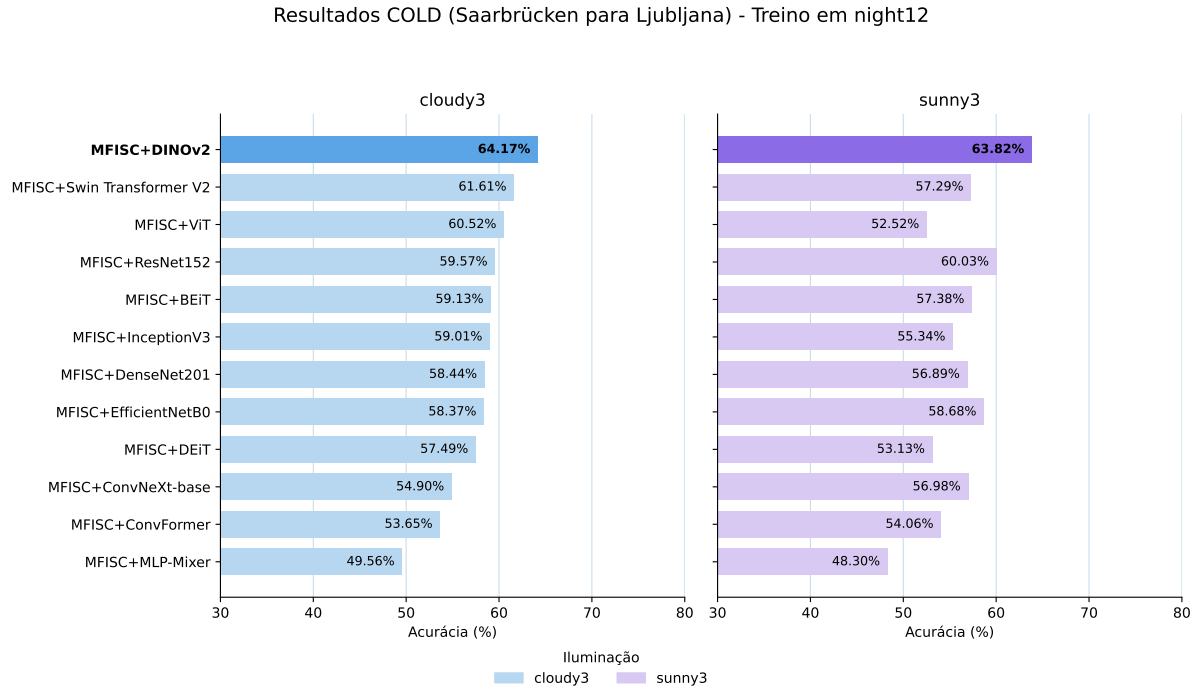


Figura 55 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-night12* e teste em *Ljubljana-cloudy3* e *Ljubljana-sunny3*.

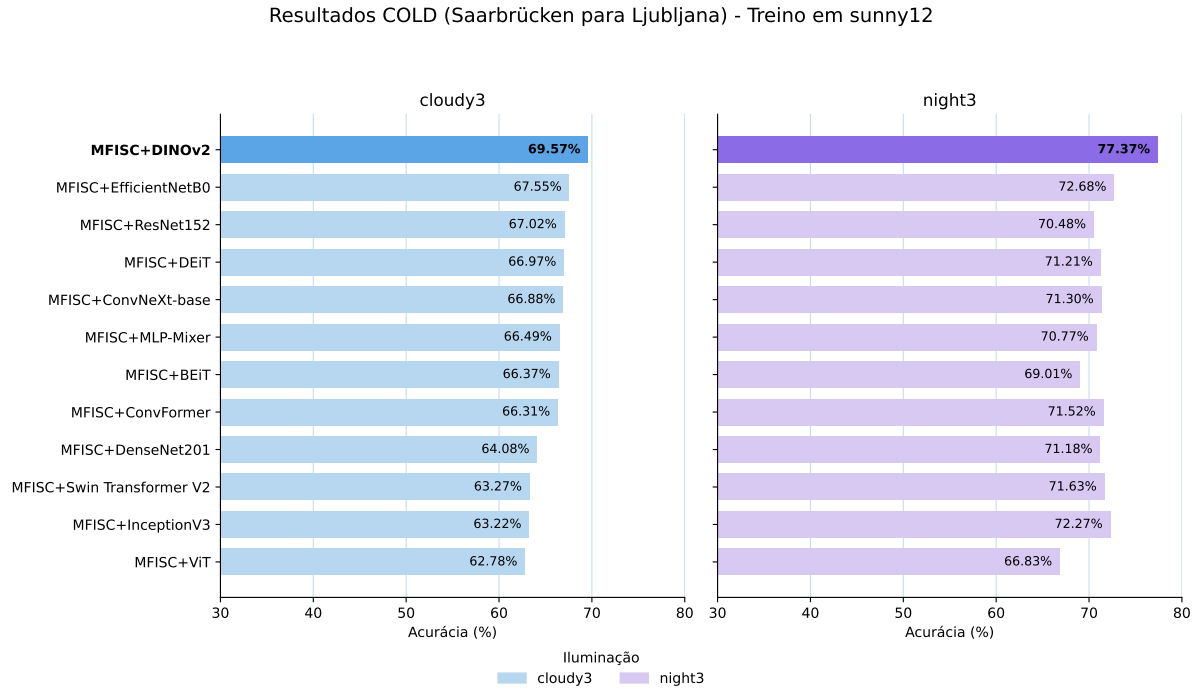


Figura 56 – Esquema 4 no *COLD* (*Saarbrücken* → *Ljubljana*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-sunny12* e teste em *Ljubljana-cloudy3* e *Ljubljana-night3*.

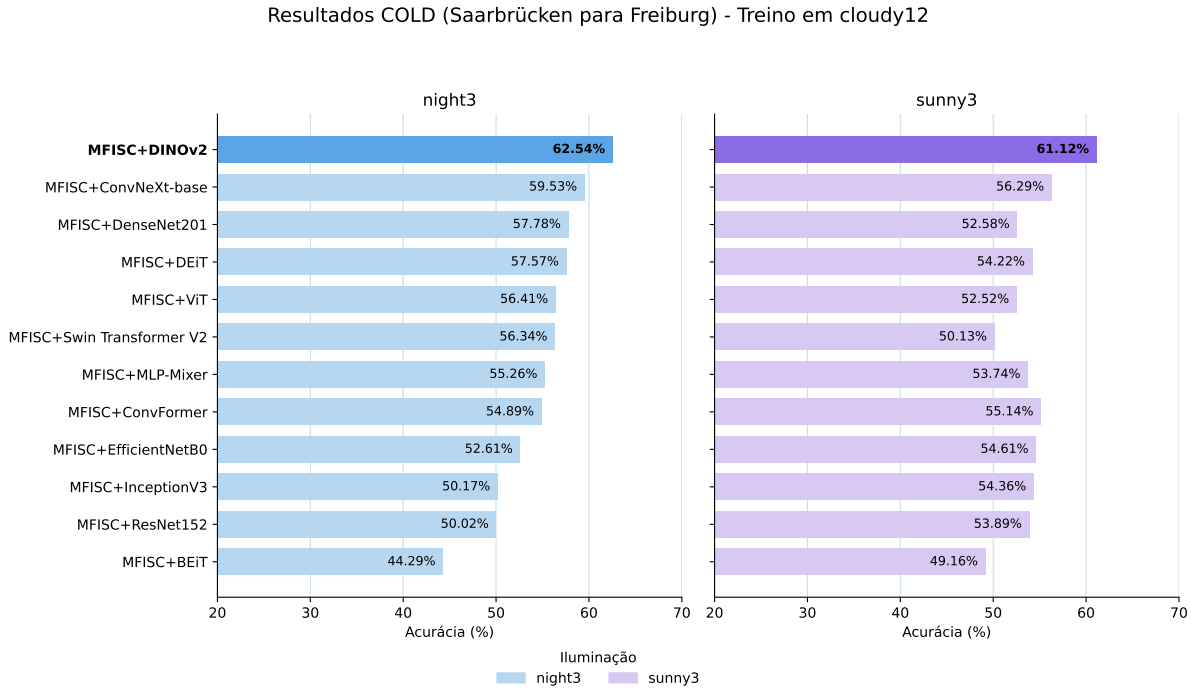


Figura 57 – Esquema 4 no *COLD* (*Saarbrücken* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-cloudy12* e teste em *Freiburg-night3* e *Freiburg-sunny3*.

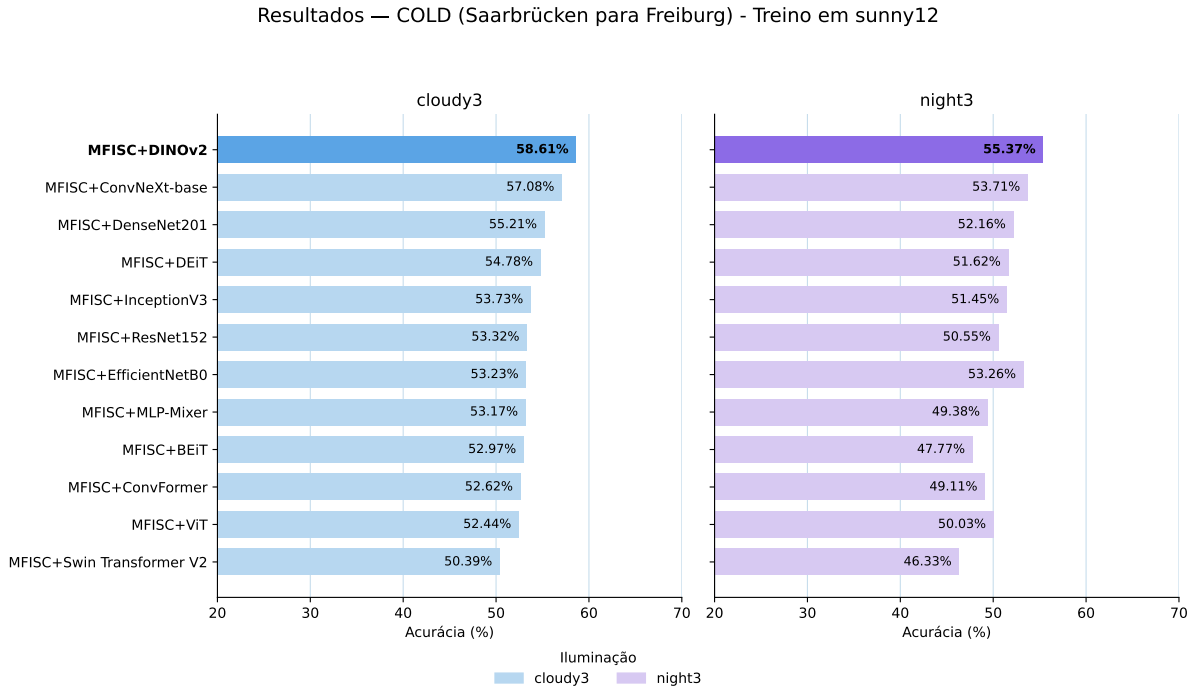


Figura 58 – Esquema 4 no *COLD* (*Saarbrücken* → *Freiburg*). Acurácia do *MFISC* variando apenas o *backbone*; treino em *Saarbrücken-sunny12* e teste em *Freiburg-cloudy3* e *Freiburg-night3*.