UNIVERSIDADE FEDERAL DO MARANHÃO

Curso: Engenharia da Computação

Hudson Costa Diniz

ESTIMATIVA DE PROFUNDIDADE MONOCULAR EM DINÂMICA DE FLUIDOS: UMA ANÁLISE COMPARATIVA DE ARQUITETURAS DE VISÃO COMPUTACIONAL.

São Luís

2026

Hudson Costa Diniz

ESTIMATIVA DE PROFUNDIDADE MONOCULAR EM DINÂMICA DE FLUIDOS: UMA ANÁLISE COMPARATIVA DE ARQUITETURAS DE VISÃO COMPUTACIONAL.

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia da Computação da Universidade Federal do Maranhão, como pré-requisito para obtenção do título de Engenheiro da Computação.

Orientador: Prof. Dr. Haroldo Gomes Barroso Filho

São Luís

2026

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Costa Diniz, Hudson.
ESTIMATIVA DE PROFUNDIDADE MONOCULAR EM DINÂMICA DE FLUIDOS: UMA ANÁLISE COMPARATIVA DE ARQUITETURAS DE VISÃO COMPUTACIONAL / Hudson Costa Diniz. - 2026.
55 p.

Orientador(a): Prof. Dr. Haroldo Gomes Barroso Filho.
Curso de Engenharia da Computação, Universidade Federal do Maranhão, São Luís - Ma, 2026.

1. Visão Computacional. 2. Estimativa de Profundidade Monocular. 3. Dinâmica dos Fluidos. 4. Modelos de Difusão. 5. Inteligência Artificial. I. Gomes Barroso Filho, Prof. Dr. Haroldo. II. Título.

FOLHA DE APROVAÇÃO

HUDSON COSTA DINIZ

ESTIMATIVA DE PROFUNDIDADE MONOCULAR EM DINÂMICA DE FLUIDOS: UMA ANÁLISE COMPARATIVA DE ARQUITETURAS DE VISÃO COMPUTACIONAL

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia da Computação da Universidade Federal do Maranhão, como pré-requisito para obtenção do título de Engenheiro da Computação.

Aprovado em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. Dr. Haroldo Gomes Barroso Filho
Universidade Federal do Maranhão
(Orientador)

_____

Prof.
(1º Examinador)

_____

Prof.
(2º Examinador)

RESUMO

A estimativa de profundidade monocular apresentou avanços significativos com o uso de Redes Neurais Profundas, contudo, a esmagadora maioria das arquiteturas estado-da-arte (SOTA) é treinada sob a hipótese do mundo rígido, baseando-se na detecção de bordas estruturais sólidas. Esse viés geométrico gera vulnerabilidades arquiteturais severas quando os modelos são submetidos a cenários amorfos e dinâmicos, como escoamentos de fluidos. O presente trabalho tem como objetivo avaliar o desempenho de nove modelos de visão computacional — englobando paradigmas de Redes Convolucionais, *Vision Transformers* e Modelos de Difusão — aplicados a simulações de Dinâmica dos Fluidos Computacional (CFD). A metodologia consistiu na análise de escoamentos não-rígidos ao redor de quatro geometrias distintas de bloqueio aerodinâmico (cilindro, quadrado, diamante e aerofólio). O desempenho foi quantificado analisando o compromisso métrico entre a precisão volumétrica espacial (Correlação de Pearson) e a coerência cinemática (Inconsistência Temporal). Os resultados empíricos comprovaram que modelos clássicos sofrem inversão matemática de profundidade devido à ausência de quinas. Demonstrou-se também que a eficácia da rede neural é estritamente dependente da topologia do fluxo: modelos de difusão latente (Marigold) apresentaram superioridade estatística absoluta em cenários de acúmulo frontal denso, enquanto modelos fundacionais de atenção espaço-temporal (Video Depth Anything - VDA) dominaram o rastreamento tridimensional em esteiras de turbulência e fluxos bifurcados. Conclui-se que existe um *trade-off* paradoxal inerente à percepção artificial contemporânea de fluidos, exigindo que a seleção da arquitetura de inteligência artificial seja guiada pela dinâmica física esperada no experimento.

Palavras-chave: Visão Computacional. Estimativa de Profundidade Monocular. Dinâmica dos Fluidos. Modelos de Difusão. Inteligência Artificial.

ABSTRACT

Monocular depth estimation has shown significant advancements with the use of Deep Neural Networks; however, the vast majority of state-of-the-art (SOTA) architectures are trained under the rigid-world hypothesis, relying heavily on the detection of solid structural edges. This geometric bias creates severe architectural vulnerabilities when models are subjected to amorphous and dynamic scenarios, such as fluid flows. This study aims to evaluate the performance of nine computer vision models—encompassing Convolutional Neural Networks, Vision Transformers, and Diffusion Models paradigms—applied to Computational Fluid Dynamics (CFD) simulations. The methodology consisted of analyzing non-rigid flows around four distinct aerodynamic block geometries (cylinder, square, diamond, and airfoil). Performance was quantified by analyzing the metric trade-off between spatial volumetric precision (Pearson Correlation) and kinematic coherence (Temporal Inconsistency). Empirical results proved that classical models suffer from mathematical depth inversion due to the lack of sharp edges. It was also demonstrated that the neural network's efficacy is strictly dependent on the flow topology: latent diffusion models (Marigold) showed absolute statistical superiority in scenarios of dense frontal accumulation, whereas spatio-temporal foundational models (Video Depth Anything - VDA) dominated the three-dimensional tracking in turbulent wakes and bifurcated flows. We conclude that there is a paradoxical trade-off inherent in the contemporary artificial perception of fluids, requiring the selection of the artificial intelligence architecture to be guided by the expected physical dynamics of the experiment.

Keywords: Computer Vision. Monocular Depth Estimation. Fluid Dynamics. Diffusion Models. Artificial Intelligence.

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A estimativa de profundidade monocular consolidou-se como um dos pilares mais desafiadores e fundamentais da visão computacional moderna. O objetivo desta tarefa é inferir a geometria tridimensional contínua de um cenário a partir de uma única projeção bidimensional, superando a carência de informações espaciais diretas inerente a sensores de lente única. Historicamente, a obtenção de profundidade dependia de algoritmos de geometria epipolar e estereoscopia, metodologias que exigem calibração rigorosa, hardware redundante e elevado custo computacional, limitando sua escalabilidade (HARTLEY; ZISSERMAN, 2004). O panorama científico sofreu uma ruptura paradigmática com a ascensão do Aprendizado Profundo (*Deep Learning*), quando Eigen, Puhrsch e Fergus (2014) comprovaram a viabilidade de deduzir mapas de profundidade utilizando redes neurais convolucionais (CNNs) multiescalas, estabelecendo que uma máquina pode aprender heurísticas espaciais a partir de texturas, sombreamentos e perspectivas em imagens estáticas.

Ao longo da última década, a evolução cronológica dessas arquiteturas foi marcada por saltos arquiteturais expressivos. A primeira geração de modelos robustos, exemplificada pela arquitetura MiDaS, baseou-se fortemente em Redes Convolucionais para extrair estimativas relativas com capacidade de transferência *zero-shot* em múltiplos domínios (RANFTL et al., 2020). Posteriormente, o campo presenciou a introdução dos *Vision Transformers* (ViTs) e de otimizações de agrupamento adaptativo (*bins*), presentes em modelos como o ZoeDepth (BHAT et al., 2023). Essa transição permitiu que as redes compreendessem o contexto global da imagem de forma muito mais eficiente do que os filtros locais e matematicamente estáticos das CNNs clássicas.

Mais recentemente, o estado-da-arte (SOTA) foi redefinido pela emergência dos Modelos Fundacionais Universais. A visão computacional atual explora o paradigma generativo, utilizando Modelos de Difusão (*Latent Diffusion Models*), como o Marigold, para tratar a profundidade espacial como um processo de síntese iterativa baseada em densidade (KE et al., 2024). Em paralelo, o desenvolvimento de arquiteturas espaço-temporais (*Video-based Foundation Models*), focadas em mecanismos de atenção contínua, tem buscado solucionar a coesão cinemática em sequências de vídeo. Trata-se, portanto, de um ecossistema tecnológico altamente sofisticado, mas cujas premissas matemáticas fundacionais ainda carecem de estresse em fronteiras físicas extremas.

1.2 Problema

Apesar do expressivo avanço das arquiteturas SOTA, o treinamento e a validação da esmagadora maioria desses sistemas baseiam-se incondicionalmente na hipótese do "mundo rígido". Os modelos contemporâneos são alimentados por conjuntos de dados (datasets) de larga escala extraídos predominantemente de ambientes urbanos, tráfego veicular e espaços indoor. Nestes contextos canônicos, a métrica espacial é estritamente definida por objetos sólidos, superfícies ortogonais, linhas de fuga demarcadas e oclusões rígidas. Consequentemente, a inteligência artificial aprende a ancorar sua inferência de profundidade na detecção de quinas e bordas nítidas que delimitam o objeto e o plano de fundo.

Essa dependência histórica do contorno estrutural introduz uma vulnerabilidade arquitetural crítica: a cegueira ou confusão algorítmica perante elementos dinâmicos não estruturados. O escoamento de fluidos, como os sintetizados em simulações de Dinâmica dos Fluidos Computacional (CFD), representa a antítese do mundo rígido. Fluidos em movimento caracterizam-se por contínuas deformações morfológicas, dissipação volumétrica gradual, semi-transparência e a ausência absoluta de delimitações geométricas duras. A física da turbulência e dos vórtices exige que a percepção espacial seja guiada pela densidade do volume e pela cinemática temporal, fatores que fogem ao escopo original das redes treinadas convencionalmente.

Quando submetidas a estes escoamentos não-rígidos, constata-se que arquiteturas clássicas sofrem colapsos interpretativos mensuráveis. Modelos com forte dependência de bordas tendem a inverter matematicamente o plano espacial, interpretando a dissipação suave do fluido como recuo em direção ao plano de fundo, resultando em profundidade negativa. Paralelamente, tentativas de mapear o fluxo contínuo frequentemente geram instabilidade severa entre quadros consecutivos, um fenômeno métrico conhecido como flickering. Esse erro temporal agrava-se sob oclusões aerodinâmicas complexas, evidenciando uma falha sistêmica das redes neurais em manter o rastreamento espaço-temporal da matéria amorfa.

1.3 Justificativa

A investigação do comportamento de redes neurais profundas diante de escoamentos de fluidos transcende a abstração teórica, possuindo ramificações diretas na aplicação industrial e na evolução dos sistemas autônomos. A engenharia contemporânea caminha para a automação de análises complexas, e compreender como as inteligências artificiais reagem a diferentes topologias e perfis de bloqueio aerodinâmico — como pontos de estagnação em

cilindros ou a formação de esteiras turbulentas em aerofólios — é uma etapa importante para avaliar o desempenho e as limitações dessas arquiteturas SOTA frente a escoamentos em cenários dinâmicos específicos, extrapolando as métricas tradicionais obtidas em ambientes urbanos controlados.

Do ponto de vista tecnológico, embora este estudo tenha sido desenvolvido em ambiente computacional simulado, os resultados obtidos podem subsidiar futuras aplicações e o refinamento algorítmico em múltiplos setores críticos. A capacidade de mensurar volumetria fluida com precisão e coerência temporal viabiliza a automação de inspeções visuais em túneis de vento aeroespaciais e a quantificação não-invasiva de fluxos térmicos em sistemas de dissipação. Além disso, sistemas robóticos que operam sob condições climáticas adversas ou em exploração submarina demandam algoritmos perceptivos imunes à confusão causada por suspensões de partículas e correntes fluidas densas.

Nesse contexto, a relevância prática desta pesquisa encontra aplicabilidade crítica no setor de segurança industrial. A viabilização da Estimativa de Profundidade Monocular (MDE) em escoamentos amorfos oferece um mecanismo avançado para o monitoramento automatizado e não-intrusivo de plantas de alta periculosidade. Historicamente, a detecção visual de vazamentos de gases tóxicos, vapores ou anomalias no escoamento de tubulações é severamente limitada pela natureza translúcida e pela rápida dissipação geométrica desses fluidos no ambiente. Ao capacitar sistemas de visão computacional baseados em câmeras convencionais (2D) para segmentar a profundidade volumétrica dessas emissões, torna-se possível não apenas identificar a presença do gás, mas mapear o seu vetor espacial de expansão e isolar a sua fonte de origem tridimensional. Consequentemente, superar as vulnerabilidades das arquiteturas de inteligência artificial perante estruturas não-rígidas configura-se como um avanço tecnológico fundamental para a automação da segurança, viabilizando respostas preditivas contra acidentes operacionais.

Finalmente, sob a ótica da pesquisa acadêmica em Ciência da Computação, justifica-se a relevância de estabelecer um *benchmark* quantitativo independente — caracterizado pela utilização de cenários fluidodinâmicos que não integraram os conjuntos de dados de treinamento originais de nenhuma das arquiteturas avaliadas, garantindo uma avaliação livre de vieses. Há uma escassez na literatura a respeito do compromisso matemático (*trade-off*) exigido na modelagem de fluidos: a escolha entre arquiteturas generativas, que priorizam a perfeição do volume estático, e modelos espaço-temporais, voltados para a estabilidade do fluxo cinemático. Este trabalho busca fornecer um referencial empírico estruturado para

auxiliar na compreensão dessa lacuna, cruzando paradigmas temporais e espaciais para expor as limitações e os potenciais dos modelos fundacionais da atualidade.

1.4 Objetivos

1.4.1 Objetivo Geral

Avaliar o desempenho e a robustez de modelos estado-da-arte (SOTA) de estimativa de profundidade monocular quando aplicados a cenários de dinâmica de fluidos computacional (CFD), analisando o compromisso entre a precisão volumétrica espacial e a estabilidade temporal em escoamentos não-rígidos.

1.4.2 Objetivos Específicos

Para viabilizar o alcance do objetivo geral e garantir uma avaliação metódica do comportamento das redes neurais perante a dinâmica de fluidos, o desdobramento desta pesquisa requer o cumprimento de etapas analíticas e estatísticas bem definidas. Desse modo, estabelecem-se os seguintes objetivos específicos para este trabalho:

- Comparar diferentes paradigmas de inteligência artificial (Redes Convolucionais, *Vision Transformers* e Modelos de Difusão) quanto à capacidade de interpretar volumes fluidos desprovidos de bordas rígidas.
- Analisar o impacto de quatro geometrias distintas de bloqueio aerodinâmico (cilindro, quadrado, diamante e aerofólio) na percepção espacial das arquiteturas avaliadas.
- Quantificar a precisão espacial de cada modelo por meio do Coeficiente de Correlação de Pearson, utilizando a densidade real do fluido como referencial analítico.
- Mensurar a inconsistência temporal (*flickering*) das estimativas de profundidade, determinando a viabilidade de modelos de vídeo e memória de curto prazo para o rastreamento contínuo de fluidos.

1.5 Contribuições

Este trabalho busca contribuir para o entendimento e avaliação prática de modelos de visão computacional estado-da-arte (SOTA) quando aplicados a cenários dinâmicos e não-rígidos, com foco em simulações de mecânica dos fluidos. As principais contribuições incluem:

- **Benchmarking Quantitativo Inédito:** Estabelecimento de um referencial empírico rigoroso que expõe as limitações e capacidades de múltiplas gerações de redes neurais na estimativa de profundidade de oclusões translúcidas e fluidos em movimento. O critério de seleção das arquiteturas, compreendendo o recorte temporal de 2019 a 2024, foi estabelecido para mapear a evolução dos principais paradigmas do estado da arte: abrangendo desde as fundacionais Redes Neurais Convolucionais (CNNs), passando pela adoção de *Vision Transformers* (ViTs) e módulos de atenção espaço-temporal, até as recentes abordagens generativas baseadas em difusão latente.
- **Identificação do Viés Geométrico:** Comprovação prática da falha sistêmica em modelos clássicos (como MiDaS e Monodepth2), demonstrando como a dependência de bordas nítidas e linhas de perspectiva resulta na inversão matemática do plano espacial em ambientes amorfos.
- **Análise de Desempenho por Perfil Aerodinâmico:** Mapeamento de como diferentes topologias de escoamento impactam a inferência das redes, revelando que geometrias de bloqueio frontal favorecem arquiteturas de difusão, enquanto geometrias de fluxo rápido exigem modelos de atenção espaço-temporal.
- **Definição do Trade-off Espaço-Temporal:** Documentação do compromisso matemático paradoxal enfrentado pelas IAs atuais ao processar fluidos, evidenciando a escolha arquitetural obrigatória entre a precisão milimétrica do volume espacial e a estabilidade cinemática temporal (prevenção de *flickering*).

1.6 Organização Do Trabalho

Este trabalho está organizado de maneira a proporcionar uma compreensão clara e sequencial dos elementos discutidos e aplicados. A estrutura segue os seguintes capítulos:

- Capítulo 1 - Introdução: Esta seção apresenta uma contextualização da evolução da visão computacional, destacando os desafios da estimativa de profundidade em cenários não-rígidos. São discutidos os problemas e as justificativas da pesquisa, além de delinear os objetivos gerais e específicos deste trabalho.
- Capítulo 2 - Fundamentação Teórica: Explora os conceitos base da estimativa de profundidade monocular, a transição dos paradigmas arquiteturais (Redes Convolucionais, *Vision Transformers* e Modelos de Difusão) e os fundamentos de dinâmica dos fluidos aplicados à percepção visual. Esta seção serve como base conceitual para a compreensão das avaliações realizadas.

- Capítulo 3 - Metodologia: Descreve o delineamento experimental adotado, incluindo as quatro geometrias de bloqueio aerodinâmico simuladas (Cilindro, Quadrado, Diamante e Aerofólio) e a seleção das arquiteturas de IA. São detalhados o *pipeline* de processamento e as métricas estatísticas aplicadas (Correlação de Pearson e Inconsistência Temporal).
- Capítulo 4 - Resultados e Discussão: Detalha o desempenho quantitativo e qualitativo obtido durante as inferências. São discutidos o impacto da aerodinâmica na percepção volumétrica, o colapso das arquiteturas antigas e o paradoxo entre a estabilidade temporal e a precisão espacial.
- Capítulo 5 - Conclusão: Este capítulo oferece uma síntese das principais descobertas do trabalho, recapitulando os objetivos atingidos e as limitações encontradas. São apresentadas recomendações práticas para a escolha de modelos e perspectivas futuras para a pesquisa na área.
- Referências Bibliográficas: Lista todas as fontes bibliográficas utilizadas ao longo do trabalho, baseando a evolução tecnológica e metodológica em literatura consolidada.
- Apêndices: Incluem materiais complementares, como questionários utilizados, códigos desenvolvidos e outros documentos relevantes para a compreensão detalhada da implementação.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Dinâmica de Fluidos e Perfis Aerodinâmicos

A Dinâmica dos Fluidos Computacional (CFD - *Computational Fluid Dynamics*) consolidou-se como a principal abordagem numérica para a resolução das equações governantes de Navier-Stokes, permitindo a análise preditiva de escoamentos complexos em engenharia. Nos últimos anos, a literatura científica tem avançado significativamente na integração entre métodos numéricos tradicionais e modelos de aprendizado de máquina (*Machine Learning*) focados em otimização de simulações e modelagem de turbulência (BRUNTON; NOACK; KOUMOUTSAKOS, 2020). Contudo, a partir de um levantamento bibliográfico focado na interseção entre visão computacional profunda e fluidodinâmica, nota-se uma expressiva lacuna na literatura contemporânea no que tange à interpretação visual automatizada dessas simulações. Grande parte dos esforços foca em melhorar os *solvers* internos, mas poucos estudos investigam como modelos de visão computacional interpretam os campos de densidade e oclusão gerados por esses fluidos.

Diferente da mecânica dos sólidos, os fluidos caracterizam-se pela deformação contínua sob tensões de cisalhamento. O comportamento cinemático e volumétrico do fluido é inexoravelmente ditado pela geometria do obstáculo inserido em sua trajetória (ANDERSON, 2023). Para investigar a fronteira de percepção das redes neurais, esta pesquisa selecionou quatro perfis aerodinâmicos, cada um induzindo fenômenos físicos que desafiam a extração de características visuais estáticas:

Perfil Cilíndrico (Curvatura Contínua): O escoamento sobre um cilindro circular é o paradigma clássico da mecânica dos fluidos incompressíveis. Caracteriza-se pela formação de uma zona de estagnação de alta pressão na face frontal, seguida pela aceleração do fluido ao longo de sua curvatura. Devido ao gradiente adverso de pressão, as partículas de fluido próximas à superfície perdem energia cinética, provocando o descolamento da camada limite. Esse fenômeno gera zonas de recirculação — onde o fluxo local inverte sua direção — e culmina na formação alternada e caótica da esteira de vórtices de Von Kármán (CAI et al., 2021). É justamente a presença dessas zonas de instabilidade, que produzem gradientes visuais translúcidos e altamente dinâmicos, que desafiam os algoritmos de percepção de profundidade, justificando a escolha desta geometria geométrica para o estudo. Visualmente, gera um acúmulo denso frontal e uma dissipação oscilatória traseira.

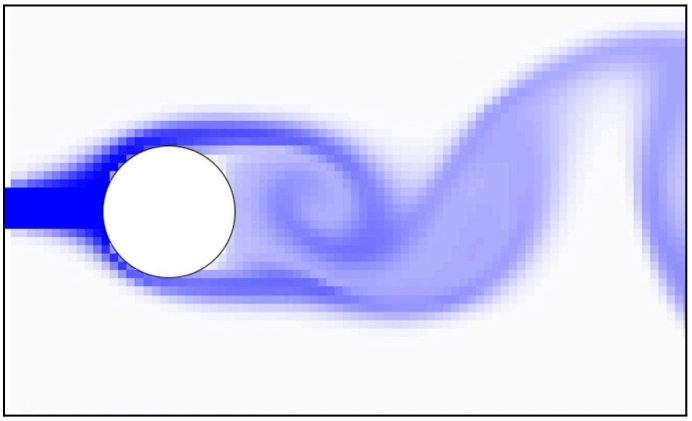


Figura 1 - Representação natural do escoamento fluido interagindo com o perfil cilíndrico, renderizada através do ambiente simulado *wxRTCFD*. Fonte: O autor (2026).

Perfil Quadrado (Bloqueio Plano): O perfil quadrado atua como um corpo rombudo extremo (*blunt body*). A face frontal plana e ortogonal ao fluxo oblitera o escoamento direto, gerando uma zona de estagnação volumétrica consideravelmente mais densa que a do cilindro. A separação do fluxo ocorre de maneira abrupta nas arestas, gerando uma esteira de recirculação turbulenta caótica e larga (WANG et al., 2022). Para algoritmos de visão, representa o maior teste de interpretação de acúmulo estático sem oclusão de bordas suaves.

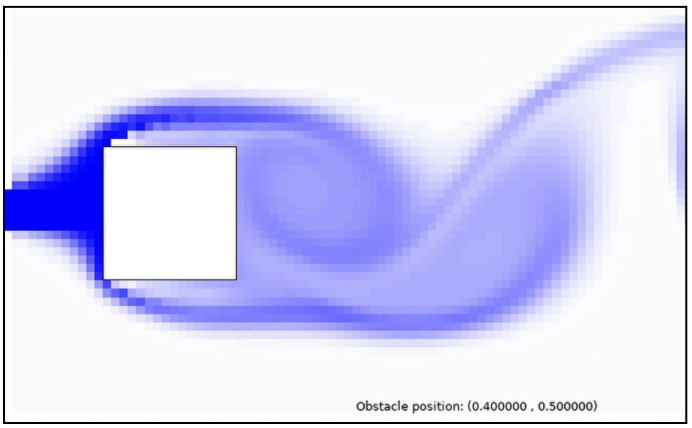


Figura 2 - Representação natural do escoamento fluido sob bloqueio plano do perfil quadrado, renderizada através do ambiente simulado *wxRTCFD*. Fonte: O autor (2026).

Perfil em Diamante (Bifurcação Angular): A topologia em diamante posiciona um vértice pontiagudo diretamente contra o vetor de velocidade do fluido, forçando uma bifurcação angular imediata. Em vez de acumular densidade na zona frontal, o perfil transfere a energia cinética, dividindo o escoamento em duas correntes simétricas de alta velocidade. Esta rápida transição testa a capacidade das redes neurais de rastrear fluidos em rápida deformação espacial sem um "paredão" de estagnação referencial.

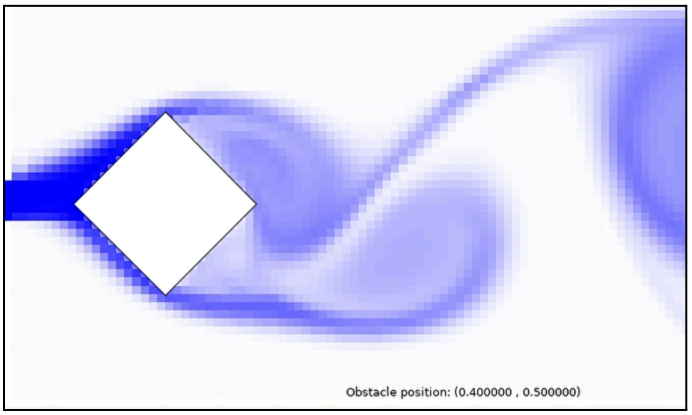


Figura 3 - Representação natural da bifurcação angular do fluxo perante o perfil em diamante, renderizada através do ambiente simulado *wxRTCFD*. Fonte: O autor (2026).

Perfil de Aerofólio (Fluxo Sustentado e Esteira): Representando a otimização aerodinâmica estrutural, o aerofólio direciona o fluido através de uma borda de ataque arredondada por uma superfície alongada e suave. O escoamento mantém-se "colado" à superfície da geometria o máximo possível para atrasar a separação, fundindo-se novamente na borda de fuga e criando uma esteira fina, alongada e de alta velocidade (BRUNTON;

NOACK; KOUMOUTSAKOS, 2020). Computacionalmente, exige dos modelos de visão a capacidade de acompanhar gradientes sutis e quase transparentes.

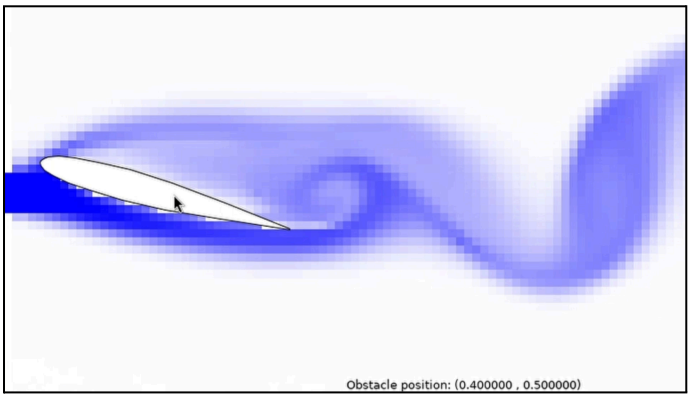


Figura 4 - Representação natural da bifurcação angular do fluxo perante o perfil em diamante, renderizada através do ambiente simulado *wxRTCFD*. Fonte: O autor (2026).

2.2 Estimativa de Profundidade Monocular (Depth Maps)

A estimativa de profundidade monocular (MDE - *Monocular Depth Estimation*) consiste na capacidade de inferir a estrutura tridimensional contínua de uma cena a partir de uma única projeção fotográfica bidimensional. Matematicamente, a tarefa propõe calcular um escalar Z para cada coordenada de pixel (u, v) de uma imagem RGB, gerando uma matriz espacial denominada Mapa de Profundidade (*Depth Map*). No contexto metodológico deste trabalho focado em fluidodinâmica, é imperativo ressaltar que a profundidade espacial real e a densidade da concentração do fluido são grandezas físicas distintas. Contudo, como a renderização bidimensional do escoamento utiliza a variação de opacidade e luminância para representar o acúmulo de partículas, as redes neurais tendem a interpretar a alta densidade visual (menor transparência) como proximidade geométrica. Esta equivalência indireta é o fenômeno que permite a segmentação estrutural do escoamento. Trata-se de um clássico "problema mal posto" (*ill-posed problem*), uma vez que a projeção óptica de um ambiente 3D para um plano 2D resulta em uma perda irreversível da escala absoluta, permitindo que uma única imagem bidimensional corresponda a infinitas configurações espaciais possíveis (ZHAO et al., 2020).

Figura 5 - Estimativa de profundidade monocular com o modelo Marigold. A fileira superior exibe as imagens originais RGB; a inferior, os respectivos mapas de profundidade, onde tons quentes indicam menor distância espacial. **Fonte:** Bing et al. (2024) - Adaptado pelo autor

Para solucionar a ambiguidade inerente à MDE, a visão computacional tradicional dependia de pistas puramente geométricas ou de múltiplos sensores (estereoscopia). Entretanto, com a maturação do Aprendizado Profundo (*Deep Learning*), as redes neurais passaram a deduzir a profundidade a partir de padrões implícitos na imagem, conhecidos como "pistas pictóricas". Essas características incluem gradientes de iluminação, textura, tamanho relativo, perspectiva linear e a sobreposição (oclusão) de objetos (MING et al., 2021).

Apesar do notável sucesso empírico dos modelos SOTA (estado-da-arte) em *benchmarks* públicos, uma análise crítica da literatura recente expõe um viés metodológico significativo. A quase totalidade das arquiteturas desenvolvidas para MDE é treinada e otimizada sobre conjuntos de dados (*datasets*) canônicos, como o KITTI (focado em navegação autônoma urbana) e o NYU Depth V2 (focado em ambientes residenciais *indoor*). Consequentemente, o aprendizado de máquina global foi condicionado sob a estrita "hipótese do mundo rígido". Dong et al. (2022) ressaltam que, embora as redes neurais atinjam elevada acurácia preditiva em cenários estruturados como superfícies opacas e texturizadas, elas sofrem degradação drástica perante materiais transparentes, superfícies espelhadas e oclusões amorfas.

É exatamente nesta vulnerabilidade que reside a principal lacuna da literatura contemporânea que este Trabalho de Conclusão de Curso visa preencher. A percepção espacial de campos volumétricos deforma-se continuamente, como os fluidos simulados em CFD, e é uma fronteira inexplorada. Escoamentos não possuem arestas definidas, quinas, textura sólida ou sobreposição nítida, o que priva as redes neurais de suas âncoras matemáticas habituais. Sem referências geométricas rígidas, arquiteturas convencionais

frequentemente apresentam severas falhas preditivas, projetando a dissipação suave do fluido como recuos espaciais irreais (inversão de profundidade). Entender como contornar essa limitação algorítmica é essencial para expandir o uso da visão monocular para domínios de monitoramento ambiental e aerodinâmico.

2.3 Arquiteturas de Redes Neurais

A evolução tecnológica da estimativa de profundidade monocular reflete mudanças profundas na forma como as redes neurais extraem, processam e correlacionam a informação espacial. Para compreender as métricas de sucesso e os colapsos estruturais documentados nos escoamentos aerodinâmicos, é imperativo analisar o funcionamento interno das principais gerações arquiteturais de visão computacional utilizadas nesta pesquisa.

2.3.1 Redes Neurais Convolucionais (CNNs) e a Dependência de Bordas

As Redes Neurais Convolucionais (CNNs - *Convolutional Neural Networks*) representam o primeiro grande marco na transição de algoritmos puramente geométricos para abordagens de aprendizado profundo na visão computacional. Durante anos, constituíram o motor base de modelos clássicos de estimativa de profundidade, como o MiDaS e o Monodepth2, estabelecendo os primeiros *benchmarks* de sucesso para a reconstrução tridimensional a partir de imagens estáticas (ALHASHIM; WON, 2018).

O funcionamento fundamental dessas redes baseia-se na aplicação de operações matemáticas de convolução. A arquitetura utiliza matrizes de pesos ajustáveis, conhecidas como filtros ou *kernels*, que deslizam sistematicamente sobre a imagem de entrada bidimensional. A cada passo, o *kernel* realiza o produto escalar entre seus próprios pesos e os valores dos pixels contidos em seu campo de visão local (o campo receptivo). O resultado dessa operação gera Mapas de Características (*Feature Maps*), que codificam padrões visuais hierárquicos. Nas primeiras camadas da rede, os filtros são treinados para identificar texturas simples, gradientes de cor e, principalmente, bordas nítidas e orientações angulares. À medida que a informação avança para camadas mais profundas, a rede combina essas bordas simples para reconhecer formas complexas e oclusões (HU et al., 2019).

Estruturalmente, as CNNs aplicadas à estimativa de profundidade adotam tipicamente uma topologia *Encoder-Decoder* (Codificador-Decodificador). O *Encoder* submete a imagem original a sucessivas convoluções e operações de redução de dimensionalidade (*pooling*), comprimindo a informação espacial em um vetor latente altamente denso que captura o contexto semântico da cena. Em seguida, o *Decoder* assume o papel de reconstrução,

utilizando convoluções transpostas (*upsampling*) para expandir esse vetor latente de volta à resolução original da imagem, atribuindo a cada pixel uma estimativa escalar de distância (ZHAO et al., 2020).

Apesar de sua altíssima eficiência computacional e da capacidade de operar em tempo real, o paradigma convolucional possui uma limitação inerente: a sua dependência absoluta de referências locais de alta frequência. Como a convolução opera sobre vizinhanças restritas de pixels, a rede necessita de transições abruptas de intensidade luminosa — como a quina de uma parede, a silhueta de um veículo ou a borda estrutural de um objeto — para delimitar o fim de um plano espacial e o início do plano de fundo.

A literatura aponta que arquiteturas estritamente convolucionais apresentam alta taxa de erro (suavização excessiva e indefinição) quando aplicadas a regiões amorfas ou materiais transparentes, uma vez que o *kernel* não encontra o gradiente rígido esperado para ancorar a profundidade (DONG et al., 2022). Em cenários de Dinâmica dos Fluidos Computacional, essa vulnerabilidade transforma-se em um colapso sistêmico. A natureza dissipativa do fluido, a ausência de fronteiras sólidas e a transição translúcida das esteiras de turbulência impedem que a CNN separe o volume do fundo do simulador, induzindo o modelo a gerar mapas com inversões espaciais profundas e rastreamento errático.

2.3.2 Vision Transformers (ViTs) e o Contexto Global

A limitação estrutural imposta pelo campo receptivo local das Redes Neurais Convolucionais motivou a comunidade científica a buscar novos horizontes interpretativos. O marco dessa transição ocorreu com a adaptação da arquitetura *Transformer*, originalmente concebida para o Processamento de Linguagem Natural (PLN), para o domínio espacial. Introduzido por Dosovitskiy et al. (2021), o *Vision Transformer* (ViT) redefiniu o paradigma da visão computacional ao propor uma alternativa robusta à dependência exclusiva das convoluções tradicionais. Ao introduzir um mecanismo de atenção global para o processamento de imagens, essa arquitetura abriu caminho tanto para modelos puramente atencionais quanto para o desenvolvimento de redes híbridas contemporâneas, que combinam a eficiência local das convoluções com o vasto campo receptivo dos *Transformers*.

A arquitetura ViT não processa a imagem como uma matriz contínua de pixels iterados localmente. Em vez disso, a imagem de entrada bidimensional é fatiada em uma grade de fragmentos não sobrepostos, denominados *patches* (frequentemente matrizes de $16x16$ pixels). Cada *patch* é então linearmente achatado e projetado em um vetor unidimensional (vetor de características). Para garantir que a rede não perca a noção da estrutura geométrica

original, vetores de codificação posicional (*positional embeddings*) são somados a essas projeções antes de serem processadas (RANFTL et al., 2021).

O núcleo matemático que confere poder aos ViTs é o mecanismo de Auto-Atenção (*Self-Attention*). Esse mecanismo permite que a rede calcule a relevância mútua entre todas as partes da imagem simultaneamente. Matematicamente, as características de entrada são mapeadas em três matrizes distintas: Busca ($Q - Query$), Chave ($K - Key$) e Valor ($V - Value$). A matriz de atenção é obtida calculando-se o produto escalar entre Q e K, escalonado pela dimensão dos vetores (d_k), submetido a uma função de ativação para normalização probabilística e, finalmente, multiplicado por V:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Na prática, essa equação significa que cada *patch* da imagem "presta atenção" e extrai contexto de todos os outros *patches* da cena de uma só vez, independentemente da distância euclidiana que os separa. Se um *patch* contém a ponta de um obstáculo, o ViT correlaciona essa informação com a base do obstáculo no extremo oposto da imagem na primeira camada de processamento. Essa capacidade de abstrair o "contexto global" permitiu o desenvolvimento de arquiteturas de estimativa de profundidade densa (DPT - *Dense Prediction Transformers*), servindo de fundação para modelos SOTA como o **ZoeDepth** e o **UniDepth**.

Para aprofundar a precisão, modelos como o ZoeDepth introduziram o conceito de agrupamento métrico adaptativo (*Metric Bins*). Em vez de tentar calcular a profundidade contínua absoluta instantaneamente, a rede utiliza o contexto global do *Transformer* para classificar as probabilidades espaciais em faixas discretas (intervalos de distância), realizando o refinamento intrapíxel posteriormente (BHAT et al., 2023).

Apesar do ganho expressivo na coerência global da cena, a literatura estabelece que o paradigma ViT estático possui limitações inerentes de percepção, uma vez que sua formulação original pressupõe a existência de superfícies definidas e processa a cena fatiada em um único instante (T), carecendo de consciência temporal. Avançando sobre essa premissa teórica, as análises conduzidas neste estudo constataram empiricamente como essa vulnerabilidade se materializa no contexto da Dinâmica dos Fluidos. Os resultados obtidos demonstraram que escoamentos aerodinâmicos — especialmente as esteiras turbulentas formadas nos perfis de

aerofólio e diamante — atuam como gradientes semi-transparentes que quebram a lógica de correlação da rede. Verificou-se na prática que a arquitetura falha em capturar o volume do fluido amorfo, comprovando que as propriedades tridimensionais dessas estruturas aerodinâmicas só se tornam calculáveis para a inteligência artificial quando analisadas por meio do movimento contínuo, o que justifica a necessidade de uma expansão temporal na arquitetura.

2.3.3 Modelos de Difusão Latente (Latent Diffusion) e a Geração Volumétrica

A mais recente e revolucionária quebra de paradigma na estimativa de profundidade monocular é representada pela adaptação dos Modelos de Difusão Latente (*Latent Diffusion Models* - LDMs). Diferente das CNNs e ViTs — que tratam a profundidade como um problema de regressão discriminativa, tentando mapear pixels diretamente para distâncias matemáticas —, a difusão aborda a tarefa como um processo estocástico e generativo. O estado-da-arte desta categoria é materializado por arquiteturas como o **Marigold**, que reaproveitam o conhecimento prévio de geradores de imagem fundacionais (como o *Stable Diffusion*) para inferir a geometria tridimensional (KE et al., 2024).

A base teórica dos modelos de difusão é inspirada na termodinâmica de não-equilíbrio. O treinamento de um modelo de difusão clássico (DDPM - *Denoising Diffusion Probabilistic Models*) é dividido em dois processos de cadeias de Markov. O processo direto (*Forward Process*) consiste em pegar uma amostra de dados reais (um mapa de profundidade perfeito) e injetar ruído Gaussiano gradativamente, ao longo de T instantes de tempo, até que o dado se degenere completamente em um tensor de ruído branco isotrópico. Matematicamente, a transição de um estado para o próximo é descrita por:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \quad (2)$$

Onde x_t é a amostra ruidosa no instante t, β_t representa o cronograma de variância do ruído inserido, e N denota a distribuição Normal (HO; JAIN; ABBEEL, 2020).

A inovação computacional dos LDMs, introduzida por Rombach et al. (2022), foi transferir essas operações do espaço de pixels (que exige poder de processamento massivo) para um espaço latente altamente comprimido, gerenciado por *Autoencoders* Variacionais (VAEs). O núcleo de inteligência da rede reside no processo reverso (*Reverse Process*). Uma arquitetura de rede neural robusta, tipicamente uma U-Net massiva acoplada a mecanismos de

atenção cruzada (*Cross-Attention*), é treinada para prever e subtrair o ruído adicionado, revertendo o processo passo a passo.

Quando o Marigold é acionado para inferir a profundidade de uma nova imagem, o processo inicia-se com um campo latente de ruído puro. O modelo iterativamente "esculpe" o mapa de profundidade, condicionando a remoção do ruído às características da imagem RGB de entrada (o fluido). Em vez de procurar deterministicamente por bordas rígidas, o LDM avalia a variação cromática e a opacidade gradativa da imagem. Cabe ressaltar que a rede não realiza uma reconstrução volumétrica tridimensional (3D) completa do escoamento, mas gera uma representação espacial 2.5D (um mapa topológico bidimensional). A geometria inferida por meio de síntese probabilística reflete estritamente a distância escalar da camada fluida visível em relação à lente da câmera.

Apesar da excelência na precisão volumétrica estática, o paradigma LDM carrega limitações operacionais. A natureza iterativa de sua inferência (exigindo dezenas de iterações da U-Net por quadro) impõe um custo computacional restritivo, inviabilizando aplicações em tempo real. Ademais, como o processo parte de um ruído estocástico aleatório a cada nova imagem, mapas de profundidade gerados em quadros consecutivos de um vídeo podem apresentar inconsistências microscópicas independentes, resultando em um fenômeno de cintilação óptica temporal (*flickering*), exigindo metodologias auxiliares para estabilização cinemática.

2.3.4 Modelos Fundacionais de Vídeo e Atenção Temporal

A aplicação de modelos de visão profunda baseados em quadros isolados (*frame-by-frame*) para a análise de sequências contínuas esbarra em uma limitação física inerente: a desconsideração da cinemática. Em simulações de Dinâmica dos Fluidos Computacional, o escoamento é um evento contínuo no espaço-tempo. Uma rede neural que avalia uma imagem isolada ignora o histórico de oclusão do ambiente, o que frequentemente resulta em estimativas de profundidade que oscilam microscopicamente entre quadros consecutivos. Esse fenômeno métrico, conhecido como *flickering* ou inconsistência temporal, prejudica severamente a confiabilidade da reconstrução tridimensional em aplicações de engenharia (LUO et al., 2020).

Para mitigar a instabilidade temporal, o estado-da-arte evoluiu para os Modelos Fundacionais de Vídeo (*Video-based Foundation Models*), representados neste estudo pelas arquiteturas **Video Depth Anything (VDA)** e **ViTA**. Estes modelos expandem o paradigma

dos *Vision Transformers* (ViTs) através da introdução de módulos de Atenção Espaço-Temporal (*Spatio-Temporal Attention*).

Matematicamente, em vez de processar apenas as coordenadas espaciais (X, Y) de um tensor atual, o mecanismo de atenção temporal projeta a correlação entre os vetores de características do quadro presente (T) e os tensores de memória extraídos dos quadros imediatamente anteriores ($T - 1$, $T - 2$, etc.). O objetivo computacional é impor restrições de coesão óptica geométrica: se um aglomerado volumétrico possui uma determinada profundidade escalar no instante $T - 1$, a restrição de continuidade espaço-temporal imposta pelo modelo assume que sua profundidade não sofrerá transições descontínuas extremas no instante T, a menos que haja uma oclusão visual abrupta. Modelos como o VDA utilizam essa atenção contínua para mapear campos de movimento (fluxo óptico latente) em vídeos longos, garantindo uma estimativa cinematograficamente suave e consistente (KANG et al., 2024).

Entretanto, o desempenho dessas arquiteturas temporais frente à mecânica de fluidos não-rígidos revela um limite arquitetural crítico. O rastreamento visual bem-sucedido por essas arquiteturas pressupõe a conservação da forma e a correspondência de aparência visual do objeto rastreado (hipótese da rigidez temporal). Escoamentos fluidos violam essa premissa visual continuamente: embora obedeçam ao princípio físico de conservação da massa, eles não preservam sua morfologia ou textura visual ao longo do tempo. Escoamentos fluidos violam essa premissa continuamente. Nas geometrias aerodinâmicas de Diamante e Aerofólio analisadas neste estudo, o fluido sofre separação de camada limite, bifurcação ultrarrápida e desenvolve esteiras de vórtices altamente translúcidas que se dissipam no espaço vazio.

Arquiteturas desenhadas para reforçar a memória de curto prazo por meio do pareamento (*matching*) de características visuais rígidas enfrentam desafios teóricos severos nesse cenário. Matematicamente, tentar forçar a correspondência exata de um fragmento (*patch*) de fluido entre quadros consecutivos pode gerar divergências nos mecanismos de atenção, visto que o fluido deforma-se ou torna-se transparente subitamente. Por outro lado, modelos que flexibilizam essa restrição geométrica, baseando-se em campos de movimento latente em vez de formas estáticas, apresentam uma formulação teórica mais adequada para acompanhar a trajetória estocástica de escoamentos amorfos sem forçar correspondências ilusórias.

2.4 Análise Topológica das Arquiteturas Selecionadas

Para a execução deste estudo comparativo, foram selecionadas arquiteturas que representam o estado-da-arte em seus respectivos anos de lançamento e paradigmas

matemáticos. A seguir, detalha-se a estrutura interna, as funções de otimização e a lógica de inferência de cada rede neural avaliada.

2.4.1 MiDaS: Transferência Zero-Shot e Invariância de Escala

O MiDaS (*Robust Monocular Depth Estimation*), especificamente em sua versão 2.1, representa um marco significativo das arquiteturas baseadas estritamente em Redes Neurais Convolucionais (CNNs) para a inferência de profundidade monocular. Desenvolvido por Ranftl et al. (2020), o modelo foi projetado para resolver um problema crônico da visão computacional clássica: a incompatibilidade estrutural entre diferentes conjuntos de dados (*datasets*).

Até a concepção do MiDaS, redes treinadas em *datasets indoor* (como NYU) falhavam catastroficamente em cenários *outdoor* (como KITTI) devido a ambiguidades de escala absoluta. O núcleo teórico do MiDaS resolve isso descartando a tentativa de prever a distância métrica real. Em vez disso, a rede foca na profundidade **relativa** (qual objeto está à frente do outro) utilizando uma função de perda (*loss function*) invariante à escala e ao deslocamento espacial (*scale-and-shift invariant loss*).

A topologia do MiDaS emprega uma espinha dorsal (*backbone*) de extração de características poderosa, tipicamente baseada na arquitetura ResNeXt-101. A imagem original é processada através de blocos residuais convolucionais que extraem representações em múltiplas resoluções. Em seguida, um módulo de fusão (*fusion module*) baseado em convoluções piramidais agrega esses mapas de características locais e globais.

Matematicamente, a função de perda fundamental da arquitetura alinha a predição d e a verdade terrestre (*ground truth*) d^* aplicando uma transformação linear para que ambas tenham média zero e variância unitária. A métrica de erro é dada pela diferença absoluta média combinada a um termo de suavidade de bordas espacial:

$$L_{info}(\theta, \theta^1) = -\frac{1}{N} \sum_{i=1}^{N} \|\theta_i - \theta_i^1\|^2 \tag{3}$$

Onde $\widehat{d}$ representa a profundidade normalizada no espaço topológico. Embora essa abordagem tenha garantido ao MiDaS uma robustez *zero-shot* sem precedentes para imagens gerais, a literatura estabelece que a sua dependência de convoluções locais para detectar 'bordas relativas' constitui uma vulnerabilidade teórica em cenários desprovidos de geometria rígida. Corroborando essa premissa teórica, as análises empíricas conduzidas neste estudo

constataram que essa limitação arquitetural se traduz em uma falha crítica de percepção, evidenciada pela incapacidade do modelo em processar a dissipação volumétrica suave e translúcida encontrada na aerodinâmica de fluidos.

2.4.2 ZoeDepth: Combinação de Profundidade Relativa e Métrica

O ZoeDepth (Bhat et al., 2023) surgiu como uma resposta direta à limitação do MiDaS: a incapacidade de fornecer escalas métricas precisas (profundidade real em metros) mantendo a capacidade de generalização. O modelo marca a transição da hegemonia convolucional para a adoção híbrida de *Vision Transformers* (ViTs).

A arquitetura do ZoeDepth utiliza um *design* de dois estágios. O primeiro estágio atua como um decodificador de profundidade relativa (herdando o conhecimento do MiDaS), focado em manter a integridade da cena global e das silhuetas. O diferencial absoluto reside no segundo estágio, denominado "Cabeça Métrica" (*Metric Head*), que emprega a inovadora técnica de Agrupamentos Métricos Adaptativos (*Metric Bins*).

O modelo abandona a regressão contínua em favor da classificação ponderada. Ele divide o espaço de profundidade possível em *N* caixas (*bins*), onde os centros dessas caixas não são predefinidos, mas sim ajustados dinamicamente pelos mecanismos de Auto-Atenção do *Transformer* de acordo com a imagem de entrada. A profundidade final *Z* de um pixel é calculada através da combinação linear das probabilidades preditas para cada *bin*:

$$Z(u,v) = \sum_{k=1}^{K} p_k(u,v) \cdot c_k \tag{4}$$

Onde $(u,\ v)$ é a probabilidade do pixel $(u,\ v)$ pertencer ao centro de caixa c_k. Ao introduzir o *Transformer* para alocar os *bins* dinamicamente, baseando-se no contexto global, o ZoeDepth conseguiu estimar o espaço com muito mais precisão. Como os escoamentos apresentam regiões com gradientes contínuos de concentração, avaliar o comportamento de arquiteturas capazes de inferir a profundidade métrica (e não apenas relativa) torna-se estritamente relevante. Na presente pesquisa, essa arquitetura híbrida demonstrou flexibilidade para se adaptar a geometrias de bloqueio fluido. Na presente pesquisa, essa arquitetura híbrida demonstrou flexibilidade para se adaptar a geometrias de bloqueio fluido, mas a ausência de uma "cabeça temporal" ainda o penaliza no rastreamento de esteiras turbulentas de alta velocidade.

2.4.3 Marigold: Estimativa de Profundidade por Difusão Latente

O Marigold (Ke et al., 2024) quebra completamente os paradigmas regressivos anteriores, transferindo a tarefa de estimativa de profundidade para o domínio da Inteligência Artificial Generativa. Ao invés de construir uma rede neural do zero, o Marigold reutiliza o conhecimento vasto e universal embutido em modelos de geração de imagem texto-para-imagem (*Text-to-Image*), especificamente o *Stable Diffusion*.

A arquitetura não opera no espaço original de pixels. A imagem de entrada é comprimida por um Codificador Variacional (VAE) em um tensor latente z. O processo de difusão latente (*Latent Diffusion Model* - LDM) funciona baseando-se nas equações de Langevin da termodinâmica. No processo reverso, o modelo inicia com um tensor de ruído puro z_T e utiliza uma U-Net massiva, rica em mecanismos de *Cross-Attention* (Atenção Cruzada), para remover o ruído em T iterações discretas, condicionando a limpeza às características visuais da imagem RGB do fluido.

O objetivo de otimização não é minimizar o erro direto de profundidade, mas sim o erro de previsão do ruído adicionado ϵ:

$$L_{LDM} := \mathbb{E}_{\epsilon(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\left|\left|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\right|\right|_2^2\right] \quad (5)$$

Onde z_T é o mapa de profundidade ruidoso no instante t, y é a imagem RGB condicionante e ϵ_θ é a rede U-Net preditora. A inovação metodológica do Marigold reside na sua capacidade de realizar uma síntese probabilística iterativa. A rede neural reconstrói o campo de profundidade realizando um processo de remoção gradual de ruído, no qual a geometria é inferida a partir da correlação entre a opacidade cromática da entrada e os padrões de densidade visual observados durante a otimização estatística do processo reverso.

2.4.4 Video Depth Anything (VDA): Atenção Espaço-Temporal

O *Video Depth Anything* (VDA) é uma arquitetura fundacional lançada em 2024 (Kang et al.) que visa solucionar o problema crônico da inconsistência e do *flickering* (cintilação) em vídeos de longa duração. Ao passo que as arquiteturas anteriores encaram a física do tempo como imagens desconexas, o VDA assume a continuidade espacial como uma restrição imperativa.

O diferencial arquitetural do VDA é a ausência de módulos complexos de alinhamento óptico (como fluxo óptico explícito) que costumavam deixar as redes antigas lentas e frágeis.

Em vez disso, o modelo se baseia na arquitetura *Depth Anything* original (treinada em mais de 60 milhões de imagens) e introduz blocos de Atenção Espaço-Temporal (*Spatio-Temporal Attention Blocks*) integrados diretamente nos *encoders* DINOv2.

Para um vídeo de entrada representado como um tensor $V \in R^{TxCxHxW}$, a rede extrai representações latentes de múltiplos quadros simultaneamente. A matriz de atenção A é expandida para abranger não apenas os *patches* adjacentes do quadro atual t, mas também os *patches* correspondentes em $t - 1\ e\ t + 1$.

O modelo é otimizado utilizando uma função de perda mútua de vídeo que penaliza oscilações temporais na derivada de primeira e segunda ordem da profundidade entre quadros. Segundo a literatura, o mecanismo de atenção espaço-temporal do VDA deveria, teoricamente, permitir a suavização de transições cinemáticas e o rastreamento consistente de estruturas volumétricas ao longo do tempo (KANG et al., 2024). Nos experimentos conduzidos neste trabalho, essa premissa teórica foi confirmada de forma prática: em simulações onde os perfis em Diamante e Aerofólio induzem descolamento dinâmico ultrarrápido, observou-se que o VDA mantém a estabilidade da inferência ao processar a esteira, rastreando a cinemática do escoamento de forma mais robusta e com menor taxa de *flickering* do que as arquiteturas baseadas em modelos de difusão estática.

2.4.5 Monodepth2: Aprendizado Auto-Supervisionado e Reprojeção Fotométrica

O Monodepth2, proposto por Godard et al. (2019), introduziu um marco na estimativa de profundidade ao consolidar o paradigma do aprendizado auto-supervisionado (*self-supervised learning*). Diferente de modelos que dependem de dados reais de profundidade obtidos por sensores LiDAR (dados caros e esparsos), o Monodepth2 aprende a estimar a profundidade utilizando apenas sequências de vídeos monoculares estéreo ou temporais durante o treinamento.

O núcleo teórico desta arquitetura baseia-se na síntese de visualização (*view synthesis*). O modelo é composto por duas redes neurais operando em conjunto: uma Rede de Profundidade (uma CNN U-Net padrão) e uma Rede de Pose. A Rede de Pose estima a rotação e translação da câmera entre quadros temporais adjacentes ($t - 1,\ t,\ t + 1$).

A otimização matemática não ocorre comparando a profundidade predita com uma profundidade real, mas minimizando o Erro de Reprojeção Fotométrica. Matematicamente, a rede tenta usar a profundidade predita no quadro t e a pose estimada para "reprojetar" os

pixels do quadro t' para formar uma imagem sintetizada $\widehat{I}_t$. A função de perda penaliza a diferença de cor (fotometria) entre a imagem real e a sintetizada:

$$L_p = \min_{t'} pe(I_t, I_{t' \to t}) \tag{6}$$

Onde pe é a função de erro fotométrico (geralmente uma combinação de erro L1 e SSIM - *Structural Similarity Index Measure*). A inovação crítica do Monodepth2 foi a introdução do *Auto-Masking*, que desativa o cálculo de perda para pixels estáticos (como o céu ou objetos que se movem na mesma velocidade da câmera).

Apesar de brilhante para navegação urbana autônoma, o modelo possui uma vulnerabilidade teórica severa para fluidos: a formulação matemática da reprojeção assume a "hipótese de brilho constante" e a "geometria estática" da cena. Como os fluidos sofrem deformação estrutural instantânea e mudanças na opacidade, a função de perda fotométrica torna-se mal condicionada, resultando em uma divergência na otimização do erro de reprojeção. Consequentemente, a rede é incapaz de realizar a correspondência de pixels entre quadros consecutivos.

2.4.6 NeWCRFs: Campos Aleatórios Condicionais e Bordas Rígidas

O modelo NeWCRFs (*Neural Window Fully-connected CRFs*), proposto por Yuan et al. (2022), foi desenvolvido para mitigar a imprecisão na delimitação de fronteiras frequentemente observada em arquiteturas convolucionais. A premissa central da arquitetura reside na imposição de uma restrição geométrica explícita, na qual a profundidade predita deve alinhar-se rigorosamente às descontinuidades de intensidade e textura da cena, utilizando para isso a formulação de Campos Aleatórios Condicionais (CRFs).

A arquitetura utiliza um *Swin Transformer* como espinha dorsal (*backbone*) de extração de características e integra os CRFs diretamente no grafo de aprendizado através de janelas locais (*windows*). A minimização de energia do modelo é regida por potenciais emparelhados (*pairwise*) que impõem uma penalidade matemática para variações de profundidade entre pixels de textura similar. Teoricamente, essa estrutura é otimizada para a preservação de limites estruturais e contornos de objetos sólidos, visto que o modelo é condicionado a forçar a coerência espacial onde detecta fronteiras semânticas na imagem RGB.

A formulação dos CRFs baseia-se na minimização de uma função de energia que equilibra potenciais unários (a estimativa de profundidade local) e potenciais emparelhados

(*pairwise*), que forçam pixels com cores e texturas semelhantes a terem a mesma profundidade:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p(x_i, x_j) \quad (7)$$

O potencial ψ_p atua como um penalizador matemático pesado contra gradientes suaves onde não deveria haver, forçando o modelo a "recortar" objetos do plano de fundo. Em cenários de mundo rígido (como cadeiras e mesas), o NeWCRFs produz mapas de profundidade incrivelmente definidos.

2.4.7 Metric3D V2: Recuperação Geométrica Zero-Shot

O Metric3D V2 (Hu et al., 2024) é uma arquitetura que integra a estimativa de profundidade relativa e métrica absoluta por meio de um protocolo de normalização em um espaço de câmera canônico (*canonical camera space*). A inovação técnica do modelo consiste na capacidade de inferir parâmetros intrínsecos de câmeras arbitrárias — eliminando a ambiguidade de escala decorrente de distâncias focais desconhecidas — e otimizar a reconstrução geométrica utilizando restrições de normais de superfície (*surface normals*) em um ambiente de treinamento *zero-shot* . O problema histórico superado pelo Metric3D é o da "ambiguidade focal": imagens retiradas de câmeras desconhecidas possuem distâncias focais e parâmetros intrínsecos variáveis, o que matematicamente impossibilita deduzir a distância métrica absoluta sem conhecimento prévio do *hardware*.

O modelo resolve isso introduzindo o Modelo de Câmera Canônica (*Canonical Camera Space*). A arquitetura converte todas as imagens de entrada, independente de suas características reais, em um espaço de câmera virtual padronizado durante a fase de pré-processamento. A rede (apoiada por um *Transformer* robusto) treina exclusivamente nesse espaço canônico para inferir a profundidade com rigor geométrico 3D estrito.

A conversão dimensional segue a relação focal:

$$d_{canonical} = d \cdot \frac{f_{canonical}}{f} \quad (8)$$

Onde f é a distância focal estimada da imagem de entrada. O modelo acopla decodificadores de densidade de volume com funções de otimização de normais de superfície (*surface normals*), forçando a rede a construir superfícies planas com vetores normais

matematicamente coerentes. Embora brilhante para reconstrução de engenharia (como digitalizar ruas ou peças mecânicas), a imposição rigorosa de normas de superfície é incompatível com estruturas translúcidas como o fluido.

2.4.8 UniDepth V2: Universalização com Pseudo-Intrínsecos

Paralelo ao desenvolvimento do Metric3D, o UniDepth V2 (Piccinelli et al., 2024) segue a mesma ambição de alcançar uma profundidade métrica universal, mas utilizando um paradigma arquitetural distinto de projeção.

O núcleo central do UniDepth baseia-se na predição conjunta da profundidade e dos parâmetros intrínsecos pseudocâmera (*pseudo-intrinsics*). A rede utiliza mecanismos de auto-atenção massivos não apenas para extrair as *features* da imagem, mas para deduzir o campo de visão da lente (FoV) a partir de pistas de fuga de perspectiva.

O decodificador do UniDepth V2 é desenhado para projetar a profundidade latente e aplicar refinações iterativas baseadas em restrições epipolares aprendidas do *dataset*. Contudo, assim como o Metric3D, a arquitetura do UniDepth opera sob a premissa de que a cena pode ser decomposta em geometrias regidas por perspectivas lineares sólidas. A ausência de linhas de fuga e de planos geométricos definidos nos simuladores de escoamento fluido limita a eficácia do módulo de predição de intrínsecos e da projeção epipolar do modelo.

2.4.9 ViTA: Colapso Analítico da Memória Temporal Rígida

Por fim, o ViTA (*Video Temporal Attention* - 2024) foi concebido para lidar com a estabilidade de vídeos através da inserção de memórias espaço-temporais determinísticas. Enquanto a arquitetura VDA adota uma abordagem mais fluida, o ViTA foca no pareamento (*matching*) explícito e estruturado de características visuais em janelas temporais curtas.

A estrutura do ViTA adiciona um módulo decodificador temporal que tenta rastrear o "mesmo pixel" ou "mesmo objeto" através de diferentes quadros. Isso é feito utilizando uma variação de Fluxo Óptico (*Optical Flow*) implícito, onde a matriz de atenção temporal calcula os custos de deslocamento de padrões texturizados rígidos de $T - 1$ para T.

Para aplicar a suavização (*smoothing*), a função limitadora do ViTA impõe uma pesada barreira matemática contra deslocamentos e desintegrações abruptas no eixo Z. No entanto, o fluxo cinemático em um simulador de CFD (como a formação alternada dos vórtices de Von Kármán no perfil em Diamante) desafia abertamente o conceito de preservação de padrão do ViTA. As partículas se dissolvem, misturam-se e se separam caóticamente, quebrando o algoritmo de correspondência visual do modelo.

3 METODOLOGIA

3.1 Delineamento Experimental e Caracterização da Pesquisa

Esta pesquisa caracteriza-se como um estudo experimental computacional, quantitativo e comparativo, aplicável à avaliação de sistemas de visão computacional. O objetivo metodológico central consiste em submeter modelos de estimativa de profundidade monocular (*State-of-the-Art* - SOTA) a cenários de escoamento de fluidos não-rígidos, gerados por meio de simulações de Dinâmica dos Fluidos Computacional (CFD).

O rigor metodológico desta abordagem fundamenta-se no estrito isolamento de variáveis. Para garantir que as discrepâncias de desempenho sejam causadas exclusivamente pelas limitações ou capacidades matemáticas de cada arquitetura, foram mantidas como variáveis controladas todas as condições de síntese do ambiente virtual: o ângulo de perspectiva da câmera, a escala do cenário, o sistema de iluminação/renderização do volume estocástico e a taxa temporal constante de atualização (30 *frames* por segundo).

Em contrapartida, atuaram como variáveis independentes (alteradas durante os ensaios) as nove arquiteturas de inteligência artificial selecionadas e as quatro geometrias de bloqueio aerodinâmico (cilindro, quadrado, diamante e aerofólio). Dessa forma, isola-se e afere-se com precisão a capacidade interpretativa das redes neurais perante as características inerentes à dinâmica de fluidos — como descontinuidades espaciais, ausência de bordas rígidas e dissipação volumétrica contínua —, diferenciando radicalmente este ensaio das validações convencionais realizadas em ambientes estáticos ou urbanos com geometria geométrica e previsível.

3.2 Geração dos Dados de Entrada e Simulação Fluidodinâmica

Para garantir a geração de escoamentos com alta fidelidade visual e dinâmica turbulenta contínua, os cenários foram projetados e executados utilizando o ambiente computacional *wxRTCFD* (*Real-Time Computational Fluid Dynamics*). Este software implementa solucionadores numéricos otimizados para a integração das equações bidimensionais de Navier-Stokes em tempo real, permitindo a interação direta com as condições de contorno topológicas durante a execução do solver.

A visualização do fluido no ambiente simulado ocorre através da injeção contínua de um campo escalar de densidade passiva (atuando como um fluido traçador ou fumaça virtual). A evolução espacial e temporal da concentração deste traçador mapeia o comportamento

macroscópico do escoamento, evidenciando fenômenos físicos como pontos de estagnação, descolamento de camada limite e formação de vórtices.

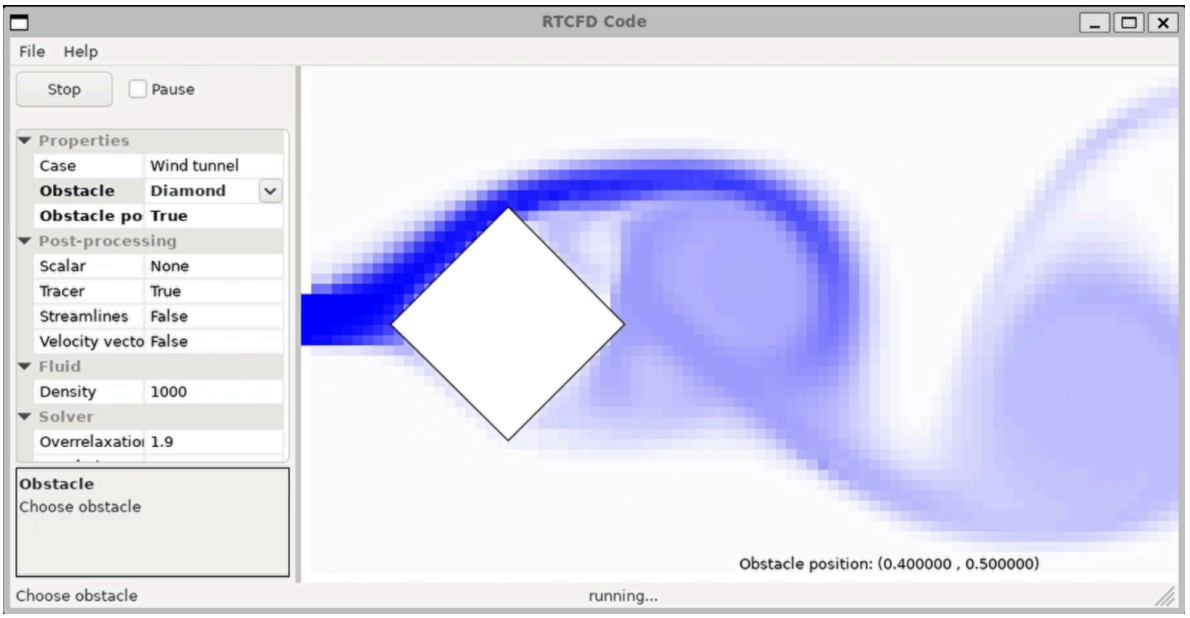


Figura 6 - Interface gráfica do solucionador *wxRTCFD* utilizada para a simulação bidimensional do escoamento e renderização do campo de densidade passiva. Fonte: O autor (2026).

Um diferencial metodológico adotado para estressar a capacidade de rastreamento temporal das arquiteturas de Inteligência Artificial foi a introdução de perturbações dinâmicas nas geometrias de bloqueio. Durante a integração temporal do fluido, os obstáculos (cilindro, quadrado, diamante e aerofólio) foram submetidos a deslocamentos translacionais manuais ao longo dos eixos cartesianos x e y.

Essa perturbação geométrica quebra intencionalmente a simetria do escoamento laminar contínuo, induzindo respostas não-lineares severas no fluido, como a aceleração abrupta do fluxo lateral e a dissipação assimétrica da esteira de vórtices. Esse procedimento assegura que as matrizes de entrada representem um ambiente altamente dinâmico, impossibilitando que as redes neurais dependam de padrões visuais estáticos para a estimativa de profundidade.

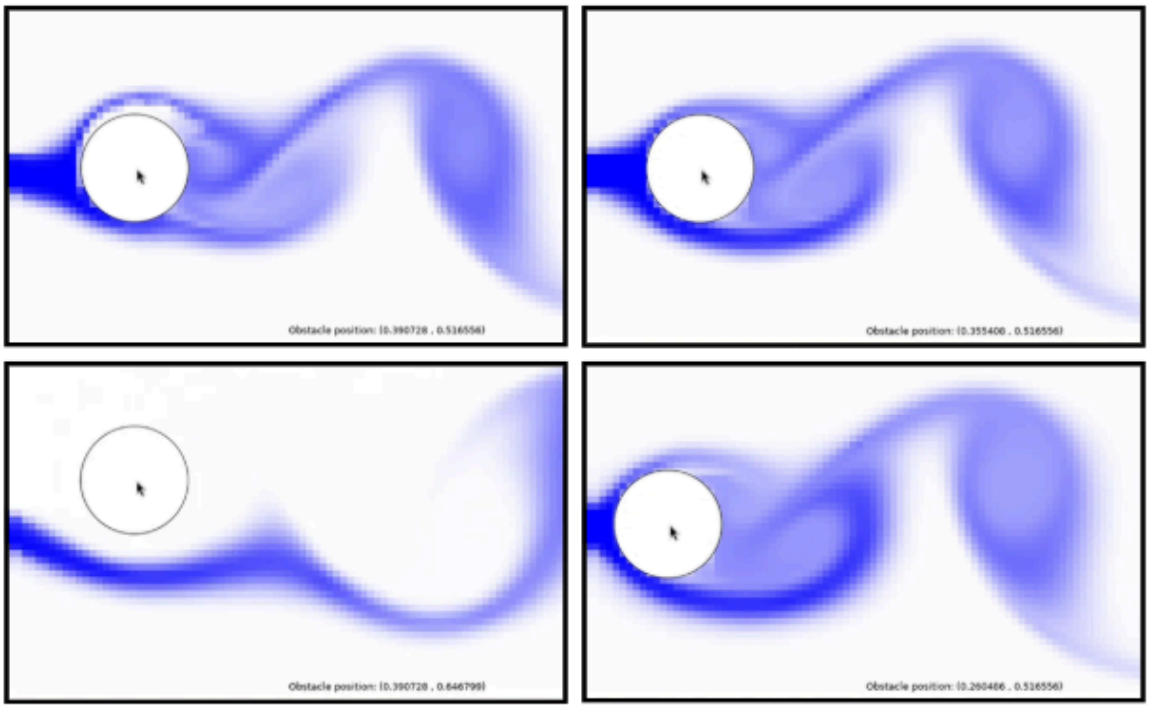


Figura 7 - Evolução cinemática da esteira de turbulência mediante perturbação translacional do obstáculo geométrico nos eixos x e y, evidenciando a resposta não-linear do campo de densidade. Fonte: O autor (2026).

A captura da resposta visual do escoamento foi realizada através da gravação em tempo real da área de renderização (*viewport*) do software. As saídas gráficas brutas foram exportadas na forma de sequências temporais de vídeo, mantendo uma taxa de atualização constante de 30 quadros por segundo (fps). Posteriormente, essas sequências foram segmentadas em matrizes bidimensionais independentes pelo *pipeline* em Python no Google Colab, convertendo as variações de intensidade de cor dos pixels do vídeo na representação da concentração volumétrica do fluido, que serviu como matriz de entrada (RGB) e Verdade Terrestre (*Ground Truth*) visual para as inferências das redes neurais.

3.3 Cenários de Simulação e Geometrias de Bloqueio

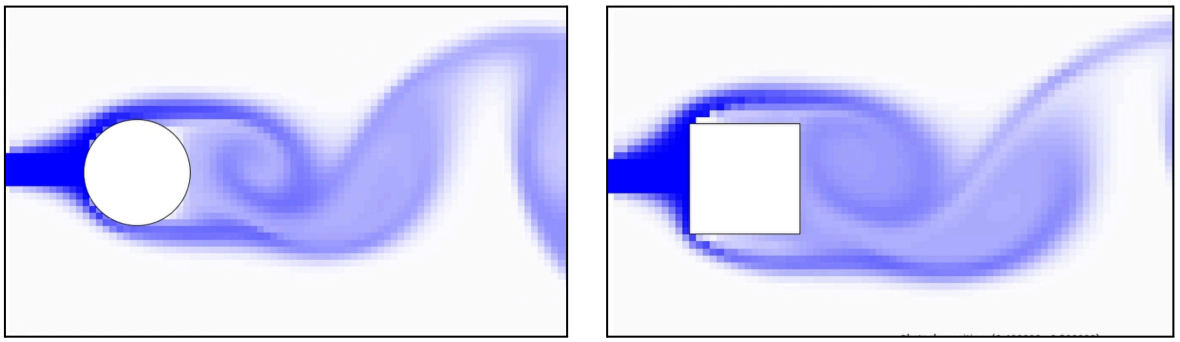
Para avaliar o impacto da aerodinâmica na capacidade de percepção das arquiteturas, o escoamento do fluido foi projetado em torno de quatro barreiras geométricas distintas. Cada geometria induz um padrão físico particular de dissipação, acúmulo e turbulência:

- **Perfil Cilíndrico (Curvatura Suave):** Promove um escoamento laminar inicial com desprendimento de vórtices subsequentes, gerando uma transição suave de densidade volumétrica ao redor do obstáculo.

- **Perfil Quadrado (Bloqueio Plano):** Induz um ponto de estagnação frontal imediato, resultando em um acúmulo denso de fluido na face anterior da barreira e vórtices caóticos nas quinas de noventa graus.
- **Perfil em Diamante (Bifurcação Angular):** Atua como uma cunha geométrica que força a separação do fluido em duas correntes diagonais rápidas, reduzindo a densidade frontal e acelerando o escoamento lateral.
- **Perfil de Aerofólio (Fluxo Sustentado e Esteira):** Direciona o fluido ao longo de uma superfície hidrodinâmica contínua, minimizando a resistência frontal e gerando uma esteira de turbulência complexa (vórtices de Von Kármán) na borda de fuga.

Os dados brutos de entrada consistem em sequências temporais de vídeo com duração exata de 27,3 segundos e taxa de atualização constante de 30 quadros por segundo (*fps*). O cruzamento desse intervalo temporal com a taxa de captura resultou na extração de 819 matrizes bidimensionais independentes por cenário avaliado. Para garantir a validade do delineamento comparativo e a equidade estatística, a exata mesma sequência de 819 quadros foi utilizada como entrada para todos os nove modelos de visão computacional testados, assegurando que as arquiteturas fossem avaliadas sob condições fluidodinâmicas estritamente idênticas.

Para materializar a complexidade visual imposta às redes neurais e ilustrar as características fluidodinâmicas descritas, a representação espacial de cada cenário simulado é apresentada a seguir. A Figura 5 compõe um painel visual consolidado com recortes exatos dos tensores de entrada (imagens RGB) extraídos das simulações, evidenciando as diferenças práticas de densidade, estagnação e formação de esteira que as arquiteturas precisaram processar durante as inferências volumétricas.



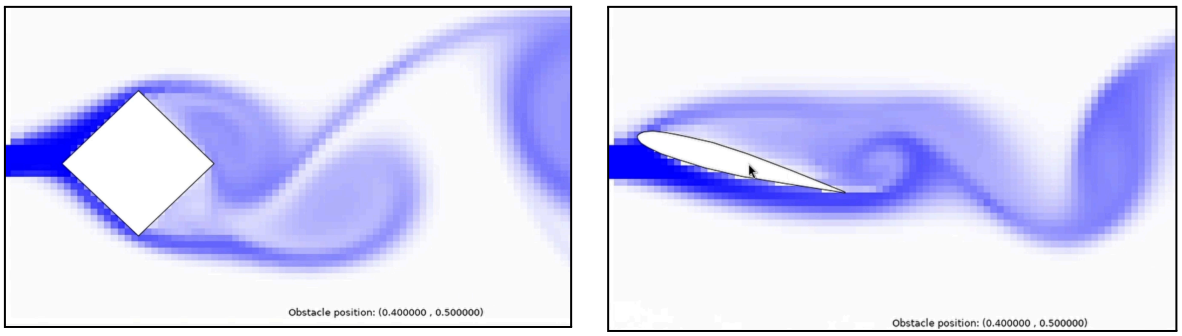


Figura 8 - Painel visual dos cenários de simulação em dinâmica de fluidos utilizados como entrada para os modelos: (a) Cilindro, (b) Quadrado, (c) Diamante e (d) Aerofólio. Fonte: O autor (2026).

3.4 Arquiteturas de Visão Computacional Selecionadas

O ecossistema experimental foi composto por nove modelos representativos de diferentes paradigmas e gerações da visão computacional, permitindo traçar uma análise cronológica e arquitetural do desempenho:

- **MiDaS (2019):** Modelo pioneiro na estimativa de profundidade relativa pura, baseado em redes convolucionais e treinado em múltiplos *datasets* heterogêneos.
- **Monodepth2 (2019):** Arquitetura auto-supervisionada projetada sob a hipótese do mundo estático, que utiliza o movimento da câmera para inferir descontinuidade espacial.
- **NeWCRFs (2022):** Estrutura que incorpora Campos Aleatórios Condicionais (*Neural Conditional Random Fields*) com foco na preservação e rigidez de bordas nítidas na cena.
- **ZoeDepth (2023):** Arquitetura de transição que introduz o conceito de caixas adaptativas (*bins*), combinando premissas de profundidade relativa e métrica através de *Vision Transformers* (ViTs).
- **Metric3D V2 (2024):** Modelo fundacional que adota uma abordagem de geometria canônica para mapeamento estrito de profundidade métrica absoluta.
- **UniDepth V2 (2024):** Modelo universal capaz de estimar, simultaneamente, os parâmetros intrínsecos da câmera e a profundidade da cena a partir de uma única imagem.
- **ViTA (2024):** Variante focada em consistência temporal de curto prazo, desenhada para aplicar restrições de memória entre frames consecutivos.

- **Video Depth Anything - VDA (2024):** Modelo fundacional espaço-temporal projetado especificamente para vídeos, que emprega mecanismos de atenção temporal densa para garantir a estabilidade do fluxo cinemático.
- **Marigold (2024):** Modelo de vanguarda fundamentado em Modelos de Difusão (*Latent Diffusion Models*), que aborda a inferência de profundidade como uma tarefa generativa baseada em texturas e densidades visuais.

3.5 Pipeline de Processamento e Extração de Dados

A execução experimental foi conduzida no ambiente de nuvem Google Colaboratory (*Colab*), utilizando a linguagem Python 3. Visando garantir a reprodutibilidade do experimento e o processamento eficiente das arquiteturas de aprendizado profundo, o *pipeline* de inferência foi executado sob aceleração de hardware utilizando uma unidade de processamento gráfico (GPU) NVIDIA T4. O ambiente de execução operou suportado pelas instâncias de processamento padrão da plataforma, contando com CPU provisionada pelo servidor e aproximadamente 12 GB de memória RAM alocada para a sessão. O *pipeline* de inferência foi estruturado para isolar a saída matemática bruta dos modelos, evitando mascarar suas reais capacidades ou limitações frente à dinâmica de fluidos.

Para cada cenário de simulação, os quadros do vídeo foram extraídos sequencialmente e convertidos para o espaço de cor RGB, requisito de entrada padrão para a maioria dos codificadores (*encoders*) baseados em *Vision Transformers* (ViTs) e Redes Neurais Convolucionais (CNNs). A saída de cada arquitetura não foi tratada como uma imagem convencional, mas sim capturada como uma matriz matemática bidimensional de canal único, salva em formato binário otimizado (.npy). Nestas matrizes, os valores flutuantes foram normalizados no intervalo estrito de 0 a 1, onde valores próximos à unidade representam proximidade espacial com a lente da câmera, e valores próximos a zero indicam o plano de fundo ou distâncias infinitas.

Como decisão metodológica estrita para garantir a integridade da análise de instabilidade, nenhum filtro de suavização espacial ou temporal (como médias móveis, filtros Gaussianos ou Bilaterais) foi aplicado sobre os dados extraídos. Essa premissa assegura que a quantificação do *flickering* reflita exclusivamente o comportamento nativo da arquitetura avaliada.

Exclusivamente para a etapa de análise qualitativa e visualização do mapa de calor, as matrizes normalizadas foram mapeadas cromaticamente utilizando a tabela de

correspondência perceptualmente uniforme *Inferno*. Este mapeamento não altera os tensores originais submetidos à análise estatística, servindo apenas para transpor o erro da rede neural para uma escala visual compreensível ao olho humano em formato de vídeo (.mp4).

3.6 Métricas de Validação Estatística

Para quantificar a eficácia dos modelos na interpretação da física de fluidos, o estudo baseou-se em dois eixos avaliativos ortogonais: a precisão volumétrica e a coerência cinemática.

3.6.1 Precisão Espacial (Correlação de Pearson)

A capacidade das arquiteturas em capturar a variação de densidade e a morfologia do escoamento fluido é quantificada por meio do Coeficiente de Correlação de Pearson (r). Diferentemente das métricas estritas amplamente adotadas na literatura de estimativa de profundidade monocular — como o Erro Quadrático Médio (RMSE) e o Erro Relativo Absoluto (AbsRel), que penalizam severamente a ausência de escala métrica e pressupõem a existência de distâncias focais e superfícies sólidas —, a métrica de Pearson mostra-se teoricamente mais adequada para a dinâmica de fluidos. Como as nuvens aerodinâmicas são constituídas por gradientes volumétricos contínuos e muitos dos modelos avaliados (como MiDaS e Depth Anything) operam em espaços de profundidade relativa, o coeficiente de correlação permite mensurar a aderência estrutural e a proporcionalidade dos gradientes inferidos em relação à referência, isolando a avaliação geométrica da necessidade de uma calibração escalar absoluta. O coeficiente avalia a relação linear entre a profundidade inferida e o comportamento esperado no primeiro plano, sendo definido matematicamente como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}} \tag{7}$$

Onde x_i e y_i representam os valores dos pixels correspondentes entre a estimativa e o referencial de aproximação temporal, enquanto $\bar{x}$ e $\bar{y}$ denotam as médias de seus respectivos conjuntos. Valores positivos e próximos a 1 indicam que a arquitetura compreendeu a morfologia do fluido como um objeto no espaço tridimensional. Valores próximos a 0 sugerem aleatoriedade, enquanto valores negativos evidenciam falha crítica de percepção,

onde o modelo sofre uma inversão ótica e confunde a dissipação do fluido com o distanciamento do cenário (plano de fundo).

3.6.2 Inconsistência Temporal (*Flickering*)

Devido à natureza não-rígida dos fluidos dinâmicos, muitas arquiteturas sofrem oscilações drásticas em suas estimativas ao longo de quadros consecutivos. Para aferir essa instabilidade, quantificou-se a Inconsistência Temporal (IT), calculada através da variação absoluta da profundidade de um dado ponto espacial p entre instantes de tempo adjacentes:

$$IT = \frac{1}{N-1} \sum_{t=1}^{N-1} |D_t(p) - D_{t-1}(p)| \tag{8}$$

Onde D_t representa o mapa de profundidade no instante t, e N é o número total de quadros da sequência de vídeo. Uma inconsistência próxima a 0 atesta que a rede neural consegue rastrear o fluido de maneira estável, mantendo a coerência de curto prazo na geometria da cena, fator crucial em fluxos aerodinâmicos e turbulências continuadas.

4 RESULTADOS E DISCUSSÃO

4.1 O Impacto da Geometria Aerodinâmica na Percepção Espacial

A análise comparativa do desempenho dos modelos frente a diferentes perfis aerodinâmicos revelou que a topologia do obstáculo é um fator determinante para a acurácia da estimativa de profundidade monocular. Os resultados demonstram que as arquiteturas de visão computacional reagem de maneira não linear às mudanças na densidade volumétrica do fluido, categorizando-se claramente entre modelos dependentes de acúmulo estático e modelos de rastreamento de fluxo contínuo.

Nos cenários de **Perfil Cilíndrico** e **Perfil Quadrado**, a dinâmica do fluido é caracterizada por um ponto de estagnação frontal acentuado. O impacto direto do escoamento contra a barreira física (especialmente a face plana do quadrado) gera uma área de alta densidade volumétrica e acúmulo de partículas antes da formação da esteira de turbulência. Nestas condições de bloqueio severo, o modelo **Marigold** apresentou superioridade estatística absoluta, registrando a maior Correlação de Pearson da pesquisa no cilindro (**0.736**) e mantendo a liderança no quadrado (**0.538**). Este desempenho valida a hipótese de que modelos baseados em difusão (*Latent Diffusion Models*), originalmente treinados para a

geração de imagens, possuem uma capacidade intrínseca superior para interpretar texturas densas e volumes dissipados como superfícies tridimensionais coesas. O modelo **ZoeDepth**, utilizando sua abordagem de agrupamento em caixas adaptativas (*bins*), também demonstrou resiliência notável nestas geometrias, alcançando uma correlação de **0.393** no quadrado.

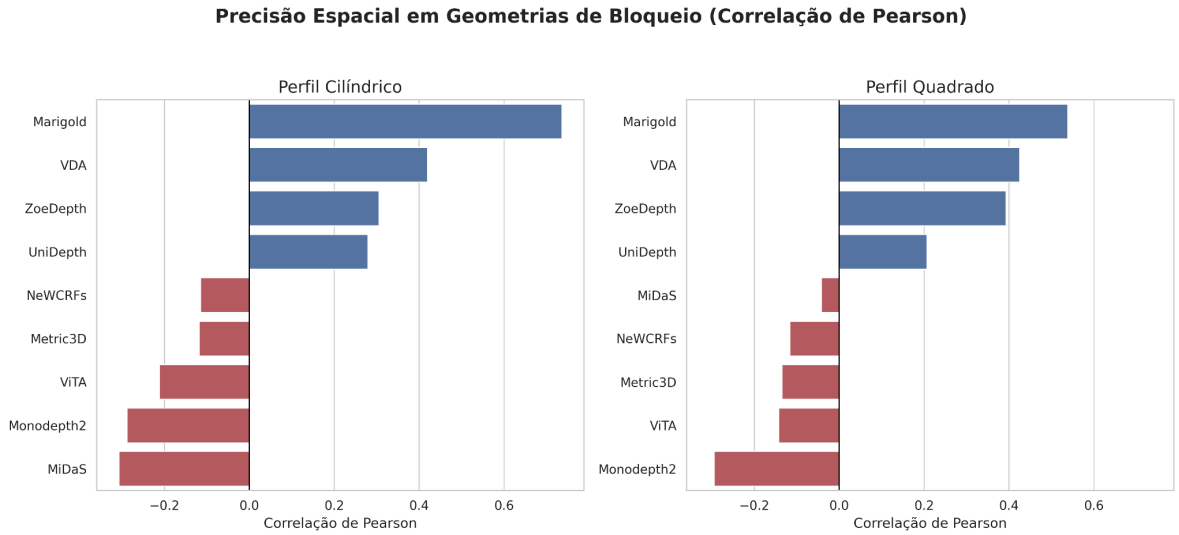


Figura 9 - Comparativo da Correlação de Pearson em geometrias de bloqueio frontal (Cilindro e Quadrado), evidenciando a liderança do modelo de difusão Marigold. Fonte: O autor (2026).

Por outro lado, os cenários de **Perfil em Diamante** e **Perfil de Aerofólio** alteram radicalmente a mecânica do fluido. O diamante impõe uma bifurcação angular severa, dividindo o escoamento em duas correntes rápidas, enquanto o aerofólio direciona o fluido ao longo de sua superfície, criando uma esteira fina e dinâmica (vórtices de Von Kármán) na borda de fuga. Ao eliminar o acúmulo frontal denso, a eficácia do Marigold despencou drasticamente para **0.169** e **0.237**, respectivamente. A ausência de uma "parede" de fluido volumétrica privou o modelo generativo das âncoras de textura necessárias para sua inferência.

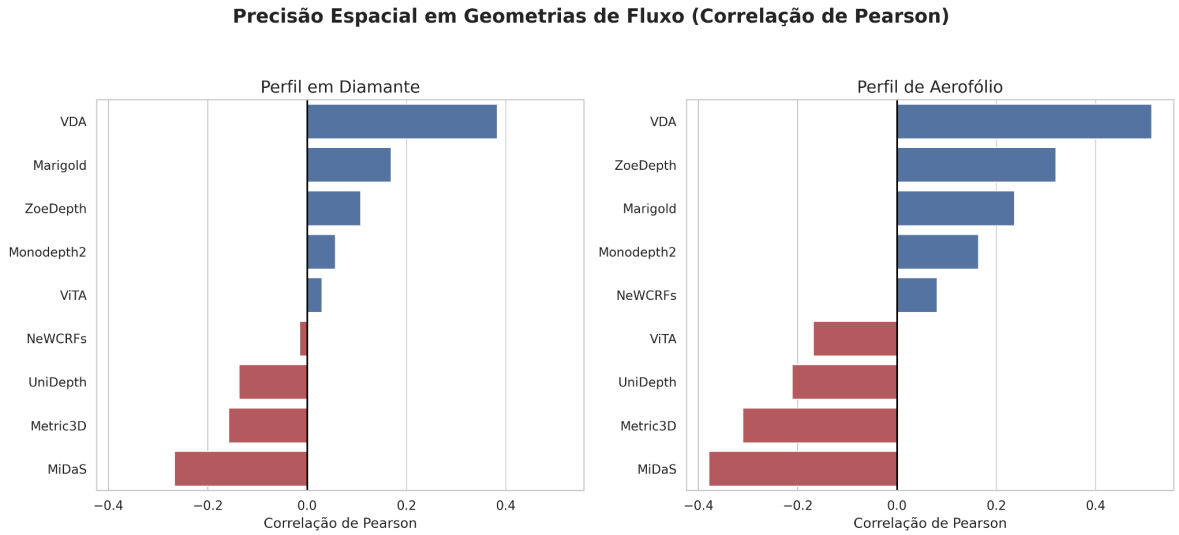


Figura 10 - Comparativo da Correlação de Pearson em geometrias de fluxo (Diamante e Aerofólio), demonstrando a superioridade da arquitetura VDA no rastreamento de esteiras cinemáticas. Fonte: O autor (2026).

Foi exatamente sob as condições de fluxo bifurcado e esteiras dinâmicas que a arquitetura **Video Depth Anything (VDA)** assumiu a liderança incontestável. Projetado como um modelo fundacional espaço-temporal, o VDA registrou coeficientes de correlação de **0.383** no diamante e **0.513** no aerofólio. Este resultado evidencia que, em cenários onde o fluido não se acumula, mas sim se divide e flui rapidamente, a capacidade da rede neural de aplicar mecanismos de atenção na dimensão temporal (*temporal attention*) sobrepõe-se à análise de densidade estática. O VDA conseguiu manter a percepção tridimensional ao "prever" a trajetória cinemática do escoamento com base no histórico dos quadros anteriores, preenchendo as lacunas de densidade com informação de movimento.

A disparidade de resultados entre as geometrias de bloqueio (cilindro e quadrado) e as geometrias de fluxo (diamante e aerofólio) comprova que não existe um modelo universal perfeito para a mecânica dos fluidos. A seleção da arquitetura de inteligência artificial deve estar inexoravelmente condicionada à aerodinâmica do experimento: obstáculos que geram acúmulo e densidade favorecem o paradigma da difusão, enquanto obstáculos que induzem separação de fluxo exigem redes neurais com memória de vídeo e rastreamento cinemático.

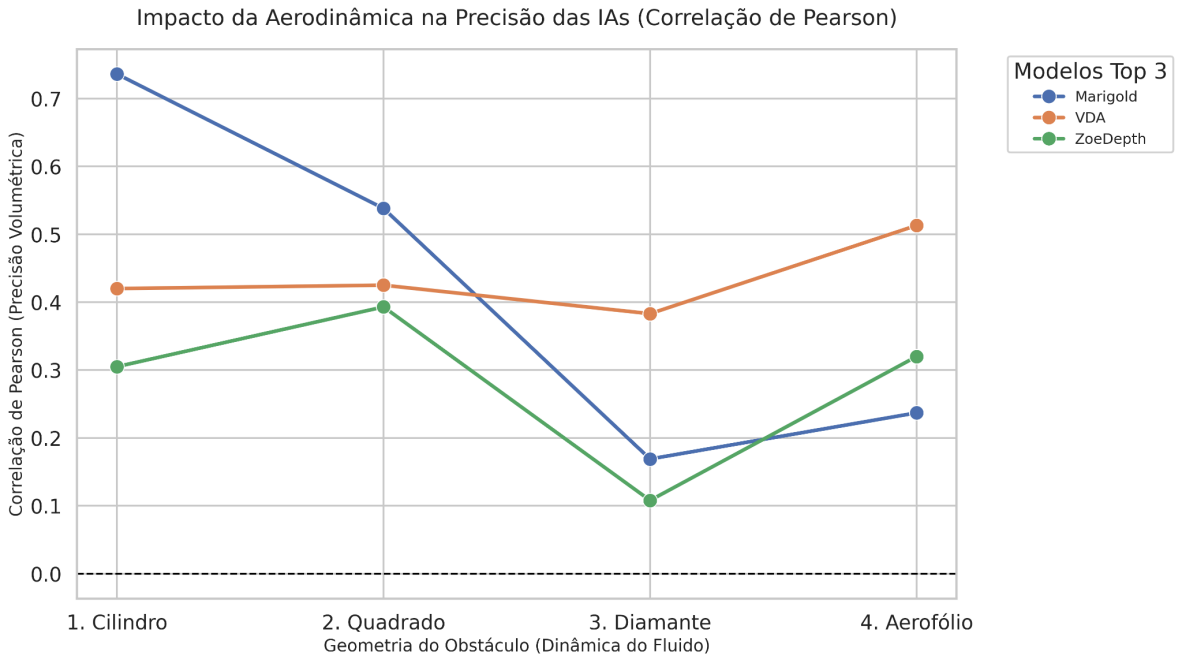


Figura 11 - Evolução da precisão espacial (Pearson) dos principais modelos através das quatro geometrias avaliadas. Fonte: O autor (2026).

4.2 O *Trade-off* da Visão em Fluidos: Estabilidade Temporal versus Acurácia Morfológica

A avaliação simultânea das métricas de Correlação de Pearson e de Inconsistência Temporal revelou um comportamento técnico que este estudo define como o *trade-off* na percepção computacional de fluidos. Ao contrário de ambientes urbanos ou de navegação estática, onde os modelos estado-da-arte (SOTA) conseguem otimizar ambas as métricas simultaneamente, os escoamentos em Dinâmica dos Fluidos Computacional (CFD) forçam as redes neurais a um compromisso (*trade-off*) matemático explícito: priorizar a morfologia estrutural da pluma aerodinâmica ou garantir a coerência cinemática entre os quadros consecutivos.

Ao projetar o desempenho das arquiteturas em um espaço bidimensional — onde o eixo das ordenadas representa a precisão espacial (Pearson) e o eixo das abscissas denota a inconsistência temporal em escala logarítmica —, a dicotomia arquitetural torna-se evidente.

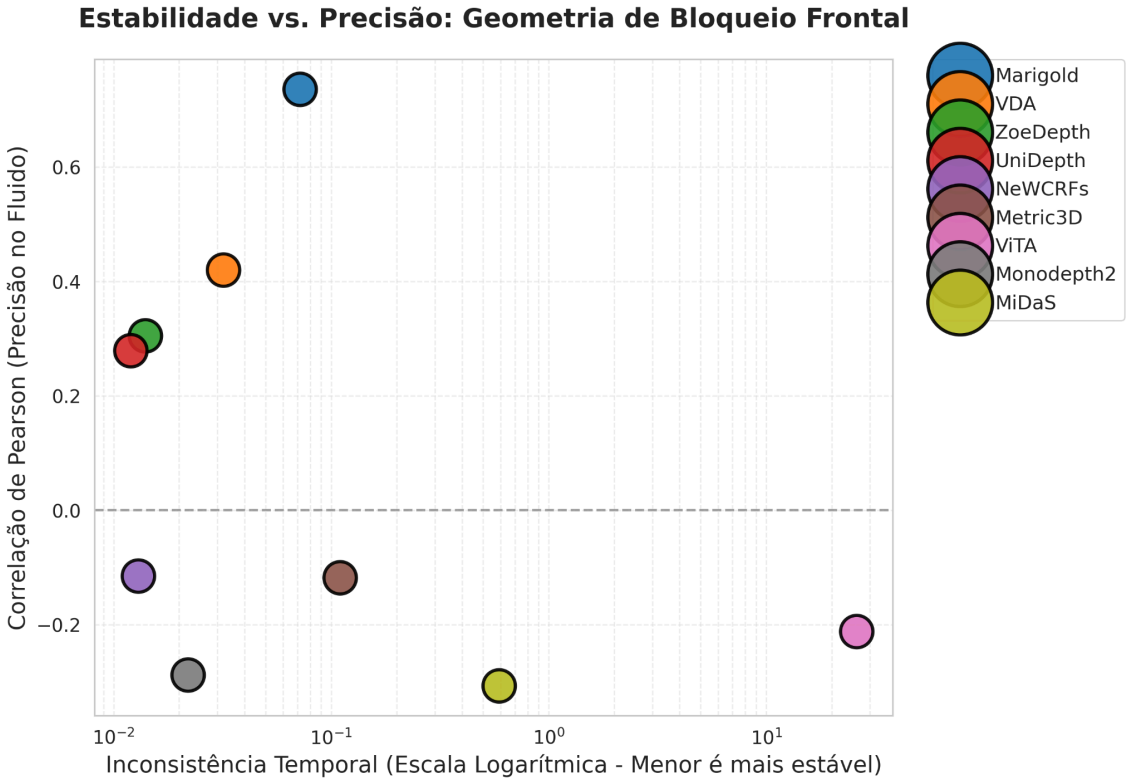


Figura 12 - Dispersão de desempenho evidenciando o compromisso entre precisão espacial e estabilidade temporal. Modelos ideais situam-se no quadrante superior esquerdo. Fonte: O autor (2026).

O modelo Marigold domina o estrato superior de acurácia espacial em cenários de alta densidade; entretanto, apresenta leves penalidades no eixo temporal. Esse fenômeno (*flickering* de baixa amplitude) é intrínseco à natureza estocástica dos Modelos de Difusão, que reconstróem o mapa de profundidade iterativamente a partir do ruído latente a cada quadro isolado, sem herdar a coerência exata do instante temporal anterior.

Em contrapartida, arquiteturas como Video Depth Anything (VDA) e ZoeDepth tendem a se agrupar no extremo esquerdo do espectro (alta estabilidade temporal, na casa de 0.015 a 0.030). Esses modelos sacrificam a percepção de variações milimétricas na densidade volumétrica para garantir que a esteira do fluido não sofra transições abruptas, entregando um rastreamento cinemático contínuo.

O achado crítico desta seção, contudo, reside no colapso arquitetural sistêmico do modelo ViTA. Projetado especificamente para impor restrições de memória de curto prazo e garantir consistência em vídeos, o ViTA falhou de forma severa ao ser confrontado com dinâmicas não-rígidas.

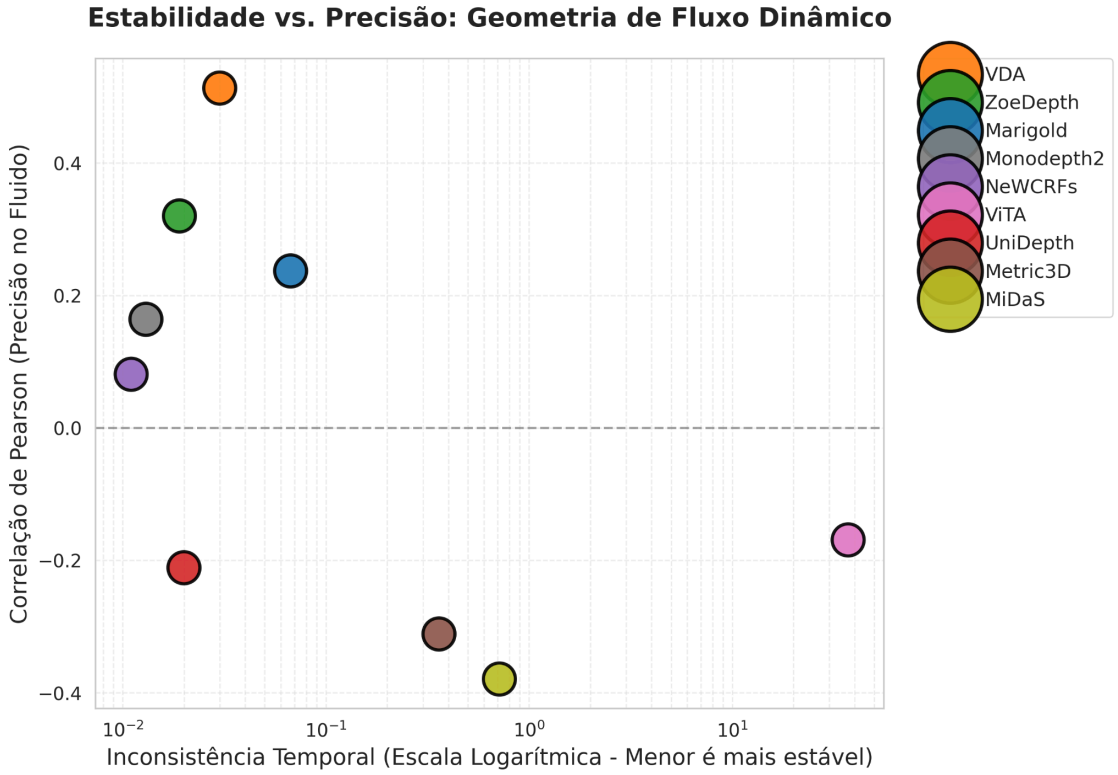


Figura 13 - Colapso temporal do modelo ViTA frente a esteiras aerodinâmicas, registrando os piores índices de inconsistência temporal da análise. Fonte: O autor (2026).

Nas geometrias de Diamante e Aerofólio, onde o escoamento sofre desprendimentos rápidos e forma vórtices de Von Kármán alternados, o módulo de atenção temporal do ViTA tentou, erroneamente, reconciliar a posição anterior do fluido com sua nova topologia instantânea. O resultado foi um *flickering* extremo, ultrapassando os índices de 36.0 e 37.0 de inconsistência temporal. Este desvio estatístico atesta que mecanismos de memória treinados estritamente para oclusões de objetos sólidos e rígidos entram em pane matemática ao tentar processar a transparência e a dissipação iterativa de um fluido em alta velocidade.

4.3 A Evolução Arquitetural Frente à Dinâmica de Escoamentos Amorfos

A cronologia das arquiteturas testadas neste estudo revela que a incapacidade de estimar a profundidade de fluidos não é uma falha aleatória, mas um viés de projeto intrínseco aos paradigmas clássicos da visão computacional. Modelos desenvolvidos entre 2019 e 2022 foram concebidos sob a hipótese do mundo rígido, sendo majoritariamente treinados em *datasets* urbanos e de navegação autônoma. Nestes cenários, a identificação de profundidade depende de indícios geométricos fortes: oclusões nítidas, linhas de fuga de perspectiva e limites duros entre objetos e o plano de fundo.

O escoamento de um fluido, por sua natureza dissipativa e translúcida, viola todas essas premissas geométricas. A ausência de bordas estruturais causa um colapso interpretativo nas Redes Neurais Convolucionais (CNNs) mais antigas.

Ao analisar os dados consolidados, constata-se que arquiteturas como MiDaS e Monodepth2 não apenas falharam em identificar o volume, mas incorreram em um erro sistemático de inversão de profundidade. Com coeficientes de Pearson consistentemente negativos (alcançando -0.307 no cilindro para o MiDaS e -0.294 no quadrado para o Monodepth2), esses modelos interpretaram o gradiente suave do fluido como um "buraco" ou plano de fundo distante, enquanto projetaram o espaço vazio do simulador para o primeiro plano.

A tentativa de forçar a rigidez geométrica em ambientes fluidos provou-se ineficaz mesmo com o avanço para módulos de otimização mais sofisticados. O modelo NeWCRFs (2022), cuja arquitetura utiliza Campos Aleatórios Condicionais (*Conditional Random Fields*) especificamente para preservar a nitidez e a integridade de bordas, foi penalizado por sua própria especialização. Ao procurar fronteiras duras que não existem no escoamento, o modelo falhou em capturar a nuvem volumétrica, resultando em correlações negativas nos perfis de bloqueio.

A ruptura desse viés geométrico e a consequente ascensão dos índices de acurácia só ocorreram com os modelos fundacionais de 2024. A evolução não se deu apenas pelo aumento da capacidade computacional, mas pela mudança do paradigma interpretativo.

O Marigold contornou a ausência de bordas ao tratar o problema de profundidade como uma tarefa de geração latente (Modelos de Difusão), utilizando a textura e a opacidade do fluido como indícios de densidade tridimensional. Simultaneamente, o Video Depth Anything (VDA) superou a dependência espacial aplicando a Atenção Temporal (*Temporal Attention*). Ao analisar fluxos, o VDA não precisa que o fluido possua bordas estáticas em um quadro isolado; ele deduz a estrutura tridimensional a partir do campo de movimento e da trajetória cinemática das partículas ao longo do tempo.

Conclui-se, portanto, que a percepção computacional da mecânica dos fluidos exige arquiteturas emancipadas da detecção clássica de bordas, dependendo fundamentalmente de abordagens que priorizem o rastreamento cinemático contínuo ou a reconstrução generativa de densidades volumétricas.

4.4 Análise Qualitativa dos Mapas de Profundidade

Embora a análise rigorosa das matrizes bidimensionais e a extração de métricas estatísticas forneçam o embasamento quantitativo necessário para ranquear o desempenho das arquiteturas, a inspeção qualitativa dos mapas de profundidade gerados é cientificamente indispensável. A representação visual das inferências permite identificar anomalias estruturais e comportamentos algorítmicos que os valores numéricos isolados podem mascarar, tais como a fragmentação da esteira aerodinâmica, o borramento de fronteiras e o crítico fenômeno de inversão espacial do plano.

Para ilustrar o impacto prático dos diferentes paradigmas matemáticos na percepção dos fluidos, esta seção apresenta uma avaliação visual comparativa entre quatro modelos representativos das arquiteturas estudadas: **Marigold** (Difusão Latente), **Video Depth Anything** (Atenção Espaço-Temporal), **MiDaS** (CNN Clássica) e **ZoeDepth** (ViT Híbrido).

A fim de isolar a variável topológica, a análise visual foi desmembrada em quatro recortes independentes, correspondentes a cada geometria de bloqueio simulada. Em todos os mapas de calor gerados, a paleta de cores reflete o gradiente de profundidade Z estimado: cores quentes (amarelo/laranja) indicam maior proximidade do observador, enquanto cores frias (roxo/azul) representam distanciamento escalar ou o plano de fundo.

Perfil Cilíndrico (Curvatura Contínua)

Neste cenário, o fluido contorna a superfície curva, gerando um acúmulo moderado na face frontal e uma esteira oscilatória traseira. A Figura 11 demonstra qualitativamente como cada rede neural interpreta essa transição suave de densidade. Crucialmente, essas representações visuais materializam e justificam os resultados quantitativos discutidos anteriormente. A instabilidade geométrica e as distorções espaciais observadas nos mapas de profundidade gerados pelo MiDaS, por exemplo, explicam fisicamente as baixas pontuações de correlação de Pearson e a alta taxa de inconsistência temporal. Em contrapartida, as arquiteturas que exibem um gradiente cromático contínuo e repousado na Figura 11, notadamente o VDA, ilustram a robustez algorítmica que lhes garantiu os melhores índices de desempenho, provando a capacidade dessas redes em manter a coerência espacial sem apresentar pane interpretativa frente à difusão gradual do vórtice.

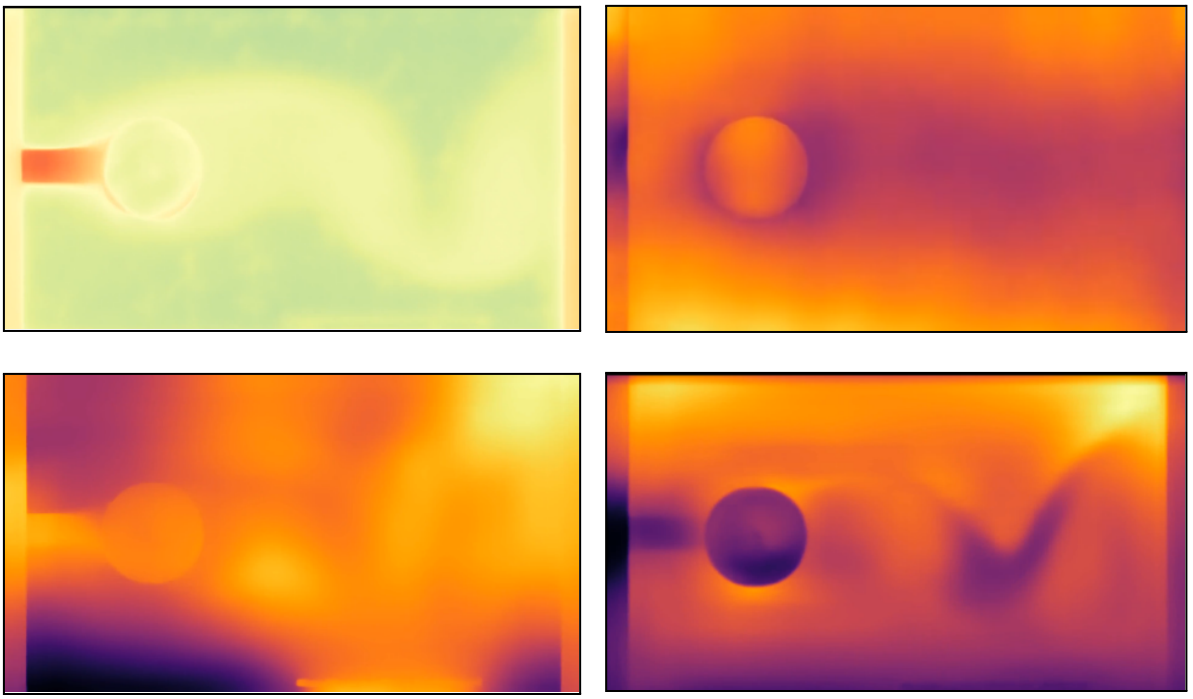
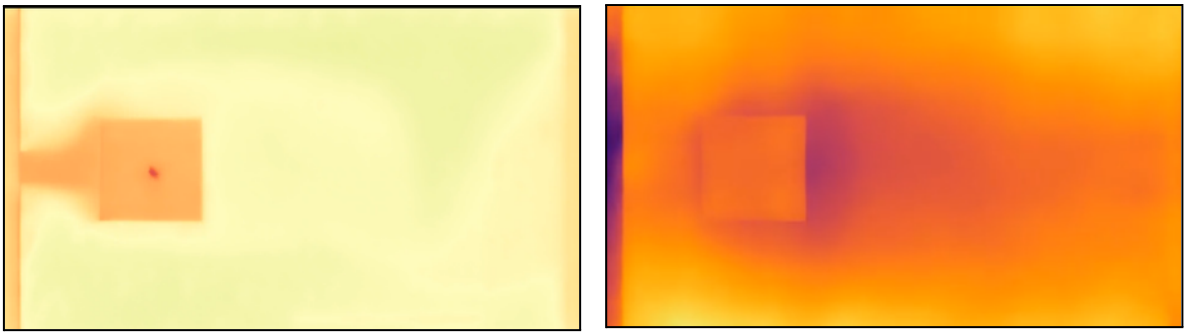


Figura 14 - Análise qualitativa dos mapas de profundidade para o escoamento no Perfil Cilíndrico. Comparativo entre Marigold, Video Depth Anything, MiDaS e ZoeDepth. Fonte: O autor (2026).

Perfil Quadrado (Bloqueio Plano e Estagnação)

A face plana oblitera o escoamento frontal, forçando uma zona de alta densidade e estagnação imediata. A Figura 12 ilustra o comportamento dos algoritmos perante essa "parede" amorfa de partículas, destacando o contraste visual gritante entre o sucesso da arquitetura generativa em mapear nuvens densas e a falha geométrica dos modelos clássicos que não encontram linhas de perspectiva para ancorar o volume.



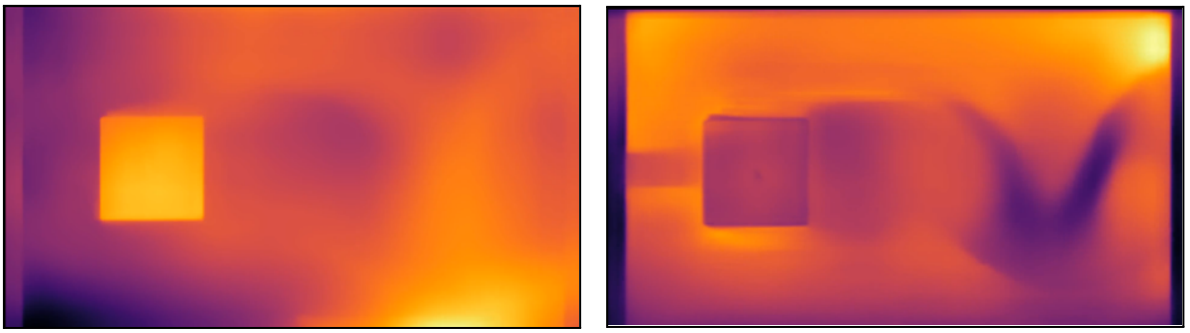


Figura 15 - Análise qualitativa dos mapas de profundidade para o escoamento no Perfil Quadrado. Nota-se o comportamento das redes perante o acúmulo denso frontal. Fonte: O autor (2026).

Perfil em Diamante (Bifurcação Angular)

A geometria afiada divide o fluxo em duas correntes rápidas e diagonais, minimizando drasticamente o acúmulo frontal. O painel da Figura 13 expõe a dificuldade arquitetural de rastrear o volume quando o fluido deforma-se e desloca-se em alta velocidade pelas laterais, testando de forma severa os mecanismos de atenção temporal e a estabilidade cinemática das predições.

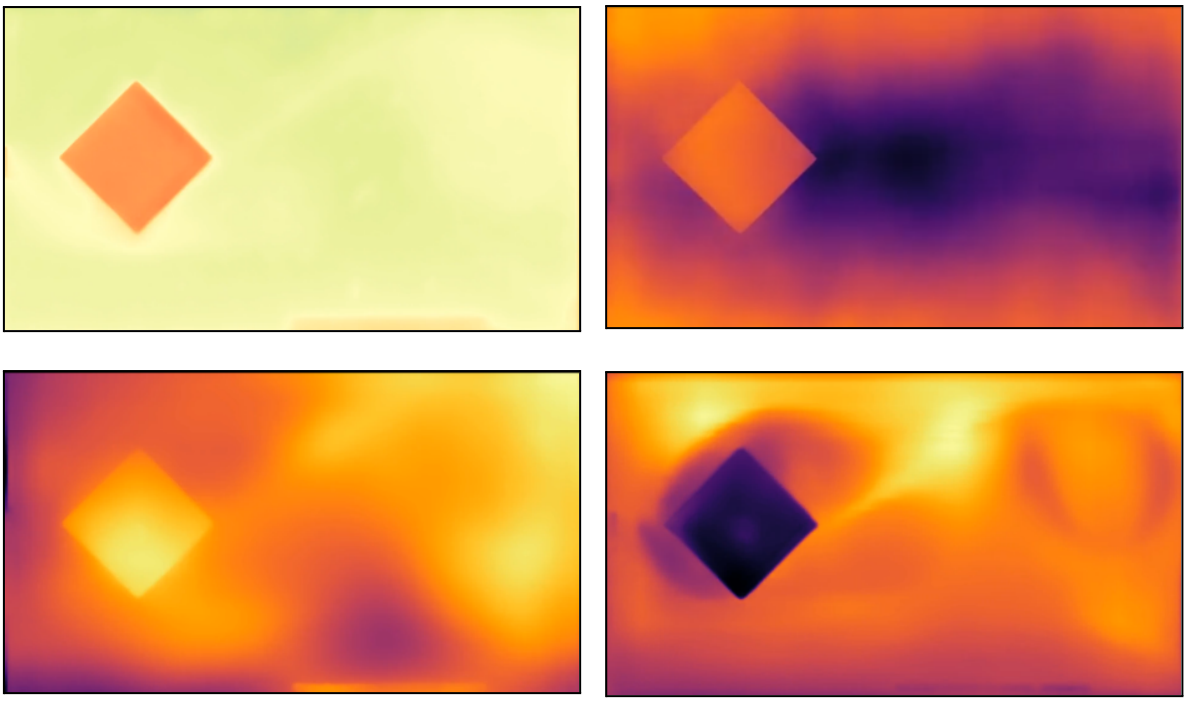


Figura 16 - Análise qualitativa dos mapas de profundidade para o escoamento no Perfil em Diamante, evidenciando o rastreamento das correntes diagonais de alta velocidade. Fonte: O autor (2026).

Perfil de Aerofólio (Esteira de Turbulência)

Com um perfil hidrodinâmico otimizado, o fluido mantém-se adjacente à superfície e dissipa-se em uma esteira alongada e translúcida na borda de fuga. A Figura 14 revela o limite extremo da estimativa de profundidade monocular contemporânea, expondo o colapso estrutural das redes que dependem de bordas sólidas perante gradientes de fumaça quase transparentes.

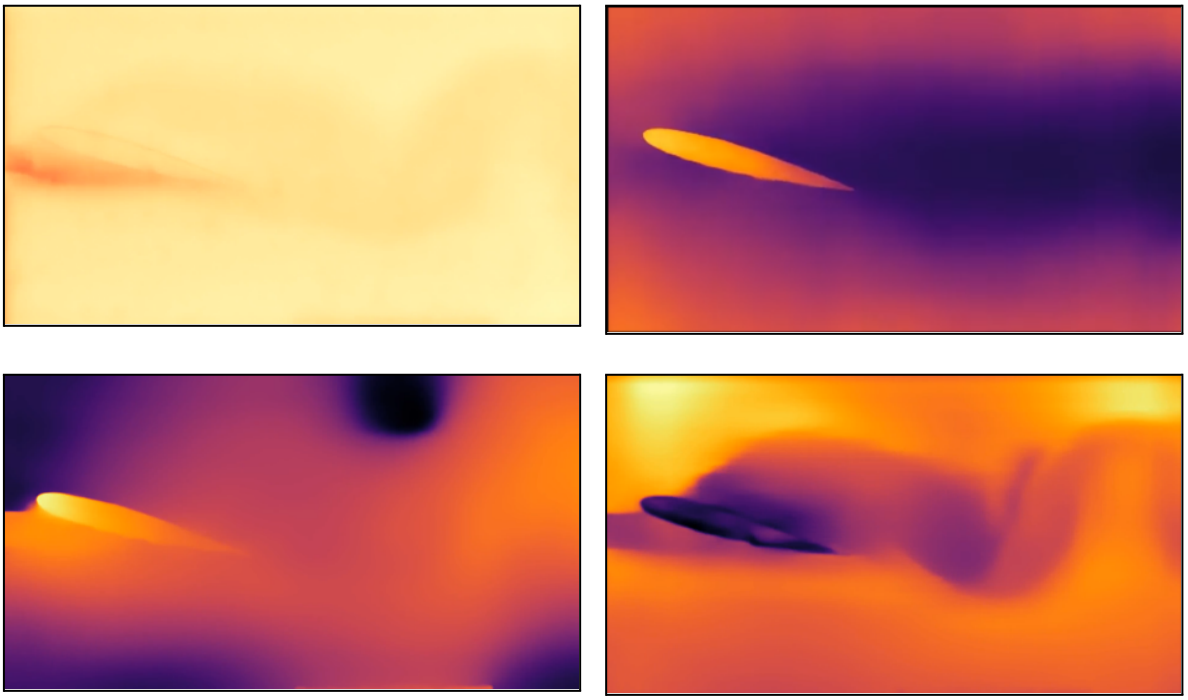


Figura 17 - Análise qualitativa dos mapas de profundidade para o escoamento no Perfil de Aerofólio. Destaque para a leitura da esteira dissipativa na borda de fuga. Fonte: O autor (2026).

Ao observar a sequência de projeções, torna-se evidente o peso do viés de treinamento de cada rede neural. Modelos estruturados sob a hipótese do mundo rígido, como o MiDaS, demonstram clara incapacidade de ancorar a profundidade na área de dissipação translúcida, gerando blocos espaciais ruidosos ou invertendo a fumaça para o plano de fundo. Em contrapartida, as saídas visuais provenientes da arquitetura generativa (Marigold) e de atenção temporal (VDA) comprovam uma capacidade superior de traduzir a variação de densidade do escoamento CFD em uma transição espacial estruturalmente coesa.

5 CONCLUSÃO

5.1 Síntese dos Resultados

Este trabalho teve como objetivo central avaliar o desempenho de modelos estado-da-arte de estimativa de profundidade monocular quando submetidos a cenários dinâmicos e não-rígidos, especificamente em escoamentos de fluidos. A pesquisa partiu da premissa de que a dependência histórica das arquiteturas de visão computacional por bordas nítidas e geometrias rígidas consistia em um gargalo limitante para a percepção tridimensional de elementos translúcidos e amorfos. Os resultados obtidos por meio do cruzamento da Correlação de Pearson com o índice de Inconsistência Temporal confirmaram não apenas a existência desse gargalo, mas delinearam o comportamento paradoxal das inteligências artificiais contemporâneas perante a aerodinâmica.

A análise demonstrou categoricamente o colapso dos modelos clássicos treinados sob a hipótese do mundo rígido. Arquiteturas como MiDaS, Monodepth2 e NeWCRFs falharam na identificação do volume do fluido, resultando em inversões matemáticas severas da percepção espacial. A ausência de linhas de fuga e limites ortogonais causou uma pane interpretativa nessas redes, evidenciando que a tentativa de impor restrições geométricas a ambientes fluidos é uma abordagem arquiteturalmente defasada.

A superação desse viés ocorreu de forma bipartida, dependendo diretamente do perfil aerodinâmico do obstáculo. Constatou-se que modelos baseados em difusão (*Latent Diffusion Models*), representados pelo Marigold, são a escolha ideal para geometrias de bloqueio frontal (como cilindros e quadrados). Nesses cenários, onde há estagnação e acúmulo denso de volume, a difusão reconstrói a profundidade com acurácia incomparável. Entretanto, quando o fluido é submetido a geometrias de bifurcação e fluxo rápido (como diamantes e aerofólios), os modelos generativos perdem sua referência estática. Para essas esteiras dinâmicas, os modelos fundacionais espaço-temporais, como o Video Depth Anything (VDA), provaram-se superiores, utilizando mecanismos de atenção contínua para mapear o volume através de sua trajetória cinemática, evitando o grave erro de *flickering* observado em arquiteturas de memória curta como o ViTA.

Em suma, a pesquisa conclui que o estado-da-arte atual ainda enfrenta um *trade-off* inegável em escoamentos de fluidos: a escolha entre a máxima precisão morfológica estática (via difusão) e a máxima coerência cinemática (via rastreamento de vídeo). A aplicação de modelos de visão profunda em simulações aerodinâmicas exige, portanto, uma seleção arquitetural rigorosamente alinhada à dinâmica física esperada para o experimento.

5.2 Limitações do Estudo

Apesar do rigor adotado no isolamento das variáveis e na extração direta dos dados, o presente estudo possui limitações inerentes à sua modelagem e ao seu escopo. Primeiramente, é imprescindível explicitar que os resultados quantitativos e qualitativos apresentados foram obtidos exclusivamente a partir de simulações de Dinâmica dos Fluidos Computacional (CFD) específicas, não havendo validação experimental empírica. As conclusões, portanto, são fundamentadas sobre dados originados dessas simulações bidimensionais projetadas no espaço, que, embora obedeçam rigorosamente às equações de Navier-Stokes quanto à pressão e velocidade, não reproduzem a complexidade óptica completa de fluidos reais no espectro visível, como o espalhamento de luz (Rayleigh/Mie) ou reflexos especulares presentes em túneis de vento físicos.

Adicionalmente, a pesquisa limitou-se à avaliação monocular puramente baseada em *software*, não contabilizando as restrições de *hardware* para aplicações em tempo real. Modelos de vanguarda que obtiveram destaque, particularmente os fundamentados em processos iterativos de difusão, demandam um custo computacional extremamente elevado, tornando inviável, no atual estágio tecnológico, sua implementação para inferência de fluidos *on-the-fly* (em tempo de execução) a altas taxas de quadros por segundo.

5.3 Trabalhos Futuros

As descobertas e limitações delimitadas por esta pesquisa abrem um vasto campo para investigações subsequentes na intersecção entre Engenharia da Computação e Mecânica dos Fluidos. Sugere-se, para trabalhos futuros, o desenvolvimento e a validação de arquiteturas híbridas que combinem o paradigma de *Latent Diffusion* com módulos de *Temporal Attention*. O objetivo seria fundir a altíssima precisão de volume estático do Marigold com a estabilidade de fluxo do VDA, visando eliminar o *trade-off* e o *flickering* sem sacrificar a percepção de densidade.

Recomenda-se também a execução de ensaios empíricos em túneis de vento físicos, utilizando injeção controlada de aerossóis acoplada a câmeras de alta velocidade. Essa transição do ambiente puramente simulado (CFD) para a captura óptica no mundo real permitirá avaliar o impacto da iluminação ambiente e do ruído de sensor na taxa de acerto espacial das redes neurais. Por fim, a exploração de técnicas de otimização de inferência (como quantização e destilação de conhecimento) aplicadas a modelos fundacionais de vídeo pode viabilizar o rastreamento autônomo de fluidos não-rígidos em tempo real para sistemas industriais críticos.

REFERÊNCIAS

ANDERSON, J. D. **Fundamentals of aerodynamics**. 7. ed. New York: McGraw-Hill Education, 2023.

ALHASHIM, I.; WON, P. High-quality monocular depth estimation via transfer learning. **arXiv preprint arXiv:1812.11941**, 2018.

BING, X. et al. Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, Seattle, p. 23674-23684, 2024. Disponível em: https://marigoldmonodepth.github.io. Acesso em: 16 jul. 2026.

BHAT, S. F. et al. ZoeDepth: zero-shot transfer by combining relative and metric depth. **arXiv preprint arXiv:2302.12288**, 2023. Disponível em: https://arxiv.org/abs/2302.12288. Acesso em: 20 jun. 2026.

BRUNTON, S. L.; NOACK, B. R.; KOUMOUTSAKOS, P. Machine learning for fluid mechanics. **Annual Review of Fluid Mechanics**, v. 52, p. 477-508, 2020.

CAI, S. et al. Physics-informed neural networks for fluid mechanics: a review. **Acta Mechanica Sinica**, v. 37, n. 12, p. 1727-1738, 2021.

DONG, X. et al. Towards robust and accurate monocular depth estimation in challenging scenes: A review. **IEEE Transactions on Intelligent Transportation Systems**, v. 23, n. 8, p. 10191-10208, 2022.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 9., 2021, Viena. **Proceedings...** Viena: ICLR, 2021.

EIGEN, D.; PUHRSCH, C.; FERGUS, R. Depth map prediction from a single image using a multi-scale deep network. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS), 27., 2014, Montreal. **Proceedings...** Montreal: Curran Associates, Inc., 2014. p. 2366-2374.

GODARD, C. et al. Digging into self-supervised monocular depth estimation. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2019, Seoul. **Proceedings...** Seoul: IEEE, 2019. p. 3828-3838.

HARTLEY, R.; ZISSERMAN, A. **Multiple view geometry in computer vision**. 2. ed. Cambridge: Cambridge University Press, 2004.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. **Advances in Neural Information Processing Systems (NeurIPS)**, v. 33, p. 6840-6851, 2020.

HU, J. et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2019, Waikoloa. **Proceedings...** Waikoloa: IEEE, 2019. p. 1043-1051.

HU, M. et al. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. **arXiv preprint arXiv:2404.15506**, 2024.

KANG, B. et al. Video Depth Anything: Consistent depth estimation for super-long videos. **arXiv preprint arXiv:2406.02595**, 2024. Disponível em: https://arxiv.org/abs/2406.02595.

KE, B. et al. Repurposing diffusion-based image generators for monocular depth estimation. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2024, Seattle. **Proceedings...** Seattle: IEEE, 2024. p. 9492-9502

LUO, X. et al. Consistent video depth estimation. **ACM Transactions on Graphics (TOG)**, v. 39, n. 4, p. 71:1-71:13, 2020..

MING, Y. et al. Deep learning for monocular depth estimation: A review. **Neurocomputing**, v. 438, p. 14-33, 2021.

PICCINELLI, L. et al. UniDepth: Universal monocular metric depth estimation. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2024, Seattle. **Proceedings...** Seattle: IEEE, 2024. p. 9503-9513.

RANFTL, R. et al. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 44, n. 3, p. 1623-1637, 2020.

RANFTL, R. et al. Vision transformers for dense prediction. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2021, Montreal. **Proceedings...** Montreal: IEEE, 2021. p. 12179-12188.

ROMBACH, R. et al. High-resolution image synthesis with latent diffusion models. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2022, New Orleans. **Proceedings...** New Orleans: IEEE, 2022. p. 10684-10695.

WANG, J. et al. Deep learning in fluid dynamics: A review. **Physics of Fluids**, v. 34, n. 8, p. 081301, 2022.

YUAN, W. et al. NeW CRFs: Neural window fully-connected CRFs for monocular depth estimation. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2022, New Orleans. **Proceedings...** New Orleans: IEEE, 2022. p. 3916-3925.

ZHAO, C. et al. Monocular depth estimation based on deep learning: An overview. **Science China Information Sciences**, v. 63, n. 11, p. 211301, 2020.

APÊNDICES

APÊNDICE A – Repositório de Códigos e *Scripts* de Execução

Visando garantir a transparência metodológica e a reprodutibilidade dos experimentos descritos nesta pesquisa, todos os *scripts* computacionais desenvolvidos e utilizados para a inferência volumétrica e validação estatística foram consolidados em um repositório digital nuvem.

O repositório contém os cadernos de execução (*notebooks* interativos da plataforma Google Colaboratory) desenvolvidos em linguagem Python. O material inclui:

- Os *pipelines* de carregamento e processamento dos tensores de vídeo (dados das simulações de Dinâmica dos Fluidos Computacional);
- Os algoritmos de inferência adaptados para cada uma das redes neurais avaliadas sob aceleração de hardware (GPU);
- Os *scripts* matemáticos responsáveis pelo cálculo do Coeficiente de Correlação de Pearson e da taxa de inconsistência temporal.

Devido à extensão dos códigos e visando facilitar a execução dinâmica por parte dos avaliadores ou de pesquisadores futuros, o material integral foi disponibilizado eletronicamente e pode ser acessado através do link abaixo:

Link de Acesso ao Repositório:

ESTIMATIVA DE PROFUNDIDADE MONOCULAR EM DINÂMICA DE FLUIDO…