CENTRO UNIVERSITÁRIO FEI

PEDRO HENRIQUE SILVA DOMINGUES

SEGMENTAÇÃO DE FACES PARCIALMENTE OCLUÍDAS PARA AVALIAÇÃO DA EXPRESSÃO DE DOR NEONATAL

São Bernardo do Campo

2024

PEDRO HENRIQUE SILVA DOMINGUES

SEGMENTAÇÃO DE FACES PARCIALMENTE OCLUÍDAS PARA AVALIAÇÃO DA EXPRESSÃO DE DOR NEONATAL

Dissertação de mestrado apresentada ao Centro Universitário da FEI para obtenção do título de Mestre em Engenharia Elétrica. Orientada pelo Prof. Dr. Carlos Eduardo Thomaz.

São Bernardo do Campo

2024

Domingues, Pedro Henrique Silva.
Segmentação de faces parcialmente ocluídas para avaliação da expressão de dor neonatal / Pedro Henrique Silva Domingues. São Bernardo do Campo, 2024.
69 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Carlos Eduardo Thomaz.

1. Segmentação de faces. 2. Dor em recém-nascidos. 3. Classificação de expressão facial. 4. Processamento de Imagens. 5. Visão Computacional. I. Thomaz, Carlos Eduardo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno(a): Pedro Henrique Silva Domingues **Matrícula:** 122103-5

Título do Trabalho: SEGMENTAÇÃO DE FACES PARCIALMENTE OCLUÍDAS PARA AVALIAÇÃO DA EXPRESSÃO DE DOR NEONATAL

Área de Concentração: Processamento de Sinais e Imagens

Orientador(a): Prof. Dr. Carlos Eduardo Thomaz

Data da realização da defesa: 22/03/2024

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

O aluno fez a apresentação oral em modo virtual por 40 minutos. Depois, foi arguido pelos examinadores sobre questões da apresentação e do texto da dissertação. Ele respondeu de forma satisfatória a essas questões levantadas mostrando conhecimento adequado sobre a pesquisa desenvolvida. Ao final, o candidato foi considerado aprovado por unanimidade.

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

MEMBROS DA BANCA EXAMINADORA

Prof. Dr. Carlos Eduardo Thomaz

Prof. Dr. Paulo Sérgio Silva Rodrigues

Profª Drª Josy Davidson Okida Vieira

Aprovação do Coordenador do Programa de Pós-graduação

Prof. Dr. Carlos Eduardo Thomaz

AGRADECIMENTOS

Após os anos de desenvolvimento desta dissertação, gostaria de agradecer àqueles cuja contribuição foi essencial para a evolução da pesquisa.

Primeiro gostaria de agradecer ao professor Carlos Thomaz, pela orientação e a Tatiany Heiderich pela contribuição fundamental na evolução de ideias e no suporte com o tema de recém-nascidos.

A Amanda Maciel por suas revisões, conselhos e cuja inspiração, junto à minha mãe Adriana Domingues, fez esta projeto se concretizar.

Por fim, agradeço ao Centro Universitário FEI e a CAPES pela oportunidade e suporte financeiro que me propiciaram, sem os quais não seria possível a realização desta pesquisa.

RESUMO

A avaliação e o tratamento corretos da dor são procedimentos clínicos importantes para o desenvolvimento saudável de recém-nascidos (RNs). Visto que RNs ainda não desenvolveram a capacidade de expressão verbal, a identificação da dor através de expressões faciais visuais é o meio alternativo mais comum para este caso, utilizada por pais e por profissionais da saúde, para estes últimos baseados em escalas como a *Neonatal Facial Coding System* (NFCS). Desse modo, a automatização do uso dessas escalas utilizando imagens de face é alvo de estudos recentes. As estratégias desenvolvidas utilizam redes neurais, modelos de classificação ou a medição de distâncias entre partes da face. No entanto, alguns equipamentos médicos utilizados por estes RNs dificultam a análise automática da dor através de imagens, pois obstruem parte da face, dificultando a detecção facial e a localização de pontos chave. Diante disso, o objetivo desta dissertação foi estudar a segmentação facial de RNs, identificando o melhor método computacional para casos com e sem a presença de oclusão parcial da face e verificando a influência desta segmentação como uma ferramenta para remoção de ruído e melhora de desempenho de modelos de classificação da expressão de dor. As oclusões aqui estudadas provêm de imagens de UTIs neonatais (UTINs), nas quais equipamentos médicos como sonda enteral ou gástrica, intubação orotraqueal e óculos de fototerapia muitas vezes impossibilitam a visualização completa de partes importantes da face como olhos e boca. Três modelos de segmentação foram testados em cenários com e sem a presença dessas oclusões e o melhor foi utilizado para segmentar faces de RNs e classificação da expressão de dor com quatro diferentes classificadores baseados em redes neurais. O SAM (Segment Anything Model) foi considerado o melhor modelo para segmentação, com alto coeficiente Dice (>0,91). No entanto, utilizar o SAM para remoção de ruidos como plano de fundo e oclusões não gerou melhora significativa no desempenho dos classificadores, que apresentaram em média 66% de acurácia em casos com oclusão.

Palavras-chave: Segmentação de faces. Dor em recém-nascidos. Classificação de expressão facial. Processamento de Imagens. Visão Computacional.

ABSTRACT

The correct identification and treatment of pain are crucial clinical procedures for the healthy development of newborns (NBs). Since NBs have not yet developed the capacity for verbal expression, identifying pain through visual facial expressions is the most common alternative method for this case, used by parents and healthcare professionals, with the latter relying on scales such as the Neonatal Facial Coding System (NFCS). Therefore, the automation of using these scales with facial images has become the subject of recent studies. The developed strategies involve the utilization of neural networks, classification models, or measuring distances between facial parts. However, some medical equipment used on these NBs complicates the automatic analysis of pain through images, as they obstruct parts of the face, hindering facial detection and key point localization. Consequently, the objective of this dissertation was to study NB facial segmentation, discovering the optimal computational method for cases with and without partial face occlusion and assessing the influence of this segmentation as a tool for noise removal and enhancement of pain expression classification model performance. The occlusions studied here stem from images captured in neonatal intensive care units (NICUs), where medical equipment such as enteral or gastric tubes, orotracheal intubation, and phototherapy goggles often prevent the complete visualization of crucial facial parts such as the eyes and mouth. Three segmentation models were tested in scenarios with and without the presence of these occlusions, and the best one was used to segment NB faces and classify pain expression with four different neural network-based classifiers. The SAM (Segment Anything Model) was considered the best model for segmentation, with a high Dice coefficient (>0.91). However, using SAM for noise removal such as background and occlusions did not result in a significant improvement in classifier performance, which averaged 66% accuracy in cases with occlusion.

Keywords: Pain classification. Infant pain. Image processing. Face segmentation. Computer vision.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A habilidade de comunicação verbal, trivial para a maioria de nós, facilita o tratamento clínico da dor, uma vez que podemos indicar sua presença e descrever fatores como intensidade, localização e origem. Porém, a inabilidade de comunicação verbal dos RNs apresenta o desafio de realizar o diagnóstico da dor nesta faixa etária, sendo possível realizar apenas observações comportamentais, como as expressões faciais e as contrações de pernas e braços, assim como as medições de índices fisiológicos, como alterações na frequência cardíaca e respiratória (SBP, 2018).

O número de tratamentos invasivos e dolorosos em bebês internados em unidades de tratamento intensivo neonatal (UTIN) pode variar de 2 a 14 procedimentos por dia, dos quais menos de um terço recebem a devida terapia analgésica (RANGER; JOHNSTON; ANAND, 2007). Este alto índice de procedimentos dolorosos e a dificuldade do diagnóstico, unido ao não tratamento de eventos dolorosos recorrentes, podem gerar consequências significativas e permanentes (STEVENS; JOHNSTON; HORTON, 1993), impactando no desenvolvimento dos componentes somatossensoriais e emocionais da resposta à dor na vida adulta (WALKER, 2019).

Atualmente, o reconhecimento da dor em neonatos é realizado utilizando escalas de dor como a *Neonatal Facial Coding System* (NFCS) (GRUNAU; CRAIG, 1987), a *Neonatal Infant Pain Scale* (NIPS) (LAWRENCE et al., 1993) e a *Premature Infant Pain Profile-Revised* (PIPP-R) (GIBBINS et al., 2014) que, a partir dos índices comportamentais e fisiológicos, geram pontuações para determinar a presença e, em alguns casos, intensidade da dor.

Pela proximidade da área de Processamento de Imagens ao meio médico, muitos métodos computacionais para a classificação automática de dor em neonatos surgiram, utilizando métricas faciais e escalas de dor (HEIDERICH; LESLIE; GUINSBURG, 2015), *Support Vector Machines* (BRAHNAM et al., 2005), ou redes neurais convolucionais (ZAMZMI et al., 2019) (BUZUTI, 2020), (HEIDERICH et al., 2023). Estes métodos alcançam valores de acurácia de até 91% em cenários ideais, porém enfrentam o mesmo desafio para a aplicação prática em UTINs, a presença constante de equipamentos médicos obstruindo parcialmente a face.

1.1 OBJETIVO

O objetivo geral desta dissertação é avaliar métodos computacionais de segmentação de faces parcialmente ocluídas para avaliação de expressão da dor neonatal. Mais especificamente objetiva-se:

a) Avaliar o desempenho de modelos de segmentação facial em cenários com e sem a presença de oclusão aplicados no contexto de unidades de tratamento intensivo neonatal;
b) Verificar a influência da segmentação da face, visando a remoção de plano de fundo e artefatos clínicos obstruindo a face, quando aplicada previamente a modelos de classificação da expressão de dor.

1.2 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em 7 capítulos. No próximo capítulo, Capítulo 2, são apresentados os conceitos fundamentais diretamente ou indiretamente vinculados ao objetivo da dissertação. O Capítulo 3 apresenta trabalhos de temas relacionados a este, separados entre trabalhos de classificação de imagens, segmentação de imagens e dor em RNs. O Capítulo 4 contém os materiais (bases de imagens) e métodos (modelos de segmentação e classificação) utilizados nos experimentos. O Capítulo 5 apresenta os experimentos propostos juntamente aos respectivos resultados obtidos. O Capítulo 6 possui uma discussão sobre os resultados obtidos no Capítulo anterior. Por fim, o Capítulo 7 traz as conclusões do trabalho.

2 CONCEITOS FUNDAMENTAIS

Neste capítulo, serão apresentados os conceitos e termos fundamentais com enfoque em escalas, técnicas de processamento de imagens e métricas relacionados a pesquisa.

2.1 DETECÇÃO DE FACES

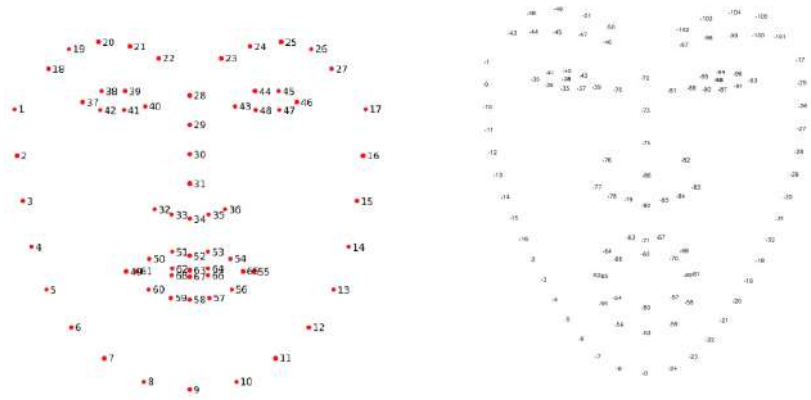
A detecção de faces é um subcampo da visão computacional amplamente estudado e é definido pela criação de modelos e algoritmos capazes de, dada uma imagem qualquer, encontrar a posição de uma ou mais faces, gerando como saída dados que possam ser utilizados para localizar cada face na imagem. Os dados representando a posição da face são usualmente encontrados como: um conjunto de pontos-chave, uma máscara de segmentação ou uma caixa retangular representando a posição estimada.

Estratégias para detecção de faces evoluíram de algoritmos utilizando técnicas de redução de dimensionalidade, como Turk e Pentland (1991) e Belhumeur, Hespanha e Kriegman (1997), para busca de características como em Viola e Jones (2001), e atualmente convergiram para o uso de redes neurais convolucionais, a exemplo de Deng et al. (2019a) e Zhang et al. (2016b).

2.1.1 Pontos-chave

Os pontos-chave, ou pontos fiduciais, utilizados para representar uma face são um conjunto de coordenadas de pontos descritivos da face em duas ou três dimensões, consistindo, mas não se restringindo, do contorno de regiões, como: boca, olhos, sobrancelhas, nariz e face propriamente dita.

Múltiplas configurações de pontos e localizações podem ser encontradas na literatura, a exemplo de Zhang et al. (2016a) para 5 pontos, Zhu e Ramanan (2012) 68 pontos para faces frontais ou 39 para faces de perfil, Deng et al. (2019b) 68 pontos, Insightface (2023) 106 pontos, e Kazemi e Sullivan (2014) 194 pontos. A Figura 1 ilustra dois possíveis conjuntos de pontos-chave com 68 e 106 pontos, respectivamente. As coordenadas destes pontos são encontradas por meio de modelos de regressão que, portanto, estimam a posição de pontos mesmo que estes estejam oclusos por objetos.



(a) Conjunto de 68 pontos (b) Conjunto de 106 pontos

Figura 1 – Exemplos de conjuntos de pontos fiduciais.

Fonte: (a) Deng et al. (2019b); (b) Insightface (2023).

2.2 SEGMENTAÇÃO DE IMAGENS

Segmentação de imagens, segundo H.D. Cheng et al. (2001), é o processo de dividir uma imagem em várias regiões de forma que cada região seja homogênea, mas a união de pelo menos duas regiões adjacentes não seja homogênea.

O processo de segmentação de uma imagem resulta na geração de uma segunda imagem, denominada máscara, a qual delimita diferentes regiões de interesse, atribuindo valores para cada região, ilustrada na Figura 2b.

A seguir estão descritos alguns exemplos de métodos para segmentação de regiões em imagem utilizando estratégias comuns no meio de processamento de imagens:

a) *Clustering*: Utilizando algoritmos como o K-Means (MACQUEEN, 1967), pode-se separar a imagem em K regiões distintas agrupadas por semelhança em cor;
b) Detecção de bordas: Utilizando algoritmos de detecção de bordas, como o Canny Edge (CANNY, 1986), pode-se determinar a delimitação entre regiões distintas e consequentemente, segmentá-las;
c) Histograma: Através do histograma de intensidades da imagem, pode-se dividi-la em grupos de forma manual ou definida por picos deste histograma;
d) Redes neurais: Diversas arquiteturas para segmentação de imagens estão disponíveis atualmente, um exemplo sendo a arquitetura convolucional U-Net de Ronneberger, Fischer e Brox (2015).

(a) Imagem original

(b) Máscara, resultado do processo de segmentação

Figura 2 – Segmentação da face de um recém-nascido. *Background* indicado na cor preta e *foreground* (face) na cor branca.

Fonte: Autor.

2.2.1 Segmentação Semântica

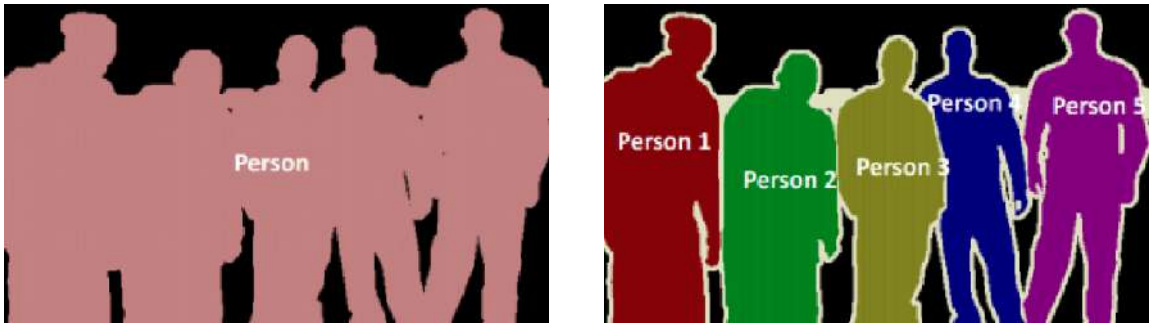
A segmentação semântica atribui a cada pixel da imagem uma classe semântica de um conjunto de N classes. Um caso particular da segmentação semântica ocorre para N =2, em que temos uma divisão binária, na qual a região de valor 0 (preto) simboliza o plano de fundo enquanto a região de valor 255 (branco) representa um objeto de interesse, conforme demonstrado na Figura 2. A Figura 3 apresenta um exemplo de segmentação semântica para N >2.

Figura 3 – Exemplo de segmentação semântica com cores indicando as diferentes classes.

Fonte: Adaptado de Wang et al. (2018).

2.2.2 Segmentação de Instância

A segmentação de instância diferencia objetos (instâncias) diferentes de uma mesma classe, portanto existindo apenas para objetos contáveis como pessoas, casas ou carros, mas não para objetos não contáveis, como céu ou chão. Na Figura 4 é possível notar que, para ambas as formas de segmentação, todas as pessoas possuem a mesma classe (*Person*), mas na segmentação de instância as pessoas são segmentadas independentemente umas das outras.



(a) Segmentação semântica (b) Segmentação de instância

Figura 4 – Exemplo de segmentação semântica e de instância aplicadas a mesma imagem.
Fonte: Adaptado de Varatharasan et al. (2019).

2.2.3 Segmentação Panóptica

A segmentação panóptica, proposta por Kirillov et al. (2019), é uma forma de unificar a segmentação semântica (Seção 2.2.1) e a segmentação de instância (Seção 2.2.3). Na segmentação panóptica a classificação é realizada por pixel, assim como na segmentação semântica, porém cada pixel recebe uma classe e um identificador de instância. A Figura 5 apresenta uma comparação entre a segmentação panóptica, semântica e de instância.

2.2.4 Métricas de Avaliação

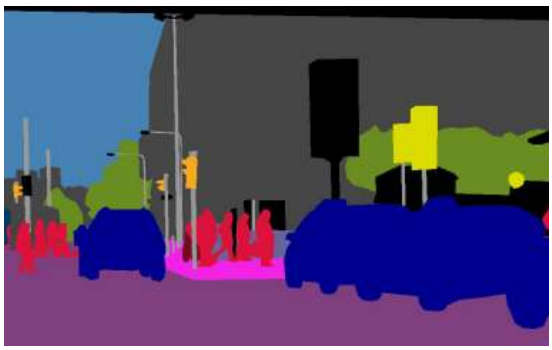
Para avaliar máscaras de segmentação, uma máscara é criada de forma manual, com os pixels corretamente classificados. Esta é conhecida como padrão ouro, ou, *ground truth* (GT). Uma vez que uma máscara de segmentação pode ser vista como uma matriz de ordem $N \times M$, qualquer métrica capaz de comparar duas matrizes de mesma ordem pode ser utilizada para avaliar uma máscara qualquer e seu respectivo padrão ouro.

A seguir serão apresentadas algumas das principais métricas utilizadas para avaliação de segmentações, incluindo as mencionadas anteriormente.

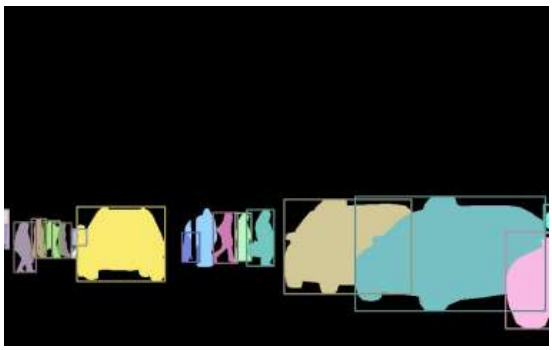
2.2.4.1 Pixel Accuracy (PA)

A métrica mais simples para comparação de máscaras e encontrada em alguns locais da literatura como PLA (*Pixel Label Accuracy*), ou simplesmente *Accuracy*, consiste no cálculo da acurácia na classificação de cada pixel comparado ao padrão ouro, em que TP, TN, FP e FN são,

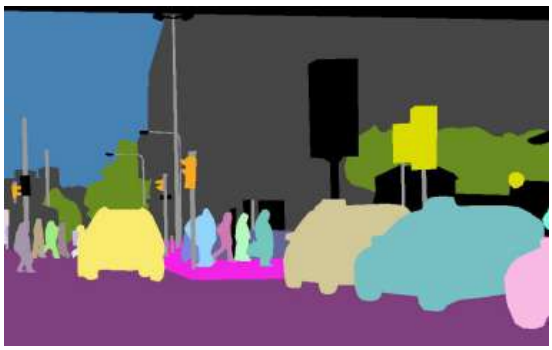
(a) Imagem original



(b) Segmentação semântica



(c) Segmentação de instância



(d) Segmentação panóptica

Figura 5 – Exemplo de segmentação panóptica, semântica e de instância aplicadas a mesma imagem.

Fonte: Adaptado de Kirillov et al. (2019).

respectivamente, *True Positives* (quantidade de predições positivas corretas), *True Negatives* (quantidade de predições negativas corretas), *False Positives* (quantidade de predições positivas incorretas), *False Negatives* (quantidade de predições negativas incorretas), tal que

$$PA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

O valor $TP+TN$ pode ser interpretado como o número total de predições corretas, portanto os pixels iguais na máscara resultado da segmentação e na máscara padrão ouro, enquanto $TP + TN + FP + FN$ pode ser interpretado como o número total de pixels na imagem.

A métrica não contempla o problema de balanceamento de classes, logo, para casos como a segmentação de regiões pequenas em uma imagem, máscaras sem nenhuma predição (todos os pixels com classe de plano de fundo) atingem valores altos de PA. Para ilustrar este problema, a Figura 6 apresenta um padrão ouro e uma máscara que não segmenta nenhuma região. Para este exemplo, o valor de $PA \approx$97,16%.

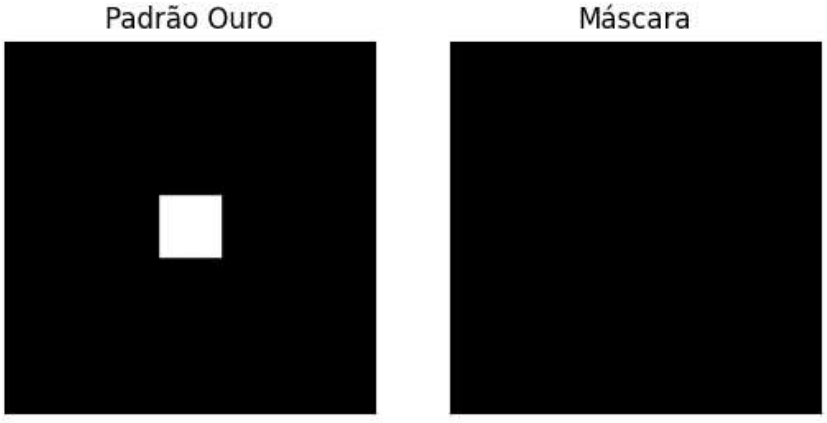


Figura 6 – Exemplo em que a segmentação de nenhuma região apresenta valor de $PA \approx$97,16%.

Fonte: Autor.

2.2.4.2 Coeficiente de Similaridade de Jaccard (IoU)

O coeficiente de Jaccard, também conhecido como IoU, um acrônimo para a descrição direta da fórmula (*Intersection over Union*), é calculado com base apenas na região de interesse da segmentação, diferente da PA (Seção 2.2.4.1), que calcula um valor com base em todos os pixels da imagem. A fórmula verifica a porcentagem da região segmentada comum em relação ao padrão ouro, conforme descrito na Equação 2 e ilustrado na Figura 7, ou seja

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{(TP + FP + FN)}, \tag{2}$$

Para casos com a presença de múltiplas classes, a exemplo da segmentação de instância, é comum o uso da média entre o IoU de cada classe. Esta métrica pode ser encontrada na literatura com o acrônimo mIoU.

2.2.4.3 Average Precision (AP)

Para avaliação de segmentação de instância uma métrica comumente encontrada é a AP. Esta é calculada a partir da área abaixo da curva de precisão e *recall*. A curva por sua vez é gerada utilizando múltiplos *thresholds* de IoU para definição de predições como positivas ou negativas.

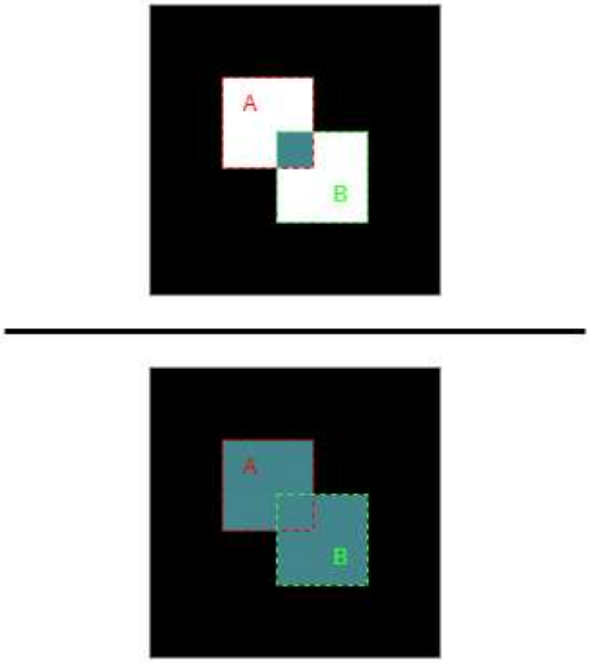


Figura 7 – Equação 2 entre duas regiões segmentadas (A e B) representado de forma gráfica.
Fonte: Autor.

2.2.4.4 Coeficiente de Similaridade Dice

O coeficiente Sørensen-Dice, conhecido também como coeficiente, ou pontuação, F1, é uma métrica de similaridade que pondera precisão e *recall*. Para isso, a área de intersecção entre a região de interesse segmentada e a região de interesse padrão-ouro é comparada com a soma das duas áreas e multiplicada por dois, fazendo com que a métrica possua limite inferior e superior iguais a 0 e 1 respectivamente. O coeficiente é descrito na Equação 3 e ilustrado de forma gráfica para o caso de segmentação de imagens na Figura 8.

$$Dice = 2 * \frac{|A \cap B|}{|A| + |B|} = 2 * \frac{TP}{(TP + FP) + (TP + FN)} = 2 * \frac{\text{Precisão} * \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (3)$$

2.2.4.5 Panoptic Quality (PQ)

Visto que a segmentação panóptica propõe uma forma de unificar a segmentação de instância e semântica, os autores de Kirillov et al. (2019) criam também uma métrica para avaliação. A métrica, segundo os autores, envolve dois passos principais, o *segment matching* e o cálculo do PQ. O primeiro passo, visando garantir que cada região segmentada pelo modelo sendo avaliado está associada a uma única região do padrão ouro, especifica que: duas regiões

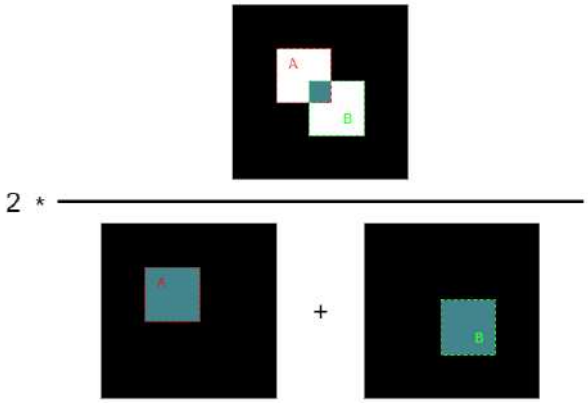


Figura 8 – Equação 3 de forma gráfica.

Fonte: Autor.

(uma no padrão ouro e outra na segmentação proveniente do modelo) estão relacionados apenas se o valor de IoU entre elas é superior a 0,5.

Segundo Kirillov et al. (2019) o segundo passo, referente ao cálculo da PQ, é realizado para cada classe de forma independente e a média é utilizada para obter a pontuação final. A fórmula de PQ está descrita na Equação 4 e seus componentes podem ser visualizados de forma gráfica na Figura 9, tal que

$$PQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}, \tag{4}$$

em que p e g são siglas para *predicted* e *ground truth*, referenciando-se a segmentação de cada região presente.

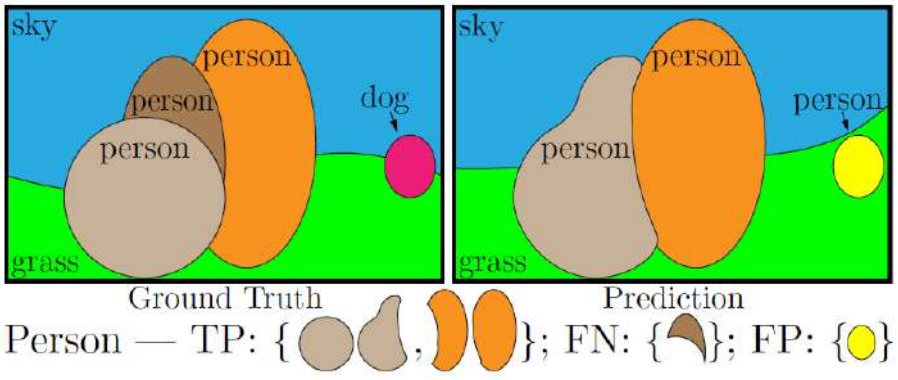


Figura 9 – Termos da Equação 4 ilustrados de forma gráfica considerando GT à esquerda e máscara segmentada a direita.

Fonte: Kirillov et al. (2019).

2.3 MÉTRICAS DE PERFORMANCE

As seções subsequentes apresentarão quatro métricas para avaliação de classificadores. Nestas seções serão utilizadas as seguintes siglas: TP (Verdadeiros Positivos), FP (Falsos Positivos), TN (Verdadeiros Negativos) e FN (Falsos Negativos).

2.3.1 Precisão

A precisão, calculada a partir da Equação 5, mede a porcentagem de classificações positivas corretas feitas pelo modelo (TP) em função de todas as classificações positivas feitas por este ($TP + FP$), ou seja

$$Precisão = \frac{TP}{TP + FP}. \tag{5}$$

2.3.2 Recall

O *recall*, calculado a partir da Equação 6, mede a porcentagem de classificações positivas presentes na base de dados ($TP + FN$) classificadas corretamente pelo modelo (TP), tal que

$$Recall = \frac{TP}{TP + FN}. \tag{6}$$

2.3.3 Pontuação F1

A pontuação F1, ou F1 *score*, calculada a partir da Equação 7, consiste na média harmônica entre a precisão e o *recall*, gerando o seguinte valor em porcentagem:

$$F1 = \left(\frac{Precisão^{-1} + Recall^{-1}}{2} \right)^{-1} \tag{7}$$

$$= \frac{2}{Precisão^{-1} + Recall^{-1}} \tag{8}$$

$$= 2 * \frac{Precisão * Recall}{Precisão + Recall} \tag{9}$$

$$= 2 * \frac{TP}{(TP + FN) + (TP + FP)}. \tag{10}$$

2.3.4 Acurácia

A acurácia, diferente das métricas mencionadas anteriormente, considera tanto classificações positivas corretas (TP) quanto classificações negativas corretas (TN), desta forma medindo a porcentagem de classificações corretas ($TP+TN$) em relação a todas as classificações feitas pelo modelo ($TP + TN + FP + FN$), conforme apresentado na seguinte equação:

$$F1 = \frac{TP + TN}{TP + TN + FP + FN)}. \tag{11}$$

2.4 ESCALAS DE DOR

Escalas de dor para neonatos são ferramentas criadas para avaliação de dor através da medição de indicadores fisiológicos ou comportamentais. Exemplos de parâmetros fisiológicos são: frequência cardíaca, pressão arterial e saturação de oxigênio. Exemplos de parâmetros comportamentais são: choro, mímicas (expressões) faciais e atividade motora.

2.4.1 NEONATAL FACIAL CODING SYSTEM (NFCS)

A escala *Neonatal Facial Coding System* (NFCS), de Grunau e Craig (1987), utiliza apenas métricas faciais visuais. Originalmente foram propostas 9 ações faciais, porém, tabelas com apenas 8 ações podem ser encontradas na literatura, visto que estas unem as ações **boca esticada (vertical)** e **boca esticada (horizontal)** do artigo original como uma única ação nomeada apenas de boca esticada. Utilizando a Tabela 1 como referência, classifica-se como dor caso a soma da pontuação resultante seja superior a 3 pontos.

Movimento facial	0 pontos	1 ponto
Fronte saliente	Ausente	Presente
Olhos espremidos	Ausente	Presente
Sulco nasolabial aprofundado	Ausente	Presente
Lábios entreabertos	Ausente	Presente
Boca esticada	Ausente	Presente
Lábios franzidos	Ausente	Presente
Língua tensa	Ausente	Presente
Tremor de queixo	Ausente	Presente

Tabela 1 – Escala de dor NFCS.

Fonte: SBP (2018)

A seguir serão descritas as características dos indicadores apresentado na Tabela 1 exatamente conforme descrito por SBP (2018):

a) **Fronte saliente:** abaulamento e sulcos acima e entre as sobrancelhas;
b) **Olhos espremidos:** compressão total ou parcial da fenda palpebral;
c) **Sulco nasolabial aprofundado:** aprofundamento do sulco que se inicia em volta das narinas e se dirige à boca;
d) **Lábios entreabertos:** qualquer abertura dos lábios;
e) **Boca esticada:** vertical (com abaixamento da mandíbula) ou horizontal (com estiramento das comissuras labiais);
f) **Lábios franzidos:** parecem estar emitindo um "úúúú";
g) **Língua tensa:** em protrusão, esticada e com as bordas tensas;
h) **Tremor de queixo**.

2.4.2 NEONATAL INFANT PAIN SCALE (NIPS)

A *Neonatal Infant Pain Scale* (NIPS), de Lawrence et al. (1993), funciona de forma semelhante a NFCS (Seção 2.4.1), mas realiza a classificação entre sem dor, dor moderada ou dor severa caso a soma total da pontuação seja, respectivamente, entre 0 e 2, entre 3 e 4 ou maior do que 4. Os parâmetros utilizados são: expressão facial, choro, respiração, braços, pernas e estado de alerta. As pontuações atribuídas a cada expressão encontram-se na Tabela 2.

Parâmetro	0 pontos	1 ponto	2 pontos
Expressão Facial	Relaxada	Contraída	-
Choro	Ausente	"Resmungo"	Vigoroso
Respiração	Regular	Diferente da basal	-
Braços	Relaxados	Fletidos ou estendidos	-
Pernas	Relaxados	Fletidas ou estendidas	-
Estado de Alerta	Dormindo / Acordado	Agitado	-

Tabela 2 – Escala de dor NIPS.
Fonte: SBP (2018)

2.4.3 PREMATURE INFANT PAIN PROFILE-REVISED (PIPP-R)

A escala *Premature Infant Pain Profile-Revised* (PIPP-R), de Gibbins et al. (2014), inspirada na escala PIPP, de Stevens et al. (1996), utiliza uma pontuação gerada a partir de indicadores fisiológicos, comportamentais e idade gestacional. Segundo SBP (2018), esta é uma

escala consolidada no meio clínico para casos de dor aguda especialmente para prematuros. A pontuação é gerada a partir do seguinte procedimento utilizando a Tabela 3, conforme descrito por SBP (2018):

Indicador	**Pontuação**				**Pontuação**
	0	**+1**	**+2**	**+3**	
Mudança na Frequência Cardiaca (bpm) Basal: _____	0-4	5-14	15-24	>24	
Mudança na Saturação de O_2 (%) Basal: _____	0-2	3-3	6-8	>8 ou ↑O_2	
Testa franzida (seg)	Nada (<3)	Min (3-10)	Mod (11-20)	Max (>20)	
Olhos espremidos (seg)	Nada (<3)	Min (3-10)	Mod (11-20)	Max (>20)	
Sulco nasolabial profundo (seg)	Nada (<3)	Min (3-10)	Mod (11-20)	Max (>20)	
				*Subtotal:	_____
Idade gestacional (semanas + dias)	$\geq$36	32-35$^{6/7}$	28-31$^{6/7}$	$<$28	
Estado de alerta basal	Ativo e acordado	Quieto e acordado	Ativo e dormindo	Quieto e dormindo	
				**Total:	_____

Tabela 3 – Escala de dor PIPP-R.

Fonte: SBP (2018)

a) Observar o recém-nascido por 15 segundos, em repouso e avaliar os sinais vitais (frequência cardíaca mais alta, saturação de oxigênio mais baixa e estado de alerta);

b) Observar o recém-nascido por 30 segundos após o procedimento e avaliar a mudança dos indicadores (frequência cardíaca mais alta, saturação do oxigênio mais baixa e duração das ações faciais). Se o recém-nascido precisar de aumento da oferta de O_2 em qualquer momento, antes ou durante o procedimento, ele recebe +3 pontos no indicador de saturação de oxigênio;

c) Pontuar idade gestacional e estado de alerta se o *Subtotal > 0;

d) Calcular a pontuação **Total = *Subtotal + idade gestacional + estado de alerta.

3 TRABALHOS RELACIONADOS

Este capítulo está dividido em seções, cada uma referente a um tópico desta pesquisa e apresentando os trabalhos em ordem cronológica de publicação de cada seção. As duas primeiras Seções (3.1 e 3.2), referentes a segmentação e classificação, estão relacionadas a tarefas de visão computacional e foram divididas entre trabalhos envolvendo a criação de bases de dados (Seções 3.1.1 e 3.2.1) e de modelos (Seções 3.1.1 e 3.2.1) para cada um dos tópicos em questão. A Seção 3.3, dor em recém-nascidos, traz algumas das abordagens criadas em função dos anos, para contextualização sobre as técnicas disponíveis na literatura.

3.1 SEGMENTAÇÃO DE IMAGENS

Esta seção apresenta a evolução das técnicas de segmentação de imagens com base em redes neurais entre os anos de 2015 e 2023, agrupando segmentação de instância, semântica e panóptica, assim como as bases de dados mais utilizadas no meio.

3.1.1 CONJUNTOS DE DADOS

Everingham et al. (2012) introduzem o *Pascal Visual Object Classes Challenge*, uma competição que ocorre desde o ano de 2005 e com essa a base de imagens e classes para segmentação, detecção e classificação de objetos: Pascal VOC, ou VOC 2012. A base é atualizada anualmente e possui, na sua versão de 2012, um total de 20 classes e 9.993 imagens anotadas.

Cordts et al. (2016) criaram o Cityscapes, um conjunto de dados para segmentação semântica que contém 30 classes e utiliza imagens de cenários urbanos de 50 cidades durante o dia, totalizando 5.000 imagens com anotações refinadas (precisão ao nível de pixel) e 20.000 imagens com anotações aproximadas.

Zhou et al. (2017) produziram o ADE20K, um *dataset* para segmentação semântica, com 25.000 imagens de cenários cotidianos anotadas com 150 classes. Segundo os autores, cada imagem possui em média 19,5 instâncias de, em média, 10,5 classes.

3.1.2 MODELOS

Ronneberger, Fischer e Brox (2015) introduzem a U-Net, uma arquitetura de rede neural para segmentação que faz uso de contração e expansão da imagem e aumento de dados com o objetivo de reduzir a necessidade de bases com milhares de exemplos pre-anotados.

Chen et al. (2017) trazem três contribuições. Primeiro, criam as convoluções *atrous* (dilatadas), para aumentar o campo de visão dos filtros sem aumentar o custo computacional. Além disso, também é proposto o *atrous spatial pyramid pooling* (ASPP), em que as convoluções *atrous* são aplicadas com múltiplos espaçamentos para aumentar a robustez na tarefa de segmentação em múltiplas escalas. Por fim, é proposto o modelo DeepLab, que atinge estado da arte para o VOC 2012 com 79.7% mIoU no conjunto de teste.

Zhao et al. (2017) trazem duas inovações para o campo da segmentação semântica de imagens, a primeira sendo uma nova forma de *pooling*, denominada *pyramid pooling module*, capaz de melhorar a capacidade de segmentação em cenários complexos que possuem muitas classes e objetos de tamanhos variados. A segunda inovação foi a rede que incorpora este *pooling*, chamada de *Pyramid Scene Parsing Network* (PSPNet), que atingiu estado da arte para segmentação semântica na época de publicação por obter acurácia de 85,4% na base de dados PASCAL VOC 2012 e 80,2% na Cityscapes.

Chen et al. (2018) incorporam e estendem o módulo de *Spatial Pyramid Pooling* com a estrutura *encoder-decoder* para a tarefa de segmentação semântica, estendendo o modelo DeepLabV3 e renomeando para DeepLabV3+. Os resultados obtidos atingem valores de acurácia 89,0% para o conjunto de dados PASCAL VOC 2012 e 82,1% para o Cityscapes.

Cheng, Schwing e Kirillov (2021) unificam a segmentação a nível de instância e semântica em um modelo único, com um único procedimento de treinamento e função de erro, permitindo assim que o modelo opere em segmentação panóptica. O modelo nomeado MaskFormer gera um conjunto de máscaras binárias, associadas a uma classe cada. Os autores reportam pontuações de 55,6 mIoU para segmentação semântica no conjunto de dados ADE20K e 52,7 PQ para segmentação panóptica no conjunto de dados COCO.

Xie et al. (2021) apresentam o SegFormer, um *framework* para segmentação semântica de imagens que unifica um *encoder* hierárquico com base na arquitetura Transformer a um *decoder* simples, utilizando Perceptron multi-camada. Os autores disponibilizam o *framework* em múltiplas quantidades de camadas, nomeados SegFormer-B0 a SegFormer-B5, balanceando performance e custo computacional. O resultado para o SegFormer-B4 é de 50,3% mIoU para o

conjunto de dados ADE20K, o que, segundo os autores, representa uma melhora de 2,2% com cinco vezes menos parâmetros do que o estado da arte contemporâneo.

Bowen Cheng et al. (2022) evoluem o MaskFormer ajustando parâmetros de treino e modificando o *decoder* da camada Transformer pela *masked attention*. O novo modelo, nomeado Mask2Former, opera em todas as tarefas de segmentação (instância, semântica e panóptica) e supera tanto o seu predecessor quanto modelos especialistas para cada tipo de segmentação, obtendo os resultados de 57,7 mIoU para a tarefa de segmentação semântica no *dataset* ADE20K, 57,8 PQ para segmentação panóptica no COCO e 50,1 AP para segmentação de instância no COCO.

Kirillov et al. (2023) apresentam o projeto *Segment Anything* (SA), que constitui, em parte, do SA-1B, o conjunto de dados para segmentação de imagens já criado e do novo modelo de segmentação *Segment Anything Model* (SAM). A base SA-1B é constituída por um bilhão de máscaras relativas a onze milhões de imagens e foi utilizada para o treinamento do SAM, um modelo para segmentação de objetos sem classificação, que recebe como entradas de dado um *prompt*, podendo ser textual, uma coordenada ou um retângulo referentes a imagem e a imagem propriamente dita. Os autores reportam alta capacidade do modelo proposto de realizar segmentação de objetos sem a necessidade de ajuste fino ou treinamento adicional do modelo.

3.2 CLASSIFICAÇÃO DE IMAGENS

Nesta seção encontram-se alguns dos principais modelos para classificação de imagens, assim como os conjuntos de dados mais utilizados no meio para treinamento e comparação destes modelos.

3.2.1 CONJUNTOS DE DADOS

Krizhevsky (2009) cria os conjuntos de dados CIFAR-10, que contém 10 classes e 6000 imagens e o CIFAR-100, que possui 100 classes e 600 imagens. Todas as imagens em ambos os conjuntos são de baixa resolução (32×32) e cada imagem pertence a uma, e apenas uma, classe. Exemplos de classes nestes dataset são: Cachorro, gato, avião, automóvel, entre outros.

Russakovsky et al. (2015) organizam a competição *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), que ocorre desde o ano de 2010 e envolve uma ou mais das seguintes categorias: classificação de imagens, localização de um único objeto e detecção de

objetos. Junto a competição é publicado o conjunto de dados ImageNet, que no ano de 2023 possui 1.000 classes de objetos, 1.281.167 imagens de treino, 50.000 imagens de validação e 100.000 imagens para teste. A versão do *dataset* utilizada para a competição é comumente referida como ImageNet-1k, porém a coleção completa de dados deste conjunto, nomeada de ImageNet-21k, possui mais de 21 mil classes e 14 milhões de imagens.

Liu et al. (2015) constroem dois conjuntos de dados para detecção de faces, o CelebA e o LFWA. CelebA é anotado a partir da base de imagens de faces Celeb-Faces, possui 10.000 pessoas distintas, cada uma com 20 imagens. LFWA é anotado a partir da base de imagens de faces LFW e possui 13.233 imagens de 5.749 identidades diferentes. Para ambos os conjuntos, são anotadas 40 características faciais e 5 pontos fiduciais.

3.2.2 MODELOS

LeCun et al. (1989) pesquisam sobre o reconhecimento de caracteres em documentos escritos à mão, criando a arquitetura LeNet, uma rede que consiste em camadas convolucionais seguidas por camadas de *pooling*, e finalizando com camadas totalmente conectadas. Os autores introduziram o *backpropagation* a redes convolucionais, permitindo o aprendizado automático e substituindo o uso de técnicas para escolha manual de coeficientes. Lecun et al. (1998), ainda estudando a classificação de caracteres, expandem a LeNet para a LeNet-5, uma rede que firmou a superioridade das redes convolucionais na tarefa de classificação de imagens comparada a outros métodos da época.

Krizhevsky, Sutskever e Hinton (2012) criam a AlexNet, uma arquitetura que consiste em oito camadas, das quais cinco são camadas convolucionais seguidas por camadas de *pooling* e três camadas totalmente conectadas. Os autores introduzem a utilização de funções de ativação ReLU, técnicas de regularização, como o *dropout*, e a aplicação de aumento de dados.

Szegedy et al. (2014) propõem uma otimização de uma rede neural, de forma a expandir a rede em largura e profundidade, enquanto mantém o custo computacional constante. Os autores criam os blocos *inception*, que consistem em múltiplas operações convolucionais paralelas de diferentes tamanhos de filtro e *pooling*, permitindo que a rede capture características em várias escalas de forma eficiente. A arquitetura criada foi nomeada Inception, mas recebeu também o nome de GoogLeNet na submissão para a competição ILSVRC14.

Simonyan e Zisserman (2014) estudam a relação entre profundidade e acurácia em redes para classificação de imagens. Os autores utilizam o empilhamento de camadas convolucionais

com filtros pequenos (3×3) a fim de gerar redes muito profundas e reportam que a rede criada, nomeada *Visual Geometry Group* (VGG), apresenta um ganho significativo em comparação ao até então estado da arte.

Szegedy et al. (2015) evoluem a arquitetura do modelo Inception, otimizando a eficiência computacional e o desempenho de classificação de imagens. Os autores propõem dois novos modelos: o Inception-v2, que utiliza convoluções de diferentes tamanhos e estruturas com o foco em simplificar e otimizar a arquitetura, e o Inception-v3, que possui as mesmas características do predecessor e adicionalmente introduz a normalização por lote e fatorização espacial.

He et al. (2016) trabalham no problema de treinamento de redes neurais muito profundas. Os autores melhoram a capacidade de treino destas redes através do uso de camadas que aprendem funções residuais em relação a sua entrada. As redes com camadas residuais, nomeadas ResNet, demonstram-se aptas a possuírem muito mais camadas com menor aumento de complexidade, como ocorre em outras redes. Os autores receberam o primeiro lugar em múltiplas competições de classificação de imagens e a ResNet tornou-se o novo estado da arte no meio.

Chollet (2017) cria a arquitetura Xception, a qual utiliza um módulo de convoluções separáveis em profundidade, mas com a ordem inversa ao método tradicional, com a convolução 1×1 sendo realizada primeiro, seguida das convoluções para cada canal, inspirado pela arquitetura Inception. Desta forma, foi possível atingir um ganho considerável em performance para grandes conjuntos de dados sem aumento no número de parâmetros em comparação ao modelo Inception.

Tan e Le (2019) introduzem o método *compound scaling*, substituindo a prática de modificação dos parâmetros para dimensionamento de modelos de rede neural, como largura, profundidade e resolução, por uma técnica que estima coeficientes para cada parâmetro através de uma busca em grade na versão original do modelo. Os autores relatam que a versão EfficientNet-B7 atinge o estado da arte com uma pontuação de 84,3% para a acurácia top-1 na ImageNet, ao mesmo tempo que é 8,4 vezes menor e 6,1 vezes mais rápida para o tempo de inferência do que a melhor rede convolucional concorrente.

Dosovitskiy et al. (2020) reaproveitam o modelo *Transformer*, originalmente inventado para o campo de Processamento de Linguagem Natural, realizando as modificações necessárias para adaptar o modelo para que seja possível realizar classificação de imagens. O nomeado VisionTransformer (ViT) supera todas as abordagens baseadas em convolução que eram até então

dominantes no meio. Três tamanhos diferentes são estudados para o modelo, nomeados pelos autores como ViT-B, ViT-L e ViT-H, como acrônimos para *Base*, *Large* e *Huge*, respectivamente.

3.3 DOR EM RECÉM-NASCIDOS

Brahnam et al. (2005) estudam a classificação de dor em recém-nascidos utilizando *Support Vector Machines* (SVM). Para isso, os autores desenvolvem o conjunto de dados *Infant Classification of Pain Expressions* (COPE/ICOPE), constituído de 204 imagens faciais de 26 recém-nascidos com 5 expressões provenientes de origens distintas: descanso, choro, estímulo de ar, fricção e dor.

Heiderich, Leslie e Guinsburg (2015) criam um método para classificação de dor neonatal em tempo real a partir de 16 pontos fiduciais capturados em imagens faciais de recém-nascidos, utilizados para mensurar 14 distâncias entre pontos e detectar movimentos faciais relativos a escala de dor NFCS. Para avaliação dos resultados, os autores criam um conjunto de dados com 360 imagens de recém-nascidos antes e após procedimentos dolorosos. O conjunto de dados levantado é referenciado como a base UNIFESP.

Zamzmi et al. (2019) estudam a avaliação automática de dor em recém-nascidos utilizando imagens faciais no conjunto de dados COPE. Os autores propõem e treinam uma rede neural convolucional para detecção de dor, nomeada *Neonatal Convolutional Neural Network* (N-CNN). O modelo criado alcançou acurácia de 91%.

Silva (2020) utiliza análise estatística multivariada para classificar imagens de recém-nascidos entre as classes com dor e sem dor. Os autores combinam a normalização das imagens e os métodos estatísticos PCA e MLDA para realizar a classificação, reportando uma taxa de acerto de aproximadamente 96% para a base COPE e 77% para a base UNIFESP.

Buzuti (2020) estuda a detecção de face de recém-nascidos comparando as redes N-CNN e ResNet50 propostas por Zamzmi et al. (2019) com uma versão da ResNet50. O autor relata que apesar de valores de acurácia de 87,2% e 78,7% para as bases COPE e UNIFESP respectivamente, é notável que os modelos podem aprender artefatos das imagens em vez de características faciais para realizar a classificação.

Domingues et al. (2021) criam um *pipeline* para separação de imagens de recém-nascidos em segmentos faciais relevantes para a classificação de dor. Os segmentos são inspirados na escala de dor NFCS e incluem as seguintes regiões: sobrancelhas, olhos, nariz, boca, sulco

nasolabiais, região entre olhos (fronte), testa e bochechas. As imagens geradas pelo método são recortes da face, detectadas por meio do modelo RetinaFace, normalizadas e com máscaras aplicadas para cada região.

Gkikas e Tsiknakis (2023) realizam uma revisão sistemática de métodos e bases de dados para avaliação de dor em múltiplas faixas etárias. Os autores comentam o papel promissor das técnicas de aprendizado profundo e destacam a necessidade de conjuntos de dados públicos de tamanho e qualidade suficiente para treinamento de modelos.

4 MATERIAIS E MÉTODOS

Neste capítulo serão apresentadas as bases de imagens (materiais) e as arquiteturas de redes neurais (métodos) empregados nos experimentos descritos no Capítulo 5.

4.1 MATERIAIS

As bases de imagens utilizadas neste trabalho não possuem acesso público. O acesso a base ocorreu pela parceria existente entre o Centro Universitário FEI e o grupo de pesquisa de dor da UNIFESP, junto às pesquisadoras principais do estudo desenvolvido por Heiderich, Leslie e Guinsburg (2015).

4.1.1 BASE DE IMAGENS UNIFESP 1

A base de imagens UNIFESP 1, criada por Heiderich, Leslie e Guinsburg (2015), é constituída de 360 imagens com resolução de 450x230 pixels, no formato bmp, de um grupo de 30 recém-nascidos. A base possui classificações pré e pós procedimento doloroso (punção) para cada recém-nascido. Todas as imagens possuem pontuação da escala NFCS preenchida por um grupo de profissionais da saúde. Ao todo, 122 imagens pertencem a classe com dor e 238 a classe sem dor.

A base possui aprovação do Comitê de Ética em Pesquisa da Universidade de São Paulo - UNIFESP-EPM, no ano de 2009, sob o número 1299/09.

A Figura 10 apresenta exemplos de imagens contidas na base UNIFESP1.

Figura 10 – Exemplo de imagens da base UNIFESP1.

Fonte: Heiderich, Leslie e Guinsburg (2015).

4.1.2 BASE DE IMAGENS UNIFESP 2

A base de imagens UNIFESP 2 utilizada neste trabalho é uma versão análoga da base de mesmo nome construída por Heiderich (2022). Em seu formato reduzido, é constituída de 10 imagens com resolução de 2322x4128 pixels no formato jpg. A base contem fotografias de 5 recém-nascidos diferentes, sendo que cada recém-nascido possui 1 imagem pertencente a classe com dor e 1 imagem pertencente a classe sem dor. Todas as imagens possuem equipamentos médicos obstruindo parcialmente a face.

A base possui aprovação do Comitê de Ética em Pesquisa da Universidade de São Paulo - UNIFESP-EPM, no ano de 2015, sob o número 0566/2015.

A Figura 11 apresenta exemplos de imagens contidas na base UNIFESP2.

Figura 11 – Exemplo de imagens da base UNIFESP1.

Fonte: Heiderich, Leslie e Guinsburg (2015).

4.2 MÉTODOS

4.2.1 VGG

A VGG é uma arquitetura convolucional para classificação de imagens projetada por Simonyan e Zisserman (2014), formada de múltiplas camadas convolucionais sequenciais com filtros de *kernel* 3×3, aplicadas com função de ativação ReLU e unidas por algumas camadas de *pooling*. O modelo pode possuir N camadas, e a rede resultante é usualmente nomeada como VGGN. O artigo original estuda a aplicação para N = 11, 13, 16 e 19.

A Figura 12 apresenta a arquitetura da VGG16, com valores exemplificando a inferência de uma imagem 224×224 com 3 canais. A resolução e o número de canais são decrescentes, conforme a aplicação de cada operação de convolução.

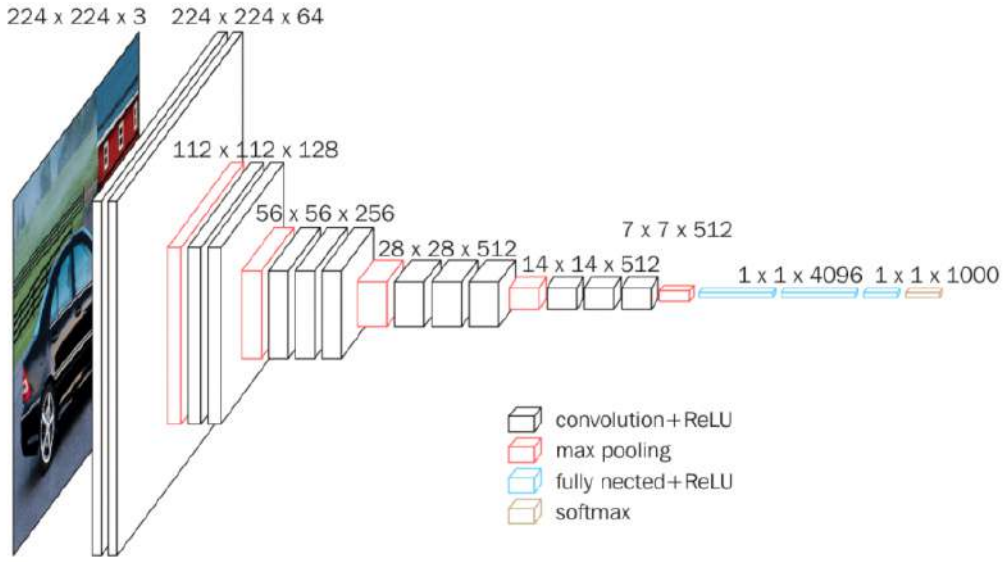


Figura 12 – Arquitetura da VGG16.

Fonte: Kaggle (2020).

4.2.2 ResNet

As redes residuais, ou ResNet, desenvolvidas por He et al. (2016), são arquiteturas convolucionais para classificação de padrões formadas pelo empilhamento dos chamados blocos residuais, seguidos de uma camada completamente conectada. Estes blocos, ilustrados na Figura 13, empregam o que os autores nomearam de *skip connections*, conexões que permitem que camadas sejam puladas durante o treino, resolvendo o problema conhecido como *vanishing gradient*, que afeta principalmente redes neurais profundas e impede o treinamento de progredir por conta de valores muito baixos de gradiente.

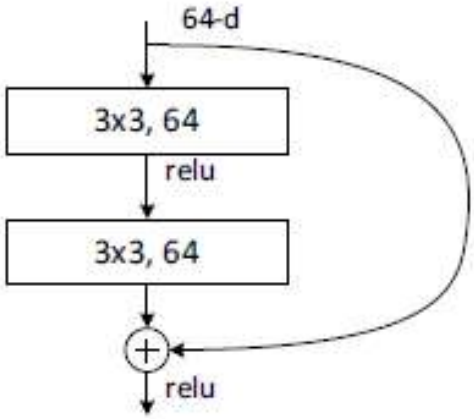


Figura 13 – Exemplo de bloco residual da ResNet.
Fonte: He et al. (2016).

A inovação trazida por este modelo permitiu a criação de redes muito mais profundas e, assim como a VGG, a ResNet não possui um número definido de camadas. Portanto, o uso de nomes como ResNet50 ou ResNet101 é uma prática comum para identificar a quantidade de camadas utilizadas no modelo.

4.2.3 InceptionV3

A InceptionV3, de Szegedy et al. (2015), é a terceira iteração da GoogLeNet (*Inception*), de Szegedy et al. (2014). O modelo possui três ideias principais, apresentadas em cada uma de suas versões.

A primeira versão origina o bloco *inception*, principal característica da arquitetura, partindo do pressuposto de que a mesma característica alvo da classificação, como um cachorro ou uma fruta, pode estar presente na imagem em diferentes tamanhos, o que é imprevisível para o modelo. Esta aplica convoluções com tamanhos de filtros (*kernels*) diferentes em paralelo, concatenando os filtros ao final, conforme ilustrado na Figura 14.

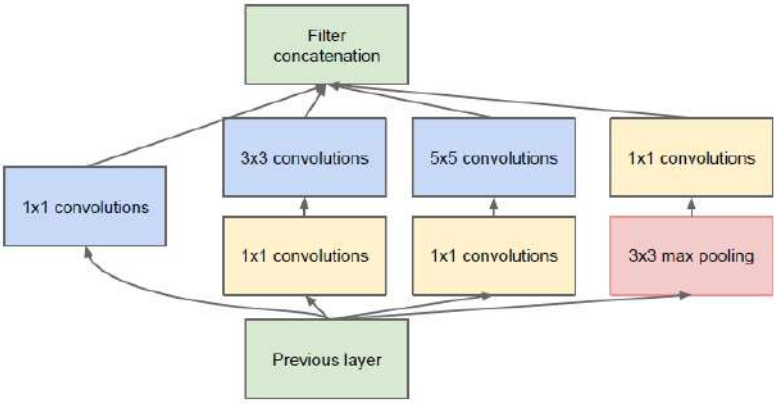


Figura 14 – Módulo *inception* da arquitetura GoogLeNet (Inception).

Fonte: Szegedy et al. (2014).

A segunda versão, InceptionV2, de Szegedy et al. (2015), aplica otimizações para a rede, com destaque para o uso da técnica de *Batch Normalization*, que acelera o tempo de treinamento e pode, segundo os autores, agir como um regularizador, ajudando a prevenir *overfitting*.

Por fim, a versão 3 do modelo, de Szegedy et al. (2015), estuda a fatorização das convoluções, para, por exemplo, transformar convoluções 5×5 em duas convoluções 3×3, o que reduz o número de parâmetros de treinamento de 25 para 18, conforme ilustrado na Figura 15, em que o bloco 5×5 visível na Figura 14 foi substituído por dois blocos 3×3. Esta modificação facilita o treinamento, por diminuir o número de parâmetros a serem aprendidos pelo modelo, e permite a criação de modelos mais profundos, com mais camadas. A Figura 16 apresenta a arquitetura InceptionV3 completa.

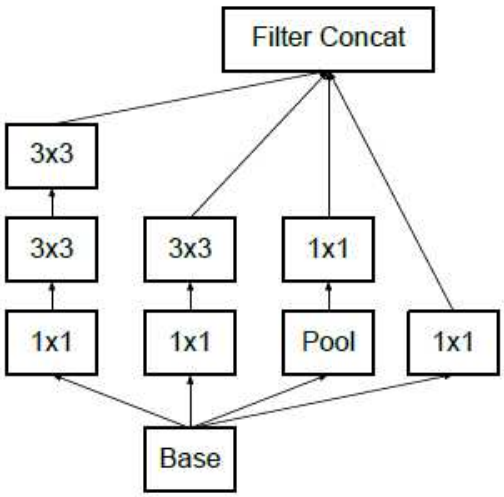


Figura 15 – Módulo *inception* da arquitetura InceptionV3.

Fonte: Szegedy et al. (2015).

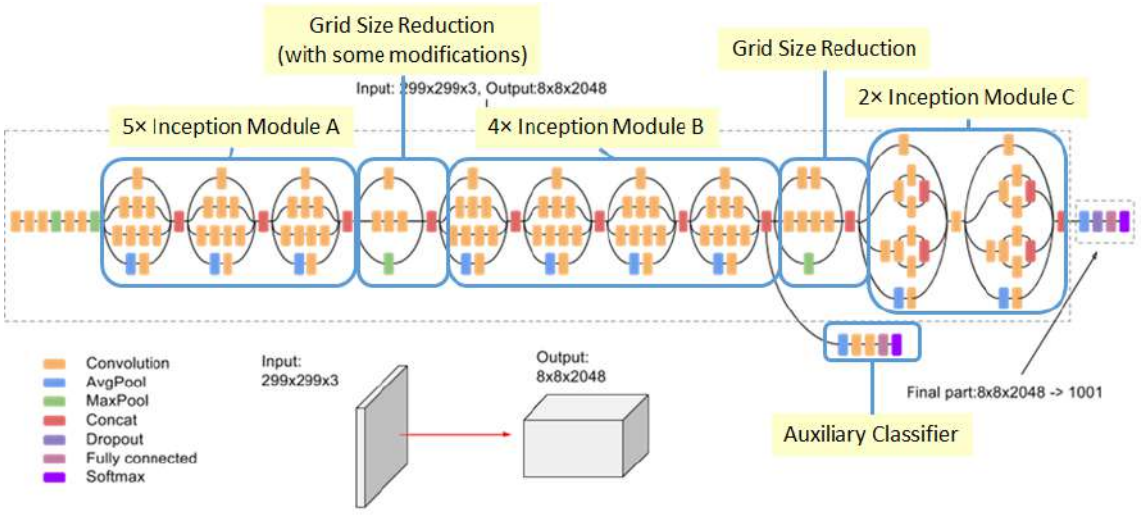


Figura 16 – Arquitetura da InceptionV3.

Fonte: Tsang (2018).

4.2.4 VisionTransformer

O VisionTransformer (ViT), de Dosovitskiy et al. (2020), utiliza como base a arquitetura *Transformer*, originalmente proposta por Vaswani et al. (2017) para Processamento de Linguagem Natural, porém adaptada para ser utilizada na classificação de imagens.

O funcionamento da arquitetura consiste na seguinte sequência de procedimentos, ilustrados na Figura 17.

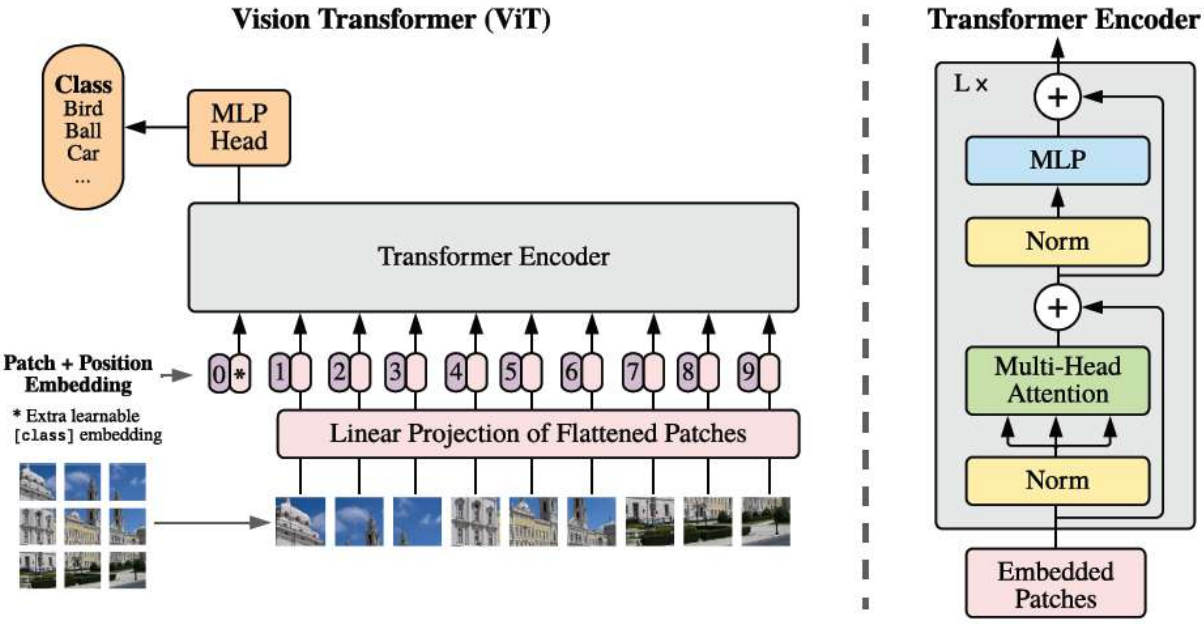


Figura 17 – Arquitetura do VisionTransformer (ViT).

Fonte: Dosovitskiy et al. (2020).

Primeiro, a imagem de entrada, de tamanho (HxWxC), em que H é a altura, W a largura e C o número de canais da imagem, é dividida em uma grade contendo N recortes quadrados (PxP), chamados *patches*, para $N = HW/P^2$. Cada recorte é achatado, de forma a obter um vetor com os mesmos valores do recorte original, porém de forma sequencial.

A seguir, de forma a obter um vetor codificado de dimensão D que será utilizado pelos estágios seguintes, os recortes achatados passam por uma projeção linear na forma de uma multiplicação por uma matriz E (acrônimo para *Embedding*) de dimensão P^2CxD, cujos valores são aprendidos pelo modelo. O resultado desta operação é adicionado a codificadores de posicionamento, vetores responsáveis por adicionar a cada vetor a informação sobre a posição do recorte na imagem original.

O conjunto de vetores resultantes da última etapa, junto a um vetor aprendido, chamado *class embeding*, equivalente ao presente no modelo Transformer original, é utilizado como entrada para o codificador (*encoder*) do Transformer, ilustrado à direita na Figura 17.

A principal característica do codificador desta arquitetura é o mecanismo de auto-atenção, responsável por calcular um valor numérico que define a relação entre cada *token* com os demais. *Tokens* são os valores resultantes do processamento de um recorte, logo a auto-atenção gera matrizes de correlação, chamadas mapas de atenção, que podem ser utilizadas para visualização da atenção do modelo sobre a imagem, conforme ilustrado na Figura 18.

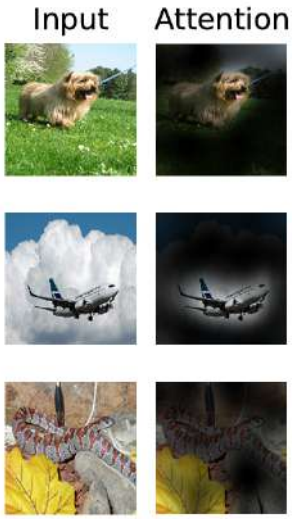


Figura 18 – Visualização do mecanismo de atenção do ViT.

Fonte: Dosovitskiy et al. (2020).

A Figura 17 apresenta o nome *Multi-Head Attention*, pois o processo mencionado no parágrafo anterior é repetido múltiplas vezes em paralelo, com cada processamento sendo de-

nominado um *attention head*. O resultado de todos estes processamentos é então concatenado como uma única saída.

O resultado do codificador é passado para um Multi-Layer Perceptron (MLP), uma arquitetura totalmente conectada, cujo resultado é a classe correspondente à imagem original.

4.2.5 RetinaFace

O RetinaFace é um modelo de rede neural proposto por Deng et al. (2020), para a tarefa de detecção de faces. Sua arquitetura, ilustrada na Figura 19, foi criada utilizando como base as pirâmides de características, de Lin et al. (2017), para analisar a imagem em 5 resoluções diferentes.

O modelo implementa a ResNet, de He et al. (2016), como *backbone*, uma rede responsável por realizar a extração de características da imagem e gerar mapas de características, uma forma de condensar e extrair informação relevante, passada a diante e utilizada pelas etapas seguintes do modelo.

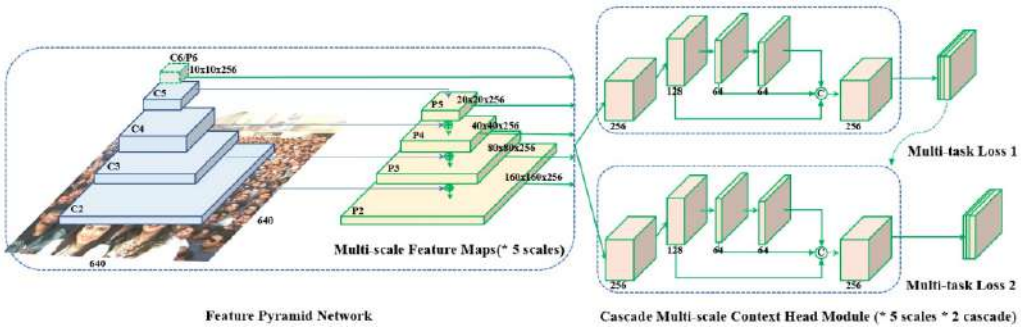


Figura 19 – Arquitetura do RetinaFace.

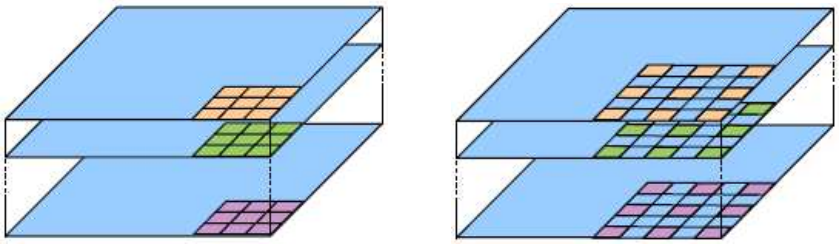
Fonte: Deng et al. (2020).

O modelo é treinado utilizando aprendizado supervisionado e apresenta a vantagem de ser categorizado como *single-shot*, ou seja, avalia a imagem uma única vez, tornando o processo de inferência mais rápido do que outros modelos.

4.2.6 DeepLabV3+

O modelo DeepLabV3+, de Chen et al. (2018), uma evolução do modelo DeepLabV3, é uma rede neural convolucional que faz uso da arquitetura codificador-decodificador (*encoder-decoder*). O codificador possui como *backbone* a arquitetura Xception, de (CHOLLET, 2017),

e explora o uso de *Spatial Pyramid Pooling* (SPP), Kaiming He Xiangyu Zhang e Sun (2014), uma estratégia de *pooling* que realiza convoluções em diferentes tamanhos paralelamente para permitir o uso de imagens de entrada com diferentes escalas e tamanhos, unida a operações de convolução dilatada (*atrous convolutions*), criada pelos autores e exemplificada na Figura 20, com diferentes distâncias entre pixels, valor indicado como *rate* na Figura 21.



(a) Operação de convolução convencional

(b) Operação de convolução dilatada com *rate*=2

Figura 20 – Exemplos de conjuntos de pontos fiduciais.

Fonte: Adaptado de Chen et al. (2018).

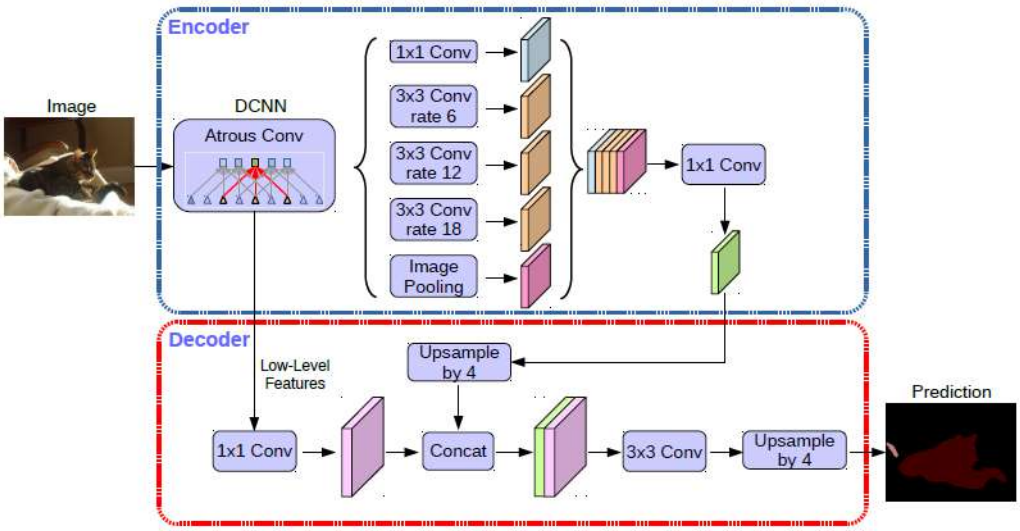


Figura 21 – Arquitetura do modelo DeepLabV3+.

Fonte: Chen et al. (2018).

O codificador é conectado a um decodificador, responsável por gerar a máscara de segmentação a partir dos mapas de características gerados. O resultado do processo é uma máscara de segmentação semântica.

O DeepLabV3+, apesar de não constituir mais o estado da arte, que atualmente é concedido ao modelo SAM, foi escolhido devido à disponibilidade pública de uma versão treinada para detecção de faces e com ajuste fino em imagens com a presença de oclusão por Voo, Jiang e Loy (2022).

4.2.7 SAM

O *Segment Anything Model* (SAM), de Kirillov et al. (2023), é um modelo desenvolvido pelo grupo Meta AI Research para realizar a tarefa de segmentação de regiões em imagens. O modelo foi treinado para diferenciar regiões e objetos sem classificá-los, gerando uma segmentação panóptica da imagem ou encontrando uma região de interesse determinada pelos dados de entrada da rede.

O treinamento do modelo, disponibilizado de forma gratuita pelos autores, foi realizado utilizando o conjunto de dados originado no mesmo artigo e nomeado SA-1B, o qual consiste de um grupo de aproximadamente 1 bilhão de máscaras e 11 milhões de imagens. O modelo apresenta boa capacidade de generalização *zero-shot*, ou seja, capacidade de segmentar objetos e regiões que não fizeram parte do treinamento sem a necessidade de ajuste fino.

A arquitetura, proposta em Kirillov et al. (2023) e ilustrada na Figura 22, pode ser dividida em 3 partes principais: codificador (*encoder*) de imagem, representado pela cor verde na Figura 22, codificador de dados de entrada, representado pela cor roxa, e o decodificador (*decoder*), representado pela cor laranja.

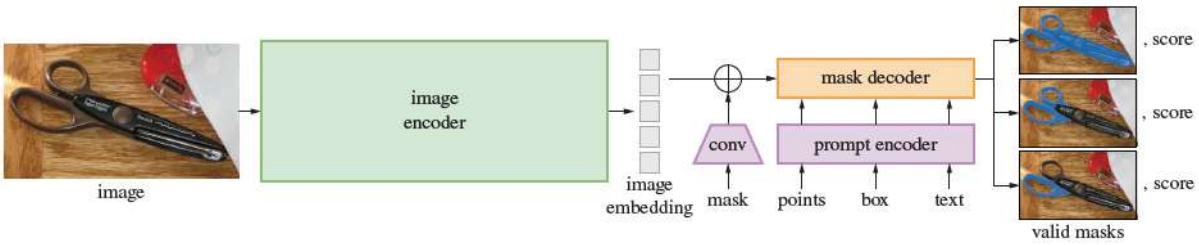


Figura 22 – Arquitetura do modelo SAM.

Fonte: Kirillov et al. (2023).

O codificador de imagem realiza a tarefa de encontrar uma representação em menor dimensão (denominada *embedding*) para a imagem de entrada. Para isso, é utilizado o codificador do *Vision Transformer* (Seção 4.2.4) pré-treinado com *Masked Autoencoder* (MAE), de He et al. (2022), e ajustado para imagens em alta resolução.

Os dados de entrada, ou *prompts*, segundo os autores, podem estar em dois possíveis formatos: esparsos ou densos. Os dados densos são máscaras, as quais são processadas pelo codificador de dados de entrada através de convoluções e o resultado é somado com o *embedding* da imagem, enquanto os dados esparsos podem ser um conjuntos de um ou mais pontos, caixas retangulares ou texto. No caso de pontos ou caixas retangulares, são utilizados codificações posicionais gerados pela técnica de Tancik et al. (2020), somados a *embeddings* aprendidos pelo

modelo para cada tipo de entrada. Para textos, o codificador CLIP, de Radford et al. (2021), é utilizado.

Por fim, o decodificador é responsável por transformar a representação dos dados em menor dimensão (*embeddings*), resultado dos codificadores, em um conjunto de máscaras. A saída do modelo é constituída por um conjunto de máscaras para resolver problemas com ambiguidade sobre o objeto alvo da segmentação.

O modelo SAM foi escolhido como uma das opções de segmentação a serem testadas por constituir o novo estado da arte para a tarefa de segmentação de imagens, incluindo a segmentação de faces (DOMINGUES et al., 2023).

4.2.8 Métricas de avaliação

Este trabalho selecionou o coeficiente de similaridade Dice, Seção 2.2.4.4, como métrica para avaliação dos métodos de segmentação estudados devido a sua capacidade de medir a sobreposição entre a máscara padrão-ouro e pois, diferente do coeficiente de Jaccard, Seção 2.2.4.2, seu resultado varia entre os valores 0 e 1, podendo ser facilmente interpretado como uma porcentagem.

As etapas que envolveram modelos de classificação coletaram resultados de precisão, *recall*, acurácia e pontuação F1, Capítulos 2.3.1, 2.3.2, 2.3.4 e 2.3.3 respectivamente. Os resultados foram discutidos a partir dos valores de pontuação F1 por esta representar um equilíbrio na consideração de falsos positivos e falsos negativos, ambos importantes no meio médico para reconhecimento de dor.

5 EXPERIMENTOS E RESULTADOS

Visando compreender a influência de artefatos obstruindo a face para modelos de classificação da expressão de dor que utilizam imagens faciais, este trabalho avaliará paralelamente duas bases de imagens, a UNIFESP 1, que possui exclusivamente imagens de face livre, e a UNIFESP 2, que possui exclusivamente imagens de faces obstruídas.

Os experimentos foram distribuídos em dois grupos e serão descritos junto a seus resultados em seções separadas. Aqueles relacionados a segmentação serão apresentados primeiro, seguidos dos experimentos que envolveram classificação da expressão de dor. O arcabouço computacional criado utilizou python com as bibliotecas pytorch, para as redes neurais, e opencv, para manipulação de imagens.

5.1 SEGMENTAÇÃO

Os experimentos de segmentação visam encontrar a melhor estratégia para segmentação facial, aplicada aqui em recém-nascidos, disponível atualmente.

A Figura 23 apresenta um diagrama ilustrando a metodologia completa de segmentação.

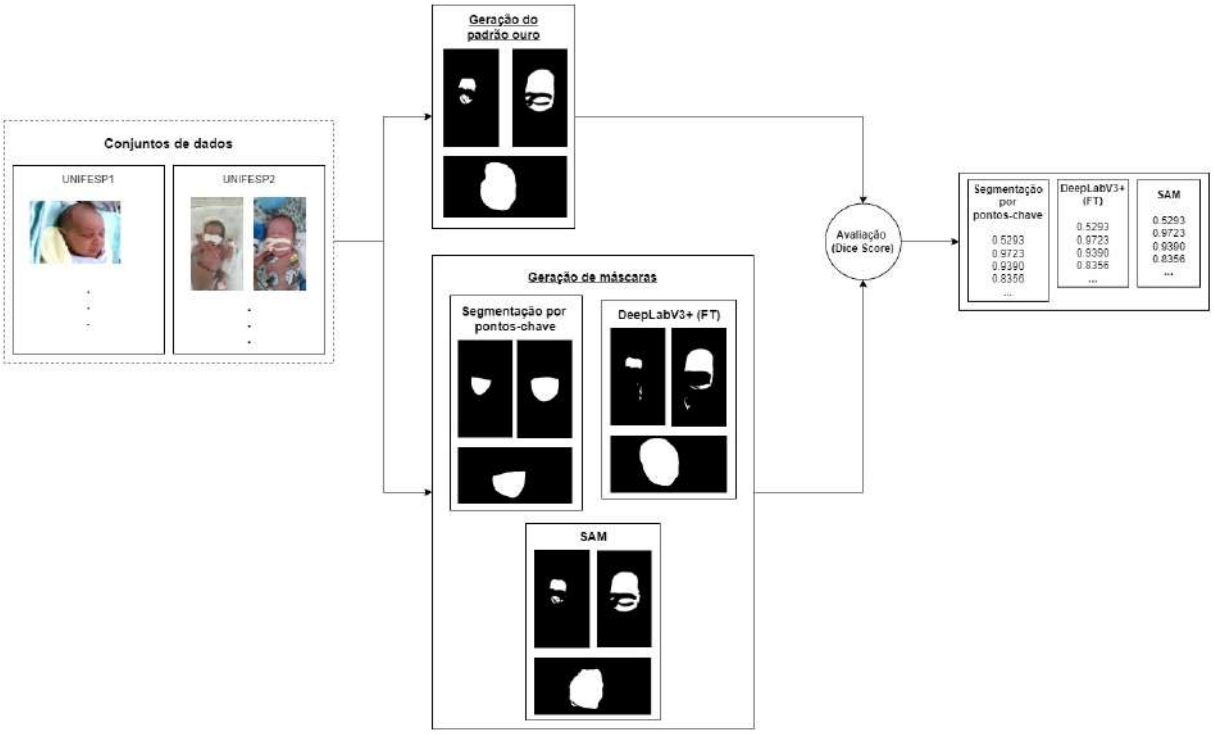


Figura 23 – Diagramas da metodologia de segmentação.

Fonte: Autor.

Três alternativas para segmentação da face do recém-nascidos foram exploradas. São estas:

a) **Segmentação por pontos-chave:** Consiste na detecção de um conjunto de pontos discriminantes da face, seguido da criação e preenchimento de um polígono convexo para gerar a máscara da face. Esta abordagem foi escolhida como uma opção simples e com resultados estáveis para servir como referência na comparação dos algoritmos subsequentes. Ilustrada na Figura 24a;

b) **DeepLabV3+ (FT):** Como um representante para os modelos convolucionais, extensamente utilizados para segmentação de imagens, o modelo DeepLabV3+ foi escolhido por ser uma publicação recente e ter passado pelo processo de ajuste fino (*Fine-Tuning*) para melhorar a segmentação facial em casos de oclusão. O modelo utilizado provêm do trabalho de Voo, Jiang e Loy (2022), o qual disponibilizou publicamente os pesos ajustados para a tarefa. Ilustrado na Figura 24b;

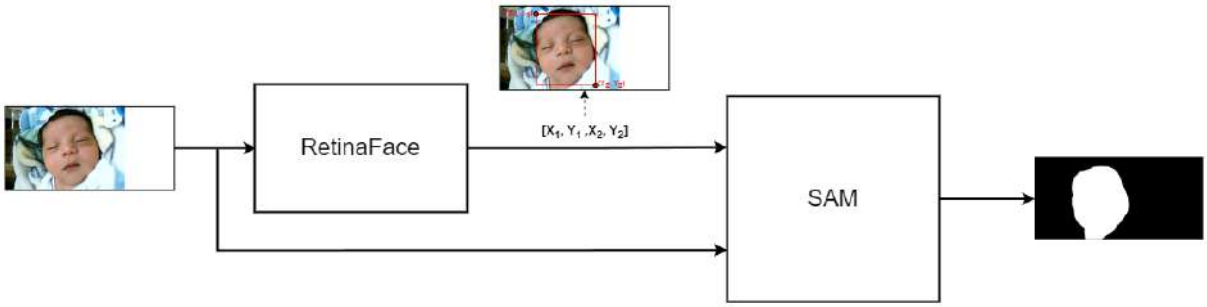
c) **SAM:** O modelo SAM foi selecionado para os experimentos por compor o atual estado da arte para segmentação de imagens. Para utilizar este modelo para a tarefa de segmentação facial, primeiro foi necessário aplicar o detector facial RetinaFace de Deng et al. (2020), pois esse detector possui resultados consistentes para imagens de recém-nascidos (BUZUTI, 2020) (DOMINGUES et al., 2021). A coordenada da caixa retangular contendo a face, resultado da inferência do modelo, juntamente com a imagem do recém-nascido, são utilizadas como entrada para o SAM. Este modelo não foi treinado para segmentação de faces especificamente, mas utilizando a caixa retangular como indicador da face como região alvo da segmentação foi possível que o SAM cumprisse a tarefa desejada. Uma vez que este modelo, para resolver problemas com ambiguidade, apresenta sempre três máscaras como saída, aquela que possuísse o maior índice de confiança era selecionada como resultado final automaticamente. Ilustrado na Figura 24c.

5.1.1 Padrão-ouro

Para geração do padrão-ouro de segmentação, o modelo SAM foi utilizado para gerar máscaras de todas as imagens de ambas as bases da UNIFESP. Estas máscaras foram então corrigidas manualmente a nível de pixel e salvas para serem utilizadas como padrão-ouro. Um

(a) Segmentação por pontos chave

(b) DeepLabV3+ (FT)



(c) SAM

Figura 24 – Diagramas de funcionamento dos métodos de segmentação avaliados.

total de 375 máscaras foram criadas, sendo 365 correspondentes a base UNIFESP 1 e 10 à UNIFESP 2.

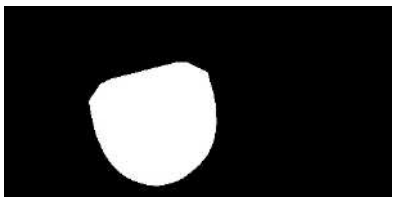
5.1.2 Resultados

As Figuras 25, 26 e 27 expõem três amostras das bases UNIFESP 1 e UNIFESP 2 junto ao padrão-ouro e aos resultados obtidos pela aplicação de todos os algoritmos de segmentação escolhidos, para realização de uma avaliação qualitativa.

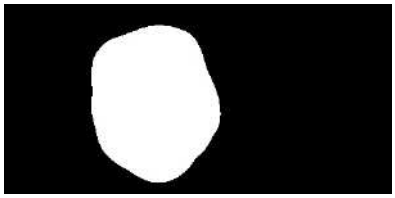
A Figura 28 apresenta uma sobreposição entre imagens da base UNIFESP 2 e as máscaras geradas por cada um dos modelos, permitindo uma inspeção mais detalhada das regiões segmentadas.

(a) Imagem original

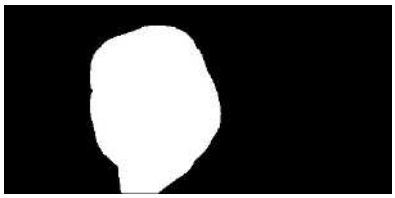
(b) Máscara *Ground truth*



(c) Segmentação por *landmarks*



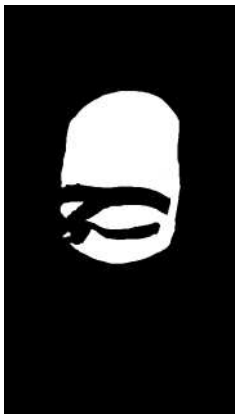
(d) DeepLabV3+ (AF)



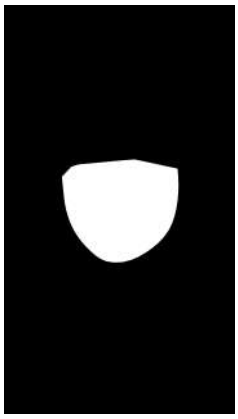
(e) SAM (máscara de maior pontuação)

Figura 25 – Segmentação de uma imagem proveniente do conjunto de dados UNIFESP 1 (sem oclusão) utilizando segmentação por pontos-chave e os modelos DeepLabV3+ e SAM.

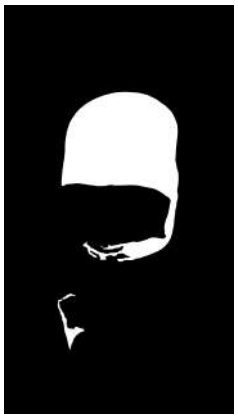
(a) Imagem original



(b) Máscara *Ground truth*



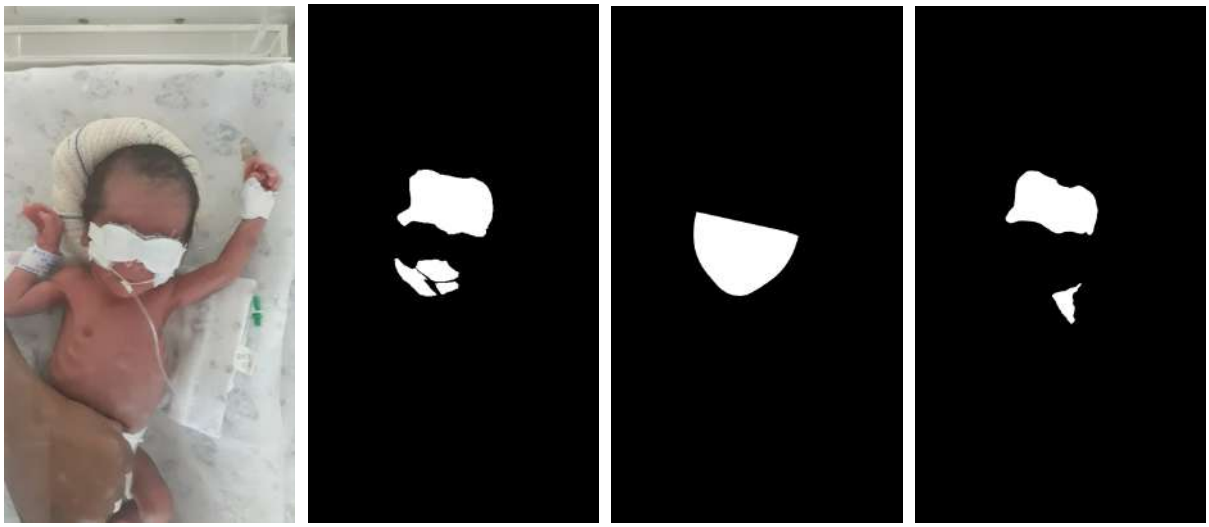
(c) Segmentação por *landmarks*



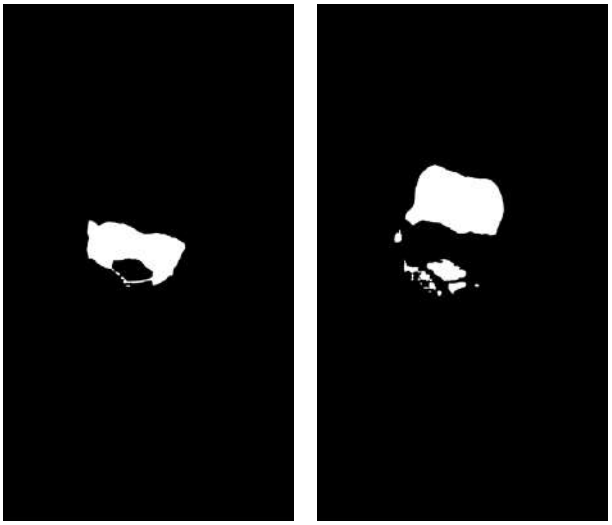
(d) DeepLabV3+ (FT)

(e) SAM (máscara de maior pontuação)

Figura 26 – Segmentação de uma imagem proveniente do conjunto de dados UNIFESP 2 (com oclusão) utilizando segmentação por pontos-chave e os modelos DeepLabV3+ e SAM.



(a) Imagem original (b) Máscara *Ground truth* (c) Segmentação por landmarks (d) DeepLabV3+ (FT)



(e) SAM (segmentação do artefato) (f) SAM (segmentação da face)

Figura 27 – Segmentação de uma imagem proveniente do conjunto de dados UNIFESP 2 (com oclusão) utilizando segmentação por pontos-chave e os modelos DeepLabV3+ e SAM.

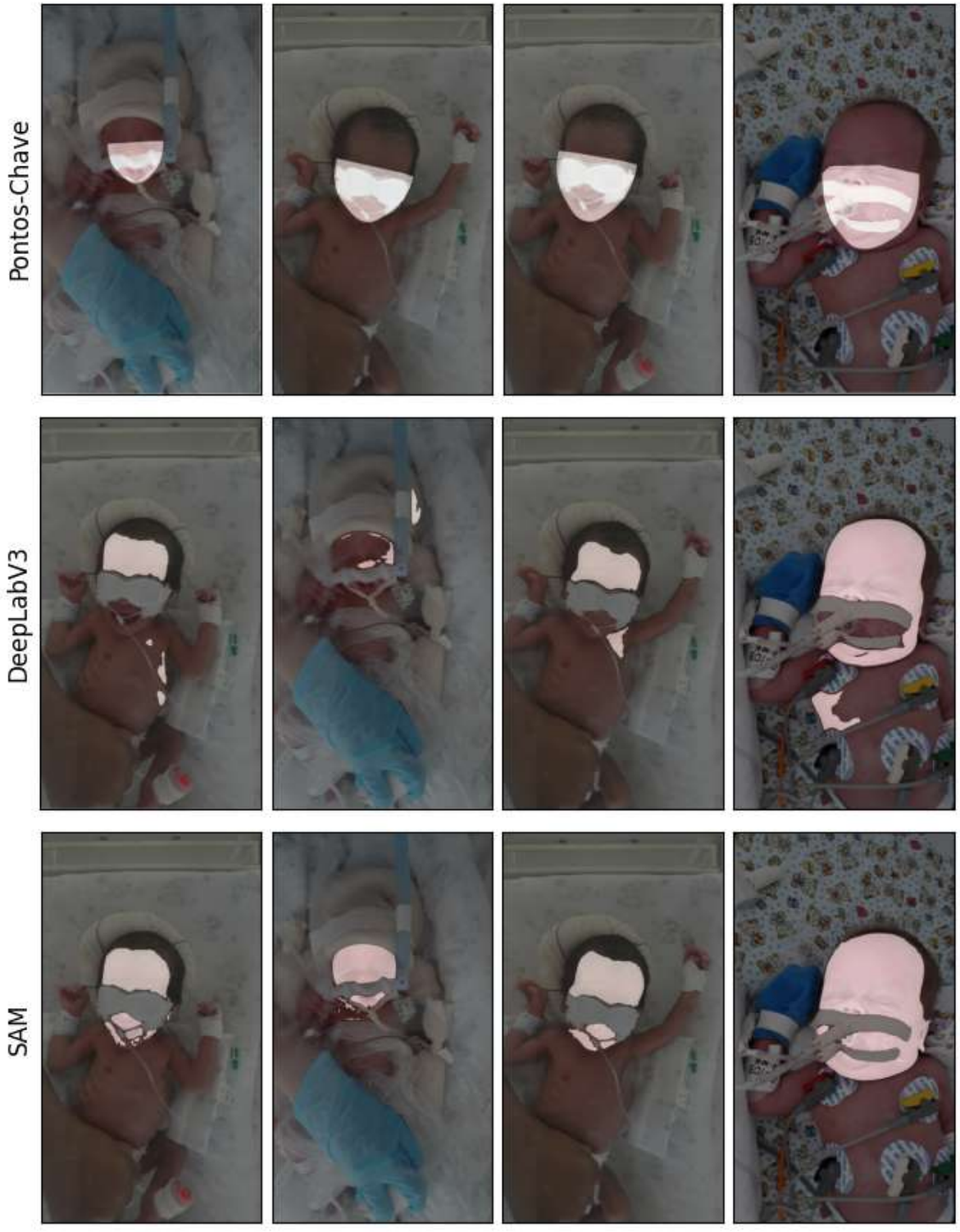


Figura 28 – Sobreposição entre máscaras geradas por cada modelo e a respectiva imagem.

Fonte: Autor.

5.2 CLASSIFICAÇÃO

Experimentos relacionados a classificação visaram determinar a influência da segmentação ao ser aplicada em imagens antes da classificação da expressão de dor por diferentes modelos.

Os modelos treinados para a tarefa de classificação da expressão de dor foram: VGG16, Resnet50, InceptionV3 e ViT. Todos os modelos mencionados seguiram o mesmo protocolo de treinamento, descrito a seguir.

a) Ajuste fino para a tarefa de classificação da expressão de dor realizado a partir de um modelo base treinado com a base ImageNet;
b) Validação cruzada utilizando KFold com K=5;
c) Treinamento por 200 *epochs*;
d) Melhor modelo determinado por meio da pontuação F1 de validação;
e) Decaimento dos pesos (*weight decay*) de 0;
f) *Batch* de treino de 10 imagens.

A taxa de aprendizado (*learning rate*) foi determinada individualmente para cada caso de acordo com o valor utilizado para treino do modelo base. Os valores finais utilizados foram: $0{,}00005$ (VGG16), $0{,}001$ (Resnet50), $0{,}0002$ (InceptionV3) e 0.0002 (ViT).

5.2.1 Resultados

Os resultados a seguir apresentam três possíveis legendas indicando o pré-processamento aplicado a imagem de entrada do modelo durante o processo de treinamento. As legendas UNIFESP1, UNIFESP1-Faces e UNIFESP1-Faces-SAM indicam respectivamente: Nenhum pré-processamento, o recorte da face (encontrada utilizando o RetinaFace, Seção 4.2.5) e recorte da face com aplicação da segmentação facial utilizando SAM. Exemplos de cada forma de pré-processamento podem ser vistos na Figura 29.

(a) Pré-processamento UNIFESP1

(b) Pré-processamento UNIFESP1-Faces

(c) Pré-processamento UNIFESP1-Faces-SAM

Figura 29 – Exemplo das formas de pré-processamento estudadas.

Fonte: Autor.

A Tabela 4 contém a média entre as pontuações F1 de classificação após o treinamento dos modelos utilizando K-Fold com k=5. Resultados completos, com as demais métricas de treinamento, incluindo precisão, recall e acurácia, não discutidas no Capítulo 6 podem ser encontrados no Apêndice A.

	Vgg16	**Resnet50**	**InceptionV3**	**ViT**
UNIFESP1	0.854±0.041	0.862±0.036	0.882±0.054	0.896±0.036
UNIFESP1-Faces	**0.889±0.042**	**0.892±0.011**	0.883±0.038	**0.912±0.030**
UNIFESP1-Faces-SAM	0.869±0.031	0.888±0.016	**0.884±0.031**	0.911±0.023

Tabela 4 – Comparação entre estratégias de pré-processamento para cada modelo utilizando a pontuação F1 média de validação.

Fonte: Autor.

A Tabela 5 também apresenta as pontuações F1, porém com destaque em negrito para os modelos que obtiveram o melhor desempenho.

Aplicando os modelos de classificação treinados no conjunto de dados UNIFESP 2, de imagens com obstrução facial, e medindo a acurácia para cada alternativa de pré-processamento foi obtida a Tabela 6. A Figura 30 apresenta um exemplo de cada pré-processamento aplicado a imagens do conjunto UNIFESP2.

	UNIFESP1	**UNIFESP1-Faces**	**UNIFESP1-Faces-SAM**
Vgg16	0.854±0.041	0.889±0.042	0.869±0.031
Resnet50	0.862±0.036	0.892±0.011	0.888±0.016
InceptionV3	0.882±0.054	0.883±0.038	0.884±0.031
ViT	**0.896±0.036**	**0.912±0.030**	**0.911±0.023**

Tabela 5 – Comparação entre modelos para cada estratégia de pré-processamento utilizando a pontuação F1 média de validação.

Fonte: Autor.

	UNIFESP1	UNIFESP1-Faces	UNIFESP1-Faces-SAM
Vgg16	55.56%	44.44%	66.67%
Resnet50	66.67%	55.56%	66.67%
InceptionV3	55.56%	55.56%	55.56%
ViT	66.67%	55.56%	66.67%

Tabela 6 – Acurácia dos modelos treinados com cada estratégia de pré-processamento aplicados ao conjunto de imagens com oclusão.

Fonte: Autor.

(a) Pré-processamento UNIFESP1

(b) Pré-processamento UNIFESP1-Faces

(c) Pré-processamento UNIFESP1-Faces-SAM

Figura 30 – Exemplo de imagem da base UNIFESP2 com as formas de pré-processamento aplicadas.

Fonte: Autor.

As Tabelas 7 e 8 contém resultados para testes de comparação de média entre as pontuações F1 dos modelos, uma prática atípica no meio de *machine learning*, mas útil para verificar a relevância das diferenças descobertas. A Tabela 7 utilizou o Teste T, com hipótese nula de duas amostras possuírem a mesma média, portanto os valores em destaque na tabela, com p valor inferior a 0,05, representam os casos em que as médias são diferentes. Como uma alternativa não paramétrica para a comparação, a Tabela 8 foi criada aplicando o Teste de Wilcoxon com a mesma hipótese nula mencionada anteriormente.

A seguir estão dispostos os tempos de inferência e tamanho do arquivo de cada modelo gerado. A inferência foi realizada em um computador com a seguinte especificação: proces-

		UNIFESP1	UNIFESP1-Faces	UNIFESP1-Faces-SAM
Vgg16	**Resnet50**	0.741	0.888	0.254
	InceptionV3	0.373	0.814	0.474
	ViT	0.126	0.354	**0.041**
Resnet50	**InceptionV3**	0.506	0.624	0.778
	ViT	0.181	0.203	0.102
InceptionV3	**ViT**	0.657	0.218	0.148

Tabela 7 – P valores do Teste T para comparação de médias entre modelos. Destacados os p valores menores do que 0,05

Fonte: Autor.

		UNIFESP1	UNIFESP1-Faces	UNIFESP1-Faces-SAM
Vgg16	**Resnet50**	0.812	0.715	0.109
	InceptionV3	0.438	0.812	0.068
	ViT	0.125	0.625	0.062
Resnet50	**InceptionV3**	0.438	0.625	0.715
	ViT	0.273	0.312	0.062
InceptionV3	**ViT**	0.812	0.312	0.144

Tabela 8 – P valores do Teste de Wilcoxon para comparação de médias entre modelos.

Fonte: Autor.

	Tamanho do arquivo do modelo (KB)	Tempo de inferência médio (s) (imagem completa)	Tempo de inferência médio (s) (recorte da face)
Vgg16	524.498	0.2137	0.2091
Resnet50	92.164	0.3042	0.2758
InceptionV3	**85.435**	**0.1442**	**0.1186**
ViT	335.224	0.2910	0.2630

Tabela 9 – Dados dos modelos.

Fonte: Autor.

sador AMD Ryzen 5 5600X, placa de vídeo Geforce RTX3070 8GB VRAM GDDR6, 32GB RAM DDR4 3200MHz e Sistema Operacional Windows 10.

5.2.2 Mapas de atenção

Visto a utilidade dos mapas de atenção, inerentes ao modelo Vision Transformer, devido a sua capacidade de demonstrar o peso de cada região em imagens processadas, algumas amostras destes mapas de imagens da base de dados UNIFESP 2 foram geradas para discussão. Os mapas, presentes na Figura 31, representam a média dos *attention heads* na última camada do decodificador do ViT.

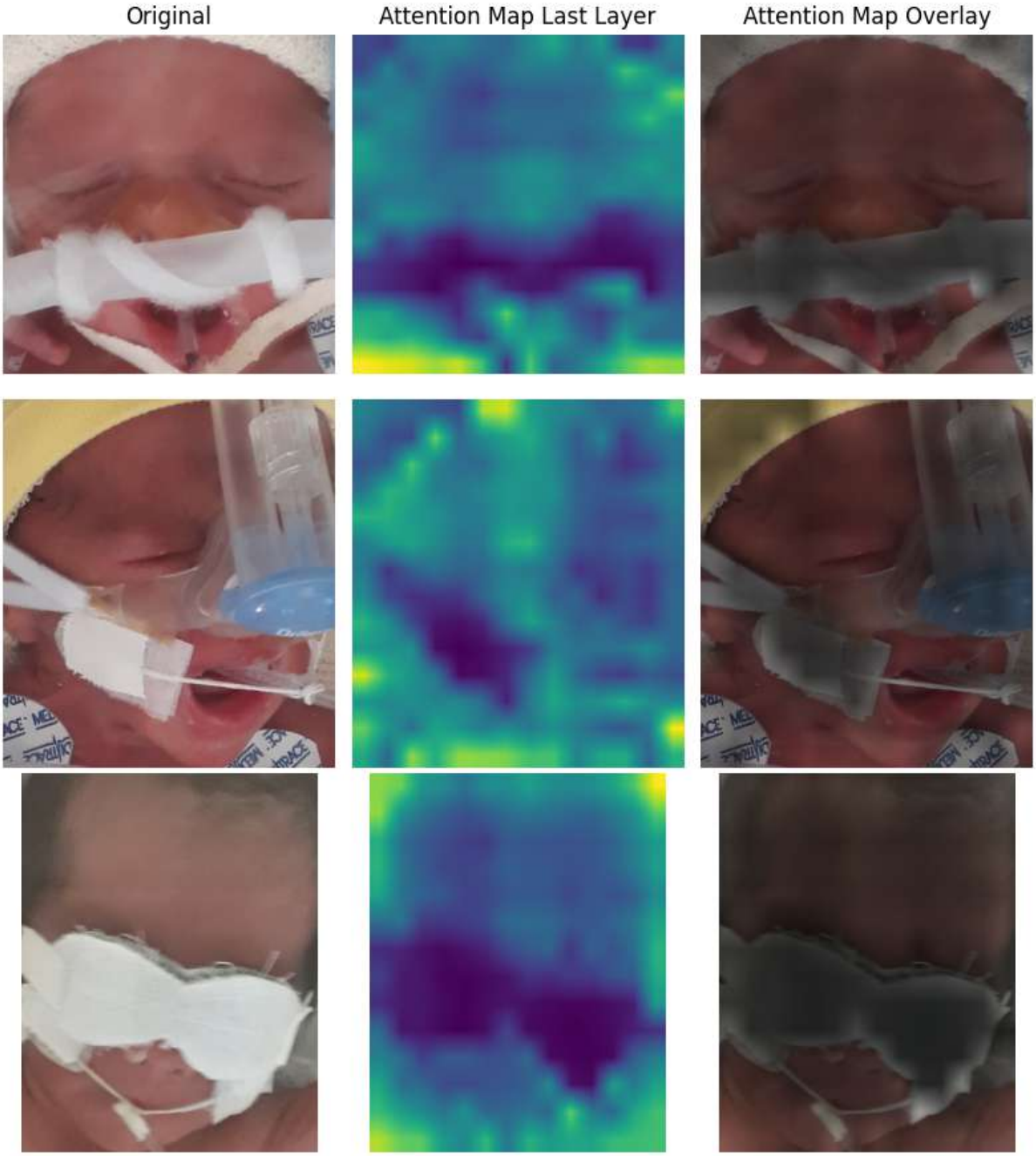


Figura 31 – Exemplos de mapa de atenção extraídos do ViT utilizando imagens da base UNIFESP2. Cores indicam o valor de atenção dada para a região, com tons variando de escuro (preto, menor atenção) para claro (amarelo, maior atenção)

Fonte: Autor.

6 DISCUSSÃO

Neste capítulo serão discutidos os resultados obtidos no Capítulo 5, separados entre segmentação (Seção 6.1) e classificação (Seção 6.2).

6.1 SEGMENTAÇÃO

A seguir serão apresentadas as análises relativas aos experimentos e comparações entre os três métodos de segmentação facial.

6.1.1 Segmentação por pontos-chave

Conforme demonstrado na Figura 25c, o método de segmentação a partir de pontos fiduciais gerou resultados consistentes e precisos para imagens sem a presença de oclusão, com dois pontos negativos para este caso. Primeiro, uma perda de detalhes no contorno na face, perceptível em todas as imagens. Além disso, a incapacidade de detectar pontos chave na parte superior da cabeça incapacita esta abordagem de detectar a fronte do recém-nascido.

Para imagens com a presença de artefatos, o método demonstra, além dos problemas mencionados anteriormente, mais dois pontos importantes. Devido à desconsideração de artefatos na face e ao uso apenas dos pontos mais externos desta, inerentes ao método, a máscara resultante é sempre uma região convexa totalmente preenchida, impedindo o contorno de tais artefatos, conforme é possível notar nas Figuras 26c e 27c, quando comparadas as Figuras 26b e 27b.

6.1.2 DeepLabV3+ (FT)

Os resultados obtidos para o conjunto de dados sem oclusão se demonstraram inconsistentes. Muitas máscaras geradas se encontraram próximas do padrão-ouro, gerando contornos extremamente precisos. Porém, alguns problemas foram comumente encontrados mesmos sem a presença de oclusão, como segmentação de partes do corpo com a pele exposta, como pescoço, braços e tórax e segmentação incompleta da face. As mesmas características foram encontradas com maior incidência no conjunto de dados com oclusão, conforme ilustrado nas Figuras 26d e 27d.

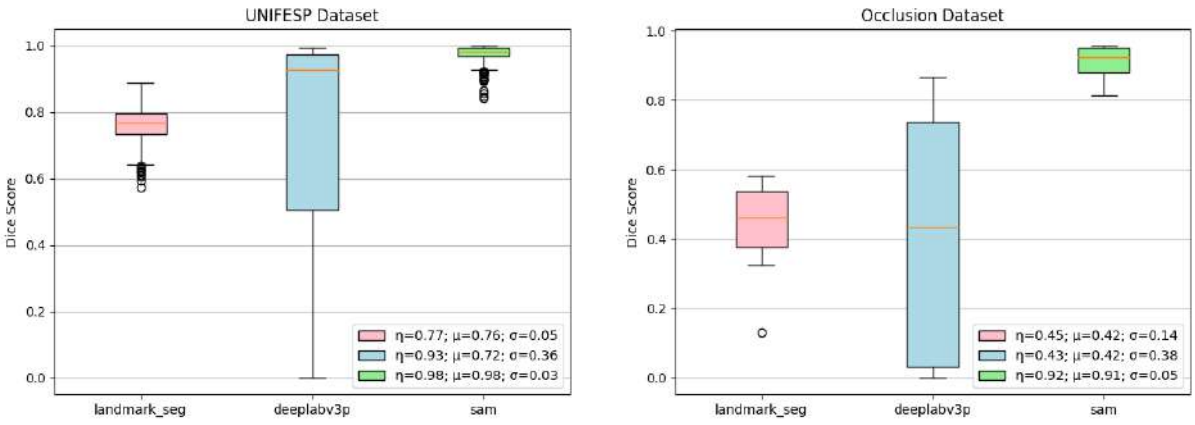
6.1.3 SAM

Ao avaliar qualitativamente as máscaras com maior índice de confiança deste modelo, o mesmo apresenta grande precisão das máscaras geradas. Porém, apesar de, em sua maioria, os resultados gerados para imagens sem oclusão se demonstrarem quase indistinguíveis do padrão-ouro, algumas máscaras apresentaram segmentação de partes expostas do corpo, de forma semelhante ao modelo DeepLabV3+ (FT), como demonstrado na Figura 25e.

O resultado para imagens com oclusão manteve-se preciso, como pode ser visto na Figura 26e, porém uma características demonstrou-se comum para este caso, a segmentação de algumas regiões apresentou um efeito de granulação, com alguns pedaços da máscara faltando, como visível na região do queixo da Figura 27f.

Em uma das imagens, uma das três máscaras geradas pelo modelo segmentou o artefato que ocluía a face do recém-nascido, em vez da face propriamente dita, conforme apresentado na Figura 27e. Este efeito pode demonstrar-se mais comum e problemático em conjuntos maiores de dados com oclusão, porém a máscara não possuía o maior índice de confiança e logo não interferiu nos resultados desta pesquisa.

6.1.4 Comparação



(a) Conjunto de dados sem oclusão (UNIFESP1) (b) Conjunto de dados com oclusão (UNIFESP2)

Figura 32 – Pontuação *Dice* para todos os métodos em ambos os conjuntos de dados. Os valores indicados são: Mediana η; Média μ; Desvio Padrão σ.

A Figura 32a apresenta os resultados quantitativos dos três métodos (DOMINGUES et al., 2023). Para o conjunto de dados UNIFESP 1, que não possui oclusão da face, podemos notar que:

a) O método de segmentação por pontos-chave, anotado como $landmark\_seg$, é consistente, apresentando desvio padrão de apenas 0,05, além de possuir média e mediana altas (0,76 e 0,77 respectivamente) que são limitadas principalmente pela incapacidade de detectar a região da fronte, reduzindo a região de intersecção entre a máscaras geradas e o padrão ouro e, por consequência, reduzindo a pontuação Dice (Seção 2.2.4.4);

b) O DeepLabV3+ (FT), anotado como $deeplabv3p$, possui uma mediana alta (0,93), apresentando uma grande quantidade de predições muito próximas ao padrão ouro, porém o método é instável, pois possui desvio padrão σ =0,36 e distribuição não-normal visto que sua média é muito inferior à mediana;

c) Por fim, o SAM demonstrou-se tanto um método acurado que segue uma distribuição normal, com média e mediana $\eta = \mu$ =0,98, quanto estável, com desvio padrão σ =0,03, inferior a segmentação por pontos-chave.

Ao modificar o *dataset* para o de imagens com oclusão (UNIFESP 2), todos os métodos apresentam o mesmo comportamento, a redução da média, redução da mediana e aumento do desvio padrão. Os aumentos são principalmente notáveis para o DeepLabV3+ (FT), por sua mediana passar a ser menor do que a do método de detecção por pontos-chave, e para o SAM, visto que a pontuação Dice média e mediana, mesmo em imagens mais desafiadoras, continuam superiores a todos os demais métodos e acima de 0,90.

Logo, após as análises realizadas, o modelo SAM foi selecionado como o mais adequado para ser utilizado em ambos os cenários (com e sem a presença de objetos obstruindo parcialmente a face).

6.2 CLASSIFICAÇÃO

Os valores em destaque na Tabela 4 demonstram que, apesar de existir um ganho pequeno em recortar a região da face antes de realizar a classificação, os modelos são capazes de aprender a diferenciar a face do fundo com robustez o bastante para que a aplicação da segmentação não gere diferença nos resultados.

Observando os valores em negrito na Tabela 5 parece evidente a superioridade da arquitetura *Vision Transformer* na classificação da dor, independente da estratégia de pré-processamento utilizada, porém partir dos resultados dos Testes T e de Wilcoxon apresentados nas Tabelas 7

e 8, pode-se concluir que não há diferença estatística relevante o suficiente para considerarmos diferenças entre os modelos.

Neste caso, outros fatores podem ser mais relevantes, como os em destaque na Tabela 9. O modelo InceptionV3 destaca-se caso a aplicação necessite de modelos que ocupem menos memória do sistema, como para o uso em hardware embarcado, ou para casos em que o tempo de inferência seja crucial. O modelo Vision Transformer, apesar de ocupar mais espaço em disco e possuir tempos de inferência maiores, contém *attention heads*, componentes que podem ser facilmente extraídos e utilizados para auxiliar na interpretação da decisão do modelo para a classificação de uma imagem qualquer.

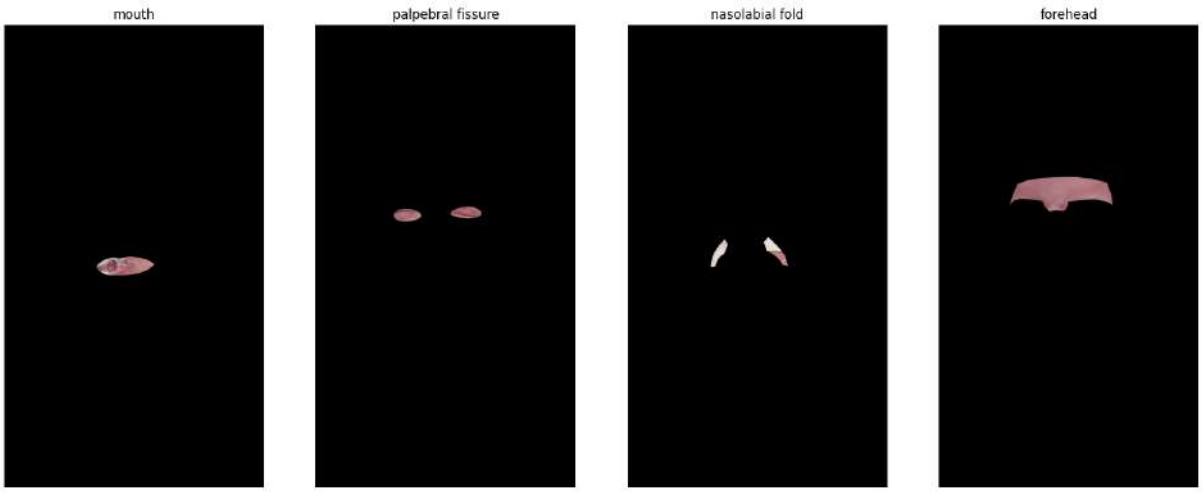
Como pode ser observado na Tabela 6, os valores de acurácia não excederam 66.67%, indicando que a técnica ainda não é suficiente para funcionar em casos de alta oclusão facial, como os de imagens de UTIs neonatais do conjunto de imagens UNIFESP 2.

Observando os mapas de atenção do Vision Transformer aplicados a imagens do UNIFESP 2, presentes na Figura 31, dois pontos são notáveis. Primeiro, os artefatos clínicos presentes na face dos recém-nascidos são vistos majoritariamente em tons mais escuros e, portanto, não recebem atenção do modelo, indicando que este é capaz de analisar a imagem sem a necessidade de segmentação, o que foi confirmado na Tabela 4 pois a segmentação destes artefatos não gerou nenhum aumento na pontuação F1. Além disso, é possível notar em todos os exemplos, que o modelo, após o processo de ajuste fino para a classificação de dor com a base de imagens UNIFESP 1, aprendeu um padrão de dados utilizando os valores próximos às bordas das imagens, pois em todos os exemplos apresentados a atenção é grande (tons mais claros, próximos de amarelo) em pontos próximos a extremidade, mesmo que estes não contenham informação relevante para a classificação.

Para melhorar o desempenho da classificação em casos de oclusão, uma possível solução a ser estudada é o uso de estratégias computacionais que apliquem diretamente escalas como a NFCS para tentar alcançar melhores acurácias nestas imagens. Uma possível alternativa para se atingir este objetivo é o aprimoramento de técnicas como o mosaico facial (DOMINGUES et al., 2021), através do uso de modelos de regressão com mais pontos-chave, conforme exemplificado na prova de conceito apresentada na Figura 33, para que funcione de forma mais robusta em imagens com oclusão, permitindo a extração, validação (verificação se a região faz mesmo parte da face) e classificação de regiões faciais independentes.

(a) 478 pontos fiduciais

(b) Mosaico facial



(c) Regiões segmentadas a partir do mosaico

Figura 33 – Prova de conceito de um refinamento do mosaico facial (DOMINGUES et al., 2021), utilizando MediaPipe (LUGARESI et al., 2019).

Fonte: Autor.

7 CONCLUSÃO

Este trabalho visou verificar as capacidades, ganhos e limitações no uso da segmentação facial destinada a tarefa de avaliação da expressão de dor neonatal. Para isso, três modelos (segmentação por pontos chave, DeepLabV3+ e SAM) foram testados em cenários com e sem a presença de oclusões faciais e o melhor foi utilizado para segmentar faces de recém nascidos utilizadas no treinamento e classificação da expressão de dor com quatro diferentes classificadores (Vgg16, Resnet50, InceptionV3 e ViT).

As bases de dados UNIFESP 1 e UNIFESP 2 foram utilizadas para imagens faciais de recém-nascidos com e sem a presença de oclusão, respectivamente. O coeficiente Dice foi utilizado para comparação entre as respostas dos modelos de segmentação e o padrão-ouro, e a pontuação F1 foi selecionada para avaliar os classificadores da expressão de dor.

Após os experimentos com e sem oclusão, o modelo SAM foi considerado o melhor ao apresentar média e desvio padrão de $\mu = 0{,}98$ e $\sigma = 0{,}03$ para o coeficiente Dice ao ser aplicado na base UNIFESP 1 e $\mu = 0{,}91$ e $\sigma = 0{,}05$ quando aplicado a base UNIFESP 2. Desta forma, o modelo mostrou-se consistentemente superior aos demais em ambos os casos, com e sem a presença de oclusão, com destaque para cenários de oclusão em que os outros modelos apresentaram média aproximada $\mu = 0{,}4$.

Utilizar o SAM para segmentação da face e, consequentemente, remoção de ruídos como plano de fundo e oclusões, não gerou melhora significativa no desempenho dos classificadores, o que indica um aprendizado robusto por todos os modelos testados, que pode ser confirmado utilizando os mapas de atenção do ViT extraídos ao observar imagens com oclusão, conforme apresentado na Figura 31, uma vez que as oclusões causadas por equipamentos médicos resultaram em pouca ou nenhuma atenção durante a inferência.

Os modelos de classificação testados também não foram capazes em nenhum cenário de gerar resultados superiores a 66% de acurácia para classificação da expressão de dor em casos com oclusão, porém, a base UNIFESP2 utilizada é uma amostra, contendo apenas 10 imagens, logo uma baixa quantidade de classificações erradas gera mudanças significativas no valor de acurácia obtido.

Para trabalhos futuros, uma vez que classificadores treinados para reconhecer a expressão de dor em imagens faciais de recém-nascidos parecem não apresentar bons resultados em imagens com alta presença de oclusão, como aquelas presentes no conjunto de dados UNIFESP 2, mesmo quando oclusões e plano de fundo são previamente segmentados para não interfe-

rirem no aprendizado e na inferência do classificador, sugere-se o estudo de segmentações de partes da face para automação direta de escalas de dor como a NFCS, conforme exemplificado ao fim da Seção 6.2.

REFERÊNCIAS

BELHUMEUR, Peter N.; HESPANHA, Joao P; KRIEGMAN, David J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 19, n. 7, p. 711–720, 1997.

BRAHNAM, Sheryl et al. SVM Classification of Neonatal Facial Images of Pain. In: p. 121–128.

BUZUTI, LF. Avaliação de dor em expressão facial neonatal por meio de redes neurais profundas. Centro Universitário FEI, São Bernardo do Campo, 2020.

CANNY, John. A Computational Approach To Edge Detection. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, PAMI-8, p. 679–698, dez. 1986.

CHEN, Liang-Chieh et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 40, n. 4, p. 834–848, 2017.

CHEN, Liang-Chieh et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In_____. **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. P. 833–851.

CHENG, Bowen; SCHWING, Alexander G.; KIRILLOV, Alexander. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In.

CHENG, Bowen et al. Masked-attention Mask Transformer for Universal Image Segmentation. In.

CHENG, H.D. et al. Color image segmentation: advances and prospects. **Pattern Recognition**, v. 34, n. 12, p. 2259–2281, 2001. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320300001497>.

CHOLLET, François. Xception: Deep learning with depthwise separable convolutions. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. 2017. P. 1251–1258.

CORDTS, Marius et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: PROC. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

DENG, Jiankang et al. Arcface: Additive angular margin loss for deep face recognition. In: PROCEEDINGS of the IEEE/CVF conference on computer vision and pattern recognition. 2019. P. 4690–4699.

DENG, Jiankang et al. Retinaface: Single-shot multi-level face localisation in the wild. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. P. 5203–5212.

DENG, Jiankang et al. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. **International Journal of Computer Vision**, Springer, v. 127, p. 599–624, 2019.

DOMINGUES, Pedro et al. Neonatal Face Mosaic: An areas-of-interest segmentation method based on 2D face images. In: ANAIS do XVII Workshop de Visão Computacional. Online: SBC, 2021. P. 201–205. Disponível em: <https://sol.sbc.org.br/index.php/wvc/article/view/18914>.

DOMINGUES, Pedro et al. Neonatal Face Segmentation with and without Clinical Devices using SAM. In: ANAIS Estendidos do XXXVI Conference on Graphics, Patterns and Images. Rio Grande/RS: SBC, 2023. P. 101–104. Disponível em: <https://sol.sbc.org.br/index.php/sibgrapi_estendido/article/view/27459>.

DOSOVITSKIY, Alexey et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **ArXiv**, abs/2010.11929, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:225039882>.

EVERINGHAM, M. et al. **The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results**. 2012. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

GIBBINS, Sharyn et al. Validation of the premature infant pain profile-revised (PIPP-R). **Early human development**, Elsevier, v. 90, n. 4, p. 189–193, 2014.

GKIKAS, Stefanos; TSIKNAKIS, Manolis. Automatic assessment of pain based on deep learning methods: A systematic review. **Computer Methods and Programs in Biomedicine**, v. 231, p. 107365, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169260723000329>.

GRUNAU, Ruth VE; CRAIG, Kenneth D. Pain expression in neonates: facial action and cry. **Pain**, Elsevier, v. 28, n. 3, p. 395–410, 1987.

HE, Kaiming et al. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778.

HE, Kaiming et al. Masked autoencoders are scalable vision learners. In: PROCEEDINGS of the IEEE/CVF conference on computer vision and pattern recognition. 2022. P. 16000–16009.

HEIDERICH, Tatiany M. et al. Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units. **Jornal de Pediatria**, v. 99, n. 6, p. 546–560, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021755723000669>.

HEIDERICH, Tatiany Marcondes. Reconhecimento automatizado da dor por regiões faciais de recém-nascidos prematuros internados em unidade de terapia intensiva neonatal. Centro Universitário FEI, São Bernardo do Campo, 2022.

HEIDERICH, Tatiany Marcondes; LESLIE, Ana Teresa Figueiredo Stochero; GUINSBURG, Ruth. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. **Acta Paediatrica**, Wiley Online Library, v. 104, n. 2, e63–e69, 2015.

INSIGHTFACE. **Coordinate Regression**. 2023. Disponível em: <https://github.com/deepinsight/insightface/tree/master/alignment/coordinate_reg>. Acesso em: 23 set. 2023.

KAGGLE. **VGGNet-16 Architecture: A Complete Guide**. 2020. Acessado em 17 de outubro de 2023. Disponível em: <https://www.kaggle.com/code/blurredmachine/vggnet-16-architecture-a-complete-guide>.

KAIMING HE XIANGYU ZHANG, Shaoqing Ren; SUN, Jian. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. **CoRR**, abs/1406.4729, 2014. arXiv: 1406.4729. Disponível em: <http://arxiv.org/abs/1406.4729>.

KAZEMI, Vahid; SULLIVAN, Josephine. One millisecond face alignment with an ensemble of regression trees. **2014 IEEE Conference on Computer Vision and Pattern Recognition**, p. 1867–1874, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:2031947>.

KIRILLOV, Alexander et al. Panoptic segmentation. In: PROCEEDINGS of the IEEE/CVF conference on computer vision and pattern recognition. 2019. P. 9404–9413.

KIRILLOV, Alexander et al. Segment Anything. **arXiv:2304.02643**, 2023.

KRIZHEVSKY, Alex. **Learning multiple layers of features from tiny images**. 2009.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, p. 84–90, 2012. Disponível em: <https://api.semanticscholar.org/CorpusID:195908774>.

LAWRENCE, Jocelyn et al. The development of a tool to assess neonatal pain. **Neonatal network: NN**, v. 12, n. 6, p. 59–66, 1993.

LECUN, Y. et al. Backpropagation Applied to Handwritten Zip Code Recognition. **Neural Computation**, v. 1, n. 4, p. 541–551, 1989.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LIN, Tsung-Yi et al. Feature pyramid networks for object detection. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. 2017. P. 2117–2125.

LIU, Ziwei et al. Deep Learning Face Attributes in the Wild. In: PROCEEDINGS of International Conference on Computer Vision (ICCV). Dez. 2015.

LUGARESI, Camillo et al. **MediaPipe: A Framework for Building Perception Pipelines**. 2019. arXiv: 1906.08172 [cs.DC].

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In. Disponível em: <https://api.semanticscholar.org/CorpusID:6278891>.

RADFORD, Alec et al. Learning transferable visual models from natural language supervision. In: PMLR. INTERNATIONAL conference on machine learning. 2021. P. 8748–8763.

RANGER, Manon; JOHNSTON, C Celeste; ANAND, KJS. Current controversies regarding pain assessment in neonates. In: ELSEVIER, 5. SEMINARS in perinatology. 2007. v. 31, p. 283–288.

RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-Net: Convolutional Networks for Biomedical Image Segmentation. In_____. **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**. Cham: Springer International Publishing, 2015. P. 234–241.

RUSSAKOVSKY, Olga et al. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision (IJCV)**, v. 115, n. 3, p. 211–252, 2015.

SBP, Sociedade Brasileira de Pediatria. **A Linguagem de dor no recém-nascido**. 2018. Disponível em: <https://www.sbp.com.br/fileadmin/user_upload/DocCient-Neonatol-Linguagem_da_Dor_atualizDEz18.pdf>. Acesso em: 1 out. 2023.

SILVA, PASO. Interpretação e reconhecimento de padrões para avaliação de dor em imagens faciais de recém-nascidos. Centro Universitário FEI, São Bernardo do Campo, 2020.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. **CoRR**, abs/1409.1556, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:14124313>.

STEVENS, Bonnie et al. Premature infant pain profile: development and initial validation. **The Clinical journal of pain**, LWW, v. 12, n. 1, p. 13–22, 1996.

STEVENS, Bonnie J; JOHNSTON, C Celeste; HORTON, Linda. Multidimensional pain assessment in premature neonates: a pilot study. **Journal of Obstetric, Gynecologic, & Neonatal Nursing**, Wiley Online Library, v. 22, n. 6, p. 531–541, 1993.

SZEGEDY, Christian et al. Going deeper with convolutions. **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 1–9, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:206592484>.

SZEGEDY, Christian et al. Rethinking the Inception Architecture for Computer Vision. **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 2818–2826, 2015. Disponível em: <https://api.semanticscholar.org/CorpusID:206593880>.

TAN, Mingxing; LE, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. INTERNATIONAL conference on machine learning. 2019. P. 6105–6114.

TANCIK, Matthew et al. Fourier features let networks learn high frequency functions in low dimensional domains. **Advances in Neural Information Processing Systems**, v. 33, p. 7537–7547, 2020.

TSANG, Sik-Ho. **Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015**. 2018. Acessado em 17 de outubro de 2023. Disponível em: <https://sh-tsang.medium.com/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c>.

TURK, Matthew A; PENTLAND, Alex P. Face recognition using eigenfaces. In: IEEE COMPUTER SOCIETY. PROCEEDINGS. 1991 IEEE computer society conference on computer vision and pattern recognition. 1991. P. 586–587.

VARATHARASAN, Vinorth et al. Improving Learning Effectiveness For Object Detection and Classification in Cluttered Backgrounds. In: p. 78–85.

VASWANI, Ashish et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VIOLA, Paul; JONES, Michael. Rapid object detection using a boosted cascade of simple features. In: IEEE. PROCEEDINGS of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. 2001. v. 1, p. i–i.

VOO, Kenny T. R.; JIANG, Liming; LOY, Chen Change. Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2022.

WALKER, Suellen M. Long-term effects of neonatal pain. In: ELSEVIER, 4. SEMINARS in Fetal and Neonatal Medicine. 2019. v. 24, p. 101005.

WANG, Panqu et al. Understanding convolution for semantic segmentation. In: IEEE. 2018 IEEE winter conference on applications of computer vision (WACV). 2018. P. 1451–1460.

XIE, Enze et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: NEURAL Information Processing Systems (NeurIPS). 2021.

ZAMZMI, Ghada et al. Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN). In: IEEE. 2019 International Joint Conference on Neural Networks (IJCNN). 2019. P. 1–7.

ZHANG, Kaipeng et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. **IEEE Signal Processing Letters**, v. 23, p. 1499–1503, 2016. Disponível em: <https://api.semanticscholar.org/CorpusID:10585115>.

_____. Joint face detection and alignment using multitask cascaded convolutional networks. **IEEE signal processing letters**, IEEE, v. 23, n. 10, p. 1499–1503, 2016.

ZHAO, Hengshuang et al. Pyramid Scene Parsing Network. In: CVPR. 2017.

ZHOU, Bolei et al. Scene Parsing through ADE20K Dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 5122–5130.

ZHU, Xiangxin; RAMANAN, Deva. Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. P. 2879–2886.

APÊNDICE A – RESULTADOS DE TREINAMENTO

A Tabela 10 apresenta os resultados de treinamento de cada modelo para classificação da expressão de dor, utilizando como dado de entrada as imagens do conjunto de dados UNIFESP1, sem nenhum pré-processamento.

	Acurácia	**Precisão**	**Recall**	**F1**
Vgg16	0.842±0.041	0.868±0.048	0.844±0.069	0.854±0.041
Resnet50	0.842±0.045	0.830±0.050	0.899±0.046	0.862±0.036
InceptionV3	0.869±0.058	0.873±0.062	0.894±0.062	0.882±0.054
ViT	**0.883±0.041**	**0.884±0.048**	**0.909±0.038**	**0.896±0.036**

Tabela 10 – Métricas de treinamento por modelo utilizando UNIFESP1.

Fonte: Autor.

A Tabela 11 apresenta os resultados de treinamento de cada modelo para classificação da expressão de dor, utilizando como dado de entrada as imagens do conjunto de dados UNIFESP1 com as faces dos recém-nascidos recortadas com auxilio do RetinaFace.

	Acurácia	**Precisão**	**Recall**	**F1**
Vgg16	0.872±0.052	0.859±0.069	**0.924±0.039**	0.889±0.042
Resnet50	0.881±0.016	0.894±0.059	0.894±0.051	0.892±0.011
InceptionV3	0.867±0.046	0.860±0.058	0.909±0.045	0.883±0.038
ViT	**0.903±0.031**	**0.905±0.022**	0.919±0.045	**0.912±0.030**

Tabela 11 – Métricas de treinamento por modelo utilizando UNIFESP1-Faces.

Fonte: Autor.

A Tabela 11 apresenta os resultados de treinamento de cada modelo para classificação da expressão de dor, utilizando como dado de entrada as imagens do conjunto de dados UNIFESP1 com as faces dos recém-nascidos recortadas com auxilio do RetinaFace e com o plano de fundo e oclusões segmentadas utilizando SAM.

	Acurácia	**Precisão**	**Recall**	**F1**
Vgg16	0.853±0.040	0.857±0.054	0.884±0.038	0.869±0.031
Resnet50	0.875±0.017	0.874±0.027	0.904±0.032	0.888±0.016
InceptionV3	0.872±0.030	0.881±0.024	0.889±0.067	0.884±0.031
ViT	**0.900±0.025**	**0.891±0.034**	**0.934±0.043**	**0.911±0.023**

Tabela 12 – Métricas de treinamento por modelo utilizando UNIFESP1-Faces-SAM.

Fonte: Autor.

As Tabelas 13, 14, 15 e 16 contem as métricas de treinamento, mas agora agrupadas por modelo para permitir a comparação entre os métodos de pré-processamento propostos.

	Acurácia	Precisão	Recall	F1
UNIFESP1	0.842±0.045	**0.868±0.048**	0.844±0.069	0.854±0.041
UNIFESP1-Faces	**0.872±0.052**	0.859±0.069	**0.924±0.039**	**0.889±0.042**
UNIFESP1-Faces-SAM	0.853±0.040	0.857±0.054	0.884±0.038	0.869±0.031

Tabela 13 – Métricas de treinamento por estratégia de pré-processamento utilizando Vgg16.

Fonte: Autor.

	Acurácia	Precisão	Recall	F1
UNIFESP1	0.842±0.045	0.830±0.050	0.899±0.046	0.862±0.036
UNIFESP1-Faces	**0.881±0.016**	**0.894±0.059**	0.894±0.051	**0.892±0.011**
UNIFESP1-Faces-SAM	0.875±0.017	0.874±0.027	**0.904±0.032**	0.888±0.016

Tabela 14 – Métricas de treinamento por estratégia de pré-processamento utilizando Resnet50.

Fonte: Autor.

	Acurácia	Precisão	Recall	F1
UNIFESP1	0.869±0.058	0.873±0.062	0.894±0.062	0.882±0.054
UNIFESP1-Faces	0.867±0.046	0.860±0.058	**0.909±0.045**	0.883±0.038
UNIFESP1-Faces-SAM	**0.872±0.030**	**0.881±0.024**	0.889±0.067	**0.884±0.031**

Tabela 15 – Métricas de treinamento por estratégia de pré-processamento utilizando InceptionV3.

Fonte: Autor.

	Acurácia	Precisão	Recall	F1
UNIFESP1	0.883±0.041	0.884±0.048	0.909±0.038	0.896±0.036
UNIFESP1-Faces	**0.903±0.031**	**0.905±0.022**	0.919±0.045	**0.912±0.030**
UNIFESP1-Faces-SAM	0.900±0.025	0.891±0.034	**0.934±0.043**	0.911±0.023

Tabela 16 – Métricas de treinamento por estratégia de pré-processamento utilizando ViT.

Fonte: Autor.